Characterization of the Iberian pig genome and transcriptome using high throughput sequencing

From genes to genomes

Anna Esteve Codina

PhD thesis

September 2012

Characterization of the Iberian pig genome and transcriptome using high throughput sequencing

From genes to genomes

Anna Esteve Codina

PhD thesis

September 2012

Thesis director: Miguel Pérez-Enciso

Thesis co-director: Josep Maria Folch Albareda

INDEX

Welcome to the “Omics” era. Genomics — understanding an organism’s entire genome, transcriptomics — understanding every RNA in an organism’s genome, proteomics — understanding all the proteins in an organism, metabolomics, understanding all the metabolites in an organism, and of course, how all this layers interact with each other– This is the future of biology

Jennifer Spindel
Cornell University
http://happygrad.wordpress.com/

SUMMARY

This thesis provides insights about the evolutionary forces that have shaped the nucleotide variability patterns of pig genome. In the first Chapter, we used the traditional gene-centered approach to characterize in detail different regions of a gene putatively associated with meat quality in pigs (*SERPINA6*). In a wide diversity porcine panel, and although we found a putative causal non-synonymous substitution at high frequencies in European breeds, we were not able to infer any conclusive signal of selection.

Next, with the advent of high throughput sequencing technologies, we studied the genome-wide nucleotide variability and expression patterns in Iberian pig. Complementary methodological approaches were employed: whole genome shotgun sequencing, reduced representation libraries, sequencing a pool of individuals and transcriptome sequencing. Overall, the estimated autosomal nucleotide diversity of the Iberian pig (Guadyerbas strain) was ~0.7 kb^{-1} after correcting for low depth, a non-negligible variability considering the high inbreeding coefficient of this line. Telomeric regions presented consistently higher levels of nucleotide variability than centromers, likely a result of increased recombination rates. Further, chromosome X was much less variable than expected under a neutral scenario, relative to autosomes, which may be explained by selection or other demographic effects.

To study putative regions which may have undergone selection during domestication or modern breeding practices, we divided the genome is non-overlapping windows and calculated different selection tests in a pool of Iberians and in a single individual. Regions with an excess of polymorphisms were enriched in olfactory receptors and swine leukocyte antigens (*SLA*) genes, suggesting that they are under balancing selection. In contrast, regions with an excess of differentiation and low variability contained genes involved in oxygen transport, keratinization, hair follicle morphogenesis, feeding behavior and lipid transport, biological processes which may be under positive selection.

We also characterized the Iberian genome in terms of structural variants. For this purpose, we used a read depth approach and detected many multi-copy regions gains with respect to the reference assembly. About 5% of annotated genes were totally comprised inside those regions and the majority belonged to gene superfamilies.

In a comparison of the gonad transcriptome of two pigs with extremes phenotypes, an Iberian pig and a Large White pig, we detected differentially expressed genes involved in spermatogenesis and lipid metabolism. This agrees with phenotypic differences between both breeds. To improve the annotation of the pig genome, we also developed a pipeline to detect long-non-coding RNAs and novel protein coding genes expressed in the male gonad tissue.

CHAPTER 1

GENERAL INTRODUCTION

1.1 General overview of next generation sequencing applications

The advent of the next generation sequencing (NGS) technologies has revolutionized biology, making it possible the thorough investigation of the genome and transcriptome in multiple species. With these new technologies, not only human or model organisms (e.g., mouse) can be studied in detail, but also non-model organisms. Compared with standard Sanger sequencing, they provide a marked improvement in sequencing speed and throughput at a reasonable cost. At the biological level, the advantages are many. Genome-wide variants catalogues, and genome-wide expression patterns in different tissues are being available for many species, which will pave the way for a new biomedicine and agricultural research. For example, in the near future, we expect to pay a few hundred euros to have our genome sequenced, which would fuel the personalized medicine. Whole-genome sequence association studies will overcome the SNP ascertainment biases inherent in the current SNP chips, since they will uncover all variants genome-wide, including population, individual or region specific variants. In cancer genomics, there are many ongoing projects (*The International Cancer Genome Consortium)*, which aim to explore different cancer types to detect somatic mutations and chromosomal rearrangements sequencing specific tumor tissues (http://www.icgc.org/icgc).

In the area of agrigenomics, these techniques will help to detect causative variants or genotypes associated with economically important traits that humans have been 'blindly' selected throughout centuries. An illustrative example is the intron mutation located in the myostatin gene, which confers an extreme musculation phenotype in the cow (See Figure 1) (Grobet *et al.* 1997). Also important, is the impact that NGS would cause in evolutionary studies, which may help us to better understand population demographic events and evolutionary forces that have shaped nucleotide variability patterns in species' genomes. These studies help to better understand human evolution (Tishkoff & Verrelli 2003) and the process of animal speciation and domestication (Diamond 2002; Trut *et al.* 2009; Amaral *et al.* 2011; Wiener & Wilkinson 2011).

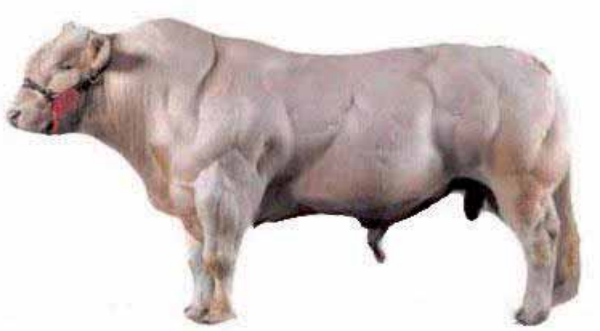

Figure 1. A causative mutation in the *MSTN* gene confers an extreme musculation phenotype in cows

There are a broad range of NGS applications in biological research; our challenge will be to intelligently integrate different data types to produce a coherent picture of the genetic bases of complex phenotypic traits (economical production traits in agrigenomics, complex diseases in biomedicine, morphological traits in forensics or tameness in domestication). These complex traits are quantitative traits; the genetic background and the environment explain the observed variance. The genetic bases are either caused by multiple single mutations in the DNA (SNPs), structural variants (inversions, insertions, deletions, duplications and translocations affecting stretches of the DNA) or epigenetic factors that change gene expression levels (e.g., modifications in histone conformation). NGS technologies make possible to study all these genetic variants genome-wide at a very high-resolution. Genome sequencing (DNA-seq) allows detecting SNPs, small indels, chromosome rearrangements and copy number variants (CNVs). transcriptome sequencing (RNA-seq) measures gene and isoforms expressions, chromatin immunoprecipitation sequencing (Chip-Seq) identifies DNA and transcription binding factor interactions and methylome sequencing (Meth-seq) detects methylated DNA regions along the genome, high-order chromatin architecture (Hi-C) determines chromatin interactions in 3D. Therefore sequencing the genome, transcriptome and epigenome of an organism, and then integrating all these data in gene networks and biological pathways may help to elucidate the complex genetic mechanisms that result in the observed phenotype diversity.

1.2 The technology and its drawbacks

Technology for biological applications has exponentially improved in a very short period of time. As a result, terabytes of biological data have been generated and the challenge would be to develop new computational infrastructures, statistical and bioinformatical tools to analyze and store such a vast amount of data. Now we can easily sequence the whole genome or transcriptome. The basis of this revolutionary technique consists in fragmenting the DNA from an organism into small pieces, sequence them in a high-throughput parallel machine obtaining millions of short sequences (reads) which then must be aligned and assembled in larger sequences using computational methods, e.g., Burrows-Wheeler algorithm (Li & Durbin 2009), until the genome sequence is fully reconstructed. If a reference genome of the target organism is available, the reads obtained can be mapped against it, facilitating the work. Otherwise, we must assemble the genome de novo; the drawback will be a dramatic increase in computational resources needed, since we need more efficient algorithms.

One major problem of the current draft or finished genomes used as reference sequences is miss-assembly (Salzberg & Yorke 2005). The fact that a high percentage of the high eukaryote genomes is composed of repetitive elements (a recent study reported 66-69% in the human genome (de Koning *et al.* 2011)), and some of them could be very long, makes it particularly difficult to know their exact location and the number of copies in the genome. The scientific community must take into consideration that if the reference genome is not well assembled, these include regions where a genome is incorrectly re-arranged as well as places where large chunks of DNA are absent; this could lead to erroneous conclusions in subsequent genome analysis, e.g., synteny comparisons or structural variants detection. Some other not fully resolved problems that we have to deal with next generation sequencing techniques are sequencing errors, read miss-alignments and due to a miss-assembled reference genome, some of the reads generated may not map to it. Sequencing errors are produced during the PCR amplification step or the sequencing process itself and tend to be 1-2%, which means that for a read of 100bp, on average there will be 2 mistakes. These mistakes can be indistinguishable from a real polymorphism and for that we have to take into account base qualities and if possible to have high depth. Alignment errors tend to occur due to the short nature of the reads, which may align in different locations if there are repetitive stretches in the reference genome. Mate-pair reads can resolve the correct genome assignment for some repetitive regions as long as one read in the pair is unique to the genome. For SNP calling, it is important to avoid wrongly mapped reads, otherwise the rate of false positives will increase, and for that reason uniquely mapped reads or mapping quality must be taken into consideration. Finally, those unmapped reads can be clustered together to try to get more data from them, but the logically option will be to improve the reference assembly.

1.3 Genomes available and current sequencing projects

A HiSeq 2500 Illumina machine produces nowadays 6 billion (6x10^9) of 2x100 bp paired-end short reads with a throughput of 600 Gb in 11 days. A human genome needs just 1 day to be fully sequenced at depth 30X. Craig Venter and James Watson were the first human genomes sequenced with next generation sequencing technologies to be publicly available (Levy *et al.* 2007; Wheeler *et al.* 2008). In livestock, many specie genomes have been also fully sequenced like chicken (International Chicken Genome Sequencing Consortium (2004)), cow (Elsik *et al.* 2009) and pig (Groenen 2012). In plants, we have the complete genome of rice (International Rice Genome Sequencing Project (2005)), potato (Xu *et al.* 2011) and melon (USDA 2010; Garcia-Mas 2012) among others. Regarding human infectious diseases, HIV and the malaria parasite *Plasmodium falciparum* genomes have also been sequenced (Gardner *et al.* 2002; Watts *et al.* 2009). Every year, the number of species sequenced increases. Many international projects are currently funded to take advantage of the speed and efficiency of next generation sequencing to sequence large amount of organisms. The 1000 Genomes Project (http://www.1000genomes.org/) had the objective of sequencing full genomes of different human populations (Asians, Europeans, Americans and Africans) and to have a resource of human genetic variation, whereas The 1001 Plant Genomes Project (http://www.1001genomes.org/) provides a broad catalog of Arabidopsis thaliana genetic variation. Also interesting is the Human Microbiome Project (http://www.hmpdacc.org/), which aims at sequencing the entire DNA that conform human microbiota (intestines, mouth, skin), which will certainly discover new microorganisms and help to interpret host-pathogen interactions in human diseases. For the fruit fly, we have the The Drosophila Genetic Reference Panel (http://dgrp.gnets.ncsu.edu/), a living library of common polymorphisms affecting complex traits, and a community resource for whole genome association mapping of quantitative trait loci.

1.4 Cost-effective strategies for SNP discovery

In a resequencing project, the goal is normally to identify variants, and a reference genome is assumed to exist to carry out the alignment. Whole-genome sequencing of many individuals is not currently affordable for small laboratories and alternative approaches are used to have a cost effective way to generate SNPs, e.g.,. sequencing fewer parts of the genome (less coverage) but at a higher depth. For this, reduced representation libraries (RRL) are a good choice. This method can be applied to either a single individual or pools of individuals. It consists of fragmenting the genomic DNA with a restriction enzyme and then sequencing only fragments of a certain size. In general, the percentage of genome sequenced varies between 1-5%, but the ratio can be adjusted approximately using in silico digestion of the assembled genome. The RRL pooling approach has been successfully employed in several animal species, like cow (Van Tassell *et al.* 2008), turkey (Kerstens *et al.* 2009) and pig (Wiedmann *et al.* 2008; Amaral *et al.* 2011). However, DNA pooling of different individuals results in a number of uncertainties: the exact number of chromosomes sequenced is not known and a given chromosome may be under or over-represented (Perez-Enciso & Ferretti 2010). This fact is more pronounced as the number of individuals on the pool increases and the depth decreases. Moreover, singletons will be very difficult to spot, which means that the site frequency spectrum will be biased towards mid-frequency alleles. In that way, as sequencing errors will also be more difficult to detect, the variance of the estimators will be higher than with individual sequencing (Perez-Enciso & Ferretti 2010). All these caveats have been addressed for the calculation of the Watterson estimator of nucleotide diversity in pools (Ferretti 2012). For organisms with large genomes (e.g., mammals), the trade-off of coverage versus cost may justify dealing with statistical complexities of low-coverage datasets at least until further sequencing improvements and cost reductions are achieved.

1.5 Structural variants

It has been reported that two human genomes can differ in 3,000,000 SNPs (Jorde & Wooding 2004; Tishkoff & Kidd 2004) and up to 300 Mb (Li *et al.* 2011) of sequence length, which highlights the high plasticity of the genome architecture, even within the same species. Note that more nucleotides are affected by structural variants than single point mutations, implying significant consequences of structural variants in phenotypic variation. Structural variants consist of many kinds of variation in the genome of one species, and usually include microscopic and submicroscopic types, such as deletions, duplications, copy-number variants, insertions, inversions and chromosomal translocations. Copy number variants are stretches of DNA ranging from 1kb to hundreds of kb that appear in multiple copies in the genome and the number of copies differs between individuals or populations. One of the consequences is different gene dosage, which could lead to more expression of a particular gene. On the other hand, segmental duplications refer to duplicated regions in the genome that are fixed in the population. This process of gene duplication originated many gene superfamilies like olfactory receptors, hemoglobins or the histocompatibility complex proteins. Over time, one of the duplicated regions has evolved conferring a different function from the original gene. Two main methods are suitable for a genome-wide detection of these paralogous and multi-copy regions using high-throughput sequencing, a read density approach (see Figure 2a) using depth information along the genome to detect if our sample has more or less copies than the reference genome (Alkan *et al.* 2011), or a paired-end distance information approach (Alkan *et al.* 2011) (see Figure 2b) , which uses aberrant distance or orientation between read pairs to detect novel insertions, deletions, inversions and translocations. Both methods can be conducted using whole-genome sequencing, but the latter needs a high-quality assembled reference genome in order to minimize false positives.

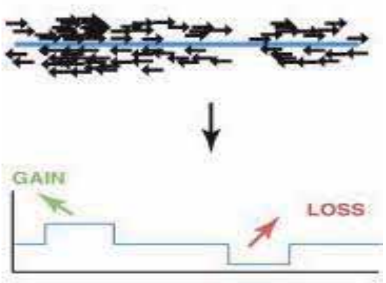

Figure 2a. Gains and losses detection with respect to the reference genome using a read depth approach. Reads are aligned against the reference (top) and then, after GC bias correction, read depth is computed to infer changes of copies (bottom).

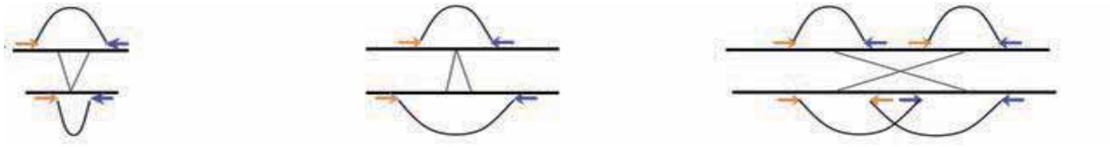

Figure 2b. Structural variants detection using aberrant paired-end distances. The first case the donor (top) has an insertion with respect to the reference (bottom), the second case a deletion and the third case an inversion.

1.6 Sequencing the transcriptome with RNA-seq

As mentioned, it is also possible to sequence the full transcriptome, which comprises the complete set of transcripts in a cell or tissue expressed under a certain physiological condition. RNA-seq, a recently deep-sequencing technology has already modified our view of the extant complexity of transcriptomes and provides us with the opportunity to characterize the functional elements of the genome including mRNAs, small non-coding RNAs and the newly discovered long-non-coding RNAs. Some RNA-seq studies have been performed in pig; for example several research groups (Huang *et al.* 2008; Li *et al.* 2010b; Xie *et al.* 2011; Lian *et al.* 2012) discovered microRNAs in different tissues, others (Chen *et al.* 2011a; Zhao *et al.* 2011) sequenced divergent breeds in terms of growth and leanness.

Regarding expression levels, this technique overcomes the hybridization-based approaches, e.g., microarrays, due to its wide dynamic range of expression detection (Wang *et al.* 2009). Microarrays lack sensitivity either at very low or high expression levels. Moreover, another limitation of microarrays is that only a portion of the transcript is analyzed and isoforms are generally indistinguishable from each other. In terms of sequence architecture, for instance, to detect alternative splicing at single-base resolution, RNA-seq is the best – although not perfect - technology. Another advantage is that RNA-seq is not limited to detecting transcripts that correspond to existing annotated genes; many new expressed regions are uncovered, which is very attractive for incomplete annotated genomes or non-model organisms. Another application is the detection of transcript chimeras (exon fusion from different genes). For example (Frenkel-Morgenstern *et al.* 2012) discovered hundreds of chimeric RNAs to be genuinely expressed in normal human cells. In addition, polymorphism (SNPs or SVs) and allele specific expression (ASE) can be revealed (Skelly *et al.* 2011; Li *et al.* 2012a). Nevertheless, some challenges need to be resolved, e.g., those related with bias in library preparation (RNA fragmentation, the GC bias amplification and PCR artifacts) or the sequencing process (sequencing errors). Also, during the mapping step one read could have multiple locations in the genome (e.g., paralogs), which could be alleviated either by assigning probabilistic approaches (Pasaniuc *et al.* 2011; Glaus *et al.* 2012), obtaining larger reads, or using paired-end sequencing. In order to construct and assemble transcripts, sufficient depth is required, which can be difficult to achieve for low expressed genes. Furthermore, for genome annotation it would be necessary to analyze different tissues and developmental stages to fish the total amount of genes in a genome, increasing the costs of the experiments.

1.7 Applications in livestock

A broad goal of this thesis was to apply these new technologies to livestock, and to pigs in particular. Livestock populations form a unique genomics resource as a result of their remarkable phenotypic diversity and their population structures. Information from the genome, and the effect of its variation on phenotype will help to clarify basis of adaptation within populations under selection pressures since domestication, ca. 7-10 KYA. Due to artificial selection, a high variety of phenotypes emerged to fulfil different production objectives. This is the case for chicken broilers (meat producing) and chicken layers (egg producing), dairy cattle (milk producing) and beef cattle (meat producing), among others. Although not a livestock species, the dog is a good example, due to an intense modern breeding, its variability in size, color and shape is without parallel without any other species. There are many interesting agricultural traits that are relevant from an industry perspective, e.g., improved growth and development, wool production, disease resistance, reproductive performance or reduced environmental impact. Recently published studies used whole genome sequencing data to detect fingerprints of artificial selection. In pigs, (Amaral *et al.* 2011) sequenced four domestic pig breeds genomes (Landrace, Large White, Duroc and Pietrain) and one wild boar and found signals of selection for behavior, coat color, growth and muscle development. Recently, the pig genome paper identified strong selection on genes related with RNA processing and regulation, since the split of the European and Asian wild boar populations 1M years ago (Groenen 2012). In chicken, Rubin et al. (Rubin *et al.* 2010) identified a selective sweep in domestic chickens at the locus for thyroid stimulating hormone receptor (*TSHR*), which has a pivotal role in metabolic regulation and photoperiod control of reproduction in vertebrates. In dairy cattle, eleven candidate genes were identified with functions related to milk-production, fertility, and disease-resistance traits (Larkin *et al.* 2012).

In this thesis, we focused in the pig, which is an important livestock species for several reasons. With production and consumption of about 100 million metric tons per annum; pork is the most widely consumed meat globally (USDA 2010). China, USA, Germany and Spain are the top producing countries worldwide. In genetic terms, it would be helpful to target the genotypes associated with economical interesting traits, such as meat quality, disease resistance and growth, to perform breeding on selected animals. Due to its extreme phenotypic diversity for several traits of interest, and the fact that the wild ancestor is still available and other outgroup species as well (see Figure 3), scanning nucleotide patterns through their genomes, genes that underwent selection during

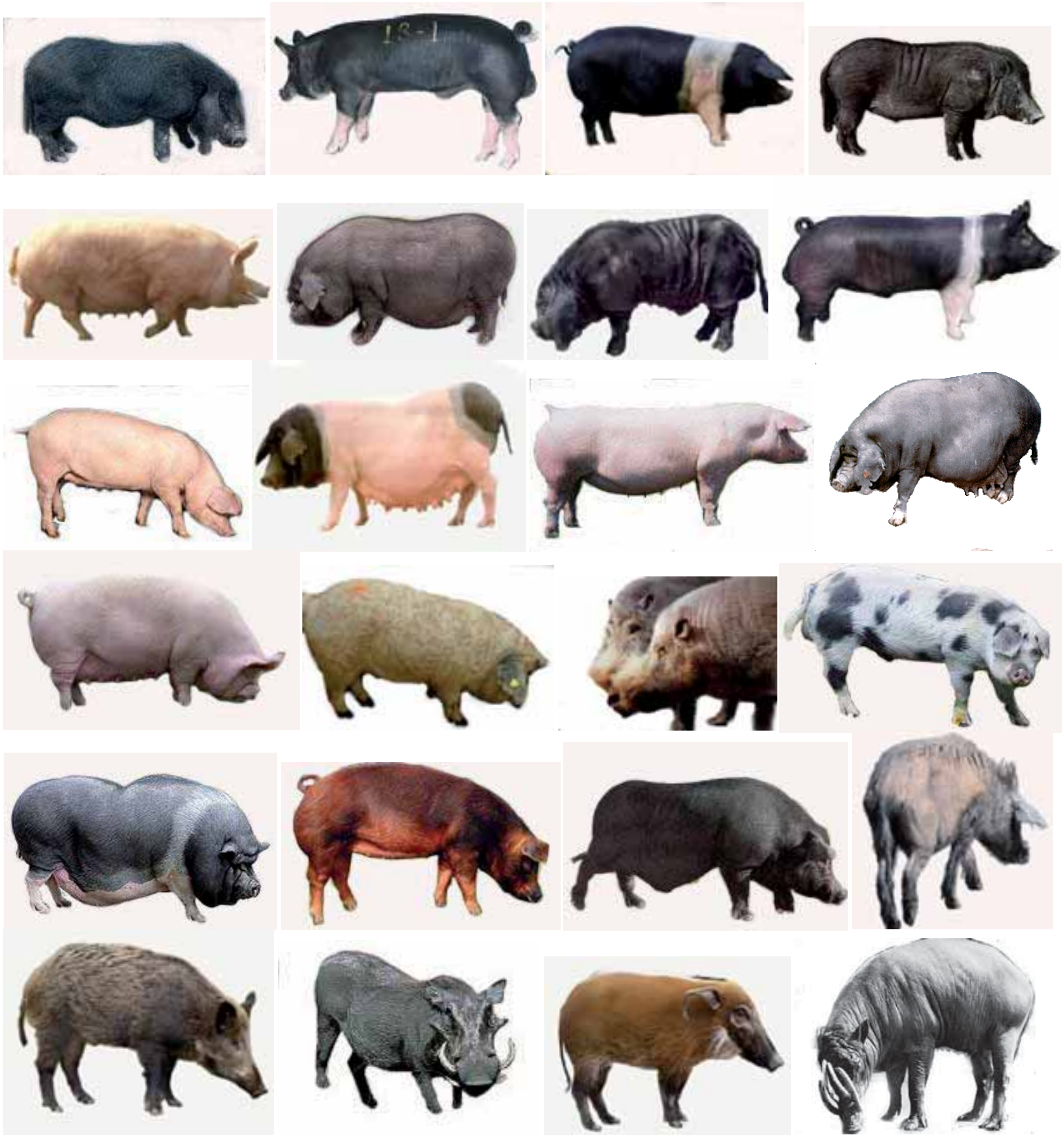

Figure 3. In order of appearance: Bamei, Berkshire, British Lop, Diannan (small ears), Tamworth, Chenghua, Erhualian, Hampshire, Large White, Jinhua, Landrace, Meishan, Middle White, Paradise, Bearded pigs, Bentheim Black Pied, Vietnamese, Duroc, Huai, Tibetan pig, Wild boar, Phacochero, Potamochero, Babyrousa.

domestication can be revealed. Thus, causal mutations responsive for adaptive processes to a new environment can be detected and pig speciation process better understood (Groenen 2012). Third, the pig provides a uniquely relevant animal model for human disease (e.g., melanoma, obesity, diabetes, wound repair, atherosclerosis), surgical research and as a potential source of organs for xenotransplantation owing to the similarities in size, anatomy and physiology (Groenen 2012).

Currently, there are about 7000 pig QTL representing almost 600 different traits at the Animal Quantitative Locus database (http://www.animalgenome.org/cgi-bin/QTLdb/SS/index). Pig meat quality, lipid deposition, growth and prolificacy are some studied and economically interesting traits. Although many QTLs associated with economical traits have been identified in different livestock species, very few have been related to its causal mutation. Examples of causative mutations identified are the *RYR1* is causative for malignant hyperthermia susceptibility in cattle (Fujii *et al.* 1991), *MSTN* with extreme muscular phenotype in cattle (Grobet *et al.* 1997), *IGF2* is associated in muscle growth in pigs (Van Laere *et al.* 2003), *PRKAG3* with excess glycogen content in pig skeletal muscle (Milan *et al.* 2000), *DGAT1* with milk production in cattle (Grisart *et al.* 2002), *CLPG* in sheep muscularity (Freking *et al.* 2002) and *FecB* (Boroola) with fecundity in sheep (Mulsant *et al.* 2001). Moreover, several CNVs have been found to be associated with phenotypic traits in livestock. In swine, dominant white color is associated with a duplication of the *KIT* gene (Giuffra *et al.* 2002), whereas in chicken a multi-copy of the *SOX5* gene causes the pea-comb phenotype (Wright *et al.* 2009). But, in general, agricultural interesting traits are very complex and multifactorial, meaning that many genes with small effect and other environmental factors interact. As a result only those QTL with big effect, explaining most of the variance of the phenotypic trait, have been discovered. Moreover, the interval of QTL regions tend to be large due to the methodology employed to detect them (classical QTL linkage mapping), although population-based association studies (GWAS) increases the precision of the QTL position estimates and reduces their confidence intervals as it uses all recombination events since the mutations occurred (Meuwissen & Goddard 2000).

However, with the availability of cost-effective whole-genome SNP panels (e.g., SNPchip arrays) for the major livestock species; one can follow the segregation of the entire genome and not merely a set of specific regions of interest, moving to the traditional marker-assisted selection (MAS) to genomic selection. The limitation of MAS for breeding programs is that it requires prior knowledge of gene alleles or markers associated with the traits of interest together with their quantitative estimates in a specific population; it must be therefore implemented within families (Eggen 2012). Furthermore, it explains only a limited part of the genetic variance. On the contrary, with tens of thousands of SNP distributed along the genome (or even better, using the whole genome sequence to avoid SNP ascertainment biases), it is not necessary to know where are specific genes located in the genome as it is expected to be one or several SNP in linkage disequilibrium with the causal mutation and, therefore, explaining a much greater variance than MAS. Finally, genomic selection can be implemented very early in life, and extended to traits with low heritability or difficult to measure (Eggen 2012).

1.8 Introduction to pig domestication and breeding

The first evidences of pig domestication trace back to 9,000 years ago, when the Neolithic farmers in the Old World and China began to tame wild boars to provide them with a source of food (meat), clothing (skin) and tools (bones). Around 1 million year ago, the ancestral South Eastern Asia wild boar population spread towards Europe, leaving behind two differentiated populations, Asian and European (Giuffra *et al.* 2000; Groenen 2012). After, multiple independent places across Eurasia originally domesticated the pig (Larson *et al.* 2005; Wu *et al.* 2007; Megens *et al.* 2008). It has been reported that pig domestication modified pig behavior, color and size (Price 1999; Diamond 2002; Fang *et al.* 2009). Soon after, traditional breeding originated local breeds (Iberian) and two centuries ago, with the development of modern breeding practices emerged current commercial pig breeds for meat production, which have excellent performance in growth and very low amounts of fat (Hampshire, Duroc, Landrace, Large White and Pietrain). It is well documented that some of these commercial breeds have been extensively introgressed with Asian germplasm

(Jones 1998) in order to achieve a higher prolificacy. But these practices have been less accentuated in China; traditional local breeding has predominated and therefore fewer local breeds have become extinct compared to Europe (Fang *et al.* 2005; Megens *et al.* 2008). As the genus Sus originated in Asia, the ancestral genetic pool of wild boars had higher diversity and therefore is where we can find more diverse pigs, which is corroborated with genetic studies (Larson *et al.* 2005; Wu *et al.* 2007; Ramirez *et al.* 2009; Luetkemeier *et al.* 2010), whereas in Europe, due to a smaller population size of the founder wild boars, pig variability is lower. A paradigm emerges when European domesticated pigs genetic diversity is compared to its wild counterparts; they have same levels of nucleotide variability (Scandura *et al.* 2008; Ramirez *et al.* 2009). Several reasons could explain this phenomenon, a recent decrease in European wild boar populations due to hunting, genetic interchange between both sub-species (Porter 1993) and the recent introgression of Asian germplasm into the European domestic pool to create commercial improved breeds (Jones 1998).

1.9 The Iberian pig

The Iberian pig is one of the European traditional swine breeds that has not been subject to human modern intensive artificial selection of pig production (Lopez-Bote 1998). Native from the Southwest Iberian Peninsula, it is a perfectly adapted breed to the Mediterranean ecosystem (Fabuel *et al.* 2004). It has been grown and maintained for centuries in large herds in *Dehesa* ecosystem, a sparse oak woodland with Mediterranean climate. Its distinctive look, small head, narrow snout, short and muscled neck, black and open hof, scarce and weak hair, dark skin and muscled legs, makes it resistant to hard climate temperatures and suitable for pasturing. It can be quickly fattened with available acorns, grass, small roots and bulbs (Lopez-Bote 1998). The Iberian pig breed is able to store a high proportion of intramuscular fat with high content of unsaturated fatty acid (oleic and linoleic) resulting from high acorn intake (Toro *et al.* 2000). These characteristics produce hams with unique and highly appreciated flavor.

Although it has excellent maternal skills, it has low prolificacy and small number of functional teats (Lopez-Bote 1998). They are typically black or red coloured or even black spotted, depending on their origin. The Torbiscal strain was generated in 1963 from four different crosses involving Negro Lampiño, Retinto and Dourado Alentejano pig varieties (Alves *et al.* 2003; Clop *et al.* 2004), whereas Guadyerbas is a highly inbred strain that was derived from a small number of Negro Lampiño pigs in 1945 (Toro 2008). Other Iberian strains are Puebla, Campanario, Ervideira and Caldeira, the two latter ones originating from Portugal (Clop *et al.* 2004). In Figure 4 are depicted some of the most emblematic Iberian varieties.

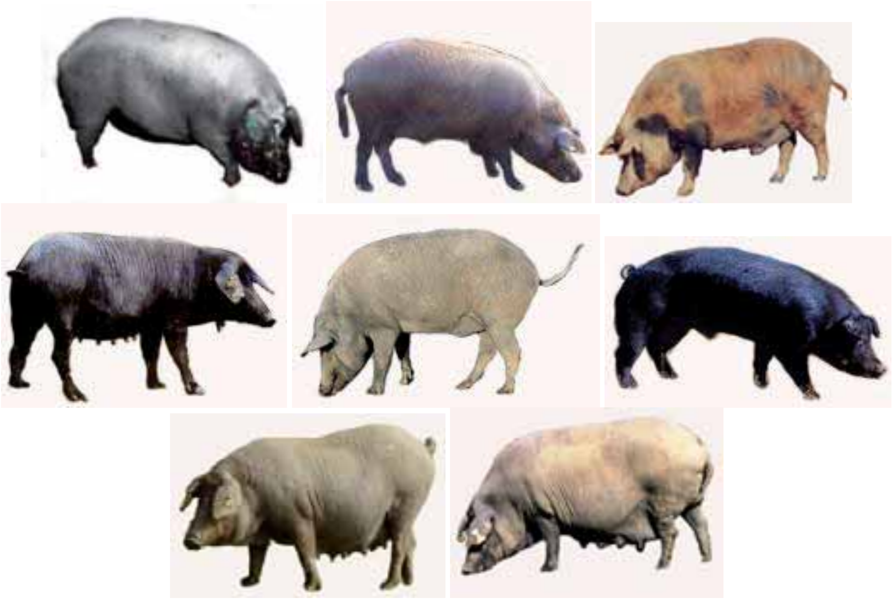

Figure 4. Guadyerbas, Dorado-Gaditano, Manchado de Jabugo, Mamellado, Retinto, Negro-Entrepelado, Torbiscal and Negro-Lampiño.

In this way, these Iberian strains have emerged as a result of the mixture of ancestral autochthonous pig populations from the Iberian Peninsula. They have not been significantly introgressed with other Chinese or European breeds, probably due to the fact that remained geographically isolated for a long time span (Clop *et al.* 2004). At the end of the 15th century, Spanish and Portuguese colonizers exported the Iberian pig to South America originating the current Creole pig breeds (Alves *et al.* 2009). Red Iberian pigs imported from Portugal and Spain in the XIX century also contributed to the origin of the Duroc-Jersey breed in the United States (Alves *et al.* 2009).

Iberian pigs suffered a strong bottleneck in 1960's due to the outbreak of the African swine fever, the lowered value of animal fat and the massive introduction of international improved pig breeds (Fabuel *et al.* 2004). The old breed structure with differentiated varieties locally distributed is disappearing; some ancient varieties are either extinct or endangered (Fabuel *et al.* 2004). In the recent years, however, Iberian pig populations increased to fulfill new demand of its high quality curated products like Iberian ham, a true and expensive gourmandize.

1.10 Genome-wide approaches to study selective fingerprints

Evolutionary forces leave a characteristic fingerprint in nucleotide patterns. Mutation and recombination are the two main forces that generate genetic variability and tend to be higher in telomeres than centromeres (Nachman 2002; Jensen-Seaman *et al.* 2004). Genetic drift is accentuated when a population suffers a bottleneck, causing a decrease in genetic diversity. Historical events where the effective population size of pigs decreased are: the domestication process, the formation of modern breeds 200 hundred years ago, the last glaciation (Scandura *et al.* 2008), and reduction of local breeds' production (e.g., Iberian pig). Migration occurs when there is a gene flow between two differentiated populations; it is the case of the American pig breeds formation (admixture of Asian and European breeds is reported in (Porter 1993; Ramirez *et al.* 2009; Souza *et al.* 2009) or the putative ancient gene flow between European wild boar and domestic pigs (Giuffra *et al.* 2000; Megens *et al.* 2008; Ramirez *et al.* 2009). But selection (natural or artificial) is the only pressure that leads to an adaptive change at the phenotypic level. Directional selection takes place when an advantageous mutation increases in frequency in a population removing variability with linked loci in the neighborhood (genetic hitchhiking). Purifying selection occurs when a deleterious mutation is removed in a populations leading to a loss of variability in linked loci (background selection). In contrast, balancing selection favors diversity maintaining heterozygote genotypes.

Traditionally, most tests for selection have compared a specific set of polymorphisms within a gene region against neutral expectations. Recently, tests have been applied to newly available genome-wide polymorphisms data, representing a turning point in the study of positive selection in many species. Genome-wide scans for evidence of selection events use either resequencing data from one or more species or populations. Generally, between-species comparisons are used to identify older events, while population-based methods reveal more recent episodes of selection. In contrast to the demographic processes acting upon the entire ensemble of genomic diversity, natural selection targets primarily functional elements in specific gene regions.

There are many different tests to detect selective pressures that act shaping nucleotide variability at specific regions of the genome. These methods assume that the selected regions display different patterns of variability than neutral regions. Tajima's *D* (Tajima 1989), Fu&Li's *D* (Fu & Li 1993), Fay&Wu' *H* (Fay & Wu 2000) neutrality tests are based on the frequency spectrum of polymorphisms. Under the neutral model, for a population at constant size at equilibrium, both moments estimates of the population genetic parameter theta, θ_S and θ_Π, are expected to be equal:

$$E[\pi] = \theta = E\left[\frac{S}{\sum_{i=1}^{n-1}\frac{1}{i}}\right] = 4N\mu$$

Where *S* is the number of segregating sites, *n* is the number of samples, *i* is the index of summation, *Ne* is the effective population size and μ is the mutation rate. But selection, demographic fluctuations and other violations of the neutral model will change the expected values of θ_S and θ_Π, so that they are no longer expected to be equal. The difference in the expectations for these two variables (which can be positive or negative) is the crux of Tajima's *D* test statistic. Tajima proposed:

$D = (\theta_\Pi - \theta_S) / s$

Where θ_S is the number of segregating sites in a sample of *n* sequences and θ_Π is the mean pairwise difference between the sequences in the sample and s is the standard deviation. A significant *D* > 0 suggests either a recent population bottleneck or some form of balancing selection (excess of intermediate frequency alleles), whereas $D < 0$ suggests either population expansion or purifying selection (excess of low frequency alleles).

In a similar way, Fu & Li also proposed:

$D = (\theta_{S>1} - \theta_{S1}) / s'$

Where S>1 are all segregating sites that affect more than one individual and $S1$ is the number of segregating sites that affect i individuals in the sample. Thus it makes a distinction between old and new mutations. In many ways it shares much information with Tajima's D statistic, a negative value indicates an excess of singletons (which would also give a negative Tajima's D), and a positive value indicates a lack of singletons (which would typically, though not necessarily, give a positive Tajima D). However, certain population genetic scenarios, particularly selective sweeps, tend to generate an excess of singletons, to which this test is more sensitive than Tajima's D.

Finally, Fay & Wu proposed a test, which is heavily influenced by high frequency derived mutations:

$H = (\theta_{\Pi} - \theta_{H}) / s$

Where Π is the mean pairwise difference between the sequences in the sample and H is sensitive to high frequency derived alleles. In this case, knowledge of the ancestral allele is needed, which can be facilitated by the availability of a close-related specie. H measures departures from neutrality that are reflected in the difference between high-frequency and intermediate-frequency alleles. In contrast, D measures departures from neutrality that are reflected in the difference between low-frequency and intermediate frequency alleles. Thus, while D is sensitive to population expansion because the number of segregating sites responds more rapidly to changes in population size than the nucleotide heterozygosity, whereas population subdivision is more of a problem for H (Holsinger 2001-2010). As a result, combining both tests (Zeng *et al.* 2006) may allow you to distinguish population expansion/ positive selection from purifying selection (Holsinger 2001-2010).

The HKA test (Hudson *et al.* 1987) uses polymorphism and divergence data from two or more loci. Under neutrality, the same ratio between polymorphism within and divergence between species in the loci under observation is expected, as they are proportional to mutation and drift. Otherwise, there is an indication of selection. The HKA test is a quite robust test to departures of the stationary model (e.g., demography). This model assumes there is free recombination between loci and no recombination within loci. Many deviations from the model, for example linkage between the loci or a population bottleneck in the past, generate correlations between the genealogies at the two loci and therefore reduce the variance of the test statistic, making the test conservative.

McDonald & Kreitman test (McDonald & Kreitman 1991) is similar to the HKA test in that it compares the levels of polymorphism and divergence at two sets of sites. Whereas for the HKA test the two sets of sites are two different loci, the McDonald-Kreitman test examines sites that are interspersed: synonymous and nonsynonymous sites in the same locus. Because the sites are interspersed, it is safe to assume that the genealogies for the two are the same. The test therefore has four statistics; *Ds*, *Dn*, *Ps* and *Pn*, corresponding to synonymous and nonsynonymous divergent and polymorphic sites. The McDonald-Kreitman test is a very robust test, because no assumption about the shape of the genealogy is made. It is therefore less sensitive to the demographic histories, geographic structuring and non-equilibrium statuses of the populations sampled. If synonymous mutations are considered neutral on a priori basis, then a significant departure from independence in the test is an indicator of selection at nonsynonymous sites. An excess of substitutions can be interpreted as evidence of adaptive evolution, and an excess of polymorphism can be interpreted as evidence of purifying selection since deleterious alleles contribute to polymorphism but rarely contribute to divergence.

Similarly, tests based on the substitution rates between nonsynonymous and synonymous sites consider the *dN/dS* ratio (number of nonsynonymous substitutions per nonsynonimous site divided by the number of synonymous substitutions per synonymous site) (Messier & Stewart 1997). *dN/dS* =1 is observed under the neutral model, *dN/dS* > 1 when positive selection is acting and *dN/dS* < 1 for purifying selection. Genetic variants that alter protein function are usually deleterious and are thus less likely to reach fixation than mutations that have no functional effect on the protein. Positive selection over a prolonged period however can increase the fixation rate of beneficial function-altering mutations.

When geographically separate populations are subject to distinct environmental pressures, positive selection may change the frequency of an allele in one population but not in another. Relatively large differences in allele frequencies between populations may therefore signal a locus that has suffered selection. Commonly used statistics for population differentiation is the Fst (Nei 1973). Nevertheless, distinguishing between genuine selection and the effect of demographic history, especially population bottlenecks, is difficult.

Both selection and population demographic history have important influences on the amount and patterns of genetic variation, and sometimes they act in the same way. Population subdivision leads an increment of intermediate frequency alleles, mimicking balancing selection, whereas population expansion mimics positive selection, since both scenarios lead to an excess of low–frequency alleles. This presents an important challenge in the analysis of population genomic data, since studies of selection should ideally incorporate the confounding effects of demographic history and viceversa. The standard neutral model assumes no population structure, constant population size and random mating, assumptions that are not always fulfilled and therefore, the mere rejection of neutrality tests does not point unambiguously to an effect of selection. The problem is therefore to generate the distribution of the statistic of interest (e.g., Tajimas' *D*) under a demographic model congruent with the observed data. Modern approaches make intensive use of simulations methods and can be applied to large amounts of data generated by NGS technologies (e.g., ABC methods) (Li & Jakobsson 2012). These methods can easily accommodate different demographic scenarios like population expansions, bottlenecks or migration and estimate population parameters (e.g., time of divergence, effective population size) that best fit the hypothesized model. Then outliers in DNA sequence data are detected comparing the distribution of the statistic under the standard neutral model against the new distribution under the simulated model.

In this way, genome-wide patterns of genetic variation will capture the effect of demography and the extreme tails of the distribution will be suggestive of regions under selection. But, as a recent genome resequencing data in Drosophila (Sella *et al.* 2009) argued, the aforementioned approach is not adequate if selection is common in the genome, which seems to be the case in Drosophila, a genus with large effective population sizes. Another difficulty in distinguishing the effects of selection and demography deals with the different population sizes and recombination rate in autosomes and sex chromosomes. The Ne for X chromosome is expected to be ¾ of that of the autosomes, and therefore the neutral prediction is that there is a reduced diversity in X chromosome. Nevertheless, under polygamy systems with biased reproductive skews, the X/A diversity ratio is expected to increase if just few males reproduce, whereas the contrary is expected if just few females reproduce (Hammer *et al.* 2008). Sex-specific demographic patterns as dispersal and philopatry also influence this ratio, for instance if males disperse more frequently than females, X genetic diversity will be reduced. Also important, are population bottlenecks or expansions. The former lead to disproportionally reduced sex-linked variation, whereas the latter have the opposite effect, leading to more equal levels of sex chromosome and autosomal diversity (Ellegren 2009). Finally, both positive and purifying selection have the effect of reducing nucleotide diversity at linked sites, and the strength of this effect is dependent of the recombination rate. Given that X chromosome do not recombine in males, with the exception of the pseudo-autosomal regions, the effect of selection becomes more pronounced because of linkage disequilibrium of adjacent loci. Therefore, X chromosome is expected to show reduced variation owing to a stronger role of selection at linked sites. The fact that the X chromosome is hemizygous in males means that recessive mutations will be exposed directly to selection, leading to a more frequent hitchhiking on the X chromosome (Vicoso & Charlesworth 2006).

CHAPTER 2

OBJECTIVES

The broad objective of this thesis was to characterize patterns of genetic variation in the Iberian pig genome in terms of SNPs and structural variants, as well as to characterize the pig transcriptome in terms of transcript composition and gene expression using parallel massive sequencing technologies (NGS). To do so, we used different bioinformatic tools and methodological approaches.
More specifically, the objectives were:

1. To characterize the nucleotide diversity of a putative causative gene for meat quality (*SERPINA6*) in different pig breeds and to ascertain whether there is evidence of a selective sweep (Chapter 3).

2. To study the genome-wide patterns of nucleotide diversity in the Iberian breed strain, combining different methodological approaches: sequencing a reduced representation library, whole genome shotgun sequencing of a highly inbred strain (Guadyerbas) and sequencing a pool of individuals (Chapters 4 and 5).

3. To characterize in detail the Iberian pig gonad transcriptome using RNA-seq, and to compare it with that from Large White breed (Chapter 6).

CHAPTER 3

NUCLEOTIDE VARIABILITY OF PORCINE *SERPINA6* GENE AND THE ORIGIN OF A PUTATIVE CAUSAL MUTATION ASSOCIATED WITH MEAT QUALITY

Anna Esteve Codina
Ana Ojeda
Lusheng Huang
Josep Maria Folch
Miguel Pérez Enciso

Animal Genetics 2010

Summary

Serpin peptidase inhibitor, clade A, member 6 (*SERPINA6*), also known as *corticosteroid binding globulin* or *CBG*, is involved in obesity and stress sensitivity. Previous studies have reported putatively causal mutations within that gene in the porcine species. In order to characterize a hypothetical selective footprint, we have resequenced ~ 6 kb of coding and non coding fragments in 20 pigs comprising domestic breeds and wild boars from Asia and Europe. Nucleotide variability was found to be far greater within Asian pig breeds than Europe (π = 1% vs. 0.05%, respectively), which is consistent with the pig evolution history. The putatively causal amino acid substitution Gly307Arg (SNP c.919G>A) associated with meat quality (drip loss) was only detected in European domestic pig breeds, suggesting a very recent mutation that appeared after domestication in Europe. No support for positive selection was detected, as no reduction in levels of diversity surrounding the mutation was found in lean breeds with respect to wild boar.

Keywords: domestication, *SERPINA6*, nucleotide variability, candidate gene, pig

Introduction

Understanding the forces that shape patterns of DNA variability is a major goal of animal population genetics. The ascertainment of these patterns is not only of academic interest, they are needed, e.g., to design optimum association studies for fine mapping or to predict the consequences of ongoing and future genomic selection schemes. However, and despite recent advances, this knowledge is to date relatively scarce except in non livestock species such as the dog. Generally speaking, livestock species must have undergone at least two bottlenecks (Bruford *et al.* 2003). The first one would correspond to the domestication process, *circa* 5 – 10 thousand years ago; the second major bottleneck occurred as a consequence of modern breeding and ensuing intense artificial selection. These two phenomena are very recent from an evolutive perspective, i.e., almost all extant DNA variability should predate domestication. Therefore, it is noticeable that the very few causal mutations that have been convincingly reported in the literature have appeared *after* domestication: *RYR1*, *IGF2* and *PRKAG3* in pigs (Fujii *et al.* 1991; Milan *et al.* 2000; Van Laere *et al.* 2003), Boorola in sheep (Mulsant *et al.* 2001) or myostatin in cattle (Grobet *et al.* 1997). This illustrates how effective artificial selection can be to increase the frequency of a favored allele. The resulting selection footprint must have been strong and clearly detectable, especially when compared to the wild ancestor or to local 'unimproved' breeds. We have indeed found such a pattern around the *IGF2* mutation in the pig (Ojeda *et al.* 2008b).

Among all phenotypic changes brought about by artificial selection in the pig, an increase in leanness is probably one of the most dramatic modifications. Among candidate genes in SSC7, a chromosome that has been consistently associated with large effect QTL, the corticosteroid binding globulin or *SERPINA6* has been studied in detail (Ousova *et al.* 2004; Geverink *et al.* 2006; Guyonnet-Duperat *et al.* 2006). This gene is a key regulator of cortisol levels and is likely associated with obesity susceptibility. It belongs to the serine protease inhibitors family (van Gent *et al.* 2003). In previous reports, Ousova *et al* (2004) and Guyonnat-Dupérat *et al*. (2006) postulated *SERPINA6* as an important positional candidate for obesity in the pig and described a non synonymous amino acid substitution (Gly307Arg), corresponding to the SNP c.919G>A in exon 4, which was associated with meat quality (drip loss) and cortisol binding capacity in Meishan x Large White intercross population. Here, we characterize the nucleotide variability of *SERPINA6* in a sample of pigs and wild boar from Asia and Europe, in order to identify putative signals of directional selection.

Materials and methods

Pig samples

The *SERPINA6* gene was partially resequenced in twenty pigs: three Duroc (DUES0304, DUUS0602, DUES0998) from Spain and USA, one Finnish Large White (LWFI0343), one Landrace (LRES0520) from Spain, one Tamworth from UK (TWGB0372), one British Lop from UK (BLGB0373), one Iberian from Guadyerbas strain (IBES0415), one unknown pig breed from Cabo Verde (NACV0908), one Vietnamese potbelly (VTES0104), one Minzhu from China (MICN0530), one Meishan from USA (MSUS0620), one Jianxin Black from China (JBCN0688), one Jinhua (JHCN0688) from China, one Huai (HUCN0692) from China, four western wild boars (WBES0007, WBIT0761, WBIT0781, WBTN0966) from Spain, Italy and Tunisia, and one from China (WBCN0698). As outgroup, we employed a Babyrousa from Madrid's zoo (BBES0280).

Sequencing

All serpina genes consist of five exons and porcine *SERPINA6* spans for 20 kb approximately (GenBank Accession: NC_010449, based on reference assembly 5). The regions resequenced were chosen after using Repeatmasker (http://www.repeatmasker.org) to avoid highly repetitive segments. Three regions were sequenced in six PCR reactions (Figure 1). The first region (975 bp) was amplified in one PCR and it covers the 5' upstream region of the gene, containing a cis-regulatory promoter, (Underhill & Hammond 1995; Zhao *et al.* 1997), exon 1 and part of intron 1. The second region (1021 bp) was amplified in one PCR and it covers part of intron 2, exon 3 and part of intron 3. The last region (3923 bp) was amplified in 4 overlapping PCRs and it spans part of intron 3, exon 4, intron 4, exon 5 and the 3' upstream region of the gene. Primers were designed using the porcine BAC of chromosome 7 (Sus_scrofa.Sscrofa2.43.dna.chromosome.7.fa.gz) available at http://www.sanger.ac.uk/Projects/S_scrofa/ . The specific coordinates of *SERPINA6* gene in assembly 9 are 123932180-123951199 (ftp://ftp.sanger.ac.uk/pub/S_scrofa/assemblies/Ensembl_Sscrofa9/). Primers and PCR conditions are in supplemental Table 1. The amplified products were sequenced using BigDye Terminator v3.1 Ready Reaction Cycle Sequencing Kit using ABI PRISM 3730 (Applied Biosystems).

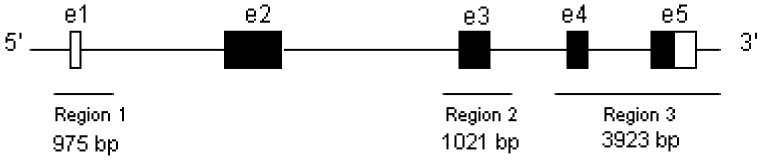

Figure 1 Scheme of the *SERPINA6* gene and the regions resequenced. 5'UTR and 3'UTR regions are in white.

Data analysis

Analysis of sequences and polymorphism identification were performed with SeqScape v2.5 (Applied Biosystems). The nucleotide diversity index (π), Tajima's D test, Fu and Li's test, differentiation statistics *Snn*, HKA test and McDonald-Kreitman test were obtained with DnaSP v5 (Rozas *et al.* 2003). To test for population structuring, we used the *Snn* test (i.e. nearest neighbour statistic) is a measure of how often the nearest neighbours of sequences are found in the same group (Hudson 2000). In principle, *Snn* has better properties than other tests, as it performs well at all levels of haplotype diversity. The McDonald-Kreitman test compares the amount of variation within a species to the divergence between species at two types of sites, one of which is putatively neutral and used as the reference (synonymous sites) to detect selection at the other types of sites (nonsynonymous sites). Under the null hypothesis, all nonsynonymous mutations are expected to be neutral and then the ratio of nonsynonymous to synonymous variation within species is expected to equal the ratio of nonsynonymous to synonymous variation between species. However, these ratios will not be equal if some nonsynonymous variation is under either positive or negative selection (McDonald & Kreitman 1991). Phase reconstruction was performed using Phase v2.1.1 (Li & Stephens 2003) with default options, except that the program ran five times and the last iteration was 10 times longer, as the authors recommend. We only retained phases with posterior probability larger than 0.90. NJ phylogenetic trees were performed with the Kimura model of two parameters using Mega4.1 (Kumar *et al.* 2004).

Results and discussion

Nucleotide variability

A total of 163 polymorphisms were detected in the 5919 bp region resequenced, i.e., one every 36 bp. These comprised 146 SNP (including 17 singletons) and 17 indels. Two SNPs were triallelic, although the third allele was found only in the outgroup. No polymorphisms were found at the 5' upstream region of the gene, 20 were localized at exons, 135 within introns and eight at the 3' downstream region of the gene. Five out of 20 coding SNP caused a non synonymous amino acid substitution (supplementary Table 2). The SNP c.622C>T and c.832G>A, not yet reported, correspond to the amino acid substitutions His208Tyr and Gly278Arg, respectively, whereas the SNP c.770C>T, c.793G>A and c.919G>A were previously described by Guyonnat-Dupérat *et al*. 2006 and they cause the aminoacid changes Thr257Met, Val265Ile and Gly307Arg.

(The last two mutations were annotated as Ile265Val and Arg307Gly by Guyonnat-Dupérat *et al*. 2006; in this study, the notation was changed to Val265Ile and Gly307Arg following the IUPAC nomenclature recommendations in order to distinguish the ancestral and derived alleles). Interestingly, among the samples tested in this study, only domestic pigs with a European origin carried the putative causal mutation (c.919G>A) associated with a higher cortisol binding capacity, low cortisol binding affinity and drip loss (Guyonnet-Duperat *et al.* 2006). Moreover, the SNP c.793G>A related with a decrease of *SERPINA6* affinity was observed only in European wild boars and in European domestic pigs, suggesting that this mutation appeared before European pig domestication but after the Asian – European lineages split. Nevertheless, the presence of these alleles in other populations cannot be ruled out given the limited number of individuals sequenced.

In addition to the high number of SNP, 17 structural variants were also detected. Most of these variants are large indels and long homopolymer fragments located within intronic sequences, which complicated the sequence analysis when the sample was heterozygous. New sequencing primers were designed to deal with these situations. The LRES0520 sample, a Landrace carrying both Asian and European haplotypes, was the most heterozygous animal both for SNP and indels, an illustration of the Asian germplasm introgression to western breeds. The MICN0530 sample, a Minzhu from China, was also highly heterozygous, carrying specific indels found only in this breed. Interestingly, a 37 nucleotide long indel was found within intron 2, which in turn carries two SNP within the insertion.

But interpreting the high number of SNP for *SERPINA6* can be misleading because the variability is primarily found in the Asian populations (Table 1). Nucleotide diversities were 0.96 vs. 0.055 % for Asian and European pigs, respectively, when the highly heterozygous Landrace animal is excluded. Interestingly, but coherent with previous results (Ojeda *et al.* 2006; Ojeda *et al.* 2008a; Ojeda *et al.* 2008b), the European wild boars are extremely uniform (π = 0.036%). Variability was also very dissimilar according to region. The first region, containing the 5'UTR, the first exon and a short part of intron 1, was almost devoid of variability even in Asia (Table 2) whereas the remaining two regions were considerably more diverse. Although intronic nucleotide diversity was higher than in exons, 0.91% and 0.48%, respectively; both values are higher than reported in the literature (Amaral *et al.* 2008). In contrast, no polymorphism was found in the 5'UTR and *π* was only 0.18% at the 3'UTR.

Table 1 Nucleotide variability, Tajima's D and FuLiD statistics per population

Population	π (%)	Tajima's D	FuLi's D
Asian animals	0.96	1.10	1.78**
European animals	0.14 (with LR) 0.055 (without LR)	-2.27** -1.05	-2.65* -1.7
European wild boars	0.036	-1.04	-0.75
European domestic pigs	0.17(with LR) 0.05 (without LR)	-2.11* -0.88	-2.02* -1.3

*: 0.01<P<0.05, **: 0.001<P<0.01

LR: sample LRES0520

Table 2 Nucleotide variability, Tajima's D and Fu-Li's D statistics per region

Domain	Length (bp)	π (%)	TajimaD	FuLiD
Region 1	975	0.06	-1.22	0.19
Region 2	1021	1.3	0.68	0.87
Region 3	3923	0.7	-0.05	1.14
5'upstream	428	0	n.d.	n.d.
3'downstream	742	0.18	-0.82	0.63
Exons	933	0.48	-0.33	0.68
Introns	3816	0.91	0.25	1.11
Total	5919	0.68	0.11	1.08

Population and haplotype structure

In the light of the highly unbalanced distribution of polymorphism between Asia and Europe, it is not unexpected that we found significant values of *Snn* statistics (Table 3). It is worth noticing that the differentiation between wild and domestic pigs was much smaller than between Asia and Europe, which is in agreement with mtDNA and microsatellite data (Scandura *et al.* 2008; Ramirez *et al.* 2009). This observation might be explained by either gene flow or a very recent split between populations. Although contributions from wild boar into European pigs appear to have occured in early European farming during prehistory (Larson *et al.* 2007), cytogenetic studies have shown little evidence of recent wild boar introgression into domestic pigs or vice versa (Ducos *et al.* 2008), although in the earlier stages of domestication introgression of wild boar seems to have been more common (Larson *et al.* 2007). The latter hypothesis may be a more plausible explanation, as domestication occurred much later than the European / Asia divergence, ca. 9000 years ago (Bökönyi *et al.* 1974) versus over 100,000 years (Giuffra *et al.* 2000; Kijas & Andersson 2001; Fang & Andersson 2006).

Table 3 Snn tests were performed comparing Asian vs European populations and European domestic pigs vs. European wild boars

Population	Region1	Region2	Region3	Total
Asian vs European	0.5	1***	1***	0.96***
European domestic vs. European wild	0.5	0.6	0.84***	0.96***

***: P< 0.001

Using reconstructed phases, NJ trees were drawn to gain a visual appraisal of how the different haplotypes are arranged in regions 2 and 3 (Figure 2). Both trees were rather similar, and display a profound Asia / Europe split; the higher Asian variability is clearly apparent, although pigs from the same breed tend to cluster together. Therefore, the within Asian breed variability is not necessarily high, although more data is needed to confirm this (Megens *et al.* 2008). Although this conclusion is tentative, it would be supported by large scale genotyping using a 60k SNP in the porcine hapmap population (Groenen *et al.* 2010). Note that the Landrace individual (LRES0520) is a hybrid made up of an European and an Asian haplotype explained by the introgression of Asian pigs to Europa. Therefore, including this animal has a considerable impact on European domestic pig variability. As for the two non synonymous SNP described previously (Guyonnet-Duperat *et al* 2006) c.793G>A (Region 2) and c.919G>A (Region 3), both are of European origin, although the former predates domestication and the latter was found only in domestic pigs, although its presence in European wild boar cannot be ruled out given the few animals sequenced. This second mutation (c.919G>A), associated with cortisol levels

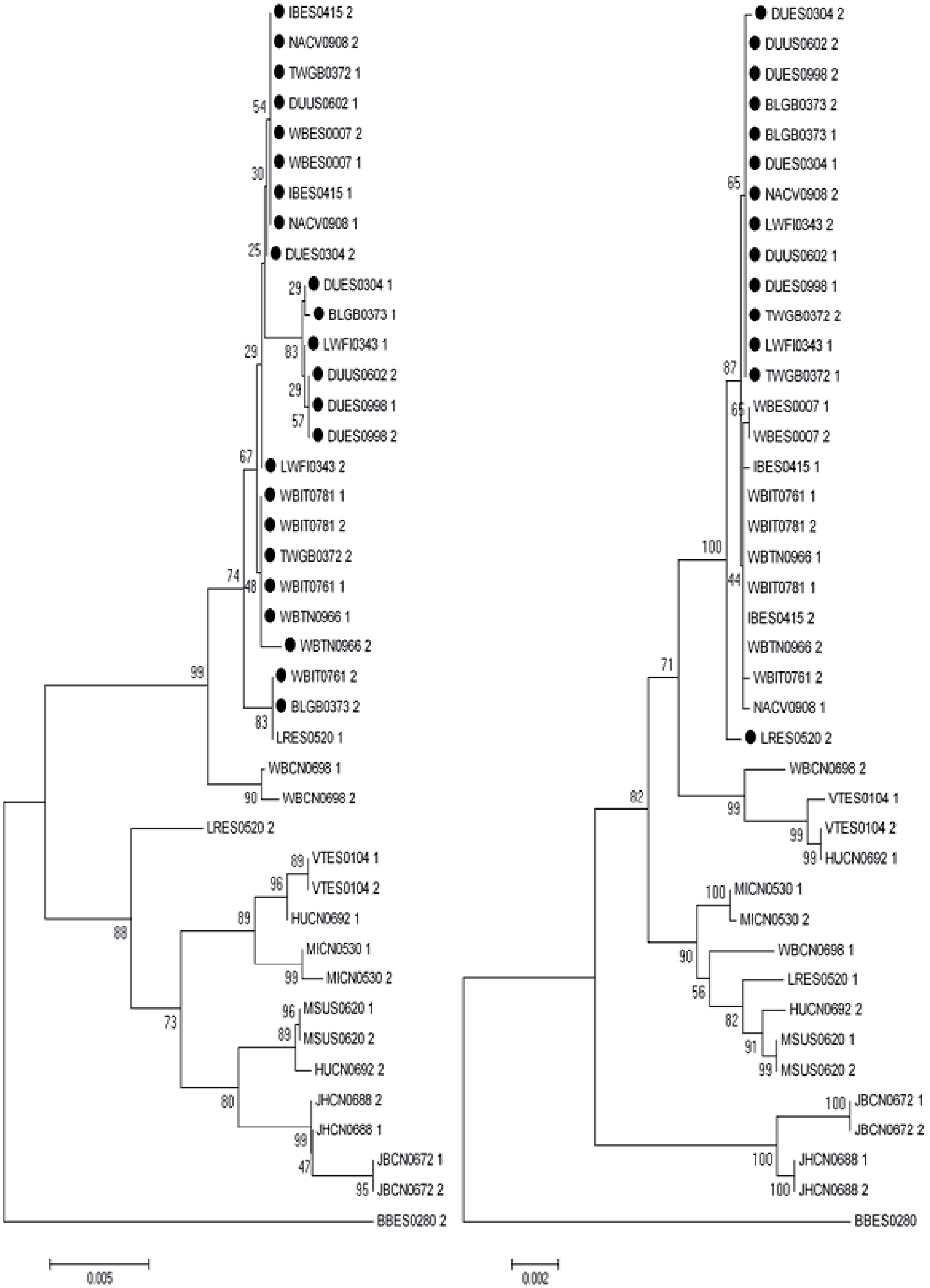

Figure 2a NJ tree of Region 2. Haplotypes carrying the nonsynonymous aminoacid change Val265Ile (SNP c.793G>A) are marked with a solid circle. **Figure 2b** NJ tree for Region 3. Haplotypes carrying the nonsynonimous aminoacid change Gly307Arg (SNP c.919G>A) are marked with a solid circle. Note that one of the LRES0520 haplotypes clusters with Asian samples.

and meat quality, seems to be at high frequency in modern domestic pigs; interestingly, the only Iberian pig sequenced (IBES0415) was homozygous for the wild type allele, whereas a Cape Verde animal of unknown breed (NACV0908) was heterozygous. Finally, note that wild boar and European domestic pigs cluster tightly, particularly for region 3.

Evidence for selection

Is there any evidence for a selection footprint on porcine *SERPINA6*? Although selection operates on the genome in many ways, animal breeding practitioners are predominantly interested in positive selection, i.e., in finding causal mutations with a beneficial effect on target selection traits. These mutations are therefore expected to be at high frequencies in animals selected for these traits as a result of a selective sweep. A distinctive selection footprint is therefore a region with low variability surrounding the causal mutation. This pattern results typically in negative Tajima's (Tajima 1989) and Fu – Li's Ds (Fu & Li 1993). In the gene analyzed here, for the whole region, these tests were only negative for the European pigs (Table 1). However, this value is primarily caused by including the highly divergent Asian haplotype of the Landrace pig. Otherwise, the tests were not significant. As for region 3, where SNP c.919G>A is located, the presence of the derived allele in the international breeds (Large White, Duroc and Landrace), but not in Iberian, is compatible with a mutation selected for in lean breeds. However, there is no evident decrease in variability, simply because the nucleotide variability was already very low, as seen by the high similarity of wild boar and domestic pig haplotypes (Figure 2).

To study whether the differences can be explained by variation under the neutral model, we used the HKA test (Hudson *et al.* 1987). We compared different regions of the resequenced gene (Table 4): Region 1, Region 2, Region 3, exons, introns and 3' downstream region of the gene. The only cases that can not be explained by the neutral model are those involving Region 1 vs. the rest: Region 1 versus Region 2 ($\chi^2 = 4.22$, P = 0.04), Region 1 versus Region 3 ($\chi^2 = 4.32$, P = 0.04) and Region 1 versus introns ($\chi^2 = 4.24$, P=0.04). Moreover, the maximum significance was found when comparing the promoter vs. the coding regions (P<0.01). It should be recalled that the HKA test assumes unlinked regions; when they are linked, as here, the test becomes over conservative. A significant test in the case of linkage becomes more credible than when unlinked regions are tested. Although all evolutionary tests have caveats, it is also relevant to mention that the identity percentage in the promoter between pig vs human and pig vs rat was 77% and 75%, respectively. In contrast, the identities between proteins were 65% and 56% between pig and human, and pig and rat. It seems plausible therefore that the promoter has been more evolutionarily constrained than the aminoacid sequence.

Table 4. Results of the HKA test

Region	χ^2 (P_{value})
Region 1 vs Region 2	4.22 (0.04)*
Region 1 vs Region 3	4.32 (0.04)*
Region 2 vs Region 3	0.06 (0.81)
Promoter vs. Coding region	6.83 (<0.01)**
Exons vs. Introns	1.82 (0.18)
3'downstream vs. Introns	0.62 (0.62)

*: 0.01<P<0.05, **:0.001<P<0.01

The comparison between the number of synonymous and nonsynonymous mutations can suggest whether, at the molecular level, natural selection is acting to promote the fixation of advantageous mutations (positive selection) or to remove deleterious mutations (purifying selection). We found 8 synonymous mutations and 5 nonsynonymous mutations; the corresponding π_a/π_s ratio was 0.16. A ratio lower than 1 is in agreement with the fact that most protein-coding genes are considered to be under the effect of purifying selection. Indeed, the majority of observed mutations are synonymous and do not affect the integrity of the encoded proteins. As a result, the number of synonymous mutations generally exceeds the number of nonsynonymous mutations. A MacDonald and Kreitman test was performed between the *Sus scrofa* samples and Babyrousa and was no significant.

Conclusion

Porcine *SERPINA6* is a highly polymorphic gene in the porcine species, except for exon 1 and the 5'upstream region. A majority of this variability is harbored by the Asian rather than the European populations (Table 1), which is agreement with other loci such as mtDNA (Larson et al. 2005; Fang and Andersson 2006). Occasionally, a highly heterozygous European animal like Landrace LRES0520 pig here has two highly divergent haplotypes as a result of Asian germplasm introgression. Also in agreement with an old divergence between Asian and European wild boars, a deep split is evident between Asian and European haplotypes (Figure 2).

Prior work (Ousova *et al.* 2004; Guyonnet-Duperat *et al.* 2006) suggested that *SERPINA6* contained causal mutations for obesity and growth related traits, specifically the SNP c.919G>A in exon 4. Our work aimed at detecting a selection footprint, i.e., a selective sweep signature. We found that the derived allele, which results in the Arg codon, is present only in European haplotypes and seemingly at high frequency. We did not find the mutation neither in Iberian nor in wild boar, although we cannot rule out that its presence in these populations because only one Iberian and a few wild boars were sequenced. None of the classical tests (Tajima's D or Fu-Li's D) were significant and overall there is not a clear evidence of a selective sweep. It is possible that the strong associated effect reported by Guyonnet-Dupérat et al. could be due to stratification because the association was reported in a synthetic Sino European line. As the Asian animals have the ancestral allele and are obese whereas the European animals have the derived allele and are lean, it is possible that the significant effect is spuriously caused by strong linkage disequilibrium between a nearby mutation(s) and the mixed background. Our data provide support instead for a constraint on proximal 5' regulatory motifs in *SERPINA6* (Table 4), a constraint that appears also when comparing sequences between pig, human and mouse.

Acknowledgements

We are grateful to all persons and institutions that generously provided samples, in particular to L. Silió and M.C. Valdovinos from INIA, Madrid, and to Madrid's zoo. AE and AO are recipients of FPI PhD fellowships from *Ministerio de Ciencia e Innovación*

(MICINN), Spain; Work funded by grants AGL2007-65563-C02-01/GAN (MEC, Spain) to MPE and the China national 973 project 2006CB102100 to LSH.

Bibliography

Amaral A.J., Megens H.J., Crooijmans R.P., Heuven H.C. & Groenen M.A. (2008) Linkage disequilibrium decay and haplotype block structure in the pig. *Genetics* **179**, 569-79.

Bruford M.W., Bradley D.G. & Luikart G. (2003) DNA markers reveal the complexity of livestock domestication. *Nat Rev Genet* **4**, 900-10.

Ducos A., Revay T., Kovacs A., Hidas A., Pinton A., Bonnet-Garnier A., Molteni L., Slota E., Switonski M., Arruga M.V., van Haeringen W.A., Nicolae I., Chaves R., Guedes-Pinto H., Andersson M. & Iannuzzi L. (2008) Cytogenetic screening of livestock populations in Europe: an overview. *Cytogenet Genome Res* **120**, 26-41.

Fang M. & Andersson L. (2006) Mitochondrial diversity in European and Chinese pigs is consistent with population expansions that occurred prior to domestication. *Proc Biol Sci* **273**, 1803-10.

Fu Y.X. & Li W.H. (1993) Statistical tests of neutrality of mutations. *Genetics* **133**, 693-709.

Fujii J., Otsu K., Zorzato F., de Leon S., Khanna V.K., Weiler J.E., O'Brien P.J. & MacLennan D.H. (1991) Identification of a mutation in porcine ryanodine receptor associated with malignant hyperthermia. *Science* **253**, 448-51.

Geverink N.A., Foury A., Plastow G.S., Gil M., Gispert M., Hortos M., Font i Furnols M., Gort G., Moisan M.P. & Mormede P. (2006) Cortisol-binding globulin and meat quality in five European lines of pigs. *J Anim Sci* **84**, 204-11.

Giuffra E., Kijas J.M., Amarger V., Carlborg O., Jeon J.T. & Andersson L. (2000) The origin of the domestic pig: independent domestication and subsequent introgression. *Genetics* **154**, 1785-91.

Grobet L., Martin L.J., Poncelet D., Pirottin D., Brouwers B., Riquet J., Schoeberlein A., Dunner S., Menissier F., Massabanda J., Fries R., Hanset R. & Georges M. (1997) A deletion in the bovine myostatin gene causes the double-muscled phenotype in cattle. *Nat Genet* **17**, 71-4.

Groenen M., Amaral A., Megens H.J., Larson G., Archibald A.L., Muir W., Malhi R., Crooijmans R.P., Ferretti L., Ramos-Onsins S.E., Perez-Enciso M. & Schook L. (2010) The Porcine HapMap Project: Genome-Wide Assessment Of Nucleotide Diversity, Haplotype Diversity And Footprints Of Selection In The Pig. In: *Plant & Anmal Genomes XVIII Conference*, San Diego.

Guyonnet-Duperat V., Geverink N., Plastow G.S., Evans G., Ousova O., Croisetiere C., Foury A., Richard E., Mormede P. & Moisan M.P. (2006) Functional implication of an Arg307Gly substitution in corticosteroid-binding globulin, a candidate gene for a quantitative trait locus associated with cortisol variability and obesity in pig. *Genetics* **173**, 2143-9.

Hudson R.R. (2000) A new statistic for detecting genetic differentiation. *Genetics* **155**, 2011-4.

Hudson R.R., Kreitman M. & Aguade M. (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**, 153-9.

Kijas J.M. & Andersson L. (2001) A phylogenetic study of the origin of the domestic pig estimated from the near-complete mtDNA genome. *J Mol Evol* **52**, 302-8.

Kumar S., Tamura K. & Nei M. (2004) MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* **5**, 150-63.

Larson G., Albarella U., Dobney K., Rowley-Conwy P., Schibler J., Tresset A., Vigne J.D., Edwards C.J., Schlumbaum A., Dinu A., Balacsescu A., Dolman G., Tagliacozzo A., Manaseryan N., Miracle P., Van Wijngaarden-Bakker L., Masseti M., Bradley D.G. & Cooper A. (2007) Ancient DNA, pig domestication, and the spread of the Neolithic into Europe. *Proc Natl Acad Sci U S A* **104**, 15276-81.

Li N. & Stephens M. (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213-33.

McDonald J.H. & Kreitman M. (1991) Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**, 652-4.

Megens H.J., Crooijmans R.P., San Cristobal M., Hui X., Li N. & Groenen M.A. (2008) Biodiversity of pig breeds from China and Europe estimated from pooled DNA samples: differences in microsatellite variation between two areas of domestication. *Genet Sel Evol* **40**, 103-28.

Milan D., Jeon J.T., Looft C., Amarger V., Robic A., Thelander M., Rogel-Gaillard C., Paul S., Iannuccelli N., Rask L., Ronne H., Lundstrom K., Reinsch N., Gellin J., Kalm E., Roy P.L., Chardon P. & Andersson L. (2000) A mutation in PRKAG3 associated with excess glycogen content in pig skeletal muscle. *Science* **288**, 1248-51.

Mulsant P., Lecerf F., Fabre S., Schibler L., Monget P., Lanneluc I., Pisselet C., Riquet J., Monniaux D., Callebaut I., Cribiu E., Thimonier J., Teyssier J., Bodin L., Cognie Y., Chitour N. & Elsen J.M. (2001) Mutation in bone morphogenetic protein receptor-IB is associated with increased ovulation rate in Booroola Merino ewes. *Proc Natl Acad Sci U S A* **98**, 5104-9.

Ojeda A., Estelle J., Folch J.M. & Perez-Enciso M. (2008a) Nucleotide variability and linkage disequilibrium patterns at the porcine FABP5 gene. *Anim Genet* **39**, 468-73.

Ojeda A., Huang L.S., Ren J., Angiolillo A., Cho I.C., Soto H., Lemus-Flores C., Makuza S.M., Folch J.M. & Perez-Enciso M. (2008b) Selection in the making: a worldwide survey of haplotypic diversity around a causative mutation in porcine IGF2. *Genetics* **178**, 1639-52.

Ojeda A., Rozas J., Folch J.M. & Perez-Enciso M. (2006) Unexpected high polymorphism at the FABP4 gene unveils a complex history for pig populations. *Genetics* **174**, 2119-27.

Ousova O., Guyonnet-Duperat V., Iannuccelli N., Bidanel J.P., Milan D., Genet C., Llamas B., Yerle M., Gellin J., Chardon P., Emptoz-Bonneton A., Pugeat M., Mormede P. & Moisan M.P. (2004) Corticosteroid binding globulin: a new target for cortisol-driven obesity. *Mol Endocrinol* **18**, 1687-96.

Ramirez O., Ojeda A., Tomas A., Gallardo D., Huang L.S., Folch J.M., Clop A., Sanchez A., Badaoui B., Hanotte O., Galman-Omitogun O., Makuza S.M., Soto H., Cadillo J., Kelly L., Cho I.C., Yeghoyan S., Perez-Enciso M. & Amills M. (2009) Integrating Y-chromosome, mitochondrial, and autosomal data to analyze the origin of pig breeds. *Mol Biol Evol* **26**, 2061-72.

Rozas J., Sanchez-DelBarrio J.C., Messeguer X. & Rozas R. (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**, 2496-7.

Scandura M., Iacolina L., Crestanello B., Pecchioli E., Di Benedetto M.F., Russo V., Davoli R., Apollonio M. & Bertorelle G. (2008) Ancient vs. recent processes as factors shaping the genetic variation of the European wild boar: are the effects of the last glaciation still detectable? *Mol Ecol* **17**, 1745-62.

Tajima F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585-95.

Underhill D.A. & Hammond G.L. (1995) cis-regulatory elements within the proximal promoter of the rat gene encoding corticosteroid-binding globulin. *Gene* **162**, 205-11.

van Gent D., Sharp P., Morgan K. & Kalsheker N. (2003) Serpins: structure, function and molecular evolution. *Int J Biochem Cell Biol* **35**, 1536-47.

Van Laere A.S., Nguyen M., Braunschweig M., Nezer C., Collette C., Moreau L., Archibald A.L., Haley C.S., Buys N., Tally M., Andersson G., Georges M. & Andersson L. (2003) A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. *Nature* **425**, 832-6.

Zhao X.F., Underhill D.A. & Hammond G.L. (1997) Hepatic nuclear proteins that bind cis-regulatory elements in the proximal promoter of the rat corticosteroid-binding globulin gene. *Molecular and Cellular Endocrinology* **126**, 203-12.

CHAPTER 4

PARTIAL SHORT-READ SEQUENCING OF A HIGHLY INBRED IBERIAN PIG AND GENOMICS INFERENCE THEREOF

Anna Esteve Codina

Robert Kofler

Heinz Himmelbauer

Luca Ferretti

Ana Vivancos

Martien Groenen

Josep Maria Folch

Maria del Carmen Rodríguez

Miguel Pérez Enciso

Heredity 2011

Abstract

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, 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 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 over represented 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. Data accession: SRP005367.

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 (Metzker, 2010). 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, e.g., 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 & Jensen 2010). Here, we decided to sequence a RRL of a single individual rather than a pool in order 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 employed 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 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 to 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.

Material and Methods

Material

The *Guadyerbas* herd was founded with four boars and ten sows in 1945, and has been maintained with controlled pedigree and minimum co-ancestry mating practices in order 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 via 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 10U 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'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC*T -3' where * denotes a phosphorothioate bond and oligonucleotide B 5' P-

GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTG -3' (Sigma). A 5 x adapter mix in water with a final concentration of 20 ⍰M of each oligonucleotide was prepared in a thermocycler by heating to 65ºC for 5 min and cooling to 20ºC with a ramp of 0.1ºC/sec. 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, i.e. in a volume of 50 ul with 10,000 units of T4 DNA ligase (New England Biolabs) in 1 x 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), eluted in 30 µl of 1 x 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 x elution buffer (500 mM ammonium acetate, 0.1% SDS, 0.1 mM EDTA) for 2 hours at room temperature with gentle agitation. We separated the crushed polyacrylamide from the eluted DNA by using a cellulose acetate column (SpinX) 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 13200 rpm 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 x 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 million base called reads were obtained. Sequences have been deposited in sequence read archive (SRA accession SRP005367).

Bioinformatic analysis

Reads were trimmed to 40 bp due to low 3' end quality. We discarded reads containing *N*s, homopolymers longer than 17 nucleotides, an average minimum phred quality smaller than 20 and reads that did not start with a CC motif (*HaeIII* 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 & 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 (4 mismatches were allowed). In the MAQ assembly, we also allowed a maximum of 4 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 4 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 2.

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 (MCLE) for *h*. If the mating is random and the population is in Hardy-Weinberg equilibrium, this is also a MCLE for the variability θ of the population. In the absence of sequencing and mapping errors, the formula for the unbiased MCLE for *h* is:

$$\hat{h}^* = \frac{S}{\sum_{nr=1}^{\infty} L(nr)P^*(S|nr)} \quad (1)$$

where *S* is the number of heterozygous sites detected in the window, $L(n_r)$ is the number of bases with depth n_r in the window and $P^*(S|n_r)$ is the probability that a heterozygous site is detected when the read depth at that site is n_r. The analytical expression is $P^*(S|n_r) = 1 - 2^{-(nr-1)}$ (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 ≥ 20% among all reads covering that position. Therefore, we applied eq. (1) using those SNPs called by two of the three softwares and summing between *nr* = 3 and 20. Therefore, equation (1) needs to be modified:

$$\hat{h} = \frac{S}{\sum_{nr=3}^{20} L(nr)P(S|nr)}, \qquad (2)$$

where

$$P(S|nr) = 1 - 2^{-nr}\left[\sum_{k=0}^{na-1}\binom{nr}{k} + \sum_{k=nr-nb+1}^{nr}\binom{nr}{k}\right], \qquad (3)$$

with *na* = max(2, 0.2×*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, i.e., 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, $P^*(S|n_r=3)$ = 0.75 *vs.* $P(S|n_r=3)$ = 0.375, whereas $P^*(S|n_r=10)$ = 0.998 *vs.* $P(S|n_r=10)$ = 0.988.

As is clear from eq. (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

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

and,

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

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 (i.e., window) *i* is:

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

and the multilocus HKA test is $\chi^2 = \sum_{\mathrm{i}} H_i$, 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

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

and

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

with *T*, the divergence time, given by

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

with approximate variances Var(•) = *E(•)* × [1 + *E(•)*]. 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, i.e., 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 theta estimator described above ($\hat{\theta}_i$) for each i-th window; as measure of goodness of fit we used the Wilcoxon 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 θ = 0.0013 for the autosomes and θ = 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.* 2009b). The whole procedure was implemented in a Perl script with calls to MaCS and R.

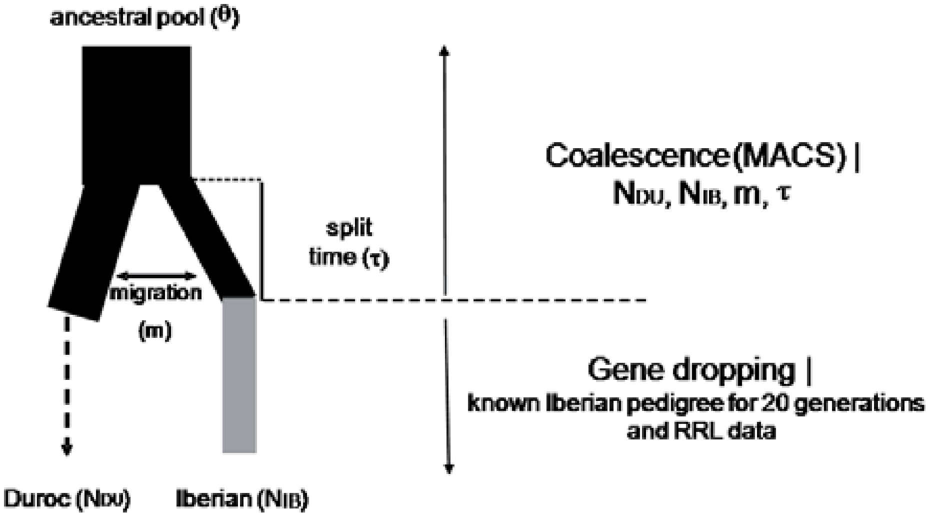
ancestral pool (θ)
split time (τ)
migration (m)
Coalescence(MACS) | NDU, NIB, m, τ
Gene dropping | known Iberian pedigree for 20 generations and RRL data
Duroc (NDU)
Iberian (NIB)

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 employed.

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 (www.biomart.org). Our Goslim (http://www.geneontology.org/GO.slims.shtml) was composed of twenty three 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. Gene ontologies statistics were calculated using the GOquick browser (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, i.e., reads matching the reference more than once, we retained five million reads for further analysis (Figure 2).

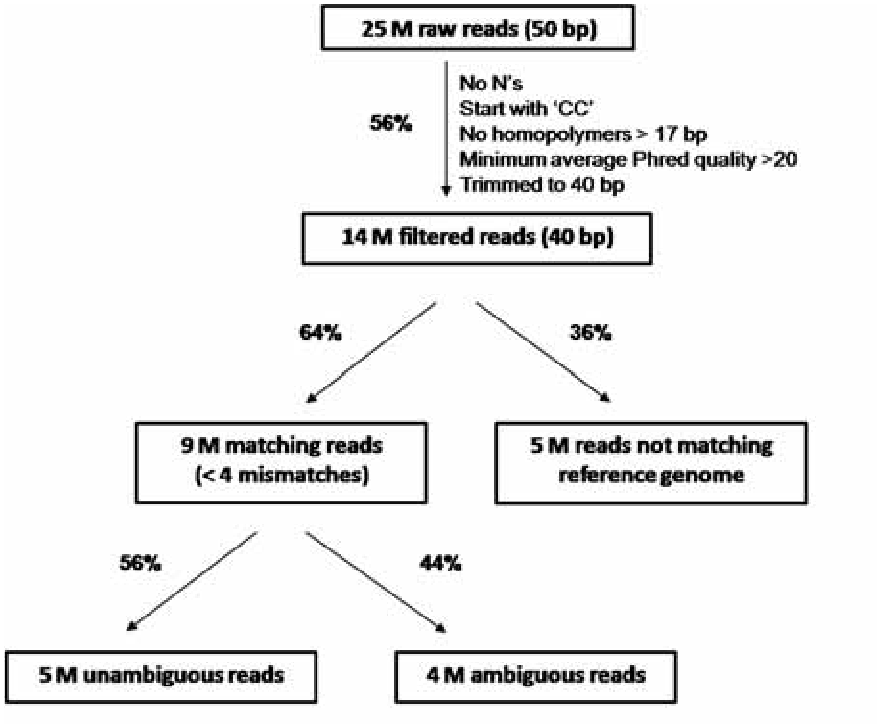

Figure 2. Bioinformatics pipeline.

The total length assembled was approximately 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.

Table 1: Statistics per chromosome

Chrom.	Total assembled ≥3× (Mb)	Average coverage (3-20×)	S[1]	F[2]	$\hat{h}$[3]	$\hat{f}$[4]
SSC1	2.45	3.97	842	5,023	0.48	1.51
SSC2	1.95	4.00	1,334	4,363	1.11	1.88
SSC3	1.90	4.01	1,517	3,519	1.15	1.48
SSC4	1.34	3.98	971	3,027	0.95	1.73
SSC5	1.05	3.96	639	2,685	1.09	2.11
SSC6	2.21	4.01	895	4,890	0.58	1.85
SSC7	1.68	3.97	471	4,509	0.60	2.49
SSC8	0.91	3.96	485	2,197	0.73	2.15
SSC9	1.33	3.96	823	3,142	0.77	1.85
SSC10	0.69	3.97	438	1,822	1.08	2.42
SSC11	0.66	3.95	406	1,770	0.95	2.44
SSC12	1.16	3.99	782	2,520	1.06	1.67
SSC13	1.40	3.96	614	2,618	0.63	1.58
SSC14	2.06	3.98	837	4,512	0.57	2.10
SSC15	1.05	4.01	653	2,607	0.67	1.77
SSC16	0.67	3.97	419	1,636	0.81	2.27
SSC17	0.87	3.99	528	1,960	0.99	2.38
SSC18	0.66	3.96	370	1,396	1.10	1.57
SSCX	1.03	4.01	297	1,261	0.37	0.92
Total autosomes	24.02	3.98	13.024	54.196	0.78	1.89

[1] Number of heterozygous sites, raw numbers
[2] Number of fixed differences, raw numbers
[3] Average estimated heterozygosity within Iberian per kb
[4] 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 / 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 4 and 5, 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, *i.e.*, 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.

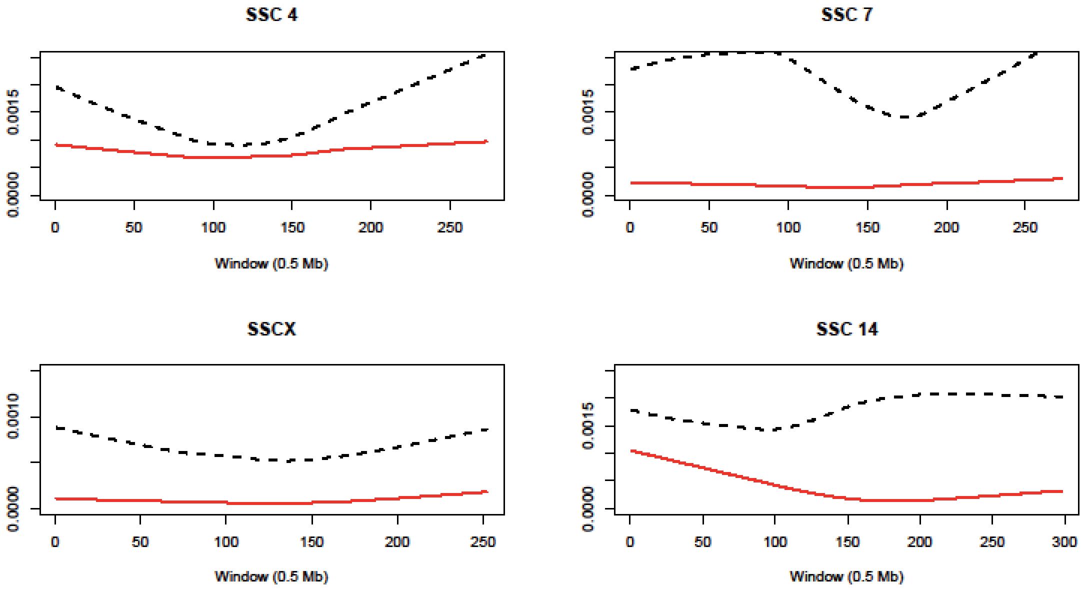

Figure 3: Lowess adjusted curves of variability in chromosomes 4, 7, 14 and X. An increased variability is observed towards the telomeres in metacentric chromosomes 4 and X, whereas the ratio is distorted in SSC7 because of high SLA variability near window 50; SSC14 is acrocentric. Solid red line, Iberian heterozygosity ($\hat{h}$); dashed black line, Iberian – Duroc heterozygosity ($\hat{f}$). Position refers to window number.

The average SNP rates per base pair for chromosome X were $\hat{f}$ = 9.2 × 10^{-4} and $\hat{h}$ = 3.7 × 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 ¾ 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, genomewide, there were no wide departures from neutrality, neither for autosomes nor for X chromosome, according to this test.

Table 2: HKA statistics

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

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

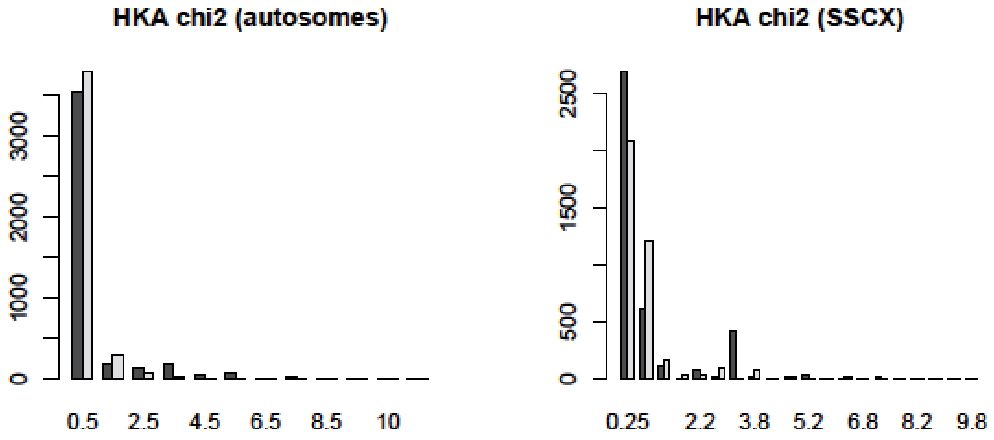

Figure 4. Histograms comparing observed (black bars) and simulated (grey bars) HKA statistics across autosomal and sex chromosome windows. The simulated results correspond to parameter values that minimized the Wilcoxon statistics.

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 over representation 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 $\widehat{f}$ was below the median. Therefore we ensured that, whereas $\hat{h}$ was extreme, $\widehat{f}$ 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 over-represented gene ontologies 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 over represented 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 over represented. The results simply suggest that these genes exhibit lower than expected variability, be it because of specific selection in livestock or because other biological constraints. More data is required to ascertain the precise cause.

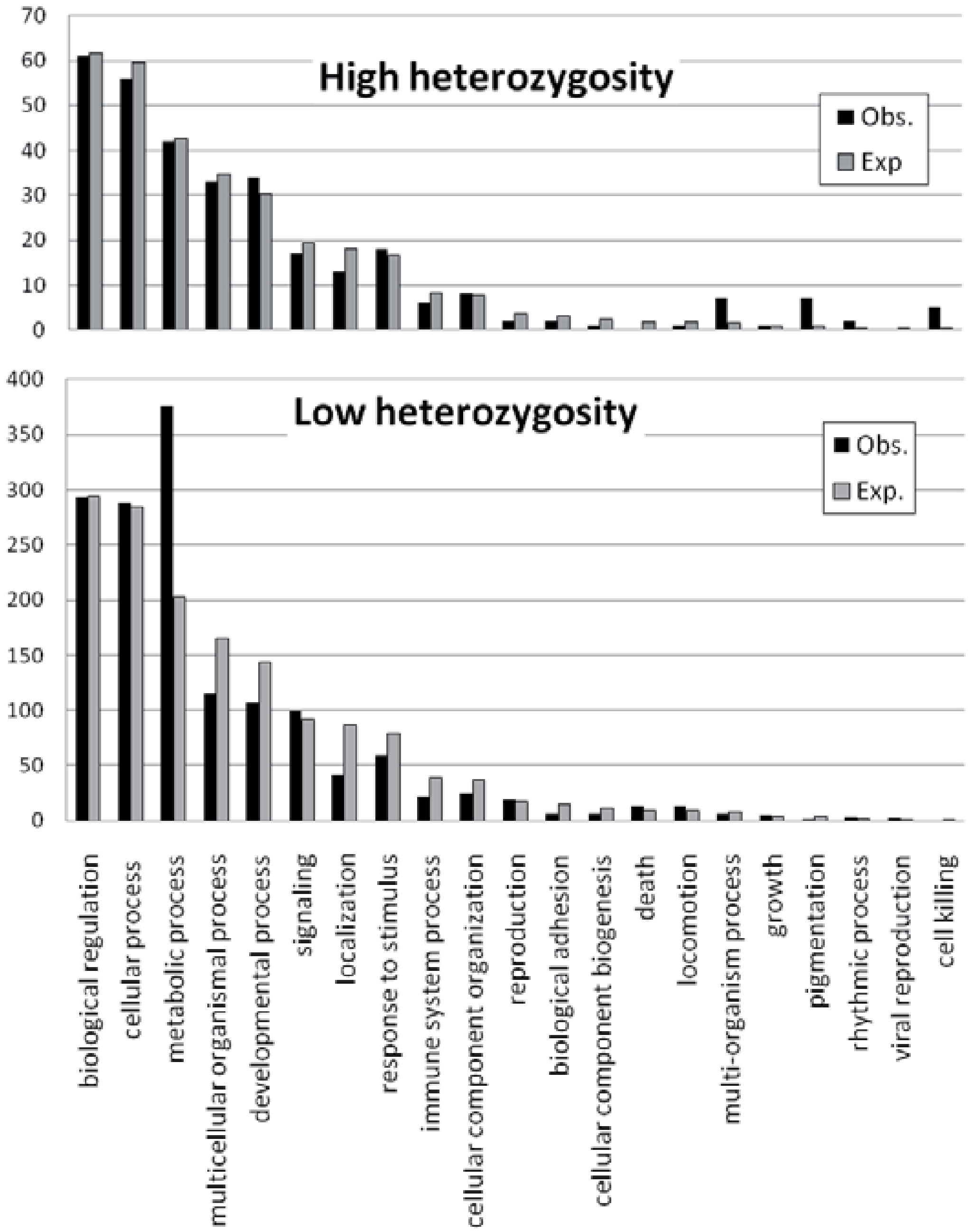

Figure 5: Expected and observed gene ontology counts among genes located in high and low variability windows. Bars with asterisk * are significant (P < 0.001) overrepresented gene ontologies.

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.* 2009b) 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 × 10^{-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 $N_e = \hat{h} / 4\mu \sim 2\times10^4$. 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 approximately. When correcting for inbreeding, this diversity is comparable to that reported in other porcine species (Amaral *et al.* 2009a; Amaral *et al.* 2009b) 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 × 10^{-4}, the average number of differences was, nonetheless, close to the genome wide mean ($\hat{f}$ = 1.7 × 10^{-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 ~ 1 Mb, the expected size of an identical by descent fragment (IBD) is ~ 2.6 Mb (SD, 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 ~ 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 to 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 hitchhiking 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 ¾ times that in autosomes, but we find a much lower value π_{SSCX} / π_{SSCA} ~ 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.* 2009b). 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 ¾ 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 & 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 ~ 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 founded is approximately (1 - 0.46) / (1 - 0.39) ~ 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 prior to the herd foundation ca. mid 20th 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, genome-wide, 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 due to 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 ~ 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 due to 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, e.g., 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".

Supplementary Files

(web link: www.nature.com/hdy/journal/v107/n3/suppinfo/hdy201113s1.html)

Supplementary File S1: Genome wide read depth obtained from Mosaik alignment; each chromosome in a different color.

Supplementary File S2: Variability by category and by chromosome; detailed statistics by window and list of extreme windows.

Supplementary File S3: Genome-wide Iberian heterozygosity $\hat{h}$ (top) and Iberian – Duroc heterozygosity $\hat{f}$ (bottom); each chromosome in a different color. The highest $\hat{f}$ window on SSC7 corresponds to SLA complex.

Supplementary File S4: Observed (solid line) vs. expected (dashed) P-values from the Chi-squared HKA test (eq. 6) for autosomes and X-chromosome.

References

Amaral A., Ferretti L., Megens H.-J., Crooijmans R., Nie H., Ramos-Onsins S.E., Perez-Enciso M., Schook L. & Groenen M. (2009a) Finding selection footprints in the swine genome using massive parallel sequencing In: *Conference on Next Generation Sequencing: Challenges and Opportunities*, Barcelona.

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

Cutler D.J. & Jensen J.D. (2010) To Pool, or Not to Pool? *Genetics* **186**, 41-3.

Chen G.K., Marjoram P. & Wall J.D. (2009) Fast and flexible simulation of DNA sequence data. *Genome Res* **19**, 136-42.

Dohm J.C., 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 R.M., Abecasis G.R., Altshuler D.L., Auton A., Brooks L.D., Gibbs R.A., Hurles M.E., McVean G.A. & Consortium G. (2010) A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73.

Ellegren H. (2009) The different levels of genetic diversity in sex chromosomes and autosomes. *Trends in Genetics* **25**, 278-84.

Haubold B., Pfaffelhuber P. & Lynch M. (2010) mlRho – a program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes. pp. 277-84. Blackwell Publishing Ltd.

Hedrick P.W. (2010) *Genetics of populations*. Jones and Bartlett, Sudbury, MA.

Hellmann I., Mang Y., Gu Z., Li P., de la Vega F.M., Clark A.G. & Nielsen R. (2008) Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. *Genome Res.* **18**, 1020-9.

Huang W. & Marth G. (2008) EagleView: A genome assembly viewer for next-generation sequencing technologies. *Genome Res*.

Hudson R.R. (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 U S A* **91**, 6815-8.

Hudson R.R., Kreitman M. & Aguade M. (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**, 153-9.

Jiang R., Tavare S. & Marjoram P. (2009) Population genetic inference from resequencing data. *Genetics* **181**, 187-97.

Kozarewa I., Ning Z., Quail M.A., Sanders M.J., Berriman M. & Turner D.J. (2009) Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat Meth* **6**, 291-5.

Li H., Ruan J. & Durbin R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research* **18**, 1851-8.

Lynch M. (2008) Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects. *Mol Biol Evol* **25**, 2409-19.

Ng S.B., Turner E.H., Robertson P.D., Flygare S.D., Bigham A.W., Lee C., Shaffer T., Wong M., Bhattacharjee A., Eichler E.E., Bamshad M., Nickerson D.A. & Shendure J. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272-6.

Noguera J.L., Rodriguez C., Varona L., Tomas A., Munoz G., Ramirez O., Barragan C., Arque M., Bidanel J.P., Amills M., Ovilo C. & Sanchez A. (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 J.M. & Perez-Enciso M. (2006) Unexpected High Polymorphism at the FABP4 Gene Unveils a Complex History for Pig Populations. *Genetics* **174**, 2119-27.

Ovilo C., Pérez-Enciso M., Barragan C., Clop A., Rodriguez C., Oliver M.A., Toro M.A. & Noguera J.L. (2000) A QTL for intramuscular fat and backfat thickness is located on porcine chromosome 6. *Mamm Genome* **11**, 344-6.

Pérez-Enciso M., Clop A., Noguera J.L., Ovilo C., Coll A., Folch J.M., Babot D., Estany J., Oliver M.A., Diaz I. & Sanchez A. (2000) A QTL on pig chromosome 4 affects fatty acid metabolism: evidence from an Iberian by Landrace intercross. *J Anim Sci* **78**, 2525-31.

Pool J.E., Hellmann I., Jensen J.D. & Nielsen R. (2010) Population genetic inference from genomic sequence variation. *Genome Res* **20**, 291-300.

Pool J.E. & Nielsen R. (2007) Population size changes reshape genomic patterns of diversity. *Evolution* **61**, 3001-6.

Quail M.A., Kozarewa I., Smith F., Scally A., Stephens P.J., Durbin R., Swerdlow H. & Turner D.J. (2008) A large genome center's improvements to the Illumina sequencing system. *Nat Methods* **5**, 1005-10.

Quinlan A.R., Stewart D.A., Stromberg M.P. & Marth G.T. (2008) Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat Methods* **5**, 179-81.

Ramos A.M., Crooijmans R.P., Affara N.A., Amaral A.J., Archibald A.L., Beever J.E., Bendixen C., Churcher C., Clark R., Dehais P., Hansen M.S., Hedegaard J., Hu Z.L., Kerstens H.H., Law A.S., Megens H.J., Milan D., Nonneman D.J., Rohrer G.A., Rothschild M.F., Smith T.P., Schnabel R.D., Van Tassell C.P., Taylor J.F., Wiedmann R.T., Schook L.B. & Groenen M.A. (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 C.J., Zody M.C., Eriksson J., Meadows J.R., Sherwood E., Webster M.T., Jiang L., Ingman M., Sharpe T., Ka S., Hallbook F., Besnier F., Carlborg O., Bed'hom B., Tixier-Boichard M., Jensen P., Siegel P., Lindblad-Toh K. & Andersson L. (2010) Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**, 587-91.

Sackton T.B., Kulathinal R.J., Bergman C.M., Quinlan A.R., Dopman E.B., Carneiro M., Marth G.T., Hartl D.L. & Clark A.G. (2009) Population genomic inferences from sparse high-throughput sequencing of two populations of Drosophila melanogaster. *Genome Biol Evol* **1**, 449-65.

Serra X., Gil F., Pérez-Enciso M., Oliver M.A., Vázquez J.M., Gispert M., Díaz I., Moreno F., Latorre R. & Noguera J.L. (1998) A comparison of carcass, meat quality and histochemical characteristics of Iberian and Landrace pigs. *Livest.Prod.Sci.* **56**, 215-23.

Toro M., Rodrigáñez J., Silió L. & Rodríguez M. (2000) Genealogical analysis of a closed herd of black hairless Iberian pigs. *Conservation Biol* **14**, 1843-51.

Van Tassell C.P., Smith T.P., Matukumalli L.K., Taylor J.F., Schnabel R.D., Lawley C.T., Haudenschild C.D., Moore S.S., Warren W.C. & Sonstegard T.S. (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat Methods* **5**, 247-52.

CHAPTER 5

Dissecting structural and nucleotide genome-wide variation in inbred Iberian pigs

Anna Esteve Codina

Yogesh Paudel

Luca Ferretti

Emanuele Raineri

Hendrik-Jan Megens

Luis Silió

Maria del Carmen Rodríguez

Martien Groenen

Sebastian Ramos Onsins

Miguel Pérez Enciso

(manuscript under development)

Abstract

Background

In contrast to international pig breeds, the Iberian breed has not been admixed with Asian germplasm. This makes it an important model to study both domestication and relevance of Asian genes in the pig. Besides, Iberian pigs exhibit high meat quality and appetite and propensity to obesity. Here we provide a genome wide analysis of nucleotide and structural diversity in a reduced representation library from a pool (n=9 sows) and shotgun genomic sequence from a single sow of the highly inbred Guadyerbas strain. In the pool, we applied newly developed tools to account for the peculiarities of these data.

Results

A total of 254,106 SNPs in the pool (79.6 Mb covered) and 643,783 in the Guadyerbas sow (1.47 Gb covered) were called. The nucleotide diversity (1.31×10^{-3} per bp in autosomes) is very similar to that reported in wild boar. A much lower than expected diversity in the X chromosome was confirmed (1.79×10^{-4} per bp in the individual and 5.83×10^{-4} per bp in the pool). A strong (0.70) correlation between recombination and variability was observed, but not with gene density or GC content. Multi copy regions affected about 4% of annotated pig genes in their entirety, and 2% of the genes partially. Genes within the lowest variability windows comprised interferon genes and, in chromosome X, genes involved in behavior like *HTR2C* or *MCEP2*. A modified Hudson-Kreitman-Aguadé test for pools also indicated an accelerated evolution in genes involved in behavior, as well as in spermatogenesis and in lipid metabolism.

Conclusions

This work illustrates how current sequencing technologies can picture a comprehensive landscape of variability in livestock species, and to pinpoint regions containing genes potentially under selection. Among genes that may have been subject to selection, we report genes involved in behavior, including feeding behavior, and lipid metabolism. The pig X chromosome is an outlier chromosome in terms of nucleotide diversity, which suggests selective constraints. Our data further confirm the importance of structural variation in the species, including Iberian pigs, and allows us to identify new paralogs from known gene families.

Keywords

Iberian pig, Next generation sequencing, Pig, Selection tests, Structural variation

Background

The pig is one of the most important sources of meat worldwide, as well as a relevant biomedical model for some diseases like metabolic syndrome or obesity. Current high throughput sequencing technologies, together with the recent completion of porcine's genome and its annotation (Groenen 2012), makes it possible to study the genomic variability of specific breeds with a detail that was not possible until now. Here, we present a thorough genome-wide analysis of the Iberian breed. Commercial pig breeds that are today exploited internationally, e.g., Landrace, Large White or Duroc, are the result of introgressing local primigenious European breeds with Asian germplasm, a process that is now well documented (Giuffra *et al.* 2000). In contrast, European wild boar, but also local Mediterranean breeds like the Iberian breed, was not affected by this admixture process. Given the high divergence between Asian and primigenious European pigs (ca. 1 MYA) (Groenen 2012) and the extent and intensity of modern selection methods, the study of Iberian pigs can illuminate both the domestication process and the influence of Asian germplasm in the shaping of current international pig breeds. Besides, Iberian pigs are important economically because of their high meat quality and resilience to endure harsh environmental conditions (Lopez-Bote 1998). They are very fat pigs, markedly different from modern lean pigs, and are interesting from a human biomedical perspective because they present high feed intake and tendency to obesity, compatible with high values of serum leptin (Fernandez-Figares *et al.* 2007).

Here, we carried out a genomic analysis of the Iberian breed using a mixed approach: a reduced representation library (RRL) sequencing of a pool of nine sows, and a shotgun complete genome sequencing of a highly inbred Iberian strain (Guadyerbas). The latter strain has been used in numerous QTL experiments and has been maintained in isolation for over 68 years and 25 generations. in a closed herd, *El Dehesón del Encinar*, located in Toledo, central Spain (Toro *et al.* 2000). In a previous work (Esteve-Codina *et al.* 2011), we reported a partial RRL sequencing of the same sow, 1 % of the genome approximately. The pool is made up of Iberian pigs from farms with strict pedigree control and that represent the extant diversity of Iberian varieties. The pool included as well the Guadyerbas sow that was individually sequenced.

Results

Nucleotide variability

Out of two paired-end (PE) lanes from a reduced representation library in the pool, about 3% of the current pig assembly v 10.2 was covered with depth between 3× and 30×. From one PE and one single end (SE) lanes in the Guadyerbas sow, ~ 60% of the genome was covered with depths 3× – 20×. Average depths were, respectively, 14× and 7× in the pool and in the individual.

A matter of concern in pools is the percentage of genetic variability that is actually discovered. To resolve this matter, we ran a simulation study mimicking as much as possible the pool process and the bioinformatics pipeline we used in the analyses of real data (see methods). Simulations suggested that we should detect ~ 47% of all SNPs actually segregating in the nine individuals - for the regions covered within at least 3-20× and with a low false discovery rate (0.02). Figure 1 shows expected results by minimum allele frequency (MAF) and depth. Note that most of SNPs missed are due to their low frequency: while 80% of SNPs at MAF < 10% are likely undetected, power for SNPs with MAF 0.3 is 60% and approaches one at higher MAFs. Importantly, the statistics used here to infer nucleotide variability were developed to account for the bias towards intermediate allele frequency in the pooling process (see methods). It should be noted as well that SNPs discovered in a single individual are also biased towards intermediate allele frequency SNPs, simply because the likelihood of a single individual being heterozygous for a rare allele is very low.

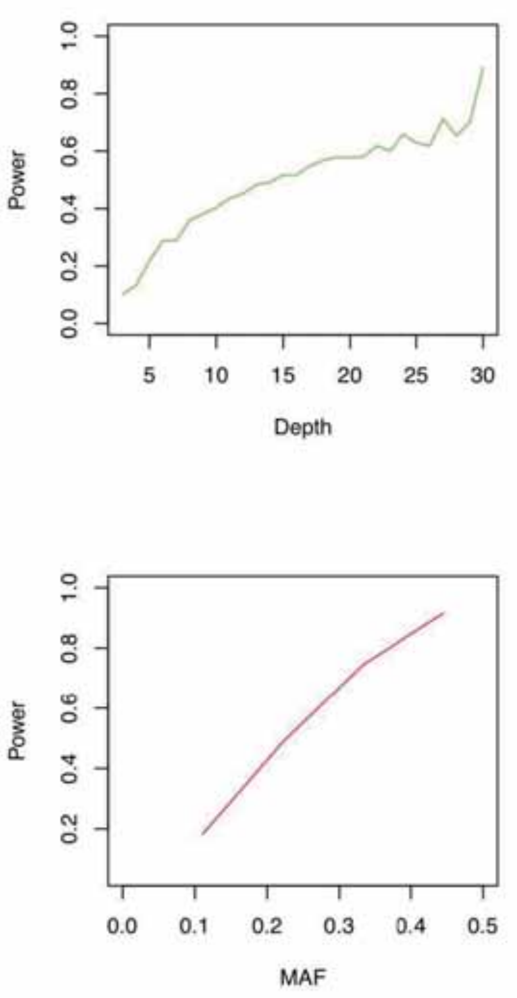

Figure 1. Top: Simulated power against depth. Power was computed as the number of SNP called by SNAPE software divided by the total number of real SNPs in the pool. Depth corresponds to the average depth in the pooled data. Bottom: Power against MAF (minimum allele frequency in the pool).

In all, the raw numbers of SNPs called (only segregating sites) were 254,106 in the pool (79.6 Mb covered) and 643,783 in the Guadyerbas sow (1.47 Gb covered). A total of 17.7 Mb of the current assembly was covered in both the pool and the individual, and 10,324 SNPs were called in both designs. The raw number of fixed differences between the assembly, primarily a Duroc female, and the Iberian pool was 152,225, and 2,503,645 for the Guadyerbas. We also detected 49,105 heterozygous indels and 316,189 fixed indels in the individual sow. We did not call indels in the pool because indel calling algorithms are not specific for pools and can be misleading. SNP annotation by autosomes, pseudoautosomal region (PAR) and non-pseudoautosomal region (NPAR) of X chromosome (SSCX) is detailed in Table 1. SNP classes are ranked in decreasing order of severity, according to ensembl pipeline (www.ensembl.org). Note that these raw numbers are not directly comparable between pool and individual because of different number of chromosomes, read depth and alignment lengths. A more meaningful observation, though, is the ratio between non – synonymous and synonymous mutations (dN/dS) within pool or individual. Interestingly, dN/dS was higher in the individual than in the pool (0.99 vs. 0.76), and this trend accentuates for chromosome X NPAR (1.21 vs. 0.82).

We computed Watterson's estimates of diversity, corrected for pooling and low depth (see methods). In general, there was a moderate correlation between both pool and individual variabilities (Pearson correlation = 0.45, Figure 2) when windows with no SNP in the Guadyerbas are removed. Nevertheless, it should be reminded that the Guadyerbas strain is highly inbred, e.g., we found that ~ 10% of the 200 kb windows were devoid of any SNP. Another factor of bias is that, while an RRL was sequenced in the pool (3% of the genome), the Guadyerbas sow was shotgun sequenced (60% genome aligned). Given that only 17.7 Mb were covered in both the pool and the individual, Figure 2 suggests then that there is a positive correlation in nucleotide diversity among nearby genome regions.

Table 1 - SNP annotation

Consequence	Autosomes Guadyerbas	Autosomes Iberian pool	NPAR Guadyerbas	NPAR Iberian pool	PAR Guadyerbas	PAR Iberian pool
ESSENTIAL_SPLICE_SITE	30	30	1	1	0	0
STOP_GAINED	44	11	2	0	0	0
STOP_GAINED,SPLICE_SITE	1	0	0	0	0	0
STOP_LOST	4	15	0	0	0	0
NON_SYNONYMOUS_CODING	2650	1222	40	28	1	0
NON_SYNONYMOUS_CODING,SPLICE_SITE	51	31	0	1	0	0
SYNONYMOUS_CODING,SPLICE_SITE	49	24	3	0	0	0
SPLICE_SITE,INTRONIC	282	254	10	8	0	1
5PRIME_UTR,SPLICE_SITE	1	1	0	0	0	0
3PRIME_UTR,SPLICE_SITE	2	0	0	0	0	0
WITHIN_NON_CODING_GENE,SPLICE_SITE	7	1	0	0	0	0
SYNONYMOUS_CODING	2676	1611	33	34	0	1
CODING_UNKNOWN	8	4	0	0	0	0
WITHIN_MATURE_miRNA	1	1	0	2	0	0
5PRIME_UTR	193	418	0	7	0	0
3PRIME_UTR	2103	1357	23	19	0	0
INTRONIC	148468	78279	1867	1204	133	147
WITHIN_NON_CODING_GENE	286	99	12	3	0	0
WITHIN_NON_CODING_GENE,INTRONIC	6	3	0	0	0	0
UPSTREAM	34314	15216	426	346	20	16
DOWNSTREAM	34395	15737	628	346	24	14
INTERGENIC	433720	150572	7161	3620	3087	1214
Total	659291	264886	10206	5619	3265	1393
dN / dS	0.99	0.76	1.21	0.82	NA	NA

Terms shown are in decreasing order of severity, as estimated by Ensembl.

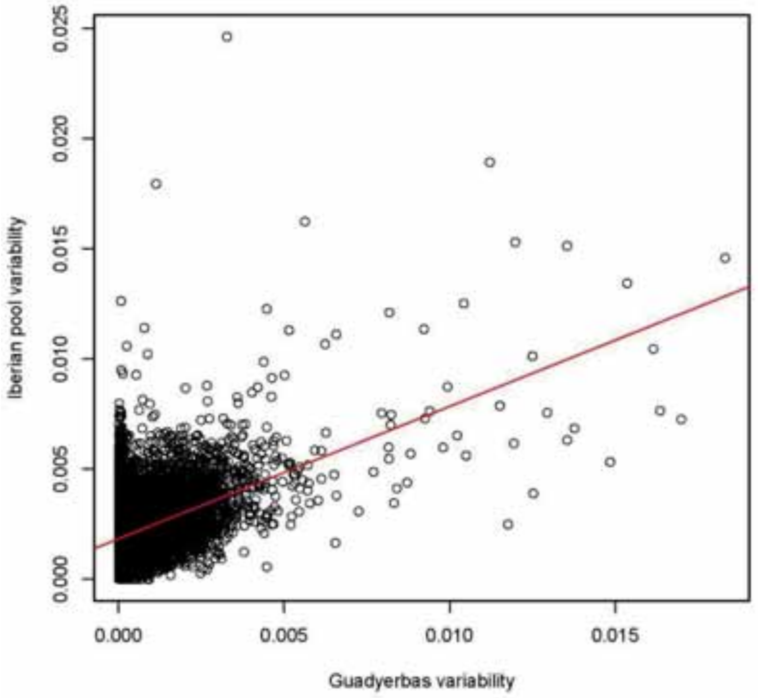

Figure 2. Correlation of Watterson's theta estimates between the individual (Guadyerbas) and the Iberian pool.

Watterson's thetas are plotted in Figure 3 in 200 kb windows for both the pool and the individual. In agreement with results from (Amaral *et al.* 2009; Amaral *et al.* 2011) and (Esteve-Codina *et al.* 2011), variability increased towards telomeric regions. This suggests a marked effect of recombination in variability. To explore this issue further, we plotted variability vs. recombination rate (obtained from) in 5Mb, 10Mb and 20Mb window sizes (Figure 4), observing a correlation of 0.53, 0.62 and 0.70, respectively. Correlation increased with window size, probably because the genetic maps were obtained from a pedigree with few generations and therefore small genetic distances are subject to large sampling errors (Muñoz 2011). We also correlated variability with other factors that have been reported to affect variability, namely GC content and gene density (Table 2). Recombination rate was still the main factor affecting variability. Although GC content was also significant, its conditional effect was slightly negative, likely because of colinearity. If a model was fitted with only GC content, the effect was positive although the model explained much lower variability than a model with only recombination rate (results not presented).

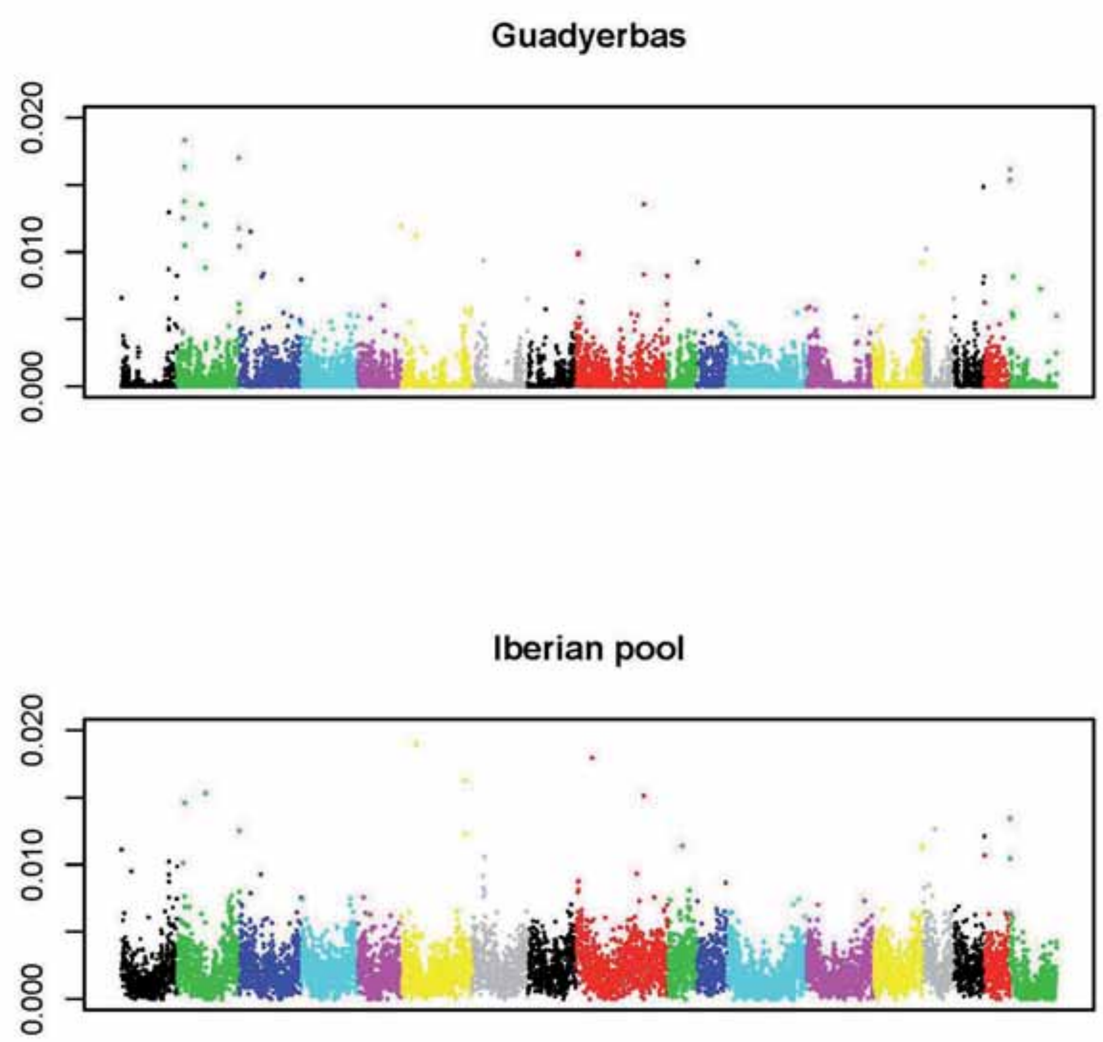

Figure 3. Watterson's theta distribution by chromosome (SSC1-SSC18, SSCX) in the pool (top) and the individual (bottom).

Also as in (Esteve-Codina *et al.* 2011), we observed a marked reduced variability in chromosome X NPAR (Table 3). Note that the SSCX is divided in PAR and NPAR regions, which exhibit quite distinct patterns of variability. The high variability regions in the telomeres correspond to the PAR. In fact, variability in PAR was over 10 times higher than in NPAR for the Guadyerbas sow. Although porcine PAR is small (~7Mb) and diversity estimates are subject to larger errors, the difference between PAR and NPAR variabilities is dramatic.

Table 2. Multiple regression estimates of recombination rate, gene and GC contents on variability estimates (20 Mb windows).

Factor	Estimate	SD	P-value
Recombination rate	4.32×10^{-4}	4.11×10^{-5}	2.00×10^{-16}
Average gene length	-2.17×10^{-9}	3.77×10^{-9}	0.57
GC content	-3.97×10^{-3}	1.12×10^{-3}	0.64×10^{-3}

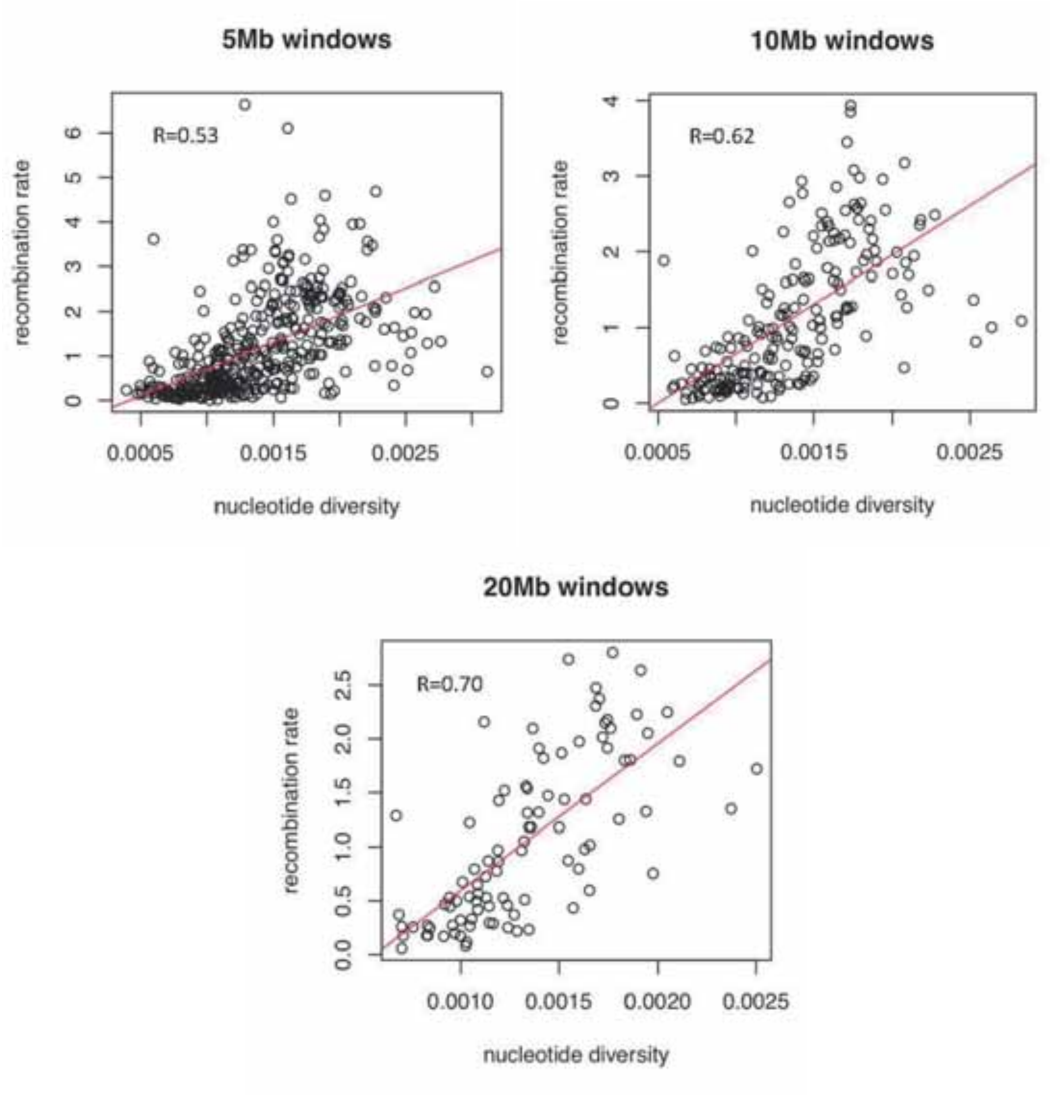

Figure 4. Correlation of Watterson's theta estimates in the pool and the recombination rate (cM/Mb) in windows of 5 Mb, 10 Mb and 20 Mb.

Table 3. Nucleotide diversity in autosomes and X chromosome.

	Guadyerbas individual	Iberian pool
Autosomes	6.55×10^{-4}	1.31×10^{-3}
Pseudo-autosomal chromosome X (PAR)	3.02×10^{-3}	2.22×10^{-3}
Nonpseudo-autosomal chromosome X (NPAR)	1.79×10^{-4}	5.83×10^{-4}

Multi-copy regions (MCR)

Given the increasing awareness of the importance of structural variants in the genome, we also sought to uncover these in the Iberian pigs. In fact, one of the advantages of resequencing vs. genotyping is that the former allows a much more reliable detection of structural variants in the genome than the latter approach. Here, employed a read density method to uncover multi copy regions (MCRs) because the current porcine assembly is still not completely reliable to ascertain other kind of variants (e.g., inversions, novel insertions, translocations) using aberrant paired-end distance methods. MCRs detection is based on read density and is therefore less sensitive to mis-assemblies in the reference genome. (We refer to MCRs rather than copy number variants because we analyzed a single individual and we do not have information on whether that multi-copy region is fixed or segregating in the population). We analyzed only the individual sow because of the uncertainty in the number of chromosomes actually sequenced for the pool in any given region. Due to limited read depth, we considered only gains with respect to reference genome rather than gains and losses.

We found 3,082 outlier regions potentially caused by MCRs in the Guadyerbas genome. They were distributed among 1,653 windows and spanned 30.5 Mb. As shown in Figure 5, the majority of the MCR are short (less than 20 kb) and only two are longer than 100 kb.

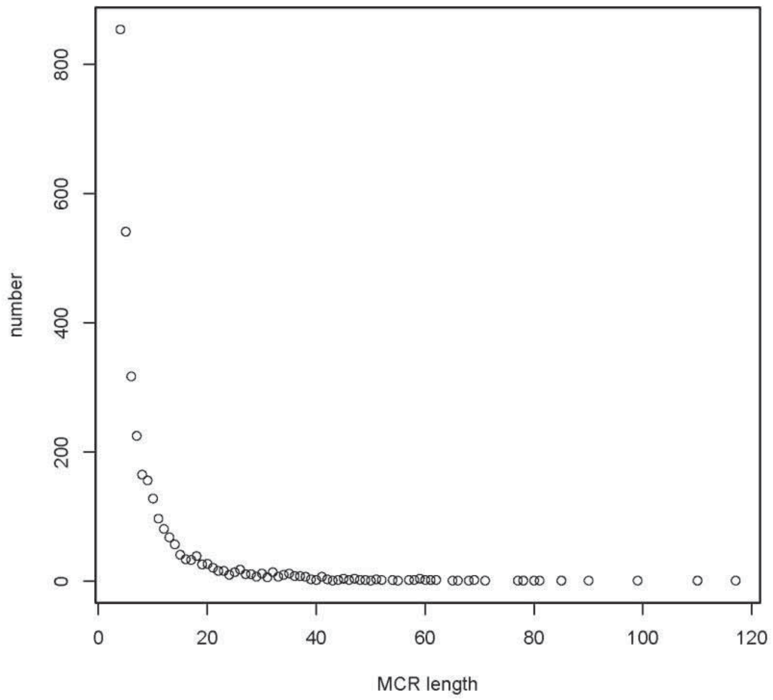

Figure 5. Distribution of MCR lengths (in kb).

These MCRs affect 4% of the annotated pig genes in their entirety (100% of the gene length) and 2% of the genes partially (>50% of their gene length). Barring for errors in the reference assembly, therefore, MCRs seem to be an important source of variability in the pig, as also observed in other species (Clop *et al.* 2012). Distribution of the MCRs along the chromosomes is represented in Figure 6. We observed a positive correlation between variability inside the MCRs and the variability within those windows (200 kb size) containing MCRs but outside MCR boundaries (Pearson correlation = 0.6, Additional File 1). Average variability inside MCRs was 1.51×10^{-3}, somewhat higher than MCR windows but outside MCRs boundaries (9.09×10^{-4}), whereas windows devoid of MCRs had the lowest average diversity (8.42×10^{-5}), suggesting that high variability windows are enriched in MCRs (Summary statistics in Table 4). On the other hand, we detected no correlation between the number of copies of a MCR and variability within MCRs.

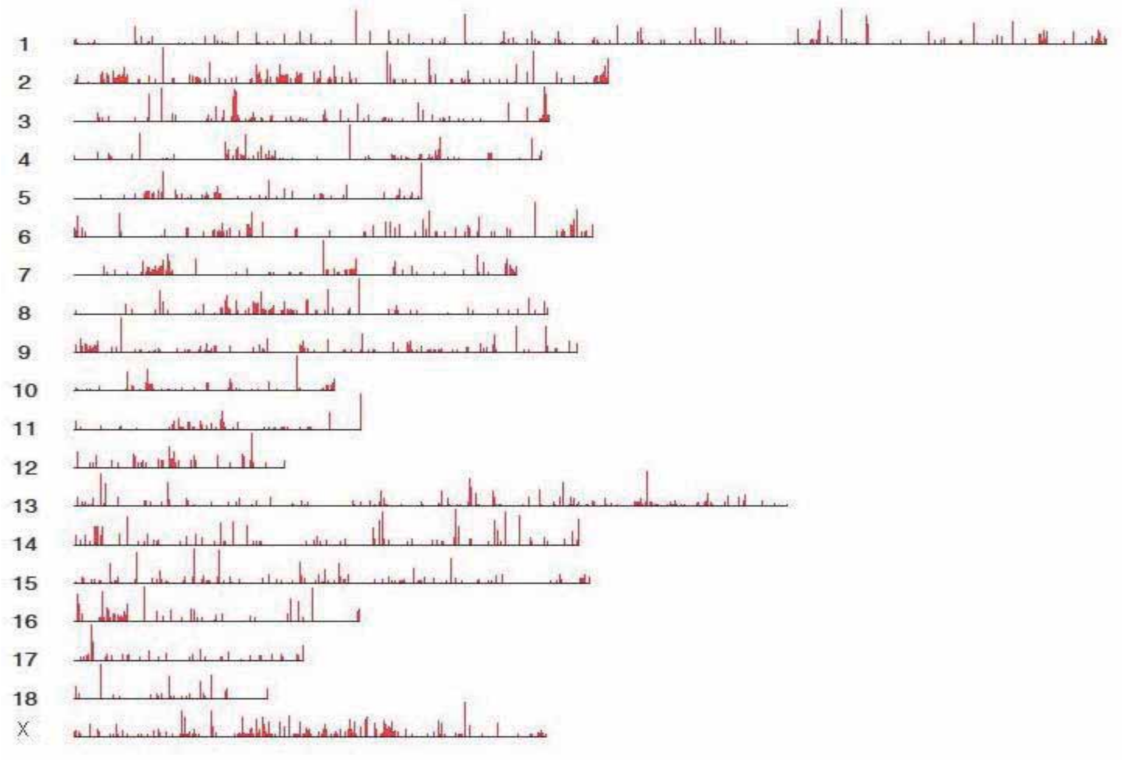

Figure 6. MCR gains with respect to the reference genome found in the Iberian genome. Each red line corresponds to a MCR location and the length of the line is proportional to the number of copies.

Table 4. Variabilities within and outside multi copy regions (MCRs)

	Median	Mean	SD
Within MCRs	$1.67\text{x}10^{-4}$	$1.52\text{x}10^{-3}$	$3.46\text{x}10^{-3}$
Outside MCRs, within windows containing MCRs	$1.83\text{x}10^{-4}$	$9.10\text{x}10^{-4}$	$1.82\text{x}10^{-3}$
Windows without MCRs	$8.43\text{x}10^{-5}$	$3.92\text{x}10^{-4}$	$6.11\text{x}10^{-4}$

A total of 696 annotated genes fully fell inside MCRs and are therefore more likely to be functional than partially duplicated genes. Our study allowed us discovering novel paralogs of annotated genes, which either are absent in the Duroc reference assembly due to a miss-assembled genome, or are Iberian specific copies. These genes primarily belonged to well known multi-genic superfamilies. By far, the most over-represented gene family was olfactory receptors, comprising a total of 476 genes. The chromosomes containing the largest number of olfactory genes were SSC2 and SSC7 (Figure 7). These results agree with data from the international consortium, who found that the pig is one of the species with the largest repertoire of olfactory receptors, likely a result of the importance of smelling in this scavenging species (Groenen 2012).

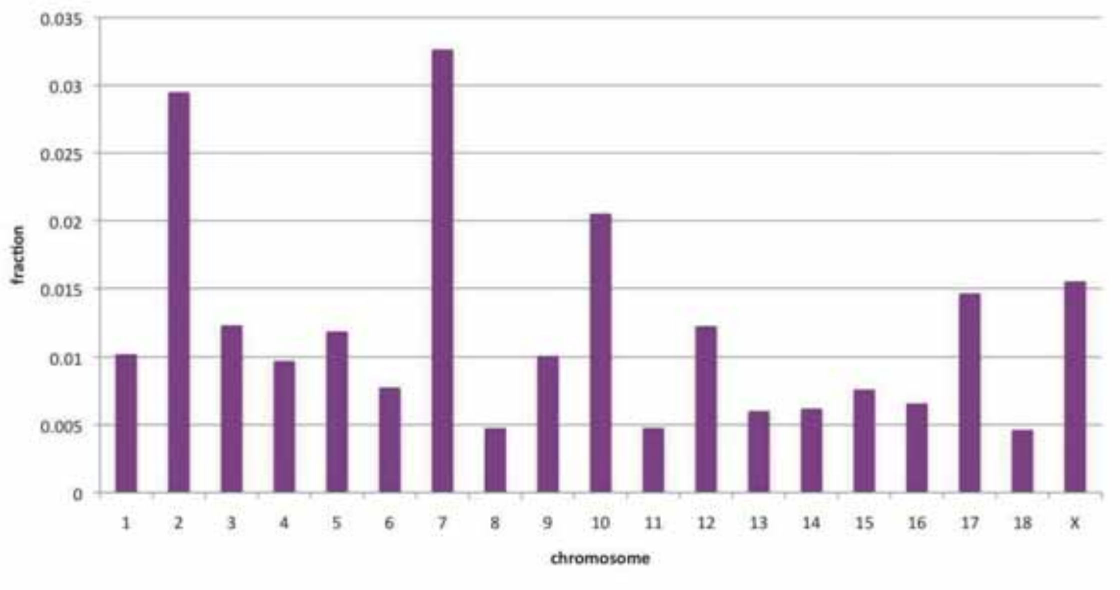

Figure 7 - Kb spent in MCRs divided by chromosome length.

Similarly, large gene families involved in defense and immune response were over-represented within MCRs; we found 8 new paralogs of annotated interferons (*IFN-α8*, *IFN-α10*, *IFNα-11*, *IFNα14*, *IFNδ2*, *IFNδ6*, *IFNω2* and *IFNω4* family under expansion genome paper), 2 interleukines (*IL1-β*, *IL1B*) and five *SLA* genes (*SLA-3*, *SLA-9*, *SLA-10*, *SLA-P1*, *SLA-DRB1*). Several tumor necrosis factor receptors (*TNFR*) and T-cell receptors (*TR*) were found as well. Others were involved in lipid (*ACOT4*, *GPAT2*) and carbohydrate metabolism, like 5 new paralogs of the *UGT2B* family and 8 salivary and pancreatic amylases, also in detoxification (*CYP2C33* and *CYP4A21*), pheromone binding (*PHEROA* and *PHEROC*), viral infectious cycle (*Gag* protein and *ENV*), perception of taste (*VN1R2*), fertilization (*SPM1*) and retinol dehydrogenase (*RDH16*). Two genes from the serpin-like clade (*Serpina 3-1* and *Serpina 3-2*), the myostatin gene (*MSTN*) and a lactase gene (*LCT*) also seem to be present in multiple copies in the pig genome. Not surprisingly, several small RNAs were also detected: two rRNAs (5S ribosomal RNA and 5.8S ribosomal RNA), one snoRNA (*SCARNA6*), one snRNA (*U1*) and two miRNAs. A complete list of genes entirely inside MCRs is shown in Additional File 3. Not surprisingly, a gene ontology (GO) enrichment analysis of biological processes (see methods) found an over-representation of sensory perception of smell (adjusted P value = 2.06×10^{-117}), response to virus (adjusted P value = 2.99×10^{-06}) and xenobiotic metabolism process (adjusted P value = 1.55×10^{-02}) (Additional File 2).

Outlier regions and potential selection targets

A matter of intense research is the study of patterns of nucleotide variability in domestic species. Outliers in these patterns with respect to the standard neutral model can be due to selection and then reveal genes of socio – economic interest, as well as helping to understand the effects of domestication and of artificial selection in the genome (Groenen 2012). A serious problem is that selection does not result in a single obvious signal (e.g., a selective sweep) but rather in a diversity of manifestations that depend on intensity and age of selective process as well as on the demographic history of the population (Li *et al.* 2012). Here, we employed a battery of tests that pinpointed a series of genome regions, hopefully enriched in non-neutral genes. We also took advantage, when possible, of the simultaneous availability of pool and individual data. Despite the fact the Guadyerbas strain only represents one of the Iberian varieties, it is conjectured that the most relevant selective sweeps will be shared across all Iberian pigs.

First, we examined extreme windows in terms of low and high variability for the Guadyerbas and the pooled data (see methods). A total of 132 genes were annotated within the lowest variability windows (Additional File 3). A window in

SSC1, was specifically enriched in interferon genes (*IFNE*, *IFN-α10*, *IFNω1*, *IFNω3* and *IFNω4*), which are involved in response to virus (adjusted enrichment GO P=1.3×10^{-04}). Note that *IFN-α10* and *IFNω4* are within MCRs, suggesting that those genes have un-annotated paralogs and putatively under positive selection. Genes within the lowest variability windows in NPAR included genes from the Ras oncogene family (*RAB33A*, *RAB39B*, *RAB39B* and *RAP2C*), the *SOX3* gene (*SRY-box3*), involved in sex determination, face development and pituitary gland development, the serotonin receptor *HTR2C*, involved in anxiety, reproductive and feeding behavior, the *MECP2*, with a role in behavioral fear response, as well as genes involved in lipid metabolism (e.g., *ACSL4*, *ALG13*, *ABCD1*, *PLP1*), in hair follicle development (*NSDGL*) and other genes related to immune response (*IL13Ra1*, *IL1RAPL2*). A complete list of these genes is in Additional File 3.

The majority (~80%) of the high variability windows contained MCRs. To ensure that the high variability found is not influenced by MCR, we removed the SNPs inside MCRs. The result was that those windows still conserved high variability levels, in agreement with results in Table 4. The majority of genes in those windows were hundreds of olfactory receptors present in gene clusters distributed among almost all chromosomes. In addition, other gene families were represented, e.g., ATP-binding cassette family, zing finger genes, T-cell receptors and *SLA* genes (mainly located in chromosome 7), transmembrane proteins (*TMEM* family), several small nucleolar RNAs, solute carrier family genes, protocadherin family genes involved in homophilic cell adhesion and cytochrome family p450 genes (*CYP*). (See Supplementary File 3 for a complete list of genes). Note that *IL1B* and other gene families are present in MCRs and also in high variability regions.

Next, we computed Tajima's *D* and Fay-Wu's *H* statistics, modified to account for the idiosyncrasy of pool data (methods). In principle, Tajima's *D* and Fay-Wu's *H* negative values can be produced by positive selection, although Tajima's *D* is particularly sensitive also to demographic effects and prone to false positives. The correlation between both statistics was positive, although moderate r = 0.28 (Figure 8). There is also an apparent number of windows with negative Tajima's

D and zero or even positive Fay-Wu's *H*. Although the interpretation of this is not clear, it might be caused by recurrent hitch hiking events (Przeworski 2002) or simply an artifact.

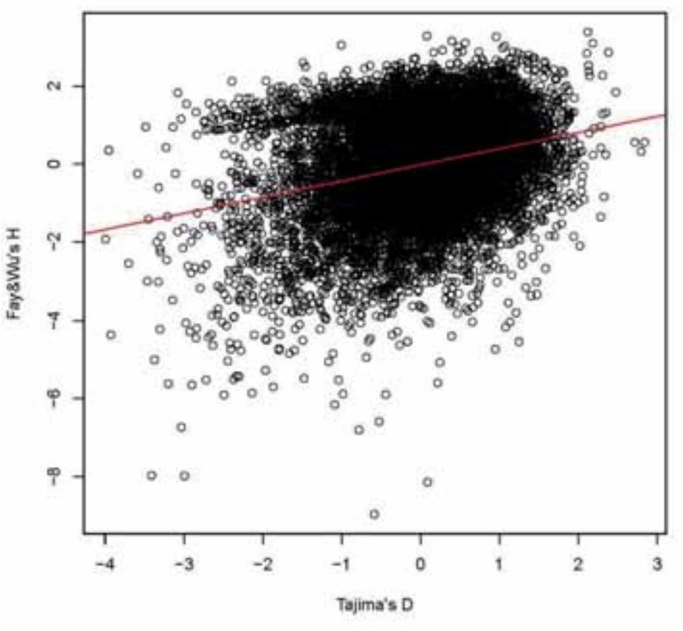

Figure 8. Correlation between Tajima's D and Fay - Wu's H statistics in pooled data.

We selected the 1% most extreme windows with combined negative Tajima's *D*, Fay-Wu's *H* and low variability (see Methods). No over-representation of GO were detected after correcting by multiple testing. Interesting candidate genes inside those windows are involved in axonogenesis and synapsis (*FOXP1*, *LRRK2*, *EHMT2*, *RAB11A*, *TEKT5*, *IGF1R*, *UNC13C*, *CNTN1*, *COL9A2*, *AXIN2*, *CADPS2*, *HTR6*, *KCND1*, *NOVA1*, *PTEN*), circadian rhythm (*HEBP1*, *ALB*), epithelial cell differentiation, keratinization and hair follicle formation (*FOXP1*, *IGF1R*, *HNF1B*, *PTEN*, *AXIN2*, *KRT81*, *KRT83*, *KRT84*, *KRT85*, *KTR86*, *PRKD1*, *AC0210066.1*), blood vessel morphogenesis (*PPAP2B*, *PRKD1*), lipid metabolism (*PPAP2B*, *VEPH1*, *RASA4B*, *ATP10B*, *NEU1*, *PTEN*, *SMPD4*, *ALB*), exploratory, locomotory, grooming and feeding behavior (*LRRK2*, *NMUR2*, *APBA2*), response to starvation (*GAS6*, *ALB*), spermatogenesis, ovulation, sex determination (*EHMT2*, *AFP*, *IGF1R*), visual/odor perception (*OR5P2*, *LRRK2*, *VSX1*, *GRK1*), immune response and inflammatory response (*CIITA*, *PRKD1*, *FOXP1*, *IGF1R*, *PTX3*).

Finally, we performed a genomewide Hudson-Kreitman-Aguadé (HKA) test in the pool data. The NPAR was analyzed separately from autosomes and PAR. After correcting for multiple testing, only 25 (0.23%) windows with an excess of differentiation were significant (adjusted Bonferroni-Hochberg P < 0.05). Although there was no over representation of any GO category, some genes are still worth mentioning. These comprise genes involved in feeding behavior (*NPW*), social behavior (*HTT*, *DVL1*), locomotory behavior (*HTT*, *SLCGA3*), pigmentation (*MC1R*), hair follicle morphogenesis (PDGFA), sensory perception of taste (*TAS1R3*, *GNG13*), male gonad development and spermatogenesis (*GFER*, *BOK*), lipid metabolism (*DECR2*), circadian rhythm (*PRKAA2*, *ADCY1*), tumor necrosis factors (*TNFSF12A*, *TNFRSF18*, *TNFRSF4*), fat cell differentiation (*SDF4*) and several genes involved in lipid transport, e.g., *ABCA3*. This gene was also reported by the pig sequencing project as being under selection. The neuropeptide *AXIN1* is also interesting as has been found differentially expressed between in brains of two extreme groups of junglefow in terms of fearfulness (Jongren *et al.* 2010).The complete gene list is in Additional File 3.

Only 39 (0.36%) windows with an excess of polymorphism vs. differentiation were significant (HKA test adjusted P < 0.05). Genes inside those windows belonged to the *ABC* superfamily (*ABCC4*), were complement activation genes (*C8A*, *C8B*), antigen processing and presentation (*SLA-DQA1*, *SLA-DQB*G01*, *SLA-DRA1*, *SLA-DRB*, *SLA-DRB1*), feeding behavior and synapsis (HCRTR2), visual, sound perception and pigment granule transport (*MYO7A*), lipid metabolism (*PPAP2A*, *PRKAA2*), viral infectious cycle (*RPS21*), defense response (*SPACA3*) and many genes from the olfactory receptor, zinc finger and *TRIM* families (full gene list in Additional File 3). Within the NPAR region of the X, only one window was significant. This window contains the *SHROOM2*, a gene involved in brain, eye and ear morphogenesis and pigment accumulation among others (Additional File 3).

Discussion

This study presents a novel combined analysis of pool and individual sequencing. Although pools biases the SNP discovery process towards common variants and have lower power than individual sequencing (Cutler & Jensen 2010), our simulation indicates that we should expect to detect almost half (47%) of all SNPs. Given that there are $\sum_{i=1,17} 1/i$ = 3.4 times more SNPs in 18 chromosomes than in a single individual, the pool process uncovers about 60% more SNPs than individual sequencing – for any given region sequenced in common and assuming an average depth 14×. Both designs, it should be noted, are biased towards common variants so they are not appropriate to detect rare mutations. Rare variants are likely to be very young and, a priori, should contribute little to total genetic variability in phenotypes of interest.

Genome-wide variability in the Guadyerbas sow was actually much lower than that in the Iberian pool, 50% and 70% lower for autosomes and NPAR, respectively (Table 3). Estimates are corrected for pooling process so the large disparity is not due to SNP calling in pools vs. individuals but, rather, to the high inbreeding of the Guadyerbas strain. Because the pedigree of the Guadyerbas is known since the foundation of the herd in 1944 (Toro *et al.* 2000), inbreeding coefficient F for the specific sow sequenced was estimated as $F_A = 0.39$ and $F_X = 0.46$ for autosomes and NPAR, respectively. This results in estimates corrected by inbreeding $\pi_{A^*} = 6.55\times10^{-4} / (1-0.39) = 1.07\times10^{-3}$ and π_{X^*} = 1.79×10-4 / (1-0.49) = 3.51×10^{-4}. These values are close to those obtained from the pool in autosomes but, intriguingly, for NPAR are still 40% lower in the Guadyerbas (Table 3). Therefore, inbreeding explains the loss in variability in the whole Iberian pig breed for autosomes but not in NPAR.

Remarkably, heterozygosities in the Iberian pool are comparable to those reported in the two European wild boars sequenced by the *International Pig Genome Sequencing Consortium*: 0.0012 and 0.0010 (Groenen 2012). In contrast, heterozygosity in international domestic breeds is higher (> 0.0016) because of introgression with Asian pigs. The fact that Iberian pigs and European wild boar diversities are comparable reinforces previous evidence showing that Iberian pigs have not been intercrossed with Asian germplasm (Alves *et al.* 2003). It also stresses the relevance of Iberian pig as a model of native Mediterranean domestic pig that should help to disentangle the effects of Asian introgression and domestication on response to selection by modern breeding.

An intriguing observation was the high dN/dS ratios that we observed, especially in the Guadyerbas sow (Table 1). Although a positive ratio is normally taken as a symptom of positive selection, this is unlikely, i.e., it is unlikely that the whole PAR is under directional selection. We believe, instead, that the most likely explanation is that the very low effective population size in the Guadyerbas strain, but also in the Iberian breed as a whole, attenuates the effect of natural selection (Charlesworth 2010). These results in more deleterious alleles found at intermediate frequencies than if the population size were very large. This effect should be more pronounced, as observed, in sex chromosomes than in autosomes.

Our data further confirm the much lower observed than expected variability in SSCX (3/4 that of autosomes) as was previously reported in the partial resequencing of the same Guadyerbas sow (Esteve-Codina *et al.* 2011). Here, because we were able to distinguish between PAR and NPAR regions, the X/A ratio is even lower than reported before: 0.27 in Guadyerbas and 0.44 in the pool. In contrast, diversity in the PAR was comparable, or even higher, than in autosomes. Although demographic effects can reduce X/A variability, the effect observed here is quite unusual, and seems to be a pervasive property of all porcine populations. Selection can be argued as an alternative explanation. Genes within the lowest variability NPAR windows included *ACSL4* (lipid metabolism), *HTR2C* (behavior), *SRY-box3* (sex determination), *MECP2* (fear response), *NSDG2* (hair follicle formation) and interleukines (immune response). Interestingly, fear response is a distinctive biological feature between wild animals and its domesticated descendants. Hair follicule formation has also evolved during domestication process, as wild pigs are furrier than domestic pigs. It should be noted that the black varieties of Iberian breed are hairless (as the Guadyerbas strain) and the red varieties present sparse hair. As for *ACSL4*, we have reported a QTL nearby this gene that affects fatty acid composition in the Iberian pig (Mercade *et al.* 2006; Corominas *et al.* 2012).

The discovery of thousands of new MCRs (>4 kb) with respect to the reference genome suggests either a mis-assembled reference genome or real copy number variants between the Iberian pig and the Duroc reference assembly. In agreement with our results, the *Pig Genome Consortium* also discovered many new paralogs of existing genes in the reference Duroc assembly. The fact that some MCRs have high values of nucleotide diversity might be caused by an artifact of the mapping (the Iberian pig presents more copies than the reference and therefore ambiguous reads map to the same locus, causing false positive SNPs). Nevertheless, the fact that variability in regions outside the MCR with respect to the assembly but within windows containing MCRs is higher than average genome-wide (Table 4) may be an indirect consequence of increased recombination, which causes MCRs as well as increased variability.

In this sense, it is not surprising that many MCRs contain genes belonging to well known gene super-families (*SLA*, *OR*, *CYP*, *IFN*, *IL*, *TNF*, *TR*), which are known to be originated by duplication events. Several studies have reported that MCRs are enriched in segmental duplications (Sudmant *et al.* 2010; van Binsbergen 2011). Besides, other genes like pheromone receptors; amylases or taste receptors seem to be new paralogs of existing genes not yet annotated. Virion assembly proteins like *gag*, *pol* and *env* belong to retroviruses inserted in the genome, whose function still needs to be characterized in detail, although some are described to be involved in gene transcription and resistance to exogenous infections. The international pig genome sequencing consortium has reported the *INF* and *OR* super-families to be under expansion (Groenen 2012), and our results support this hypothesis as many putatively functional *INF* and *OR* genes were detected inside MCRs. Many genes inside MCR overlapped with extreme high variability regions, most of them are olfactory receptors, but we also found *SLA* genes and other immunity genes. Those genes are reported to be under balancing selection in many species, since being heterozygous confer advantage in terms of distinct related odors and pathogens detection. Interestingly, common low variability regions between the individual and the pool show also over-representation of defense response, but the genes involved are different from those in high variability windows.

Extreme Tajima's *D* - Fay-Wu's *H* – variability combined test (D-H-θ_w) and HKA excess of differentiation detected some interesting genes, putatively under positive selection but at different time scales. The combined D-H-θ_w test traces selective events that happened no more than 80.000 years ago, whereas the HKA test is useful to trace very old events up to 250.000 years (Sabeti *et al.* 2006). Genes related with keratinization (D-H-θ_w test), epidermis formation (D-H-θ_w test) and hair follicle morphogenesis (D-H-θ_w test and HKA), as described by George et al (George *et al.* 2011), may be important for setting up physical barriers between the body and the outside world and could evolve rapidly in response to changing environment. Some studies in humans and primates also found adaptive signals in keratinization genes (Tennessen *et al.* 2010; Tong *et al.* 2010; George *et al.* 2011). Genes belonging to the *TNF* family (HKA test) play a role in defense response, specifically in response to wounding. They are cytokines secreted by activated macrophages and lymphocytes and exert several tumor suppressor as well as antiviral activities. Many studies found positive selection in immunity-related-responses (Zelus *et al.* 2000; Zhang & Nei 2000; O'Connell & McInerney 2005; Jiggins & Kim 2007; Carnero-Montoro *et al.* 2011; Manry *et al.* 2011), so it is not surprising to find *IL* and *INF* in low diversity regions. Considering the speed at which many pathogens, such as viruses, evolve, a coevolutionary molecular arms race between pathogens and host cells might explain the presence of strong selection favoring new mutations in these genes.

Several genes involved in feeding behavior, fear response and social behavior were inside significant windows in both D-H-θ_w test and HKA excess of differentiation. Behavior has been reported as one of the biological functions subject to selection during the process of pig domestication (Chen *et al.* 2007; Amaral *et al.* 2011; Kittawornrat & Zimmerman 2011) and feeding behavior and response to starvation are, logically, most relevant traits in domestication and breeding. The *LRRK2* gene, identified with the D-H-θ_w test, would merit special attention in future works: it is involved in exploratory behavior, odor detection and is a positive regulator of the dopamine receptor signalling pathway. Circadian rhythm (D-H-θ_w test and HKA) are additional functions that may have been affected by selection, which could be explained by a distinct biological clock between the wild ancestors and domestic pigs, due to human interference in their life habits. Perception of taste is another function that might have evolved due to novel food resources and, effectively, some genes with this GO were inside significant windows according to the HKA test. The *MCR1* gene (present in HKA test), involved in pigmentation, has been reported to be under positive selection due to human interest to cherry-pick different coat colors that would otherwise be quickly eliminated in the wild (Fang *et al.* 2009). Lipid metabolism genes (present in both combined D-H-θ_w test and HKA) might also have changed, specifically in the Iberian breed, conferring its distinctive lipid composition and deposition in the meat. Spermatogenesis genes, identified by both tests, have been reported to be rapidly evolving genes in other species (Jiang & Ramachandran 2006; Haerty *et al.* 2007; Hutter 2007). Finally, we found neurological genes involved in synapsis and axon guidance, both related to brain development and function.

Conclusion

The recent completion of the porcine sequencing project has allowed digging deeper into the complexities of the Iberian pig genomes than was possible until now. This breed is important because it represents a primigenious European breed that, while being domestic, has not been introgressed with Asian germplasm. Our data confirm the importance of structural variation in the porcine species as also observed in other species. The battery of tests we applied suggests that many and diverse selective processes have occurred, among them changes in feeding behavior. New bioinformatics tools, e.g., to deal with structural variants as well as complete annotation of the pig genome (ENCODE) projects are badly needed to improve interpretation of results.

Material and Methods

Samples and sequencing

The genome of a highly inbred Iberian pig, pertaining to the Guadyerbas strain, which has been partially sequenced (1% of the genome) in a previous study (Esteve-Codina *et al.* 2011), was shotgun sequenced using Illumina Hiseq2000, We run one 100 bp paired-end lane and one 100 bp single-end lane. In addition, we also sequenced a reduced representation library (RRL) of a pool comprising nine sows (equal concentrations of each) from the most representative Iberian varieties in Spain: Retinto, Mamellado, Torbiscal, Guadyerbas, Entrepelado and Lampiño. All sequenced sows are registered in the Iberian Herd Book and were sampled from well accredited farms that have kept purebred Iberian pigs without intercrossing with 'foreign' breeds. The method to construct the reduced representation library is described elsewhere (Esteve-Codina *et al.* 2011). For the pool, Illumina GAIIx technology of 50 bp was employed, and 2 PE lanes were available. As outgroup, we shotgun sequenced a *Potamocherus porcus* male using Hiseq2000 (three PE lanes, 100 bp long) in order to measure divergence and then gain in power to detect selection.

We were able to delineate the boundaries between PAR and NPAR because of read depth differences in males along the SSCX (unpublished data). The SSCX PAR occupies the first 6.7 Mb and the last 400 kb of SSCX, approximately. Although assembly 10.2 separates two telomeric PARs, linkage analyses using genotyping data from the 60k SNP chip in an Iberian x Landrace cross and results from Burgos-Paz (Burgos-Paz *et al.* 2012) suggest that a single PAR exists – as in most mammals. We therefore pooled the results from the two annotated PARs in the analyses reported here.

Alignment and SNP calling

Reads were mapped against the latest reference genome (assembly 10.2) using bwa (Li & Durbin 2009), allowing 8 mismatches and filtering by mapping quality of 20. *P. porcus* reads were aligned disregarding the paired end structure, i.e., they were aligned as SE. This was done to minimize the possibility that structural changes between the two species prevent alignment. A total of 345M reads were aligned, resulting in an average depth of 20× (3-50×) and 1.6 GB of the *S. scrofa* genome assembled.

SNP calling for the Guadyerbas individual was performed using Samtools mpileup option (Li *et al.* 2009) filtering by minimum depth of 3×, maximum depth of 20× and SNP quality of 20. SNP calling in the Iberian pool was done using SNAPE (http://code.google.com/p/snape-pooled/), setting divergence to 0.01, prior nucleotide diversity 0.001, folded spectrum and filtering by a posterior probability of segregation > 0.90. The SNAPE approach consists in computing the posterior probability of SNP frequency being distinct from 0 or 1, given that we observed n_A alternative alleles and C-n_A reference alleles, and given prior frequency in the population being P(f):

$$P(f|n_A) \propto P(n_A \mid f)\, P(f)$$

where

$$P(n_A \mid f) \;=\; \sum_{k=0}^{n} \binom{C}{n_A} p^{n_A}\, (1-p)^{C-n_A} \binom{n}{k} f^k (1-f)^{n-k} \;,$$

with p being the probability that an allele A is read and n, the number of chromosomes in the pool. This probability in turn depends on n, k and on whether there is a true A in the genome or whether it is the result of a sequencing error. The algorithm considers the geometric mean of sequence qualities for every allele read to compute this probability (Raineri 2012). In the equation above, we take into account the probability that k counts of the allele are present in the pool, given that its true frequency is f and that, given k, how many reads A out of n are expected. Because some quantities, notably k, is unknown, this is integrated out. For prior p(f), we considered the standard neutral model expected frequency, i.e., $f \propto 1/f$.

Simulation of pooling process

Although pools are a highly cost-effective strategy, the variability uncovered is only a fraction of the true one in the population. We sought to evaluate the power and false discovery rate of our experimental design by simulation. We simulated 18 chromosomes of 1 Mb of sequence using coalescence with ms (Hudson 2002) under a standard neutral model with nucleotide diversity π and scaled recombination rate ρ per site = 0.001. For each resulting chromosome, the program ART (Huang *et al.* 2011) was used to generate reads with the built-in profile for Illumina paired-end technology of 75 bp-long reads. To simulate the pooling process, reads were randomly selected from each sequence using an equal proportion from each individual. An average depth of 14× was simulated for the whole pool in all and reads were aligned with BWA (Li & Durbin 2009). Next, SNPs were called with SNAPE, restricting minimum and maximum depths to do the calling between 3× and 30× as in our real data analyses. Power was computed as the proportion of true SNPs in the population (i.e., before pooling) located within regions of appropriate depth that were correctly recovered. False Discovery Rate (FDR) was the proportion of SNP calls that were incorrect. A total of 100 replicates were simulated.

Multi-copy region detection

Read depth method (Sudmant *et al.*, 2010; Alkan *et al.* 2011) was applied to identify copy number of a region; mrsFAST (Hach *et al.* 2010) is an exhaustive mapping tool that allows paralog detection and was used to align reads (allowing 6 mismatches) against the repeat masked reference genome; repeat mask information was obtained from NCBI. Average read depth for each non-overlapping 1kb bin was calculated across the genome and copy number (CN) of each unit was predicted based on the average read depth across the diploid region. 1:1 orthologous genes between human, cow and pig was used to obtain read depth across diploid region. Since these regions have the same number of copies in 3 relatively distant species, we assumed these were conserved in a copy number neutral stage. Finally, chained regions in the genomes which are ≥ 4kb in length having copy number ≥3 (each bin should have CN ≥ 3 and 1 kb gap was allowed), were extracted and declared MCRs. Next generation sequencing methods introduce bias in the read depth, which is caused by the dissimilar GC content of different segments of DNA. To correct this bias, we used GC intervals and the average read depth across the diploid region to find out the correction factor and used that factor to correct depth of each 1 kb bins (Sudmant *et al.*, 2010).

Nucleotide variability estimation and selection tests

Note that, with NGS data at low depth, nucleotide diversity cannot be simply computed dividing the number of SNPs called by the length of sequence assembled. This is because, with shallow depth, the two alleles of the same SNP may not be read and because of errors in calling SNPs. For the individual, we corrected for low coverage as detailed in (Esteve-Codina *et al.* 2011):

$$\hat{\theta}_w = \frac{S}{\sum_i L(i)\, P(j|i)}$$

where S is the raw number of SNPs, *L(i)* is the length in bp of depth *i* for that window, and *P(S|i)* is the probability of reading both alleles for depth i (Esteve-Codina *et al.* 2011). In the case of pools, Watterson's theta was computed as in Amaral et al. Briefly, we correct by the expected number of chromosomes sampled for each read depth along the window:

$$\hat{\theta}_w = \frac{S}{\sum_i \; L(i) \sum_{j=2}^{\min\,(nr(i),nc)} P_c(j|nr(i),nc)\; a_j} \qquad (1)$$

(Amaral *et al.* 2011), where L(i) is the length in bp of depth i for that window, and Pc(j | nr(i), nc) is the probability that a set of nr sequences randomly extracted from nc possible chromosomes contains sequences coming from precisely j different chromosomes. Finally, a_j is Ewens constant $\sum_{i=1,n-1} 1/i$.

Definition of low and high variability windows

Given that over 10% of Guadyerbas windows had no SNP, we defined extreme low variability regions for the Guadyerbas as those windows devoid of variability and with > 10kb assembled. Among these windows, we selected those of 5%

lowest variability in the pool as well, with a minimum of 3 kb aligned. In that way, we avoid choosing fixed regions in the Guadyerbas strain due to drift. We defined extreme high variability regions as the 5% most variable windows in Guadyerbas and in the pool where at least 10 kb (Guadyerbas) and 3 kb (pool) were aligned.

Tajima's *D* and Fay-Wu's *H* tests

Tajima's *D* test (Tajima 1989) were computed as the normalized difference between the average pairwise nucleotide difference $\theta\pi$ and the Watterson estimator, divided by the theoretical variance of the same difference in the standard neutral model without recombination in pools (Ferretti 2012). The estimator of θ based on π was computed as the average pairwise nucleotide diversity across all reads for a given position, averaged over all positions and corrected by a multiplicative factor 2n/(2n-1) (Futschik & Schlotterer, 2010). This estimator is unbiased under the neutral model. The normalized Fay and Wu's *H* test (Fay & Wu 2000) was computed similarly from the standardized difference between $\theta\pi$ and the estimator θ_H based on high frequency derived alleles. For the estimator θ_H, only sites with known outgroup bases were used, and the estimator was obtained by summing all segregating sites with k derived alleles in r reads weighted by the factor $k^2/r(r-1)$ and divided by a factor correcting for the bias (Ferretti 2012). The variances in the denominators are evaluated exactly in the limit of short read for the standard neutral model without recombination following the results of (Fu 1995) and accounting for the random extraction of reads from individuals (Ferretti 2012).

In order to minimize confounding demographic effects with selection fingerprints, we calculated the empirical joint distribution combining Tajima's *D*, Fay and Wu's *H* and Watterson's θ as in (Ramos-Onsins *et al.* 2008). To do so, we sorted the normalized statistics *D*, *H* and θ, the empirical test was obtained simply by multiplying the inverse of the ranks 1/n, 2/n,… 1 of each statistic for each window 1…*n*, and normalizing. A GO enrichment analysis was performed with genes within the 1% most extreme windows.

Hudson-Kreitman-Aguadé test

Multilocus Hudson-Kreitman-Aguadé (HKA) tests were calculated in the pool using the *P. porcus* alignment as outgroup and following the original algorithm (Hudson *et al.* 1987). We applied the test dividing the genome in 200 kb windows. Then, M+1 equations were solved using a bisection algorithm to calculate the estimates of the M+1 parameters (M theta values, one per window, plus the time of split between species measured in 2Ne generations). Thus, a partial HKA test per window was obtained plus the total sum of values, where the null hypothesis (stationary neutral model) is contrasted using M-1 d.f. The approach assumes unlinked windows and it is, therefore, conservative because nearby windows are linked. The original HKA formulae require $a_n = \sum_{i=1,n-1} 1/i$ and $b_n = \sum_{i=1,n-1} 1/i^2$ constants, which in the case of pooling are unknown. Instead we used the equivalent correction to infer Watterson's theta from pools (denominator in eq. 1), whereas b_n was obtained by interpolation from a_n. The HKA function can be downloaded from http://bioinformatics.cragenomica.es/numgenomics/people/sebas. In order to identify outlier windows we performed a Bonferroni-Hochberg multiple test correction over the value of the partial Chi-square per window using a 5% nominal significance.

Annotation and Gene Ontology enrichment analysis

SNP annotation was performed using the Variant Effect Predictor perl script from Ensembl (McLaren *et al.* 2010) and the *Sus scrofa* gtf annotation file was from Ensembl release 67, the latest version and that used in the pig genome publication. Gene ontology enrichment analysis was performed using FatiGO, a module of Babelomics (Medina *et al.* 2010) using the human genome as background and converting Ensembl pig IDs to Ensembl human IDs.

Abbreviations

MAF=minimum allele frequency, GO=gene ontology, RRL=reduced representation library, PE= paired-end, SE= single-end, PAR= pseudoautosomal region, NPAR=non-pseudoautosomal region, MCR=multi-copy region, HKA=

Hudson-Kreitman-Aguadé, SLA=swine leukocyte antigens, CYP= cytochrome P450 family, OR=olfactory receptor family, TR=T-cell receptors, INF=interferon family, TNF=tumor necrosis factors, IL=interleukins

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AEC and MPE analyzed data. ER, LF, SERO, YP, HJM, and MAMG provided analytical tools and support. LS and MCR provided material. AEC and MPE wrote the manuscript with help from the rest of authors. MPE conceived and coordinated research.

Acknowledgements

This paper is dedicated to all workers at the Iberian pig farm 'Dehesón del Encinar' that has remained active and with strict pedigree control since 1944. Due to the budget cuts, the regional government has decided to close down these historical facilities. *P. porcus* blood was kindly provided by Madrid's zoo. We thank Centro Nacional de Análisis Genómico (CNAG, Barcelona, Spain) for sequencing the *P. porcus*. Conservation of the Guadyerbas strain was 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). Work funded by grants AGL2010-14822 to MPE, CGL2009-09346 (MICINN, Spain) to SERO and by a Consolider grant from Spanish Ministry of Research, CSD2007-00036 "Centre for Research in Agrigenomics".

Additional Files

Additional File 1 - Variability inside MCR vs. variability of windows containing MCR outside MCR.

Additional File 2 - Gene ontology enrichment diagram of genes within MCR.

Additional File 3 - Genes within extreme selection tests' windows.

References

Alkan C., Coe B.P. & Eichler E.E. (2011) Genome structural variation discovery and genotyping. *Nat Rev Genet* **12**, 363-76.

Alves E., Ovilo C., Rodriguez M.C. & Silio L. (2003) Mitochondrial DNA sequence variation and phylogenetic relationships among Iberian pigs and other domestic and wild pig populations. *Anim Genet* **34**, 319-24.

Amaral A.J., Ferretti L., Megens H.J., Crooijmans R.P., Nie H., Ramos-Onsins S.E., Perez-Enciso M., Schook L.B. & Groenen M.A. (2011) Genome-wide footprints of pig domestication and selection revealed through massive parallel sequencing of pooled DNA. *PLoS ONE* **6**, e14782.

Amaral A.J., Megens H.J., Kerstens H.H., Heuven H.C., Dibbits B., Crooijmans R.P., den Dunnen J.T. & Groenen M.A. (2009) Application of massive parallel sequencing to whole genome SNP discovery in the porcine genome. *BMC Genomics* **10**, 374.

Burgos-Paz W., Souza C.A., Castello A., Mercade A., Okumura N., Sheremet'eva I.N., Huang L.S., Cho I.C., Paiva S.R., Ramos-Onsins S., *et al.* (2012) Worldwide genetic relationships of pigs as inferred from X chromosome SNPs. *Anim Genet*.

Carnero-Montoro E., Bonet L., Engelken J., Bielig T., Martinez-Florensa M., Lozano F. & Bosch E. (2011) Evolutionary and functional evidence for positive selection at the human CD5 immune receptor gene. *Mol Biol Evol* **29**, 811-23.

Clop A., Vidal O. & Amills M. (2012) Copy number variation in the genomes of domestic animals. *Anim Genet*.

Corominas J., Ramayo-Caldas Y., Castello A., Munoz M., Ibanez-Escriche N., Folch

J.M. & Ballester M. (2012) Evaluation of the porcine ACSL4 gene as a candidate gene for meat quality traits in pigs. *Anim Genet*.

Cutler D.J. & Jensen J.D. (2010) To pool, or not to pool? *Genetics* **186**, 41-3.

Charlesworth B.a.C., D. (2010) *Elements of Evolutionary Genetics*.

Chen K., Baxter T., Muir W.M., Groenen M.A. & Schook L.B. (2007) Genetic resources, genome mapping and evolutionary genomics of the pig (Sus scrofa). *Int J Biol Sci* **3**, 153-65.

Esteve-Codina A., Kofler R., Himmelbauer H., Ferretti L., Vivancos A.P., Groenen M.A., Folch J.M., Rodriguez M.C. & Perez-Enciso M. (2011) Partial short-read sequencing of a highly inbred Iberian pig and genomics inference thereof. *Heredity (Edinb)* **107**, 256-64.

Fang M., Larson G., Ribeiro H.S., Li N. & Andersson L. (2009) Contrasting mode of evolution at a coat color locus in wild and domestic pigs. *PLoS Genet* **5**, e1000341.

Fay J.C. & Wu C.I. (2000) Hitchhiking under positive Darwinian selection. *Genetics* **155**, 1405-13.

Fernandez-Figares I., Lachica M., Nieto R., Rivera-Ferre M.G. & Aguilera J.F. (2007) Serum profile of metabolites and hormones in obese (Iberian) and lean (Landrace) growing gilts fed balanced or lysine deficient diets. *Livestock Science* **110**, 73-81.

Ferretti L., Ramos-Onsins, SE., Pérez-Enciso, M. (2012) Population genomics from next generation sequencing of pooled lineages. *Molecular Ecology (submitted)*.

Fu Y.X. (1995) Statistical properties of segregating sites. *Theor Popul Biol* **48**, 172-97.

Futschik A. & Schlotterer C. The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics* **186**, 207-18.

Futschik A. & Schlotterer C. (2010) The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. *Genetics* **186**, 207-18.

George R.D., McVicker G., Diederich R., Ng S.B., MacKenzie A.P., Swanson W.J., Shendure J. & Thomas J.H. (2011) Trans genomic capture and sequencing of primate exomes reveals new targets of positive selection. *Genome Res* **21**, 1686-94.

Giuffra E., Kijas J.M., Amarger V., Carlborg O., Jeon J.T. & Andersson L. (2000) The origin of the domestic pig: independent domestication and subsequent introgression. *Genetics* **154**, 1785-91.

Groenen M.A.M., Archibald, A.L., Uenishi H. (2012) Pig genomes provide insight into porcine demography and evolution. (submitted).

Hach F., Hormozdiari F., Alkan C., Birol I., Eichler E.E. & Sahinalp S.C. (2010) mrsFAST: a cache-oblivious algorithm for short-read mapping. *Nat Methods* **7**, 576-7.

Haerty W., Jagadeeshan S., Kulathinal R.J., Wong A., Ravi Ram K., Sirot L.K., Levesque L., Artieri C.G., Wolfner M.F., Civetta A., *et al.* (2007) Evolution in the fast lane: rapidly evolving sex-related genes in Drosophila. *Genetics* **177**, 1321-35.

Huang W., Li L., Myers J.R. & Marth G.T. (2011) ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593-4.

Hudson R.R. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337-8.

Hudson R.R., Kreitman M. & Aguade M. (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**, 153-9.

Hutter P. (2007) Rapidly evolving Rab GTPase paralogs and reproductive isolation in Drosophila. *Adv Genet* **58**, 1-23.

Jiang S.Y. & Ramachandran S. (2006) Comparative and evolutionary analysis of genes encoding small GTPases and their activating proteins in eukaryotic genomes. *Physiol Genomics* **24**, 235-51.

Jiggins F.M. & Kim K.W. (2007) A screen for immunity genes evolving under positive selection in Drosophila. *J Evol Biol* **20**, 965-70.

Jongren M., Westander J., Natt D. & Jensen P. (2010) Brain gene expression in relation to fearfulness in female red junglefowl (Gallus gallus). *Genes Brain Behav* **9**, 751-8.

Kittawornrat A. & Zimmerman J.J. (2011) Toward a better understanding of pig behavior and pig welfare. *Anim Health Res Rev* **12**, 25-32.

Li H. & Durbin R. (2009) Fast and accurate short read alignment with Burrows-

Wheeler transform. *Bioinformatics* **25**, 1754-60.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G. & Durbin R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-9.

Li J., Li H., Jakobsson M., Li S., Sjodin P. & Lascoux M. (2012) Joint analysis of demography and selection in population genetics: where do we stand and where could we go? *Mol Ecol* **21**, 28-44.

Lopez-Bote C.J. (1998) Sustained utilization of the Iberian pig breed. *Meat Science* **49, Supplement 1**, S17-S27.

Manry J., Laval G., Patin E., Fornarino S., Itan Y., Fumagalli M., Sironi M., Tichit M., Bouchier C., Casanova J.L., *et al.* (2011) Evolutionary genetic dissection of human interferons. *J Exp Med* **208**, 2747-59.

McLaren W., Pritchard B., Rios D., Chen Y., Flicek P. & Cunningham F. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069-70.

Medina I., Carbonell J., Pulido L., Madeira S.C., Goetz S., Conesa A., Tarraga J., Pascual-Montano A., Nogales-Cadenas R., Santoyo J., *et al.* (2010) Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. *Nucleic Acids Res* **38**, W210-3.

Mercade A., Estelle J., Perez-Enciso M., Varona L., Silio L., Noguera J.L., Sanchez A. & Folch J.M. (2006) Characterization of the porcine acyl-CoA synthetase long-chain 4 gene and its association with growth and meat quality traits. *Anim Genet* **37**, 219-24.

O'Connell M.J. & McInerney J.O. (2005) Gamma chain receptor interleukins: evidence for positive selection driving the evolution of cell-to-cell communicators in the mammalian immune system. *J Mol Evol* **61**, 608-19.

Przeworski M. (2002) The signature of positive selection at randomly chosen loci. *Genetics* **160**, 1179-89.

Raineri E., Ferretti, L., Esteve-Codina A., Nevado B., Heath S., Pérez-Enciso M. (2012) SNP calling and computing allele frequency by sequencing pooled samples. *BMC Bioinformatics (submitted)*.

Ramos-Onsins S.E., Puerma E., Balana-Alcaide D., Salguero D. & Aguade M. (2008)

Multilocus analysis of variation using a large empirical data set: phenylpropanoid pathway genes in Arabidopsis thaliana. *Mol Ecol* **17**, 1211-23.

Sabeti P.C., Schaffner S.F., Fry B., Lohmueller J., Varilly P., Shamovsky O., Palma A., Mikkelsen T.S., Altshuler D. & Lander E.S. (2006) Positive natural selection in the human lineage. *Science* **312**, 1614-20.

Sudmant P.H., Kitzman J.O., Antonacci F., Alkan C., Malig M., Tsalenko A., Sampas N., Bruhn L., Shendure J. & Eichler E.E. Diversity of human copy number variation and multicopy genes. *Science* **330**, 641-6.

Sudmant P.H., Kitzman J.O., Antonacci F., Alkan C., Malig M., Tsalenko A., Sampas N., Bruhn L., Shendure J. & Eichler E.E. (2010) Diversity of human copy number variation and multicopy genes. *Science* **330**, 641-6.

Tajima F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585-95.

Tennessen J.A., Madeoy J. & Akey J.M. (2010) Signatures of positive selection apparent in a small sample of human exomes. *Genome Res* **20**, 1327-34.

Tong P., Prendergast J.G., Lohan A.J., Farrington S.M., Cronin S., Friel N., Bradley D.G., Hardiman O., Evans A., Wilson J.F., *et al.* (2010) Sequencing and analysis of an Irish human genome. *Genome Biol* **11**, R91.

Toro M., Rodrigáñez J., Silió L. & Rodríguez M. (2000) Genealogical analysis of a closed herd of black hairless Iberian pigs. *Conservation Biol* **14**, 1843-51.

van Binsbergen E. (2011) Origins and breakpoint analyses of copy number variations: up close and personal. *Cytogenet Genome Res* **135**, 271-6.

Zelus D., Robinson-Rechavi M., Delacre M., Auriault C. & Laudet V. (2000) Fast evolution of interleukin-2 in mammals and positive selection in ruminants. *J Mol Evol* **51**, 234-44.

Zhang J. & Nei M. (2000) Positive selection in the evolution of mammalian interleukin-2 genes. *Mol Biol Evol* **17**, 1413-6.

CHAPTER 6

EXPLORING THE GONAD TRANSCRIPTOME OF TWO EXTREME MALE PIGS WITH RNA-SEQ

Anna Esteve-Codina

Robert Kofler

Nicola Palmieri

Giovanni Bussotti

Cedric Notredame

Miguel Pérez-Enciso

BMC Genomics 2011

Abstract

Background

Although RNA-seq greatly advances our understanding of complex transcriptome landscapes, such as those found in mammals, complete RNA-seq studies in livestock and in particular in the pig are still lacking. Here, we used high-throughput RNA sequencing to gain insight into the characterization of the poly-A RNA fraction expressed in pig male gonads. An expression analysis comparing different mapping approaches and detection of allele specific expression is also discussed in this study.

Results

By sequencing testicle mRNA of two phenotypically extreme pigs, one Iberian and one Large White, we identified hundreds of unannotated protein-coding genes (PcGs) in intergenic regions, some of them presenting orthology with closely related species. Interestingly, we also detected 2047 putative long non-coding RNA (lncRNA), including 469 with human homologues. Two methods, DEGseq and Cufflinks, were used for analyzing expression. DEGseq identified 15% less expressed genes than Cufflinks, because DEGseq utilizes only unambiguously mapped reads. Moreover, a large fraction of the transcriptome is made up of transposable elements (14500 elements encountered), as has been reported in previous studies. Gene expression results between microarray and RNA-seq technologies were relatively well correlated (r = 0.71 across individuals). Differentially expressed genes between Large White and Iberian showed a significant overrepresentation of gamete production and lipid metabolism gene ontology categories. Finally, allelic imbalance was detected in ~ 4% of heterozygous sites.

Conclusions

RNA-seq is a powerful tool to gain insight into complex transcriptomes. In addition to uncovering many unnanotated genes, our study allowed us to determine that a considerable fraction is made up of long non-coding transcripts and transposable elements. Their biological roles remain to be determined in future studies. In terms of differences in expression between Large White and Iberian pigs, these were largest for genes involved in spermatogenesis and lipid metabolism, which is consistent with phenotypic extreme differences in prolificacy and fat deposition between these two breeds.

Background

Understanding the mammal transcriptome architecture has proven to be a complex task (Gustincich *et al.* 2006; Jacquier 2009; Guttman *et al.* 2010; Lindberg & Lundeberg 2010). The advent of high throughput sequencing technologies, such as RNA-seq, has, yet, substantially improved our comprehension of its structure and expression patterns. By deep sequencing the poly-A RNA fraction, it is possible not only to better characterize isoforms from known genes (e.g., identifying novel exons, new transcription start sites and alternative polyadenylation sites), but also to improve the annotation by discovering novel predicted coding genes and polyadenylated processed transcripts such as long intergenic non-coding RNAs (Mortazavi *et al.* 2008). Although several surveys of the transcriptome from different tissues have been conducted in humans and model species (Ferraz *et al.* 2008; Wang *et al.* 2008; Trapnell *et al.* 2009; Gan *et al.* 2010; McManus *et al.* 2010a; Wang *et al.* 2010b; Wen *et al.* 2010; Bottomly *et al.* 2011; Daines *et al.* 2011; Graveley *et al.* 2011; Nicolae *et al.* 2011; Toung *et al.* 2011), our knowledge of livestock species remains limited. For instance, the relation between extreme phenotypic differences and their transcriptome patterns is poorly studied. The transcriptome of livestock species is, by comparison to model species, much less known despite its economic and social interest.

In this study, we used high-throughput transcriptome sequencing in two pigs from extreme breeds. Our aim was to discover and characterize novel expressed transcripts and to identify differentially expressed genes that may explain some of the phenotypic variation. We sequenced the male gonad transcriptome of a Large White and an Iberian pig, two highly divergent phenotypic breeds in terms of production traits, e.g., growth, fatness and reproductive performance. To limit the effect of enviromental influences on gene expression pattern, both pigs were housed and fed with the same conditions and were prepubescent at slaughter time. Furthermore we compared the results obtained with RNA-seq with microarray data published in a previous study (Herai & Yamagishi 2010). Finally, we also identified polymorphic sites and genes that potentially showed allele specific expression.

Results and Discussion

Mapping

We obtained about 60 M of 50 bp paired-end reads from one lane of an Illumina GAIIx machine, about 30 M was derived from each sample (Data are archived at NCBI Sequence Read Archive (SRA) under Accession SRP008516). After ambiguous mapping (allowing for multi-hits) with Tophat [17] a total of 20 M reads for each sample were mapped against the reference pig genome (assembly 9), although only 10 M were classified as proper pairs. The rest (4 M) fell into either one of these categories: reads without a mapped mate pair, mate is mapped on the same strand or mates overlap. The most likely explanations of the large amount of improperly mapped reads are the poor quality of the current pig genome assembly and the stringency of the version of Tophat used here, as this version does not allow gaps for the mapping. In addition, any situation where the distance between the mates is larger than the confidence interval of the insert size distribution could be interpreted as trans-splicing events (McManus *et al.* 2010b), structural variants or simply mapping artifacts (Trapnell *et al.* 2010). The total number fragments mapped with unambiguous mapping (1 hit per read) were 14 M for each sample; out of these, 7 M were classified as proper pairs. A comparison between ambiguous versus unambiguous mapping results obtained with Tophat is shown in Additional file 1.

Annotation of reads and transcripts assembly

To calculate the proportion of reads mapping to annotated exons, we run S-MART (see methods). Surprisingly, with a minimum overlapping of 1 nucleotide, less than half of the reads (44.1%) mapped to annotated exons; a figure that drops even further (32.9%) when considering a minimum overlapping of 50 bp (the total read length). The rest of reads mapped to annotated introns (18.7%), or either 1 kb 5'upstream or 3'downstream of the annotated gene (26.6%) (Table 1). The poor quality of the annotation of the pig genome probably explains why a majority of the mapped reads (55.9%) do not overlap with any known exons.

Table 1. Summary of reads' annotation

	Large White	**Iberian**
Exons	9238572	9141162
Introns	3833320	3943866
5' Upstream or 3'Downstream	5383546	5708057

Number of reads with at least one overlapping nucleotide mapping to either exons, introns or within 1kb of gene boundaries.

Moreover, after assembling the short reads into transcripts by Cufflinks (Stanke *et al.* 2008), only 1.2% of them matched exactly with annotated exons. The remaining reads were classified as intergenic transcripts (36.1%), intron retention events (35.6%), contained in known isoforms (12.5%), pre-mRNA molecules (6.2%), polymerase run-on fragments (3.6%), putative novel isoforms of known genes (2.9%) and others (Table 2). These results unfortunately underline the incompleteness of the current annotation of the pig transcriptome and of its complexity.

Table 2. Transcripts assembly

	=	c	e	i	j	o	p	u	Total
Large White	2178	22243	11328	67989	5557	3866	6775	72288	192224
(%)	1.13	11.57	5.89	35.37	2.89	2.01	3.52	37.61	100
Iberian	2000	21580	10349	57623	4530	3341	5752	55617	160792
(%)	1.24	13.42	6.44	35.84	2.82	2.08	3.58	34.59	100

The number of transcripts assembled with Cufflinks and the percentage they represent in each sample. The high number of assembled transcripts is probably an artifact due to truncated Cufflinks assemblies. Class codes described by Cuffcompare: "=" Exactly equal to the reference annotation, "c " Contained in the reference, annotation, "e" possible pre-mRNA molecule, "i " An exon falling into a intron of the reference, "j " New isoforms, "o" Unknown, generic overlap with reference, "p" Possible polymerase run-on fragment, "u" Unknown, intergenic transcript.

Annotating orthologs

A total of 4,124 novel transcripts (the real number of transcript units may be smaller, as less abundant transcripts receive less complete sequencing coverage resulting in numerous transfrags) were identified in intergenic regions (see methods). To investigate which of these transcripts actually encode a protein, we used Augustus (Cooper *et al.* 2003) and found 714 novel putative proteins. We identified homologous DNA sequences (see methods) in *Bos taurus* and *Homo sapiens* genomes for most (413) of these novel proteins: 362 were orthologs with both cow and human, 20 with human only and 31 with the cow genome only. This result is consistent with *Bos taurus* being closer to *Sus scrofa* than human (Shi *et al.* 2007). Interestingly, when we looked for homologous DNA regions within the *Sus scrofa* genome, 53 paralogous regions were detected (51 duplications and 2 present in three copies).

To find out whether the predicted proteins from the homologous regions were already annotated, we ran BLASTP against the *Homo sapiens*, *Bos taurus* and *Sus scrofa* protein databases (http://www.ensembl.org/info/data/ftp/index.html). Overall, we identified 38 novel computationally predicted and 344 known proteins for the human, 15 novel predicted and 378 known proteins for the cow and 653 novel predicted and 89 known proteins for the pig. The novel computationally predicted proteins found in the pig are now experimentally confirmed by RNA-seq. See Additional file 2 for the coordinates of orthologous and paralogous genes.

Transposable elements

As many previous studies reported high activity of transposable elements (TE) in germlines (Branciforte & Martin 1994; Garcia-Perez *et al.* 2007; Zamudio & Bourc'his 2010), we ran RepeatMasker to identify repetitive elements in the pig genome and in the transcriptome of the testicles. The fraction of transposable elements expressed in male gonads (SINEs, LINEs, LTR and DNA elements), compared to the total number detected in the pig genome, is less than 3%. However, approximately 20% of the expressed transcripts units harbor at least 1 transposable element (8% of the bp sequenced). The type of TE being more active in both breeds, in terms of number of elements expressed divided by the total number present in the genome, is DNA transposons, but accounting just for the number of elements expressed is SINE family for Large White and the LINE family for Iberian. LINE1 elements also have been reported to contribute to the transcriptome in human somatic cells (Rangwala *et al.* 2009). It is interesting to mention that 16% of protein-coding transcripts contain transposable elements in their sequence and they are transcribed in the same transcript unit. Apart from these interspersed repeats, hundreds of small RNA (tRNA, snRNA and rRNA) and thousands of simple and low complexity repeats were also identified in the transcriptome. The presence of non-polyadenylated RNA could be a remaining contamination as they are highly expressed molecules and difficult to remove completely. Another possible explanation is the presence of small functional RNAs embedded in the introns of polyadenylated molecules of pre-mRNA (Zamboni *et al.* 2009; Donath 2010). Detailed results of the repetitive elements detection with Repeatmasker are shown in Additional file 3a and 3b.

LncRNA annotation

In order to define a set of putative lncRNAs in the pig transcriptome, we applied several filtering criteria. Using the procedure in (Orom *et al.* 2010) for the definition of lncRNA in humans, we excluded all transcripts mapping within 1 kb of an annotated protein-coding gene in the pig genome. This makes it less likely to consider the 5' or the 3' UTR of a protein-coding gene as a non-coding RNA. Yet, this filtering may not be stringent enough when dealing with insufficiently annotated genomes. For that reason, we further refined our analysis by excluding all transcripts coding for a complete proteins (according to Augustus). A third filter was applied by removing all transcripts having a hit against NR (BlastX), against Pfam (RPS-Blast) or against Rfam (Gardner *et al.* 2009) (web-site batch search). The final filter was applied mapping all the resulting transcripts onto the human genome (the best annotated mammalian genome), and removing any transcript strongly overlapping with a protein-coding gene. The result is a dataset made of 2047 transcripts and referred in the rest of this text as the lncRNAs.

The main problem when dealing with ncRNAs is to distinguish between spurious transcripts resulting from promoter leakiness and biologically functional transcripts. In order to do so, we assessed the level of evolutionary conservation of each lncRNA across the 18 available mammalian genomes. As shown in previous work, this conservation cannot be directly inferred from reference multiple genome alignment (Orom *et al.* 2011). We therefore used a standard gene discovery strategy that relies on a combination of BlastN (version 2.0 MP, Gish, unpublished) and exonerate (Slater & Birney 2005). BlastN allows a rough identification of the location of each transcript in the considered genome while exonerate is used to precisely delineate the corresponding gene structure. We only considered as potential homologues hits for which exonerate alignments yield more than 70% coverage with the pig transcript. The results of this extensive homology based analysis are displayed on a heatmap (Figure 1).

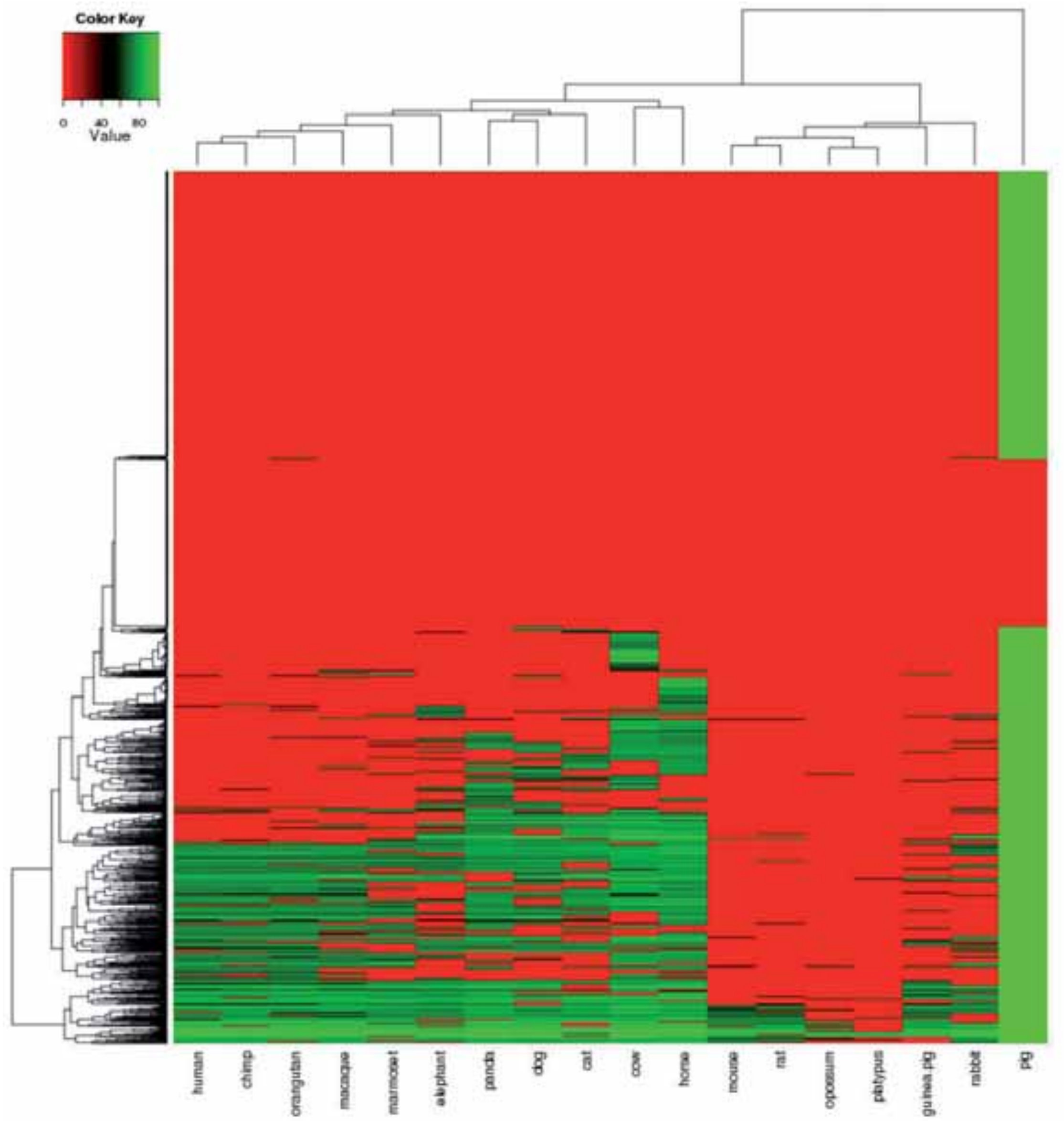

Figure 1. LncRNAs mammal conservation. The heatmap recapitulates the screening result of the new discovered 2047 pig lncRNAs versus eighteen mammal genomes. The columns represent the mammal genomes while the rows indicate the query lncRNAs. The spots indicate the result of the search of each pig lncRNA versus the different genomes. Green spots represent hits having high similarity scores. Black spots indicate low similarity scores. Red spots indicate that no homolog was detected.

In the context of this analysis, we managed to map 986 transcripts in at least one other mammal species. A sizeable number of transcripts (391) were excluded because they contain pig repeats (red block in the pig column on Figure 1). The rest of the transcripts roughly fall in three categories. The first one is made up of genes apparently conserved across most tested mammals, including human. These make up a group of 469 genes (Figure 2). In this group, 131 transcripts map onto human genomic regions with no annotation. The rest either overlap with protein-coding genes (316), with known lncRNAs (15) or pseudogenes (7). It is important to note that an overlap with a PcG is not incompatible with a transcript being a lncRNA. The second category is made up of a group of 322 transcripts conserved among Artiodactyla (pig and cow) but not found in human. The last group encompasses all the putative lncRNAs for which no homologue was found in other species. While these may be pig specific, further analysis would be needed to confirm their biological relevance (for instance by testing their differential expression across tissues).

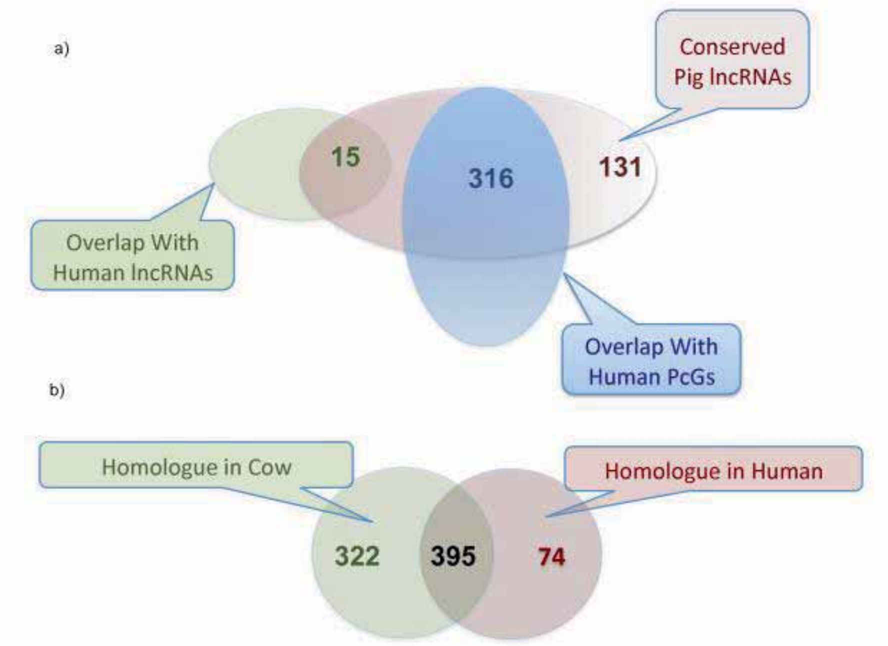

Figure 2. Ven diagrams of the predicted homologues in human and cow. a) 469 pig lncRNA presented homology with human. 15 pig lncRNA overlap with human lncRNA, 316 overlap with human PcGs annotations and 131 lncRNA presented homology with unannotated human DNA regions. **b)** Comparison of lncRNAs having a homolog in human and in cow.

It is worth mentioning that the transcripts thus identified have a gene structure significantly different from their human counterparts. 97% are single exon genes and 2.5% bi-exonic, a figure significantly different from human where a much higher portion is bi-exonic. This finding may simply reflect insufficient coverage in the RNA-seq experiment resulting in truncated cufflinks models and thus should not be taken, so far, as strong evidence of distinct lncRNA organization between species.

It is in agreement with our observation that the lncRNA we observe in pig are roughly half the size of those reported in human (456 vs 925). As a consequence, the number of independent transcripts reported here is quite likely to be an over estimation.

Gene expression analysis

In total, 12,816 annotated genes were expressed in gonads. Less than 1% of these genes were expressed more than 10000 FPKM; around 5% were expressed between 1000 -10000 FPKM, 50% between 10-1000 FPKM, 40% between 10-100 FPKM and 3% between 1-10 FPKM (Additional file 4). The rest were expressed below 1 FPKM. The maximum expression level of an annotated gene was 61,000 and 73,000 FPKM in Large White and Iberian, respectively. The gene ontologies of the 100 most expressed genes (mainly ribosomal proteins and heat-shock proteins) in both samples were related with transcription and translation, protein folding, lipid and cholesterol metabolism (apoproteins), induction of apoptosis and response to stress. These results are consistent with those observed in other mammalian species with RNA-seq (Ramskold *et al.* 2009).

The correlation of gene expression levels between both samples (Large White vs. Iberian) was very high (r = 0.85), which suggests that a large fraction of the transcriptome is conserved across individuals. This is consistent with our previous results which showed that the largest source of variability was tissue rather than sex or breed (Ferraz *et al.* 2008).

Gene expression was quantified using two different approaches: DEGseq (Wang *et al.* 2010a), which uses raw fragment counts per gene as a measure of expression, and Cufflinks (Trapnell *et al.* 2010), that uses an estimation of fragments per kilobase of exon per million reads mapped (FPKM). DEGseq's protocol recommends working only with the uniquely mapped fragments, whereas Cufflinks can deal with multiple mappable fragments. In this study, the correlation of the log2 of the fold change between both methods was 0.96 when discarding infinite values and taking expressed genes in both methods into account (see Figure 3a). Nevertheless, fragments mapping to homologous genes, which constitute 15%-20% of the expressed genes, are lost when considering fragments that map only once in the transcriptome, so it is arguable how to actually compare expression levels measured with these two programs.

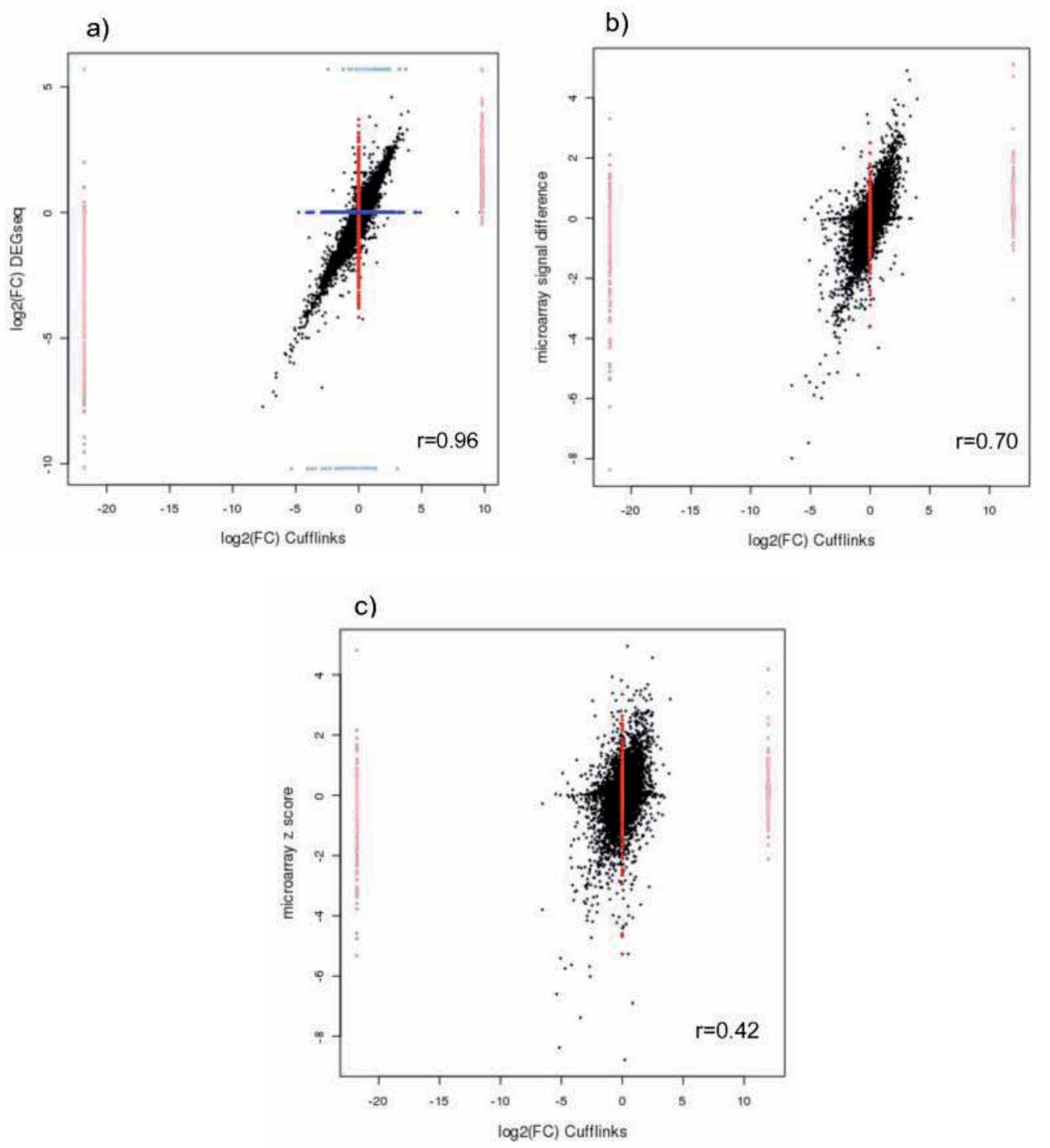

Figure 3. Measuring gene expression. a) DEGseq vs. Cufflinks estimates of log2 fold changes between Large White and Iberian expressed genes. Blue and red points correspond to not expressed genes in microarrays and Cufflinks, respectively. Light blue and light red points correspond to microarray and Cufflinks infinite values. b) Microrray vs. RNA-Seq individual measurements. The microarray data correspond to signal intensity difference between Large

White and Iberian, whereas the RNA-Seq measurement is the log2 fold change as obtained from Cufflinks. c) Microarray breed z-score values vs. RNA-Seq log2 fold change. The Pearson's correlations (r) were significant in each case (Pv < 2.2 × 10^{-16}) and calculated considering only expressed genes and no infinite values.

We also compared the RNA-seq expression results with Affymetrix microarray data obtained in a previous study (Ferraz *et al.* 2008). As many microarray probes may map to the same gene, the average probe value per gene was calculated. A total of 9,112 Ensembl ID genes could be retrieved from microarray probes data for RNA-seq comparisons. The correlation between the individual microarray signal intensity difference and the log2 of the fold change from RNAseq was quite high (r = 0.71, see Figure 3b). From the microarray study, we also had a Bayesian standardized breed score (z-score) available for each gene. When comparing the microarray breed z-score and the log2 of the fold change in RNA-seq, the correlation was also moderately high (Pearson correlation r = 0.46, see Figure 3c).

Differential expression analysis

We compared the performance of Cufflinks and DEGseq to detect differential expression between both samples (P < 0.001 and fold change > 2). Cufflinks identified 2,907 differentially expressed genes with multiple mappable fragments and DEGseq 2,330 with uniquely mapped fragments; there was a reasonable agreement between softwares, 1,830 genes (Figure 4, top). But, to be more conservative, and to try to get only differential expression due to breeds and not merely to stochastic reasons, we extracted differentially expressed genes from breed effects data, with absolute z score threshold > 1.65. Then we selected the intersection of RNA-Seq (Cufflinks) and microarrays reducing the number of differentially expressed genes to 256 (Figure 4, bottom). Out of these, 147 genes were over expressed in Large White and 109 in Iberian. Among differentially expressed genes, spermatogenesis, response to steroid hormone stimulus and sensory organ development were significantly over-represented children gene ontologies ($P < 10^{-3}$). Doing the same analysis but considering the GOslim of the pig described in the methods section, we obtained an enrichment of reproduction, developmental process and fatty acid metabolic process parental gene ontologies ($P < 10^{-3}$). Interestingly, among the significant KEGG-pathways represented, we found many differentially expressed genes in the PPAR signaling pathway, which is involved in lipid metabolism and, specifically, it has been shown to have a role in mice gonads fat deposition (Tsai *et al.* 2009).

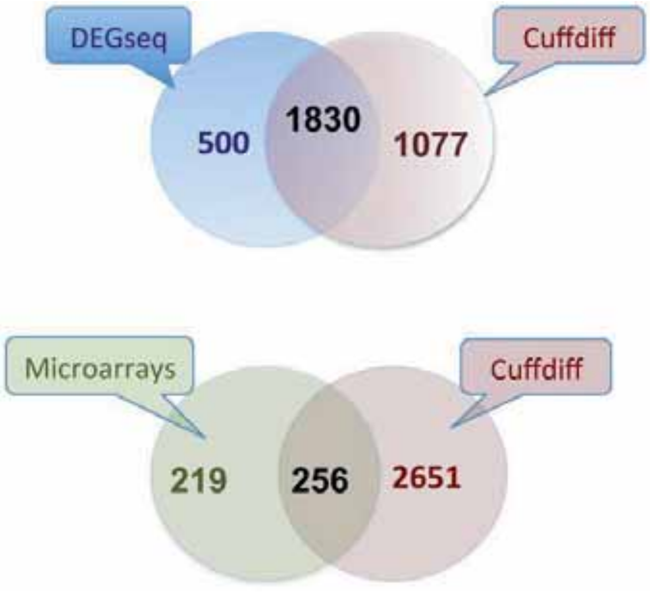

Figure 4. Overlapping of differentially expressed genes. Top: Differentially expressed genes identified by DEGseq and Cufflinks. Bottom: Differentially expressed genes identified by microarrays (breed z-scores) and RNA-Seq (Cufflinks).

Expression differences of coding and non-coding genes

We also compared the expression level of the annotated coding genes, novel coding genes, lncRNA and transcripts containing at least one transposable element (see Figure 5a). The median expression level of annotated coding genes (230.1 FPKM) was slightly lower than of the novel-coding genes (258.0 FPKM). The range of expression levels of the annotated coding genes is, however, broader than that of the novel coding. We were able to detect annotated coding genes with very low expression levels, which highlight that fact that providing the reference gene models; it is easier to detect genes even at low coverage. Simultaneously, the expression median of transcripts units with at least one TE

(111.6 FPKM) and lncRNAs (107.8 FPKM) is more than 50% lower than those of coding regions. As non-coding transcripts are probably involved in gene regulation, less number of copies is needed (Orom *et al.* 2011). The annotated coding genes are on average longer than the novel coding (Figure 5b). This may be due to several reasons, first a higher coverage is needed to fully assemble a novel gene, but, it is has been also described than novel genes tend to be shorter than annotated ones. Overall we found that the average transcript length for protein-coding gene is 1578 bp, roughly half the size of transcripts in the human transcriptome (2982 bp). Interestingly, we observed a similar ratio when comparing the average size of lncRNAs in our experiment (456 bp) with that observed in human (925 bp). This fairly constant ratio suggests a homogenous bias, most likely the result of a lack of connecting paths between exons of the same transcript unit.

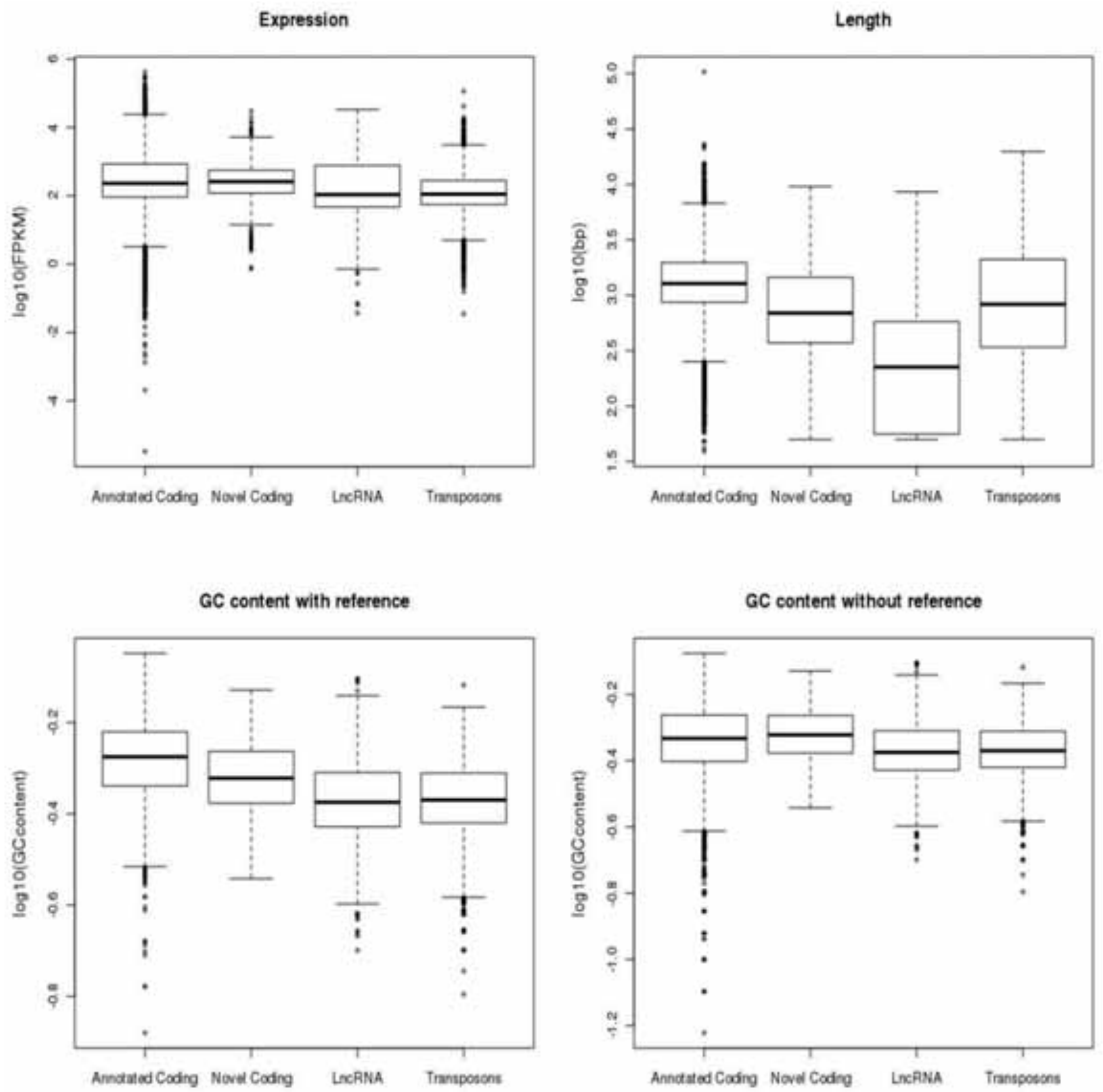

Figure 5. Expression levels according to annotation. a) Boxplots of expression level (log10 FKPM) for annotated coding genes, novel coding genes, lincRNA and transcripts with TE. The black line represents the median. b) Boxplots of the transcript unit length in base pairs (log10). c) Boxplots of the GC content (log10) using the reference annotation for transcriptome assembly. d) Boxplots of the GC content (log10) without using the reference annotation.

The GC content median of the coding genes (annotated 0.46 and novel 0.47) was the same but higher than the lncRNA (0.42) and transcripts harboring at least one TE (0.42) because coding genes tend to be rich in GC (Arhondakis *et al.* 2004). Important to notice is the fact that GC content of annotated genes differs depending on whether we provide to Cufflinks the reference gene annotations (Figure 5c) or not (Figure 5d). In the former, the GC content is much higher (0.53) than the latter (0.46), pointing to a possible bias towards AT during Illumina library preparation and sequencing workflow. Recently, a new amplification protocol has been published that solves this problem (Aird *et al.* 2011).

SNP identification

We divided the SNPs in two classes, fixed, i.e. differences with respect to the assembly, a Duroc pig, and segregating when the individual was heterozygous. The number of SNPs found per bp sequenced is shown in Table 3. In autosomes, approximately the same amount of fixed SNP with respect to the Duroc genome reference is found in both breeds, but around 30% less divergence is found in Iberian on × chromosome. Regarding the segregating SNP, in autosomes, we found 30% less variability in Iberian than in Large White and almost 50% less variability in the × chromosome. This result is in agreement with the high inbreeding level of the Iberian strains. Fixed SNP and segregating SNP annotation is shown in Table 4 introns and 3' downstream regions of annotated genes were the most polymorphic, a result of less evolutive constrains than exonic and 5'upstream regions of the genome; 3'UTR was also more variable than 5'UTR regions. As expected, more SNP were synonymous than non-synonymous in CDS.

Table 3. SNP statistics

	Fixed	Segregating	Total
	(SNP/kb)	(SNP/kb)	(SNP/kb)
Large White	29558 (0.64)	11230 (0.24)	40788 (0.88)
Iberian	25668 (0.59)	7552 (0.17)	33220 (0.76)

Number of fixed SNP with respect to the reference genome and number of segregating SNP within each breed. Within brackets, the number of SNP per kb assembled.

Table 4. SNP annotation

	Fixed		Segregating	
	Large White	Iberian	Large White	Iberian
Synonymous coding	2083	1727	1352	757
Non synonymous coding	1073	910	852	494
Synonymous coding	2083	1727	1352	757
Non synonymous coding	1073	910	852	494
5' UTR	150	77	73	39
3' UTR	1187	1029	1004	622
Stop lost	3	1	0	4
Stop gained	1	3	7	9
Intronic	15101	13020	3388	2515
5'Upstream	2983	2588	1176	628
3'Downstream	9440	8549	5579	3831
Splice site	311	259	91	86
Non synonymous coding	1073	910	852	494
Within non coding gene	262	268	14	9
Intergenic	5847	5042	1539	1076

The number of SNP is degenerated; each SNP can have more than one annotation.

Allele specific expression

A beta binomial model was applied to detect allele specific expression ASE (see methods). A total of 428 SNP (3.8%) with average coverage of 55 × and 338 SNP (4.5%) with average coverage 121 × showed allelic imbalance in Large White and Iberian samples, respectively. Coordinates and annotation of SNPs with significant results are listed in Additional file 5. Figure 6 shows the relation between coverage and the posterior mean of allele specific expression *p* (see methods). Figure 6a indicates how, although very extreme values of *p* are always significant, intermediate values (*p* between 0.3 - 0.4 and 0.6 - 0.7 approximately) are significant only if enough coverage exists. This is a result of how the prior (*p* = 0.5) is dominated by empirical evidence as data increases. Figure 6b was plotted to show that an increased coverage does not result in an average higher ASE and therefore significance is not a statistical artifact. Further, we did not observe either any consistent higher expression of the reference vs. the alternative allele (results not shown) and therefore it is not an alignment artifact either.

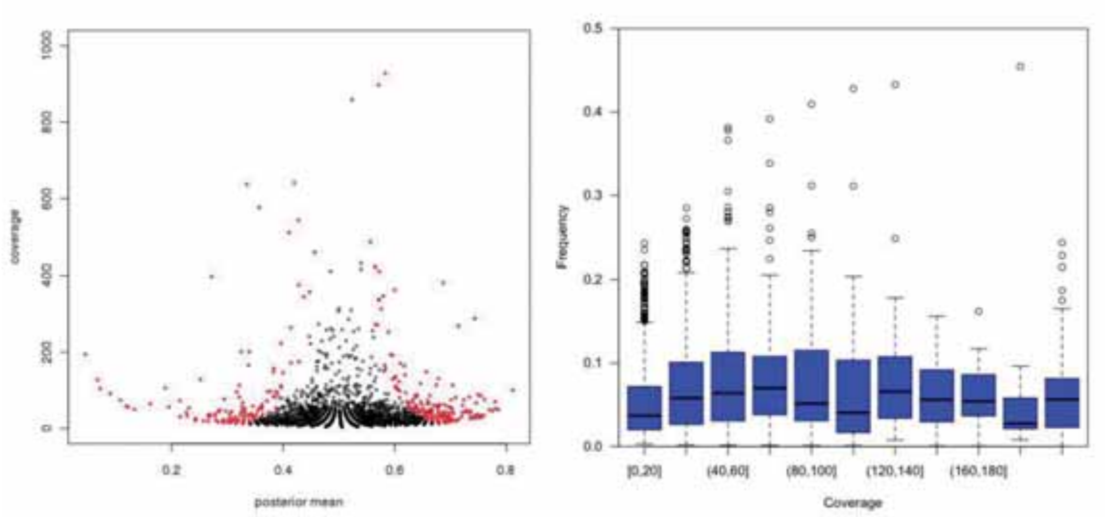

Figure 6. Allele specific expression. a) Coverage versus posterior mean of allele transcription rate (*p*); each point represents a SNP; red points are SNP showing significant ASE and black points are SNPs with no significant ASE. b) Barplot of coverage versus absolute value of *p*. It can be seen that there was not a consistent relation between ASE and coverage.

Several SNP with significant ASE are located contiguously within intergenic regions, suggesting the presence of putative functional units not yet annotated in the pig genome. There were not many genes with ASE shared between the two samples, likely due to different genotypes at the regulatory motif of the two breeds. There were only 22 common SNPs exhibiting ASE in both animals, but in three of the SNPs we observed over expression of different alleles in each breed. Logically, these results should be taken as statistical evidence, genotyping or sequencing the cis-regulatory motives and linkage disequilibrium information are, however, needed to confirm whether these SNPs show genuine ASE.

General Discussion

We present the first, to our knowledge, comprehensive exploration of the pig gonad transcriptome carried out with RNA-seq, a technology that offers critical advantages over microarray. Importantly, RNA-seq allows us to improve dramatically the annotation of the species and the discovery of new splicing events. Here, we have confirmed that a large part of the transcription effort in the cell is spent on TE sequences. A recent RNA-seq study in human and primate brain transcriptomes also found high proportion of reads mapping to repetitive elements, mainly from the Alu family (Xu *et al.* 2010). Previous works in mice also indicated high expression of TE in germlines. The number of TE is probably an over-estimation as we did an ambiguous mapping reporting only the best alignments. On the other side, Cufflinks down weights the expression level taking into account mapping uncertainty (Trapnell *et al.* 2010).

Unfortunately, we also confirm that current porcine annotation is incomplete, as evidenced by read mapping annotation: more than 50% of the fragments do not map to annotated exons. The fact that many reads map to introns could be explained either by intron retention (new isoforms) or pre-mRNA presence. Reads mapping outside the boundaries of annotated genes could be explained either by polymerase run-on fragments or a bad annotation of the gene endings. Many intergenic reads have been mapped to putative novel coding transcripts, some of them presenting orthology with related species. The poor status of the annotation is confirmed by the presence of 104 highly conserved transcripts, that would have been annotated as lncRNAs if we had only considered the pig annotation, but whose homologues in human show a perfect overlap with protein-coding genes.

Given that we had previously analyzed the transcriptome of a wider collection of pig and tissues with Affymetrix microarrays (Ferraz *et al.* 2008), we were able to compare both technologies. The correlation within individuals was rather high (r = 0.71) and comparable to other reported studies (Marioni *et al.* 2008; Bloom *et* *al.* 2009; Fu *et al.* 2009; Bradford *et al.* 2010; Toung *et al.* 2011). Furthermore, the correlation of expression between Iberian and Large White obtained with microarray (employing all animals) and the individuals (obtained with RNAseq) was also moderately high (r = 0.46), suggesting that transcriptome patterns are relatively stable. Among the most differentially expressed genes, those involved in spermatogenesis and lipid metabolism are over-represented, which may be a result of targeted tissue selection. It is noteworthy that Large White and Iberian breeds are phenotypically extreme for both reproduction and fat deposition traits so these data would suggest a correlated effect on the regulation of genes involved in these traits.

In general, Cufflinks has a better performance to map fragments to genes or isoforms that are physically overlapping or very similar in sequence, as it uses a statistical model to deal with multiply mapping fragments. DEGseq works with uniquely mapped reads, thus underestimating gene expression levels of homologous genes but also discarding those reads belonging to two overlapping genes; a bias in expression level is thus introduced in these cases. The algorithm behind Cufflinks is rather naïve, though. Recently, new approaches that implement improved algorithms to deal with ambiguously mapped reads data and avoid bias in downstream analysis have been published (Marioni *et al.* 2008; Pasaniuc *et al.* 2011).

Although not the main purpose of the work, we also found a lower rate of heterozygosity in Iberian than in the Large White animal, in agreement with the fact that Iberian pigs are normally inbred. Finally, we also explored ASE, a topic that has received a renewed interest recently. In this study, ~ 4% of the segregating SNP presented allelic imbalance. From these, around 40% were located inside annotated genes, the rest were located in blocks in intergenic regions, pointing to putative functional transcripts. To be able to confirm ASE, more animals should be tested because the majority of the SNPs with ASE were not common between Large White and Iberian pigs.

Conclusions

We provide a complete survey of the pig male gonad transcriptome and identified many novel elements. However, to further improve the annotation of the pig genome, a large effort from the community will be necessary by sequencing more tissues at different developmental stages. In order to detect novel splicing events and to reconstruct novel isoforms, RNA-seq studies with very high coverage are required. Here, we also have shed some light on the dark matter of the transcriptome; in particular, we remark the discovery of novel long non-coding transcripts and the fact that TE expression seems to take a large fraction of the transcriptome. Their precise roles need to be elucidated in future studies. We also show that correlation between microarray and RNAseq expression data are reasonably high (linear correlation r = 0.71). Finally, Large White and Iberian pigs seem to have diverged most for genes involved in spermatogenesis and lipid metabolism, not only in terms of gene expression but also phenotypically. Interestingly, it is well known that genes related to gametogenesis are subject often to a positive selection rate (Jagadeeshan & Singh 2005; Haerty *et al.* 2007). More work is required to investigate whether the differences in expression in these genes are adaptive.

Methods

Animal material

Animal material is fully described in (Perez-Enciso *et al.* 2009). The two animals were housed and slaughtered simultaneously. Animals were prepubescent, three months of age, and weights were 45.0 and 30.1 kg for Large White and Iberian animals, respectively.

Library preparation

Total RNA from gonads was extracted as described in (Perez-Enciso *et al.* 2009). Briefly, Total RNA was extracted from 100 mg tissue using the RiboPure™ kit (Ambion, Austin, USA). RNA integrity was assessed by Agilent Bioanalyser 2100 and RNA Nano 6000 Labchip kit (Agilent Technologies, Palo Alto, USA). Due to high variation in concentrations of the total RNA obtained in different tissues, all samples were concentrated and cleaned using the RNAeasy MiniElute Cleanup kit (Qiagen, Basel, Switzerland) obtaining final concentrations between 500 and 1000 ng/µl. Sequencing libraries were produced using the Illumina mRNA-Seq sample preparation kit, following the manufacturer's instructions. Briefly, 4 µg of total RNA were used as input for poly-A+ selection, followed by metal-catalyzed fragmentation of the selected mRNA (peak of size distribution at approx. 240 nt). After reverse transcription to cDNA using random hexamer primers, we performed end-repair and A-tailing of the double stranded cDNA. Large White and Iberian cDNA were ligated to indexed pairs of adapters, see Additional file 6. The cDNA was size selected on a 2% agarose gel, and fragments corresponding to an insert size of 237 nucleotides were excised from the gel. The DNA was recovered from the gel slice using QIAquick gel extraction kit (Qiagen). Therafter, the libraries were amplified in 15 cycles of PCR using primers Illumina 1.0 and Illumina 2.0. The libraries were quantified using Taqman, and pooled at a concentration of 10 pM. We performed paired-end sequencing of the libraries on the Genome Analyzer IIx using Illumina v4 sequencing chemistry.

Reads annotation

S-MART (http://urgi.versailles.inra.fr/Tools/S-MART) was used to count the number of reads mapping to exons, introns and 1 kb upstream/downstream of the annotated genes. A minimum overlapping of 1 nucleotide was chosen to declare an overlap.

Mapping, Assembling and Quantifying

Reads were mapped against the pig reference genome (assembly9) with Tophat v.1.0.14 (Trapnell *et al.* 2009) using the following settings: maximum of 40 hits per read (reporting best alignments), expected mean inner distance between mate pairs of 137 and a standard deviation for the distribution on inner distances between mate pairs of 100. For unambiguous mapping of the reads, the maximum alignments per read were set to 1. Sequence statistics were analyzed with FASTQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc). Base sequence qualities and proportion of bases per cycle are shown in Additional files 7a and 7b. A decrease in base quality is observed towards the end of the sequence and there is a bias in nucleotide content in the first 10 cycles of the reads due to the random hexamer primer library preparation approach (Hansen *et al.* 2010). Recently, a new statistical approach has been proposed to solve this bias (Roberts *et al.* 2011). Transcripts were assembled and quantified by Cufflinks v.0.9.0 (Trapnell *et al.* 2010). To improve the robustness of the differential expression estimates the quartile normalization was used and the contribution of the top 25 percent most highly expressed genes was excluded (-N option). The minimum alignment count per locus was set to 20 (-c option).

Orthology detection

Intergenic expressed regions not yet annotated in the pig genome were extracted with Cuffcompare (Trapnell *et al.* 2010) and custom Python and R scripts. Only those regions expressed in both samples were considered for a conservative approach. To identify putative coding transcripts, we run Augustus (Stanke *et al.* 2008) providing exon boundaries and allowing only complete proteins translations from the forward strand.

Transposable element analysis

We run RepeatMasker (http://www.repeatmasker.org/) with options 'quick search' and species 'pig' to identify repetitive and transposable elements (TE) in pig genome and male gonads transcriptome. We used RepeatMasker version open-3.2.9, rmblastn version (1.2) 2.2.23 and RepBase update 20090604.

LncRNA identification

All the transcripts not overlapping with pig protein-coding genes and falling at least 1 kb away from the closest protein annotation were considered for our analysis. A series of filtering steps were then implemented. The first one consisted in selecting the transcripts for which Augustus returned no (or just partial) coding potential. BlastX (NCBI Package version 2.2.25) was then used to search all possible translational products (the six possible reading frames) of each transcript against the NCBI non-redundant protein database (last update 05/29/2011). All the transcript queries that matched a known protein with an expectation value lower than 10^{-5} were discarded. Likewise, RPS-Blast (NCBI Package version 2.2.25) was used to search the possible translational products of each transcript against a database of Pfam profiles (Finn *et al.* 2008) and the transcripts returning an expectation value lower than 10^{-5} were removed. In order to filter the transcripts belonging to known classes of RNAs (snoRNAs, tRNAs, etc...), all the sequences were sought against Rfam (Release 10.0) using the Rfam searching facility available at: http://rfam.sanger.ac.uk/search#tabview=tab0.

Finally the remaining transcripts were remapped against the human genome and the homologous positions were intersected with protein-coding gene annotations (GENCODE version 3c). The screening was performed using a combination of BlastN and exonerate (as described in the screening pipeline in the methods). The transcripts whose human homologue resulted to be fully included in protein-coding exons were removed.

Screening pipeline

The screening pipeline was composed by three phases. The first consisted in seeking each query against the target genomes with a version of BlastN optimized for ncRNAs discovery [53]. Secondly, using exonerate each query was realigned versus the genomic regions pointed by Blast. For each query and for each genome was kept just the best hit that was successfully realigned. The exonerate alignments spanning for at least the 70% of the pig queries were retained. Finally, each query was compared versus all the putative discovered homologs by realigning the transcripts sequences with T-Coffee (Notredame *et al.* 2000) and measuring the query/homolog pairwise similarity.

Differential expression (DE) analysis

To test DE with unambiguous mapping data DEGseq was used (Wang *et al.* 2010a). MA plot-based method (where M is the log ratio of the counts between two experimental conditions for gene g, and A is the two group average of the log concentrations of the gene) with a random sampling method (MARS) was selected. To count the number of fragments that uniquely map to an exon, HTseq-count was used with 'union' as overlapping mode, 'gene' as feature and not strand-specific. A locus was considered as expressed if it had a minimum count of 40 fragments (summing the reads in both samples). From a total of 9 M unambiguously mapped reads for each library, 4.5 M of reads felt in the category of 'no_feature' (no annotation provided). The software discarded reads mapping to two overlapping genes (20,000 reads). Cuffdiff (Trapnell *et al.* 2010) was used to test DE using same options as discussed above for ambiguous mapping data.

In the microarray assay, we employed the GCRMA normalization method (Perez-Enciso *et al.* 2009) and a Bayesian z-score measure as detailed in (Irizarry *et al.* 2003). Briefly, normalized data were analyzed with model

y=Tissue+Breed+Sex+Probeset+PT+PB+PS+Residual,

where *PT*, *PB* and *PS* stand for the probeset × tissue, probeset × breed and probeset × sex interactions, respectively. The Bayesian breed z-score for the g-th probeset is defined as z_g = E(PB_g|y)/SD(PB_g|y), where E(PB_{gj}|y) and SD(PB_{gj}|y) are the expected and SD values of the posterior distribution of PB, respectively (Irizarry *et al.* 2003).

Gene ontology analysis

Parental gene ontology enrichment analysis was performed with the QuickGO browser (http://www.ebi.ac.uk/QuickGO/) using a GOSlim extracted from the AmiGO browser (http://amigo.geneontology.org/cgi-bin/amigo/go.cgi) and made up of 23 parental pig 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. Expected and observed GO percentages were compared with a Fisher's exact test as implemented in R. To test for an enrichment of specific ontology categories, we simply computed a two-sided *t*-test assuming a normal distribution for number of counts. The children gene ontology enrichment and KEGG pathway analyses were performed with the DAVID database (http://david.abcc.ncifcrf.gov/). Prior to GO analysis, the pig gene IDs were converted to human gene IDs with Biomart (http://www.biomart.org/) as the database had poor pig Ensembl annotations. The list of differentially expressed genes (intersection of Cufflinks and microarray breed effects) was compared against total expressed genes in male gonads (background).

SNP identification

SNPs were identified from unambiguously mapped reads using Samtools (http://samtools.sourceforge.net/). The minimum SNP quality was 10 and the minimum read depth was set to 3 × for fixed SNP with respect to the reference and 4 × for segregating SNP. As many false SNP were located at the splice sites due to the difficulties of alignments near indels (splicing sites), they were removed from the final set. Annotation of the SNP was made with custom Perl scripts using the Ensembl APIs.

Allele specific expression

To test for allele specific expression heterozygous SNP were selected from both samples using uniquely mapped reads (SNP quality > 10, minimum depth of 4x, minimum allele count of 2). Allele specific expression can be inferred when, in a heterozygous site, one allele is transcribed at significantly higher or lower rate (p) than the other allele. We used a beta - binomial model within a Bayesian framework to infer whether p was significantly different from 0.5. The posterior probability of p is given by the distribution

$\mathrm{Be}(\alpha+na,\beta+n-na)\ \mathrm{B}(na,n)\times\mathrm{Be}(\alpha,\beta)$,

where Be() is a beta distribution; B(), a binomial; n is the number of reads for that SNP; n_a , the number of reads pertaining to one arbitrary allele, and α and β are hyperparameters. The data was fitted using an empirical Bayesian approach such that the mean and variance of Be(α, β) were those observed in the real data. The obtained α and β were 4.99 and 3.84 in Large White and 6.38 and 6.20 in the Iberian data, respectively. ASE was considered when the 95% Highest Density Region (HDR) did not include p = 0.5. HDR was computed with function "HDIofICDF.R" in R (http:/ / www.indiana.edu/~kruschke/ DoingBayesianDataAnalysis/ Programs/ HDIofICDF.R).

Abbreviations

TE: Transposable elements, LncRNA: Long non-coding RNA, DE: Differential expression, UTR: Untranslated region, FPKM: Fragments Per Kilobase of exon model per Million mapped fragments, CDS: Coding sequence, ASE: Allele specific expression, GO: Gene ontology, PPAR: Peroxisome proliferator-activated receptor, KEGG: Kyoto Encyclopedia of genes and genomes, SNP: Single nucleotide polymorphism, APIs: Application programming interface, PcGs: Protein-coding genes.

Authors' contributions

MPE conceived and supervised research and provided material. AEC, RK, NP, GB and CN analyzed data. All authors discussed the results and wrote and approved the final manuscript.

Acknowledgements

Part of this work was done during a visit of AEC to Institut für Populationsgenetik, (Vienna), and we thank C. Schlötterer for hospitality. Thanks to J. Estellé for help with annotating SNPs. Sequencing was carried out at the ultra sequencing service unit of Center for Genomic Regulation (http://seq.crg.es/main/bin/view/Home/WebHome). AEC is recipient of a FPI PhD fellowship from Ministry of Education, Spain. Work funded by AGL2010-14822 grant to MPE and by a Consolider grant from Spanish Ministry of Research, CSD2007-00036 "Centre for Research in Agrigenomics". RK and NP were supported by the FWF grant P19467 awarded to C. Schlötterer. CN and GB were co-funded by the Plan Nacional BFU2008-00419, the European Commission, within the 7th Framework Programme, Grant Agreement KBBE-2A-222664 ("Quantomics"), the La Caixa PhD program and the Centre de Regulació Genomica.

Additionals Files

Additional file 1: Mapping statistics.

Comparison between ambiguous and unambiguous mapping.

http://www.biomedcentral.com/1471-2164/12/552/suppl/S1

Additional file 2. Orthologs coordinates.

Sheet 1: Coordinates of putative novel coding genes in Sus scrofa transcriptome; sheet 2: *Bos taurus* orthologs; sheet 3: *Homo sapiens* orthologs; sheet 4: *Sus scrofa* paralogs.

http://www.biomedcentral.com/1471-2164/12/552/suppl/S2

Additional file 3. RepeatMasker results.

RepeatMasker results. A) Transcriptome analysis. B) Genome analysis.

http://www.biomedcentral.com/1471-2164/12/552/suppl/S3

Additional file 4. Expression range.

Abundance of annotated genes expressed between 1-10 FPKM, 10-100 FPKM, 100-1000 FPKM, 1000-10000 FPKM or more than 10000 FPKM.

http://www.biomedcentral.com/1471-2164/12/552/suppl/S4

Additional file 5. Annotation of allelic specific expression.

Coordinates and annotation of SNP with significant ASE results. Sheet 1: Large White results; sheet 2: Iberian results; sheet 3: Shared SNPs with ASE.

http://www.biomedcentral.com/1471-2164/12/552/suppl/S5

Additional file 6. Sequence of the adapters.

Where "P" refers to a PO4 moiety and * indicates a phosphorothioate bond.

http://www.biomedcentral.com/1471-2164/12/552/suppl/S6

Additional file 7. Quality control of the reads.

A) Raw reads quality control: Base qualities per cycle. B) Library sequencing bias: Proportion of bases incorporated in each sequencing cycle.

http://www.biomedcentral.com/1471-2164/12/552/suppl/S7

References

Aird D., Ross M.G., Chen W.S., Danielsson M., Fennell T., Russ C., Jaffe D.B., Nusbaum C. & Gnirke A. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol* **12**, R18.

Arhondakis S., Auletta F., Torelli G. & D'Onofrio G. (2004) Base composition and expression level of human genes. *Gene* **325**, 165-9.

Bloom J.S., Khan Z., Kruglyak L., Singh M. & Caudy A.A. (2009) Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. *BMC Genomics* **10**, 221.

Bottomly D., Walter N.A., Hunter J.E., Darakjian P., Kawane S., Buck K.J., Searles R.P., Mooney M., McWeeney S.K. & Hitzemann R. (2011) Evaluating Gene Expression in C57BL/6J and DBA/2J Mouse Striatum Using RNA-Seq and Microarrays. *PLoS One* **6**, e17820.

Bradford J.R., Hey Y., Yates T., Li Y., Pepper S.D. & Miller C.J. (2010) A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling. *BMC Genomics* **11**, 282.

Branciforte D. & Martin S.L. (1994) Developmental and cell type specificity of LINE-1 expression in mouse testis: implications for transposition. *Mol Cell Biol* **14**, 2584-92.

Cooper G.M., Brudno M., Green E.D., Batzoglou S. & Sidow A. (2003) Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res* **13**, 813-20.

Daines B., Wang H., Wang L., Li Y., Han Y., Emmert D., Gelbart W., Wang X., Li W., Gibbs R. & Chen R. (2011) The Drosophila melanogaster transcriptome by paired-end RNA sequencing. *Genome Res* **21**, 315-24.

Donath A., Findeiβ, S., Hertel, J., Marz, M., Otto, W., Schulz, C., Stadler, P. F. and Wirth, S. (2010) Noncoding RNA. In: *Evolutionary Genomics and Systems Biology* (ed. by Caetano-Anollés G). John Wiley & Sons, Hoboken, NJ, USA.

Ferraz A.L., Ojeda A., Lopez-Bejar M., Fernandes L.T., Castello A., Folch J.M. & Perez-Enciso M. (2008) Transcriptome architecture across tissues in the pig. *BMC Genomics* **9**, 173.

Finn R.D., Tate J., Mistry J., Coggill P.C., Sammut S.J., Hotz H.R., Ceric G., Forslund K., Eddy S.R., Sonnhammer E.L. & Bateman A. (2008) The Pfam protein families database. *Nucleic Acids Res* **36**, D281-8.

Fu X., Fu N., Guo S., Yan Z., Xu Y., Hu H., Menzel C., Chen W., Li Y., Zeng R. & Khaitovich P. (2009) Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics* **10**, 161.

Gan Q., Chepelev I., Wei G., Tarayrah L., Cui K., Zhao K. & Chen X. (2010) Dynamic regulation of alternative splicing and chromatin structure in Drosophila gonads revealed by RNA-seq. *Cell Res* **20**, 763-83.

Garcia-Perez J.L., Marchetto M.C., Muotri A.R., Coufal N.G., Gage F.H., O'Shea K.S. & Moran J.V. (2007) LINE-1 retrotransposition in human embryonic stem cells. *Hum Mol Genet* **16**, 1569-77.

Gardner P.P., Daub J., Tate J.G., Nawrocki E.P., Kolbe D.L., Lindgreen S., Wilkinson A.C., Finn R.D., Griffiths-Jones S., Eddy S.R. & Bateman A. (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res* **37**, D136-40.

Graveley B.R., Brooks A.N., Carlson J.W., Duff M.O., Landolin J.M., Yang L., Artieri C.G., van Baren M.J., Boley N., Booth B.W., Brown J.B., Cherbas L., Davis C.A., Dobin A., Li R., Lin W., Malone J.H., Mattiuzzo N.R., Miller D., Sturgill D., Tuch B.B., Zaleski C., Zhang D., Blanchette M., Dudoit S., Eads B., Green R.E., Hammonds A., Jiang L., Kapranov P., Langton L., Perrimon N., Sandler J.E., Wan K.H., Willingham A., Zhang Y., Zou Y., Andrews J., Bickel P.J., Brenner S.E., Brent M.R., Cherbas P., Gingeras T.R., Hoskins R.A., Kaufman T.C., Oliver B. & Celniker S.E. (2011) The developmental transcriptome of Drosophila melanogaster. *Nature* **471**, 473-9.

Gustincich S., Sandelin A., Plessy C., Katayama S., Simone R., Lazarevic D., Hayashizaki Y. & Carninci P. (2006) The complexity of the mammalian transcriptome. *J Physiol* **575**, 321-32.

Guttman M., Garber M., Levin J.Z., Donaghey J., Robinson J., Adiconis X., Fan L., Koziol M.J., Gnirke A., Nusbaum C., Rinn J.L., Lander E.S. & Regev A. (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol* **28**, 503-10.

Haerty W., Jagadeeshan S., Kulathinal R.J., Wong A., Ravi Ram K., Sirot L.K., Levesque L., Artieri C.G., Wolfner M.F., Civetta A. & Singh R.S. (2007) Evolution in the fast lane: rapidly evolving sex-related genes in Drosophila. *Genetics* **177**, 1321-35.

Hansen K.D., Brenner S.E. & Dudoit S. (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res* **38**, e131.

Herai R.H. & Yamagishi M.E. (2010) Detection of human interchromosomal trans-splicing in sequence databanks. *Brief Bioinform* **11**, 198-209.

Irizarry R.A., Bolstad B.M., Collin F., Cope L.M., Hobbs B. & Speed T.P. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* **31**, e15.

Jacquier A. (2009) The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nat Rev Genet* **10**, 833-44.

Jagadeeshan S. & Singh R.S. (2005) Rapidly evolving genes of Drosophila: differing levels of selective pressure in testis, ovary, and head tissues between sibling species. *Mol Biol Evol* **22**, 1793-801.

Lindberg J. & Lundeberg J. (2010) The plasticity of the mammalian transcriptome. *Genomics* **95**, 1-6.

Marioni J.C., Mason C.E., Mane S.M., Stephens M. & Gilad Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* **18**, 1509-17.

McManus C.J., Coolon J.D., Duff M.O., Eipper-Mains J., Graveley B.R. & Wittkopp P.J. (2010a) Regulatory divergence in Drosophila revealed by mRNA-seq. *Genome Res* **20**, 816-25.

McManus C.J., Duff M.O., Eipper-Mains J. & Graveley B.R. (2010b) Global analysis of trans-splicing in Drosophila. *Proc Natl Acad Sci U S A* **107**, 12975-9.

Mortazavi A., Williams B.A., McCue K., Schaeffer L. & Wold B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621-8.

Nicolae M., Mangul S., Mandoiu, II & Zelikovsky A. (2011) Estimation of alternative splicing isoform frequencies from RNA-Seq data. *Algorithms Mol Biol* **6**, 9.

Notredame C., Higgins D.G. & Heringa J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**, 205-17.

Orom U.A., Derrien T., Beringer M., Gumireddy K., Gardini A., Bussotti G., Lai F., Zytnicki M., Notredame C., Huang Q., Guigo R. & Shiekhattar R. (2010) Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143**, 46-58.

Orom U.A., Derrien T., Guigo R. & Shiekhattar R. (2011) Long Noncoding RNAs as Enhancers of Gene Expression. *Cold Spring Harb Symp Quant Biol*.

Pasaniuc B., Zaitlen N. & Halperin E. (2011) Accurate estimation of expression levels of homologous genes in RNA-seq experiments. *J Comput Biol* **18**, 459-68.

Perez-Enciso M., Ferraz A.L., Ojeda A. & Lopez-Bejar M. (2009) Impact of breed and sex on porcine endocrine transcriptome: a bayesian biometrical analysis. *BMC Genomics* **10**, 89.

Ramskold D., Wang E.T., Burge C.B. & Sandberg R. (2009) An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol* **5**, e1000598.

Rangwala S.H., Zhang L. & Kazazian H.H., Jr. (2009) Many LINE1 elements contribute to the transcriptome of human somatic cells. *Genome Biol* **10**, R100.

Roberts A., Trapnell C., Donaghey J., Rinn J.L. & Pachter L. (2011) Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol* **12**, R22.

Shi X., Seluanov A. & Gorbunova V. (2007) Cell divisions are required for L1 retrotransposition. *Mol Cell Biol* **27**, 1264-70.

Slater G.S. & Birney E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31.

Stanke M., Diekhans M., Baertsch R. & Haussler D. (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637-44.

Toung J.M., Morley M., Li M. & Cheung V.G. (2011) RNA-sequence analysis of human B-cells. *Genome Res*.

Trapnell C., Pachter L. & Salzberg S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-11.

Trapnell C., Williams B.A., Pertea G., Mortazavi A., Kwan G., van Baren M.J., Salzberg S.L., Wold B.J. & Pachter L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28**, 511-5.

Tsai Y.S., Tsai P.J., Jiang M.J., Chou T.Y., Pendse A., Kim H.S. & Maeda N. (2009) Decreased PPAR gamma expression compromises perigonadal-specific fat deposition and insulin sensitivity. *Mol Endocrinol* **23**, 1787-98.

Wang E.T., Sandberg R., Luo S., Khrebtukova I., Zhang L., Mayr C., Kingsmore S.F., Schroth G.P. & Burge C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470-6.

Wang L., Feng Z., Wang X. & Zhang X. (2010a) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* **26**, 136-8.

Wang L., Xi Y., Yu J., Dong L., Yen L. & Li W. (2010b) A statistical method for the detection of alternative splicing using RNA-seq. *PLoS One* **5**, e8529.

Wen J., Chiba A. & Cai X. (2010) Computational identification of tissue-specific alternative splicing elements in mouse genes from RNA-Seq. *Nucleic Acids Res* **38**, 7895-907.

Xu A.G., He L., Li Z., Xu Y., Li M., Fu X., Yan Z., Yuan Y., Menzel C., Li N., Somel M., Hu H., Chen W., Paabo S. & Khaitovich P. (2010) Intergenic and repeat transcription in human, chimpanzee and macaque brains measured by RNA-Seq. *PLoS Comput Biol* **6**, e1000843.

Zamboni M., Scarabino D. & Tocchini-Valentini G.P. (2009) Splicing of mRNA mediated by tRNA sequences in mouse cells. *RNA* **15**, 2122-8.

Zamudio N. & Bourc'his D. (2010) Transposable elements in the mammalian germline: a comfortable niche or a deadly trap? *Heredity* **105**, 92-104.

CHAPTER 7

GENERAL DISCUSSION

In this Section, we discuss the different experimental approaches used, the estimated nucleotide variability in the Iberian strain, methodologies to detect structural variants and their potential phenotype effect, the different approximations to detect selective sweeps in the pig genome, provide new insights into the functional roles of the transposable elements and lncRNAs expressed in pig male gonads and, finally we refer to the problems related with mis-alignments and mis-assemblies.

Experimental design: RRL and pools

Although NGS have slashed prices, a few years ago sequencing complete genomes at high depth was costly, whereas now analyses are limited by computing costs. Here we opted for cost-effective methods to capture as many polymorphisms as possible but at a reasonable cost: sequencing a pool of individuals and / or sequencing a randomly selected fraction of the genome using reduced representation libraries. The use of a random RRL allowed performing a representative sampling of a single individual (Chapter 4) and also from a pool of individuals (Chapter 5). In doing so, the complexity of the genome is reduced and higher sequence redundancy is obtained than if whole genome sequencing was performed. But this method has some problems as described by Amaral et al (Amaral *et al.* 2009; Amaral *et al.* 2011), RRLs from different samples independently, could lead to distinct sequenced fragments between individuals due to an imprecise excision of the fragment size during library preparation, which causes that the animals share a small amount of sequenced coverage and therefore the usable data drops. Moreover, there is the possibility that some interesting region is not covered, since only a fraction of the genome is sequenced.

Discovering variants at population level and low cost can be carried out sequencing pools of individuals. This strategy has been used and discussed in previous studies (Van Tassell *et al.* 2008; Cutler & Jensen 2010; Kim *et al.* 2010; Amaral *et al.* 2011; He *et al.* 2011; Zaboli *et al.* 2011; Boitard *et al.* 2012). In Chapter 5, single individual sequencing was combined with sequencing a pool of nine Iberian pig varieties. However the pooling methodology also has its drawbacks, since we have to account for the uncertainty that a certain chromosome is included when pooling the DNA from the different samples and the uncertainty that a certain chromosome is actually being sequenced, meaning that some individual sequences are over or under-represented. To account for these peculiarities, we applied new statistical approaches, e.g., modified the Watterson' estimator for pools (Ferretti 2012) and developed a new Bayesian approach for SNP calling in pools (Raineri 2012). Another problem is that power to detect SNP is ~50%, which means that half of the SNP are lost when pooling. The majority of the variants in a population are at low frequencies (e.g., singletons) and they are lost during the process of pooling and SNP calling, simply because rare alleles are less likely to be sampled and difficult to distinguish from sequencing errors (Perez-Enciso & Ferretti 2010). All these caveats introduce uncertainty in the estimators that must be considered when interpreting the results. Yet, the advantage over individual sequencing is that although the power to detect SNPs in pools is lower, the total number of SNPs uncovered is much higher, since in pooled data there are more chromosomes.

Nucleotide and structural diversity in the Iberian strain

The recent completion of the pig genome has greatly facilitated the study of porcine genome variability (Groenen 2012). In this thesis, we centered our analysis in a highly inbred Iberian pig with extreme phenotypic characteristics and no evidence of introgression of Asian genes, the Guadyerbas strain. Genetic diversity was estimated using whole genome sequencing from a single individual using a modified Watterson's theta estimator that corrects by low depth (Chapter 4 and 5). This correction takes into account that there is an underestimation of heterozygous sites in low depth sequenced samples. Estimated variability in autosomes was ~ 0.7 kb^{-1}, a non-negligible value for a highly inbred line, but still 50% less variable than the pooled data containing distinct Iberian varieties. After correcting for inbreeding, however, both estimates are similar. In contrast, for the NPAR region of the X chromosome, still 40% less variability was observed in the individual after correcting by inbreeding. In fact, we also report that the Iberian nucleotide variability in X chromosome NPAR / autosome ratio is much lower than expected under a neutral scenario. This observation could be explained by several factors as described in the Introduction section. Nevertheless, given that this low variability has been observed in several pig populations (Amaral *et al.* 2009; Amaral *et al.* 2011), selection seems the most likely explanation. We studied in detail the NPARX low variability regions and found some interesting genes: MECP2 involved in fear response, NSDGL in hair follicle morphogenesis, the interleukines IL13Ra1 and IL1RAPL2 in immune response, the ACSL4 gene in lipid metabolism, and, interestingly, it has been reported as a positional candidate gene for the quantitative trait loci (QTL) related to growth and oleic fatty acid composition in pigs (Mercade *et al.* 2006) and liver expressions were studied in detail in (Corominas *et al.* 2012). Also important is the HTR2C, a serotonin receptor involved in anxiety and feeding behavior and maps to a region previously described to be potentially associated to maternal infanticide in pigs (Chen *et al.* 2011b; Quilter *et al.* 2012).

In agreement with other species like rabbits (Carneiro *et al.* 2009), chicken (Fang *et al.* 2008), Drosophila (Begun & Aquadro 1992) and human (Hellmann *et al.* 2005; Spencer *et al.* 2006) genetic diversity was higher in telomeric regions than centromeres, probably due to a higher recombination rates in telomeres. In fact, we observed a high correlation of variability and recombination rate in the pig genome (Chapter 5). This observation could be explained either by a higher mutagenic effect in recombination spots (Lercher & Hurst 2002; Hellmann *et al.* 2003) or by selection (reviewed in (Eyre-Walker & Hurst 2001). Background selection removes all variants linked to a deleterious mutation and this effect is more pronounced in regions with low recombination (e.g., centromeres or NPAR regions of X chromosome) and the same pattern happens with positive selection. It has been postulated that the former option is more likely, since positive selection does not seem to be pervasive in the pig, (Groenen 2012), unpublished), reports that approximately 1 % of the pig genome to be under selective pressures. On the contrary, this is not true for some Drosophila species with big population sizes. Sella et al. (Sella *et al.* 2009) reported that the major part of the Drosophila genome underwent positive selection. It has been reported that GC content is positively correlated with recombination rates due to biased GC gene conversion in recombination hotspots (Marais 2003; Duret & Galtier 2009). Under this hypothesis, mismatch repair systems would preferentially insert G or C at sites where strand breakage occurs during meiosis and mitosis (Genereux 2002). Therefore, indirectly, GC content is expected to correlate with variability. Here, however, we did not find such a strong correlation with GC content and nucleotide diversity.

Multi-copy regions

In this thesis, we used a sequenced-based approach to identify multi-copy regions at a genome-wide scale. This technique, which is becoming more popular due to the ongoing algorithm improvements and cost decreases in NGS, overcomes the array comparative genome hybridization (aCGH) and SNP chip genotyping methods in terms of specificity, sensitivity and resolution. The read depth approach used in our study, determines the exact number of copies of each MCR and it has much better resolutions, thus not inflating the length of the MCR. A clear example is that CNVs detected by the pig SNPchip array (Ramayo-Caldas *et al.* 2010) had a median of 754.6 kb (minimum length of 44.7 kb and maximum of 10.7 Mb), whereas we report a median of 6 kb (minimum length of 4 kb and maximum of 117 kb). The advantage over aberrant paired-end distance approaches is that it does not require a very good quality genome build, but the drawback is that it can not detect inversions, translocations, novel insertions and other complex structural variants, just duplications or deletions with respect to the reference genome (e.g., new paralogs of annotated genes). Furthermore, it is difficult to spot the exact breakpoints of the SVs due to read depth fluctuations at fine scales. Therefore, the read depth approach is complementary to the paired-end distance methods; in fact, there are some programs to detect CNV from NGS data integrating both approaches (Medvedev *et al.* 2010).

As stated in the Introduction, structural variants (e.g., CNV), although being less frequent than SNPs, they affect a higher percentage of genomic sequence and potentially have greater impacts phenotype diversity changing gene structures and dosage, altering gene regulation and exposing recessive alleles (Zhang *et al.* 2009). Remarkably, they have been associated with several diseases in human such as autism, intellectual disability, dyslexia and schizophrenia (Sebat *et al.* 2007; Bassett *et al.* 2010; Girirajan & Eichler 2010; Vacic *et al.* 2011). In livestock, there has been recently a major interest, since CNV may contribute to evolutionary adaptation and to agriculturally important traits. A representative sample of phenotypes produced by structural variants in domestic animals is shown in Table 1. Most of the associations discovered in livestock are related to Mendelian traits, but the next challenge is to elucidate their implications in complex phenotypes, e.g., growth, prolificacy or disease resistance, to incorporate them into animal genomic selection systems.

Table 1

Species	Phenotype	Gene	Reference
Cow	Anhidrotic ectodermaldysplasia	*EDA*	(Drogemuller *et al.* 2001)
	Renal tubular dysplasia	*CLDN16*	(Ohba *et al.* 2000)
	Osteoporosis	*SLC4A2*	(Meyers *et al.* 2010)
	Abortions and stillbirths	*MIMT1*	(Flisikowski *et al.* 2010)
Sheep	White and grey coat color	*ASIP*	(Norris & Whan 2008)
Goat	Polled intersex syndrome	*PISRT1*	(Pailhoux *et al.* 2001)
	White coat color	*ASIP*	(Fontanesi *et al.* 2009)
Pig	White coat color	*KIT*	(Giuffra *et al.* 1999)
Horse	Hair depigmentation and susceptibility to melanoma	*STX17*	(Rosengren Pielberg *et al.* 2008)
Dog	Copper toxicosis	*COMMD1*	(van De Sluis *et al.* 2002)
	Cone degeneration	*CNGB3*	(Sidjanin *et al.* 2002)
	Dorsal hair ridge and susceptibility to dermoid sinus	*ORAOV1*	(Salmon Hillbertz *et al.* 2007)
	Collie eye anomaly	*NHEJ1*	(Parker *et al.* 2007)
	Cone-rod dystrophy 3	*ADAM9*	(Goldstein *et al.* 2010)
	Wrinkled skin and periodic fever	*HAS2*	(Olsson *et al.* 2011)
	Startle disease	*SLC6A5*	(Gill *et al.* 2011)
Chicken	Feather growth	*PRLR*, *SPEF2*	(Elferink *et al.* 2008)
	Pea-comb phenotype	*SOX5*	(Wright *et al.* 2009)

Information extracted from (Clop *et al.* 2012)

Interestingly, in our study, genes that fully overlapped with multi-copy regions (MCR) were enriched sensory perception of smell, virus response and xenobiotic metabolism. Similar results were found in cattle (Fadista *et al.* 2010; Liu *et al.* 2010; Hou *et al.* 2011; Bickhart *et al.* 2012), horses (Doan *et al.* 2012a; Doan *et al.* 2012b) goats and sheep (Fontanesi *et al.* 2010a; Fontanesi *et al.* 2010b). One step further in this MCR analysis would be to test more Iberian pigs or pigs from other breeds to see if these MCR are in fact, copy number variants (CNVs) or they are fixed in the population. The main goal would be to find breed-specific CNV associated with production traits. Moreover, as we were able to detect only MCR gains with respect to the reference Duroc genome due to a relatively low average read depth in our sample; it would be possible as well to detect deletions with confidence if the sequenced sample had higher average read depth.

Selection fingerprints in the pig genome

The traditional method to study adaptive evolution is to investigate a small number of loci that one hypothesizes priori to have been under selection. However, an inherent limitation to single locus approaches is that population demographic history confounds natural selection. But scanning the entire genome provides the opportunity to begin to disentangle demography and selection effects; the former affects the whole genome and the latter acts upon specific loci. Therefore, for the genome-wide approach, empirical distributions of tests statistics can be performed and genes under selective pressures can be identified as outlier loci. The drawback of this approximation is that it will depend on the strength of selection and the fraction of all loci subject to selection, parameters that are difficult to estimate (Kelley *et al.* 2006).

To infer regions of the Iberian pig genome that might underwent selection during domestication or breeding, we applied different approaches: i) covering just a single target, ii) sequencing the whole genome and iii) investigating differentially expressed genes in a specific tissue.

In Chapter 3, we used the classical Sanger sequencing, a high-resolution method to study in detail the nucleotide variability of unique putative target under selection. To date, several works have shown the usefulness of this approach in humans and other species. For example, Inomata et al. (Inomata & Yamazaki 2002) studied nucleotide diversity of *AMY* gene, a digestive enzyme that breaks down starch in Drosophila, Ojeda et al 2008 (Ojeda *et al.* 2008b) focused the study at the causative gene *IGF2* related to muscle growth and leanness in pigs, (Li *et al.* 2010a) detected positive selection at the *MC1R* gene in Chinese pig breeds, or (Gilad *et al.* 2002) who detected positive selection at the *MAOA* gene associated with aggressiveness in humans. We centered our attention to the *SERPINA6* gene, which has been reported to be putatively associated with meat quality in pigs (Ousova *et al.* 2004; Guyonnet-Duperat *et al.* 2006). For that, we characterized the nucleotide variability patterns of this gene in different pig breeds originated in Asia and Europe. Although we detected a nonsynonymous mutation only present in European domestic pigs (except for the Iberian strain) we could not observe a clear selective sweep at the *SERPINA6* gene because there was no indication of an overall reduction in genetic diversity compared to the European wild ancestors. In fact, the wild boars presented even lower levels of diversity than domestic pigs. This observation is corroborated by other studies (Ojeda *et al.* 2006; Ojeda *et al.* 2008a; Ojeda *et al.* 2008b) and might be explained either by a recent split European wild boar-domestic, small founder effects or gene flow. Conversely, in maize (Wright *et al.* 2005), the wild ancestors still maintain high amounts of variability compared to the domestic. On the other hand, Asian pigs posses very high levels of diversity in agreement with other studies (Zhang & Plastow; Larson *et al.* 2005; Fang & Andersson 2006). Not surprisingly, the hybrid Landrace European domestic pig presented both Asian and European haplotypes, which reinforces the evidence of recent Asian germplasm introgression among commercial lines (Jones 1998; Giuffra *et al.* 2000). The fact that none of the neutrality tests applied were not significant when testing the whole gene could be explained by a reduced number of samples, so further studies should screen more animals. However, we were able to detect an evolutionary constrain at the 5' end of the gene (promoter region) compared to the coding sequence.

In Chapter 4 and 5, we characterized genome-wide nucleotide diversity patterns in the Iberian genome to infer regions under selection. This method has less resolution than studying a single target region, e.g., *SERPINA6* study, but provides a general overview of diversity patterns distribution. Moreover, it is a reasonable study to confine our search for positive selection to a small set of candidate genes. To carry out this study, we divided the genome in fixed window sizes and calculated different statistics, as reported in other studies (Hellmann *et* *al.* 2008; Amaral *et al.* 2011); . Watterson's theta high variability outlier regions may indicate balancing selection, slightly deleterious mutations segregating in the population or mis-assemblies of the reads (Hellmann *et al.* 2008). We found an enrichment of sensory perception of smell (*OR* genes) and defense response (*SLA* genes) gene ontologies. These findings have been also previously reported in human and pig (Markow *et al.* 1993; Black & Hedrick 1997; Alonso *et al.* 2008; Hellmann *et al.* 2008; Luetkemeier *et al.* 2009; Tong *et al.* 2010; Amaral *et al.* 2011; Cagliani *et al.* 2011). Conversely, extreme low variability regions may suggest directional selection or background selection (Hellmann *et al.* 2008). In agreement with Hellman et al., we also found an enrichment of genes related to the immune response; specifically we found many *INF* genes, which modulate B cells proliferation in response to virus.

An approach to avoid the confounding effects of demography is to define test statistics sensitive only to selection (reviewed in (Li *et al.* 2012b)). For instance, in our case, Tajima's *D* and Fay&Wu' *H* tests were combined to capture departures from the expected SFS (Zeng *et al.* 2006). The joint test is more powerful because *D* and *H* are sensitive to different demographic factors. The sensitivity of *D* to population expansion may be counterbalanced by the insensitivty of H to the same factor. Grossman et al 2010 (Grossman *et al.* 2010) proposed another join test to detect positive selection, the composite of multiple signals (*CMS*), that takes into account long haplotypes, high-frequency derived alleles and high differentiation among population. In this way, this test captures a more extense temporal range. Nevertheless, for strong bottlenecks, both tests fail to identify the target of selection. The alternative is to estimate both demography and selection in a single analysis. As discussed in the Introduction, the ABC method (Tavare *et al.* 1997; Pritchard *et al.* 1999; Beaumont *et al.* 2002) is a promising approach. Mainly it has been used to infer a variety of demographic parameters (Bertorelle *et al.* 2010; Csillery *et al.* 2010) and to a lesser extent to infer properties of selection (Jensen *et al.* 2008), but it has the potential to incorporate both at the same time. The most demanding need is to develop an ABC framework to be applied genome-wide.

In Chapter 6, we restricted our study to a specific tissue. Previous studies in Drosophila reported sex-related genes to be rapidly evolving genes (Haerty *et al.* 2007), so the male pig gonads is a good tissue to start with. In that case, we did not focus on nucleotide changes in the DNA to infer selection, but rather in gene expression patterns of two extreme breeds' gonads. We selected a local fat non-improved breed, the Iberian pig, and a commercial lean breed, a Large White pig. Both animals showed a very high and positive correlation in terms of genes expression, suggesting that tissue expression is conserved between the two distinct breeds. This is consistent with our previous results, which showed that the largest source of variability was tissue rather than breed (Ferraz *et al.* 2008). Interestingly, differentially expressed genes in gonads were enriched in spermatogenesis and lipid metabolism gene ontologies. This observation is in agreement with the extreme phenotype characteristics of the two breeds, the Iberian pig being very fat and not very prolific, and the Large White, a very lean and prolific pig. However, to confirm that changes in expression of these two biological processes are adaptive is not a trivial task. Phenotypic evolution can occur through changes in the coding sequence affecting the protein function or changes in regulation of expression altering the amount of protein produced (e.g., beak morphology in Darwin's finches (Abzhanov *et al.* 2004)). In the first case, there exist many approaches to detect outlier regions that can not be explained by the neutral model, but in the second case, although the NGS revolution has facilitated the generation of large transcriptome datasets, still we do not have a consensus about the neutral model to use to disentangle if changes in gene expression are due to drift or positive selection (Harrison *et al.* 2012)**.**

In all the aforementioned approaches, once having identified candidate loci, it is important to confirm results on independent data, functionally characterize suspected candidates, and ultimately correlate adaptive genetic variation with phenotype variation (Kelley *et al.* 2006).

Pig genome annotation

The advantage of RNA-seq over microarrays, as mentioned in the Introduction section, is that it allows detecting novel genes not annotated in the reference genome. In this thesis (Chapter 6), we were able to detect novel isoforms from annotated genes and unnanotated protein coding genes with orthology in cow and human and to experimentally confirm novel computationally predicted proteins. Important to mention is the fact that an alternative splicing could not be performed due to many inaccurately constructed transcripts partially due to low coverage and partially due to transcript assembly algorithm artefacts. Moreover, we also detected many transposable elements (TE) expressed in male gonads and long-non-coding RNAs (lncRNAs). Up to 16% of the transcripts contained a TE in their sequence. The fact that these elements are still active in germ cells suggests that they might perform functional roles, e.g., generation of innovative ways to alter gene expression and genomic structures (Muotri *et al.* 2007). Several studies reported that TE contribute to novel exon acquisition (Vaknin *et al.* 2009), formation of pseudo-genes and regulation of splicing and gene expression (Muotri *et al.* 2007). This shows that genomes are not static but rather dynamic. LncRNAs are strikingly similar to mRNAs: they are RNA polymerase II transcripts that are capped, spliced and polyadenylated, yet do not function as templates for protein synthesis (Moran *et al.* 2012). They seem to be involved also in many structural and functional roles, e.g., regulation of gene transcription, splicing, translation, imprinting, X-inactivation in females and also in cancer and neurological disorders (Rinn *et al.* 2007; Mercer *et al.* 2009; Gupta *et al.* 2010; Huarte *et al.* 2010; Orom *et al.* 2010; Guttman *et al.* 2011; Hung *et al.* 2011). In this thesis, we develop a new pipeline to detect lncRNAs in the pig transcriptome. Noteworthy, some presented homology with human, others were found only in Artiodactyla order (even-toed ungulates) and other were pig specific. To complement our analysis and to confirm they are not artefacts differential expression analysis should be performed in distinct tissues. It is important to mention that, in a recent liver RNA-seq study that we performed with extreme phenotypic pigs in terms of fat deposition (Ramayo-Caldas et al, unpublished) we confirmed the expression of hundred lncRNAs already detected in our previous gonad RNA-seq analysis, suggesting that, at least some of them, are not tissue-specific. In future RNA-seq assays more read depth is needed, as our analysis showed that many lncRNAs were truncated due to lack of connectivity between exons. In addition, further analysis should be performed including lncRNA overlapping with protein-coding genes, as it seems really common that gene regulation is carried out by protein coding genes anti-sense transcription (Faghihi & Wahlestedt 2009).

These findings in the pig transcriptome shed some light in the dark matter of the transcriptome, where many non-coding regions of the transcriptome seem to be pervasively expressed. Nevertheless, their precise function needs to be investigated further. A conservative study estimates 23,000 lncRNAs in the human genome, rivaling the 20,000 protein coding genes (Gibb *et al.* 2011). It is worth to mention, that nowadays, many tools for gene and functional annotation are emerging. It is the case of Blast2GO (Conesa *et al.* 2005; Aparicio *et al.* 2006; Conesa & Gotz 2008; Gotz *et al.* 2008; Gotz *et al.* 2011), which automatically annotates thousands of novel transcripts providing their sequences as input and doing an internal connection to the ncbi blast database. This tool not only reports the homologous sequences encountered in the database, but also retrieve the GO terms and the most reliable function of the target sequence. Another approach we applied in Chapter 6 is to use a gene predictor tool like Augustus (Stanke & Waack 2003; Stanke *et al.* 2004; Stanke & Morgenstern 2005; Stanke *et al.* 2006a; Stanke *et al.* 2006b; Stanke *et al.* 2006c; Stanke *et al.* 2008) to predict if novel discovered sequences encode proteins. In the near future it is also planned to develop non-coding RNA annotation tools, although the characterization of those might be more complex as the ncRNAs' sequence is not as conserved as protein coding genes. All these bioinformatic tools may greatly help to better characterize poorly annotated genomes. Although there were major advances in the pig annotation last year, at the time that the RNA-seq study was performed only 40% of the reads generated were mapped to annotated exons, showing the incompleteness of the annotation in the pig genome.

Alignment artefacts and reference mis-assemblies

An important point to have in consideration, as discussed in Chapter 6, is the mapping strategy, unambiguous mapping discards reads aligning to several locations in the genome, thus not covering repetitive regions and paralog genes, whereas ambiguous mapping might be so liberal that the reads are incorrectly mapped. Derrien et al (Derrien *et al.* 2012) reported that 'mappability' is an important concept to be taken into account when one is trying, for instance, to re-sequence a particular genomic region, or to produce quantitative estimates of transcript abundance from RNA-seq experiments. They provide a method to calculate the 'mappability' of different parts of a genome a priori and broadly discuss its implication in SNP calling, Chip-Seq and RNA-seq analysis, gene families and pseudo-genes discovery and its relation to paired-end sequencing schemes. The short nature of the reads produced by NGS techniques and the repetitive complex structure of many eukaryote genomes hinder and limit the analysis and interpretation of the data. All these pitfalls might be solved in the forthcoming months with the arrival of next next generation sequencing techniques (e.g., Nanopore sequencing), larger reads will be produced (100 kb of length) and the mapping ambiguity problem resolved.

In addition to the aforementioned caveats, the reference assembly must be seen with an skeptical eye. Multi-copy regions detected with read depth methodologies uncover many un-annotated gene paralogs belonging to large gene families, thus evidencing the difficulties to assembly those repetitive regions in reference genomes. These findings highlight the need to improve them; otherwise, the results must be interpreted with caution.

CHAPTER 8

PERSPECTIVES IN THE GENOMICS ERA

New miniaturized technologies producing longer reads at a cheaper price are about to emerge, so sequencing will be affordable and accessible to everyone. This will transform tomorrow's society to the point that the classical blood analysis will be replaced by sequencing one's genome or even, sequencing one's RNA in vivo. For example, we will be able to sequence infectious virus and bacteria's RNA in real time in a human's body, detecting low numbers of circulating tumor cells, take personalized preventive disease measures, track novel nucleic acid therapeutics and gene therapies, discover new microorganisms living in extreme environments and apply novel enzymes to ameliorate problems of disease, pollution, energy production and industrial processes (Kahvejian *et al.* 2008). Sequencing everything is yet only the first step, as we then need to be able to manage this huge amount of data and process it into information that can be used broadly to benefit human health and productivity.

8.1 Data processing and storage

The rate of information coming from current-generation DNA sequencers is increasing exponentially, faster than the computational power and storage size (even faster than the Moore's law, Eggen 2012). Thus, in this new scenario, data processing and analysis is becoming a bottleneck and demands more advance computer solutions. New faster and efficient software needs to be developed using new technologies like High Performance Computing clusters (HPC) or cloud-based applications. HPC are multi-processor computer architectures which work in parallel to solve complex problems in a very short period of time compared to serial computing. Similarly, cloud computing allows scientists to rent both storage and processing power virtually by accessing remote servers as they are needed. This technology is even more appealing to institutes without a vast computer infrastructure. NGS software for read alignment, assembly and variation detection, which are computationally intensive and time-consuming, can benefit from these ultra-fast computational resources. Basically, these new software must be written so that tasks can be fragmented and performed in parallel. In fact, some promising programs that work in the cloud have been recently developed, e.g., Crossbow (Langmead *et al.* 2009) for human resequencing and genotyping or Myrna (Langmead *et al.* 2010) to calculate differential gene expression from large RNA-seq datasets.

8.2 Beyond next generation sequencing

Newer technologies, the so called next next generation sequencing techniques, do not include any amplification step and are able to sequence a single molecule of DNA. A promising technology is to sequence a long single molecule of DNA using a protein nanopore (Oxford Nanopore technologies, http://www.nanoporetech.com/). The DNA passes through a nanoscale pore in a membrane and then each base is read off, using the ion current passing through the pore. As the length of the DNA molecule will be ultra long (hundreds of kb), much more than the conventional short read lengths, it would facilitate mapping repetitive regions, structural variants as well as resolving haplotypes. At present this technology was able to sequence to whole genome of lambda phage at a stretch (~54 kb) with 4% of sequencing error rate. In the next months it is planned to sequence 100 kb and all the sequencing process will take part inside a USB memory stick with thousands of nanopores (just for a few hundred of euros); the user only needs to load directly the blood sample and connect to own computer.

8.3 Systems biology: An integrative approach

The traditional approach of isolating individual genes or proteins is being replaced by the interrogation of multiple components of a cell on a genome-wide scale. The ultimate goal of biology is to integrate 'omics' data: genomics (sequence and structural variation), transcriptomics (gene expression, allele specific expression, novel functional transcripts), epigenomics (chromatin conformation, regulatory elements), interactomics (protein-protein interactions, RNA/DNA-protein interactions) and functional annotation datasets (gene ontologies, signaling pathways) to be able to answer complex biological questions about the fundamental mechanisms of genome function and disease in a unified global view. In this sense, the aim of systems biology is to model life processes with this large amount of data from different sources and discover emergent properties, of cells, tissues and organisms functioning as a system.

CHAPTER 9

CONCLUSIONS

No clear patterns of a selective sweep could be detected for the Serpina6 gene putatively associated with meat quality using a diverse panel of pigs, but our data support instead a constraint on proximal 5' regulatory region larger than the in coding sequence.

We observed much less nucleotide variability that expected under neutrality in the NPAR region of chromosome X compared to autosomes in the Iberian pig, which could be explained by the action of selection.

Strong correlations between recombination rates and variability were observed in the Iberian pig genome in agreement with other species. This finding could be explained by either a higher mutagenic effect in recombination hotspots or by selection.

More than 36Mb of the Iberian pig genome were multi-copy regions gains with respect to the reference genome, which stress their importance in genome's structures and phenotype diversity.

We detected several regions putatively under positive selection in the Iberian pig genome and presented interesting candidates genes related with feeding behaviour, immune response, lipid metabolism, hair follicle morphogenesis, epidermis formation, circadian rhythm, which need to be further investigated.

Differential expression analysis in male gonads of two extreme phenotypic pigs in terms of prolificacy and fat deposition, showed an enrichment of lipid metabolism and spermatogenesis gene ontologies, which may be a result of targeted tissue selection.

A high fraction of the pig gonad transcriptome is made up of transposable elements and long-non-coding RNAs, their functional roles must still be elucidated. We also detected several unnanotated protein coding genes that presented homologies with human and cow genes.

BIBLIOGRAPHY

The International Chicken Genome Sequencing Consortium (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695-716.

The International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature* **436**, 793-800.

Abzhanov A., Protas M., Grant B.R., Grant P.R. & Tabin C.J. (2004) Bmp4 and morphological variation of beaks in Darwin's finches. *Science* **305**, 1462-5.

Alkan C., Coe B.P. & Eichler E.E. (2011) Genome structural variation discovery and genotyping. *Nat Rev Genet* **12**, 363-76.

Alonso S., Lopez S., Izagirre N. & de la Rua C. (2008) Overdominance in the human genome and olfactory receptor activity. *Mol Biol Evol* **25**, 997-1001.

Alves E., Fernandez A.I., Fernandez-Rodriguez A., Perez-Montarelo D., Benitez R., Ovilo C., Rodriguez C. & Silio L. (2009) Identification of mitochondrial markers for genetic traceability of European wild boars and Iberian and Duroc pigs. *Animal* **3**, 1216-23.

Alves E., Ovilo C., Rodriguez M.C. & Silio L. (2003) Mitochondrial DNA sequence variation and phylogenetic relationships among Iberian pigs and other domestic and wild pig populations. *Anim Genet* **34**, 319-24.

Amaral A.J., Ferretti L., Megens H.J., Crooijmans R.P., Nie H., Ramos-Onsins S.E., Perez-Enciso M., Schook L.B. & Groenen M.A. (2011) Genome-wide footprints of pig domestication and selection revealed through massive parallel sequencing of pooled DNA. *PLoS One* **6**, e14782.

Amaral A.J., Megens H.J., Kerstens H.H., Heuven H.C., Dibbits B., Crooijmans R.P., den Dunnen J.T. & Groenen M.A. (2009) Application of massive parallel sequencing to whole genome SNP discovery in the porcine genome. *BMC Genomics* **10**, 374.

Aparicio G., Gotz S., Conesa A., Segrelles D., Blanquer I., Garcia J.M., Hernandez V., Robles M. & Talon M. (2006) Blast2GO goes grid: developing a gridenabled prototype for functional genomics analysis. *Stud Health Technol Inform* **120**, 194-204.

Bassett A.S., Scherer S.W. & Brzustowicz L.M. (2010) Copy number variations in schizophrenia: critical review and new perspectives on concepts of genetics and disease. *Am J Psychiatry* **167**, 899-914.

Beaumont M.A., Zhang W. & Balding D.J. (2002) Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025-35.

Begun D.J. & Aquadro C.F. (1992) Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. *Nature* **356**, 519-20.

Bertorelle G., Benazzo A. & Mona S. (2010) ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Mol Ecol* **19**, 2609-25.

Bickhart D.M., Hou Y., Schroeder S.G., Alkan C., Cardone M.F., Matukumalli L.K., Song J., Schnabel R.D., Ventura M., Taylor J.F., *et al.* (2012) Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Res* **22**, 778-90.

Black F.L. & Hedrick P.W. (1997) Strong balancing selection at HLA loci: evidence from segregation in South Amerindian families. *Proc Natl Acad Sci U S A* **94**, 12452-6.

Boitard S., Schlotterer C., Nolte V., Pandey R.V. & Futschik A. (2012) Detecting Selective Sweeps from Pooled Next-Generation Sequencing Samples. *Mol Biol Evol*.

Cagliani R., Riva S., Pozzoli U., Fumagalli M., Comi G.P., Bresolin N., Clerici M. & Sironi M. (2011) Balancing selection is common in the extended MHC region but most alleles with opposite risk profile for autoimmune diseases are neutrally evolving. *BMC Evol Biol* **11**, 171.

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

Clop A., Amills M., Noguera J.L., Fernandez A., Capote J., Ramon M.M., Kelly L., Kijas J.M., Andersson L. & Sanchez A. (2004) Estimating the frequency of

Asian cytochrome B haplotypes in standard European and local Spanish pig breeds. *Genet Sel Evol* **36**, 97-104.

Clop A., Vidal O. & Amills M. (2012) Copy number variation in the genomes of domestic animals. *Anim Genet*.

Conesa A. & Gotz S. (2008) Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int J Plant Genomics* **2008**, 619832.

Conesa A., Gotz S., Garcia-Gomez J.M., Terol J., Talon M. & Robles M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674-6.

Corominas J., Ramayo-Caldas Y., Castello A., Munoz M., Ibanez-Escriche N., Folch J.M. & Ballester M. (2012) Evaluation of the porcine ACSL4 gene as a candidate gene for meat quality traits in pigs. *Anim Genet*.

Csillery K., Blum M.G., Gaggiotti O.E. & Francois O. (2010) Approximate Bayesian Computation (ABC) in practice. *Trends Ecol Evol* **25**, 410-8.

Cutler D.J. & Jensen J.D. (2010) To pool, or not to pool? *Genetics* **186**, 41-3.

Chen C., Ai H., Ren J., Li W., Li P., Qiao R., Ouyang J., Yang M., Ma J. & Huang L. (2011a) A global view of porcine transcriptome in three tissues from a full-sib pair with extreme phenotypes in growth and fat deposition by paired-end RNA sequencing. *BMC Genomics* **12**, 448.

Chen C., Yang Z., Li Y., Wei N., Li P., Guo Y., Ren J., Ding N. & Huang L. (2011b) Association and haplotype analysis of candidate genes in five genomic regions linked to sow maternal infanticide in a white Duroc x Erhualian resource population. *BMC Genet* **12**, 24.

de Koning A.P., Gu W., Castoe T.A., Batzer M.A. & Pollock D.D. (2011) Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet* **7**, e1002384.

Derrien T., Estelle J., Marco Sola S., Knowles D.G., Raineri E., Guigo R. & Ribeca P. (2012) Fast computation and applications of genome mappability. *PLoS One* **7**, e30377.

Diamond J. (2002) Evolution, consequences and future of plant and animal domestication. *Nature* **418**, 700-7.

Doan R., Cohen N., Harrington J., Veazy K., Juras R., Cothran G., McCue M.E., Skow L. & Dindot S.V. (2012a) Identification of copy number variants in horses. *Genome Res* **22**, 899-907.

Doan R., Cohen N.D., Sawyer J., Ghaffari N., Johnson C.D. & Dindot S.V. (2012b) Whole-genome sequencing and genetic variant analysis of a Quarter Horse mare. *BMC Genomics* **13**, 78.

Drogemuller C., Distl O. & Leeb T. (2001) Partial deletion of the bovine ED1 gene causes anhidrotic ectodermal dysplasia in cattle. *Genome Res* **11**, 1699-705.

Duret L. & Galtier N. (2009) Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet* **10**, 285-311.

Eggen A. (2012) The development and application of genomic selection as a new breeding paradigm. *Animal Frontiers* **2**, 10-5.

Elferink M.G., Vallee A.A., Jungerius A.P., Crooijmans R.P. & Groenen M.A. (2008) Partial duplication of the PRLR and SPEF2 genes at the late feathering locus in chicken. *BMC Genomics* **9**, 391.

Elsik C.G., Tellam R.L., Worley K.C., Gibbs R.A., Muzny D.M., Weinstock G.M., Adelson D.L., Eichler E.E., Elnitski L., Guigo R., *et al.* (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324**, 522-8.

Ellegren H. (2009) The different levels of genetic diversity in sex chromosomes and autosomes. *Trends Genet* **25**, 278-84.

Eyre-Walker A. & Hurst L.D. (2001) The evolution of isochores. *Nat Rev Genet* **2**, 549-55.

Fabuel E., Barragan C., Silio L., Rodriguez M.C. & Toro M.A. (2004) Analysis of genetic diversity and conservation priorities in Iberian pigs based on microsatellite markers. *Heredity (Edinb)* **93**, 104-13.

Fadista J., Thomsen B., Holm L.E. & Bendixen C. (2010) Copy number variation in the bovine genome. *BMC Genomics* **11**, 284.

Faghihi M.A. & Wahlestedt C. (2009) Regulatory roles of natural antisense transcripts. *Nat Rev Mol Cell Biol* **10**, 637-43.

Fang L., Ye J., Li N., Zhang Y., Li S., Wong G. & Wang J. (2008) Positive correlation between recombination rate and nucleotide diversity is shown under domestication selection in the chicken genome. **53**, 746-50.

Fang M. & Andersson L. (2006) Mitochondrial diversity in European and Chinese pigs is consistent with population expansions that occurred prior to domestication. *Proc Biol Sci* **273**, 1803-10.

Fang M., Hu X., Jiang T., Braunschweig M., Hu L., Du Z., Feng J., Zhang Q., Wu C. & Li N. (2005) The phylogeny of Chinese indigenous pig breeds inferred from microsatellite markers. *Anim Genet* **36**, 7-13.

Fang M., Larson G., Ribeiro H.S., Li N. & Andersson L. (2009) Contrasting mode of evolution at a coat color locus in wild and domestic pigs. *PLoS Genet* **5**, e1000341.

Fay J.C. & Wu C.I. (2000) Hitchhiking under positive Darwinian selection. *Genetics* **155**, 1405-13.

Ferraz A.L., Ojeda A., Lopez-Bejar M., Fernandes L.T., Castello A., Folch J.M. & Perez-Enciso M. (2008) Transcriptome architecture across tissues in the pig. *BMC Genomics* **9**, 173.

Ferretti L., Ramos-Onsins, SE., Pérez-Enciso, M. (2012) Population genomics from next generation sequencing of pooled lineages. *Molecular Ecology (submitted)*.

Flisikowski K., Venhoranta H., Nowacka-Woszuk J., McKay S.D., Flyckt A., Taponen J., Schnabel R., Schwarzenbacher H., Szczerbal I., Lohi H., *et al.* (2010) A novel mutation in the maternally imprinted PEG3 domain results in a loss of MIMT1 expression and causes abortions and stillbirths in cattle (Bos taurus). *PLoS One* **5**, e15116.

Fontanesi L., Beretti F., Martelli P.L., Colombo M., Dall'olio S., Occidente M., Portolano B., Casadio R., Matassino D. & Russo V. (2010a) A first comparative map of copy number variations in the sheep genome. *Genomics* **97**, 158-65.

Fontanesi L., Beretti F., Riggio V., Gomez Gonzalez E., Dall'Olio S., Davoli R., Russo V. & Portolano B. (2009) Copy number variation and missense mutations of the agouti signaling protein (ASIP) gene in goat breeds with different coat colors. *Cytogenet Genome Res* **126**, 333-47.

Fontanesi L., Martelli P.L., Beretti F., Riggio V., Dall'Olio S., Colombo M., Casadio R., Russo V. & Portolano B. (2010b) An initial comparative map of copy number variations in the goat (Capra hircus) genome. *BMC Genomics* **11**, 639.

Freking B.A., Murphy S.K., Wylie A.A., Rhodes S.J., Keele J.W., Leymaster K.A., Jirtle R.L. & Smith T.P. (2002) Identification of the single base change causing the callipyge muscle hypertrophy phenotype, the only known example of polar overdominance in mammals. *Genome Res* **12**, 1496-506.

Frenkel-Morgenstern M., Lacroix V., Ezkurdia I., Levin Y., Gabashvili A., Prilusky J., Del Pozo A., Tress M., Johnson R., Guigo R., *et al.* (2012) Chimeras taking shape: Potential functions of proteins encoded by chimeric RNA transcripts. *Genome Res* **22**, 1231-42.

Fu Y.X. & Li W.H. (1993) Statistical tests of neutrality of mutations. *Genetics* **133**, 693-709.

Fujii J., Otsu K., Zorzato F., de Leon S., Khanna V.K., Weiler J.E., O'Brien P.J. & MacLennan D.H. (1991) Identification of a mutation in porcine ryanodine receptor associated with malignant hyperthermia. *Science* **253**, 448-51.

Garcia-Mas J. (2012) The genome of melon (Cucumis melo L.). *Proceedings of the National Academy of Sciences USA*.

Gardner M.J., Hall N., Fung E., White O., Berriman M., Hyman R.W., Carlton J.M., Pain A., Nelson K.E., Bowman S., *et al.* (2002) Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature* **419**, 498-511.

Genereux D. (2002) Evolution of genomic GC variation. *Genome Biology* **3**, reports0058.

Gibb E.A., Brown C.J. & Lam W.L. (2011) The functional role of long non-coding RNA in human carcinomas. *Mol Cancer* **10**, 38.

Gilad Y., Rosenberg S., Przeworski M., Lancet D. & Skorecki K. (2002) Evidence for positive selection and population structure at the human MAO-A gene. *Proc Natl Acad Sci U S A* **99**, 862-7.

Gill J.L., Capper D., Vanbellinghen J.F., Chung S.K., Higgins R.J., Rees M.I., Shelton G.D. & Harvey R.J. (2011) Startle disease in Irish wolfhounds associated with a microdeletion in the glycine transporter GlyT2 gene. *Neurobiol Dis* **43**, 184-9.

Girirajan S. & Eichler E.E. (2010) Phenotypic variability and genetic susceptibility to genomic disorders. *Hum Mol Genet* **19**, R176-87.

Giuffra E., Evans G., Tornsten A., Wales R., Day A., Looft H., Plastow G. & Andersson L. (1999) The Belt mutation in pigs is an allele at the Dominant white (I/KIT) locus. *Mamm Genome* **10**, 1132-6.

Giuffra E., Kijas J.M., Amarger V., Carlborg O., Jeon J.T. & Andersson L. (2000) The origin of the domestic pig: independent domestication and subsequent introgression. *Genetics* **154**, 1785-91.

Giuffra E., Tornsten A., Marklund S., Bongcam-Rudloff E., Chardon P., Kijas J.M., Anderson S.I., Archibald A.L. & Andersson L. (2002) A large duplication associated with dominant white color in pigs originated by homologous recombination between LINE elements flanking KIT. *Mamm Genome* **13**, 569-77.

Glaus P., Honkela A. & Rattray M. (2012) Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics* **28**, 1721-8.

Goldstein O., Mezey J.G., Boyko A.R., Gao C., Wang W., Bustamante C.D., Anguish L.J., Jordan J.A., Pearce-Kelling S.E., Aguirre G.D., *et al.* (2010) An ADAM9 mutation in canine cone-rod dystrophy 3 establishes homology with human cone-rod dystrophy 9. *Mol Vis* **16**, 1549-69.

Gotz S., Arnold R., Sebastian-Leon P., Martin-Rodriguez S., Tischler P., Jehl M.A., Dopazo J., Rattei T. & Conesa A. (2011) B2G-FAR, a species-centered GO annotation repository. *Bioinformatics* **27**, 919-24.

Gotz S., Garcia-Gomez J.M., Terol J., Williams T.D., Nagaraj S.H., Nueda M.J., Robles M., Talon M., Dopazo J. & Conesa A. (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res* **36**, 3420-35.

Grisart B., Coppieters W., Farnir F., Karim L., Ford C., Berzi P., Cambisano N., Mni M., Reid S., Simon P., *et al.* (2002) Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. *Genome Res* **12**, 222-31.

Grobet L., Martin L.J., Poncelet D., Pirottin D., Brouwers B., Riquet J., Schoeberlein A., Dunner S., Menissier F., Massabanda J., *et al.* (1997) A deletion in the bovine myostatin gene causes the double-muscled phenotype in cattle. *Nat Genet* **17**, 71-4.

Groenen M.A.M., Archibald, A.L., Uenishi H. (2012) Pig genomes provide insight into porcine demography and evolution. (submitted).

Grossman S.R., Shlyakhter I., Karlsson E.K., Byrne E.H., Morales S., Frieden G., Hostetter E., Angelino E., Garber M., Zuk O., *et al.* (2010) A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* **327**, 883-6.

Gupta R.A., Shah N., Wang K.C., Kim J., Horlings H.M., Wong D.J., Tsai M.C., Hung T., Argani P., Rinn J.L., *et al.* (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464**, 1071-6.

Guttman M., Donaghey J., Carey B.W., Garber M., Grenier J.K., Munson G., Young G., Lucas A.B., Ach R., Bruhn L., *et al.* (2011) lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* **477**, 295-300.

Guyonnet-Duperat V., Geverink N., Plastow G.S., Evans G., Ousova O., Croisetiere C., Foury A., Richard E., Mormede P. & Moisan M.P. (2006) Functional implication of an Arg307Gly substitution in corticosteroid-binding globulin, a candidate gene for a quantitative trait locus associated with cortisol variability and obesity in pig. *Genetics* **173**, 2143-9.

Haerty W., Jagadeeshan S., Kulathinal R.J., Wong A., Ravi Ram K., Sirot L.K., Levesque L., Artieri C.G., Wolfner M.F., Civetta A., *et al.* (2007) Evolution in the fast lane: rapidly evolving sex-related genes in Drosophila. *Genetics* **177**, 1321-35.

Hammer M.F., Mendez F.L., Cox M.P., Woerner A.E. & Wall J.D. (2008) Sex-biased evolutionary forces shape genomic patterns of human diversity. *PLoS Genet* **4**, e1000202.

Harrison P.W., Wright A.E. & Mank J.E. (2012) The evolution of gene expression and the transcriptome-phenotype relationship. *Semin Cell Dev Biol* **23**, 222-9.

He D., Zaitlen N., Pasaniuc B., Eskin E. & Halperin E. (2011) Genotyping common and rare variation using overlapping pool sequencing. *BMC Bioinformatics* **12 Suppl 6**, S2.

Hellmann I., Ebersberger I., Ptak S.E., Paabo S. & Przeworski M. (2003) A neutral explanation for the correlation of diversity with recombination rates in humans. *Am J Hum Genet* **72**, 1527-35.

Hellmann I., Mang Y., Gu Z., Li P., de la Vega F.M., Clark A.G. & Nielsen R. (2008) Population genetic analysis of shotgun assemblies of genomic sequences from multiple individuals. *Genome Res* **18**, 1020-9.

Hellmann I., Prüfer K., Ji H., Zody M.C., Pääbo S. & Ptak S.E. (2005) Why do human diversity levels vary at a megabase scale? *Genome Research* **15**, 1222-31.

Holsinger K.E. (2001-2010) *Lecture Notes in Population Genetics*. Department of Ecology & Evolutionary Biology.

Hou Y., Liu G.E., Bickhart D.M., Cardone M.F., Wang K., Kim E.S., Matukumalli L.K., Ventura M., Song J., VanRaden P.M., *et al.* (2011) Genomic characteristics of cattle copy number variations. *BMC Genomics* **12**, 127.

Huang T.H., Zhu M.J., Li X.Y. & Zhao S.H. (2008) Discovery of porcine microRNAs and profiling from skeletal muscle tissues during development. *PLoS One* **3**, e3225.

Huarte M., Guttman M., Feldser D., Garber M., Koziol M.J., Kenzelmann-Broz D., Khalil A.M., Zuk O., Amit I., Rabani M., *et al.* (2010) A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell* **142**, 409-19.

Hudson R.R., Kreitman M. & Aguade M. (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**, 153-9.

Hung T., Wang Y., Lin M.F., Koegel A.K., Kotake Y., Grant G.D., Horlings H.M., Shah N., Umbricht C., Wang P., *et al.* (2011) Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. *Nat Genet* **43**, 621-9.

Inomata N. & Yamazaki T. (2002) Nucleotide variation of the duplicated amylase genes in Drosophila kikkawai. *Mol Biol Evol* **19**, 678-88.

Jensen-Seaman M.I., Furey T.S., Payseur B.A., Lu Y., Roskin K.M., Chen C.F., Thomas M.A., Haussler D. & Jacob H.J. (2004) Comparative recombination rates in the rat, mouse, and human genomes. *Genome Res* **14**, 528-38.

Jensen J.D., Thornton K.R. & Andolfatto P. (2008) An approximate bayesian estimator suggests strong, recurrent selective sweeps in Drosophila. *PLoS Genet* **4**, e1000198.

Jones G.F. (1998) Genetic aspects of domestication, common breeds and their origin. In: (pp. 17-50. CAB INTERNATIONAL, Wallingford.

Jorde L.B. & Wooding S.P. (2004) Genetic variation, classification and 'race'. *Nat Genet* **36**, S28-33.

Kahvejian A., Quackenbush J. & Thompson J.F. (2008) What would you do if you could sequence everything? *Nat Biotechnol* **26**, 1125-33.

Kelley J.L., Madeoy J., Calhoun J.C., Swanson W. & Akey J.M. (2006) Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Res* **16**, 980-9.

Kerstens H.H., Crooijmans R.P., Veenendaal A., Dibbits B.W., Chin A.W.T.F., den Dunnen J.T. & Groenen M.A. (2009) Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: applied to turkey. *BMC Genomics* **10**, 479.

Kim S.Y., Li Y., Guo Y., Li R., Holmkvist J., Hansen T., Pedersen O., Wang J. & Nielsen R. (2010) Design of association studies with pooled or un-pooled next-generation sequencing data. *Genet Epidemiol* **34**, 479-91.

Langmead B., Hansen K.D. & Leek J.T. (2010) Cloud-scale RNA-sequencing differential expression analysis with Myrna. *Genome Biol* **11**, R83.

Langmead B., Schatz M.C., Lin J., Pop M. & Salzberg S.L. (2009) Searching for SNPs with cloud computing. *Genome Biol* **10**, R134.

Larkin D.M., Daetwyler H.D., Hernandez A.G., Wright C.L., Hetrick L.A., Boucek L., Bachman S.L., Band M.R., Akraiko T.V., Cohen-Zinder M., *et al.* (2012) Whole-genome resequencing of two elite sires for the detection of haplotypes under selection in dairy cattle. *Proc Natl Acad Sci U S A* **109**, 7693-8.

Larson G., Dobney K., Albarella U., Fang M., Matisoo-Smith E., Robins J., Lowden S., Finlayson H., Brand T., Willerslev E., *et al.* (2005) Worldwide phylogeography of wild boar reveals multiple centers of pig domestication. *Science* **307**, 1618-21.

Lercher M.J. & Hurst L.D. (2002) Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet* **18**, 337-40.

Levy S., Sutton G., Ng P.C., Feuk L., Halpern A.L., Walenz B.P., Axelrod N., Huang J., Kirkness E.F., Denisov G., *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biol* **5**, e254.

Li G., Bahn J.H., Lee J.H., Peng G., Chen Z., Nelson S.F. & Xiao X. (2012a) Identification of allele-specific alternative mRNA processing via transcriptome sequencing. *Nucleic Acids Res*.

Li H. & Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-60.

Li J., Li H., Jakobsson M., Li S., Sjodin P. & Lascoux M. (2012b) Joint analysis of demography and selection in population genetics: where do we stand and where could we go? *Mol Ecol* **21**, 28-44.

Li J., Yang H., Li J.R., Li H.P., Ning T., Pan X.R., Shi P. & Zhang Y.P. (2010a) Artificial selection of the melanocortin receptor 1 gene in Chinese domestic pigs during domestication. *Heredity (Edinb)* **105**, 274-81.

Li M., Xia Y., Gu Y., Zhang K., Lang Q., Chen L., Guan J., Luo Z., Chen H., Li Y., *et al.* (2010b) MicroRNAome of porcine pre- and postnatal development. *PLoS One* **5**, e11541.

Li S. & Jakobsson M. (2012) Estimating demographic parameters from large-scale population genomic data using Approximate Bayesian Computation. *BMC Genet* **13**, 22.

Li Y., Zheng H., Luo R., Wu H., Zhu H., Li R., Cao H., Wu B., Huang S., Shao H., *et al.* (2011) Structural variation in two human genomes mapped at single-nucleotide resolution by whole genome de novo assembly. *Nat Biotechnol* **29**, 723-30.

Lian C., Sun B., Niu S., Yang R., Liu B., Lu C., Meng J., Qiu Z., Zhang L. & Zhao Z. (2012) A comparative profile of the microRNA transcriptome in immature and mature porcine testes using Solexa deep sequencing. *FEBS J*.

Liu G.E., Hou Y., Zhu B., Cardone M.F., Jiang L., Cellamare A., Mitra A., Alexander L.J., Coutinho L.L., Dell'Aquila M.E., *et al.* (2010) Analysis of copy number variations among diverse cattle breeds. *Genome Research* **20**, 693-703.

Lopez-Bote C.J. (1998) Sustained utilization of the Iberian pig breed. *Meat Sci* **49S1**, S17-27.

Luetkemeier E.S., Malhi R.S., Beever J.E. & Schook L.B. (2009) Diversification of porcine MHC class II genes: evidence for selective advantage. *Immunogenetics* **61**, 119-29.

Luetkemeier E.S., Sodhi M., Schook L.B. & Malhi R.S. (2010) Multiple Asian pig origins revealed through genomic analyses. *Mol Phylogenet Evol* **54**, 680-6.

Marais G. (2003) Biased gene conversion: implications for genome and sex evolution. *Trends Genet* **19**, 330-8.

Markow T., Hedrick P.W., Zuerlein K., Danilovs J., Martin J., Vyvial T. & Armstrong C. (1993) HLA polymorphism in the Havasupai: evidence for balancing selection. *Am J Hum Genet* **53**, 943-52.

McDonald J.H. & Kreitman M. (1991) Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**, 652-4.

Medvedev P., Fiume M., Dzamba M., Smith T. & Brudno M. (2010) Detecting copy number variation with mated short reads. *Genome Res* **20**, 1613-22.

Megens H.J., Crooijmans R.P., San Cristobal M., Hui X., Li N. & Groenen M.A. (2008) Biodiversity of pig breeds from China and Europe estimated from pooled DNA samples: differences in microsatellite variation between two areas of domestication. *Genet Sel Evol* **40**, 103-28.

Mercade A., Estelle J., Perez-Enciso M., Varona L., Silio L., Noguera J.L., Sanchez A. & Folch J.M. (2006) Characterization of the porcine acyl-CoA synthetase long-chain 4 gene and its association with growth and meat quality traits. *Anim Genet* **37**, 219-24.

Mercer T.R., Dinger M.E. & Mattick J.S. (2009) Long non-coding RNAs: insights into functions. *Nat Rev Genet* **10**, 155-9.

Messier W. & Stewart C.B. (1997) Episodic adaptive evolution of primate lysozymes. *Nature* **385**, 151-4.

Meuwissen T.H. & Goddard M.E. (2000) Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics* **155**, 421-30.

Meyers S.N., McDaneld T.G., Swist S.L., Marron B.M., Steffen D.J., O'Toole D., O'Connell J.R., Beever J.E., Sonstegard T.S. & Smith T.P. (2010) A deletion mutation in bovine SLC4A2 is associated with osteopetrosis in Red Angus cattle. *BMC Genomics* **11**, 337.

Milan D., Jeon J.T., Looft C., Amarger V., Robic A., Thelander M., Rogel-Gaillard C., Paul S., Iannuccelli N., Rask L., *et al.* (2000) A mutation in PRKAG3 associated with excess glycogen content in pig skeletal muscle. *Science* **288**, 1248-51.

Moran V.A., Perera R.J. & Khalil A.M. (2012) Emerging functional and mechanistic paradigms of mammalian long non-coding RNAs. *Nucleic Acids Res*.

Mulsant P., Lecerf F., Fabre S., Schibler L., Monget P., Lanneluc I., Pisselet C., Riquet J., Monniaux D., Callebaut I., *et al.* (2001) Mutation in bone morphogenetic protein receptor-IB is associated with increased ovulation rate in Booroola Merino ewes. *Proc Natl Acad Sci U S A* **98**, 5104-9.

Muotri A.R., Marchetto M.C., Coufal N.G. & Gage F.H. (2007) The necessary junk: new functions for transposable elements. *Hum Mol Genet* **16 Spec No. 2**, R159-67.

Nachman M.W. (2002) Variation in recombination rate across the genome: evidence and implications. *Curr Opin Genet Dev* **12**, 657-63.

Nei M. (1973) Analysis of gene diversity in subdivided populations. *Proc Natl Acad Sci U S A* **70**, 3321-3.

Norris B.J. & Whan V.A. (2008) A gene duplication affecting expression of the ovine ASIP gene is responsible for white and black sheep. *Genome Res* **18**, 1282-93.

Ohba Y., Kitagawa H., Kitoh K., Sasaki Y., Takami M., Shinkai Y. & Kunieda T. (2000) A deletion of the paracellin-1 gene is responsible for renal tubular dysplasia in cattle. *Genomics* **68**, 229-36.

Ojeda A., Estelle J., Folch J.M. & Perez-Enciso M. (2008a) Nucleotide variability and linkage disequilibrium patterns at the porcine FABP5 gene. *Anim Genet* **39**, 468-73.

Ojeda A., Huang L.S., Ren J., Angiolillo A., Cho I.C., Soto H., Lemus-Flores C., Makuza S.M., Folch J.M. & Perez-Enciso M. (2008b) Selection in the making: a worldwide survey of haplotypic diversity around a causative mutation in porcine IGF2. *Genetics* **178**, 1639-52.

Ojeda A., Rozas J., Folch J.M. & Perez-Enciso M. (2006) Unexpected high polymorphism at the FABP4 gene unveils a complex history for pig populations. *Genetics* **174**, 2119-27.

Olsson M., Meadows J.R., Truve K., Rosengren Pielberg G., Puppo F., Mauceli E., Quilez J., Tonomura N., Zanna G., Docampo M.J., *et al.* (2011) A novel unstable duplication upstream of HAS2 predisposes to a breed-defining skin phenotype and a periodic fever syndrome in Chinese Shar-Pei dogs. *PLoS Genet* **7**, e1001332.

Orom U.A., Derrien T., Beringer M., Gumireddy K., Gardini A., Bussotti G., Lai F., Zytnicki M., Notredame C., Huang Q., *et al.* (2010) Long noncoding RNAs with enhancer-like function in human cells. *Cell* **143**, 46-58.

Ousova O., Guyonnet-Duperat V., Iannuccelli N., Bidanel J.P., Milan D., Genet C., Llamas B., Yerle M., Gellin J., Chardon P., *et al.* (2004) Corticosteroid binding globulin: a new target for cortisol-driven obesity. *Mol Endocrinol* **18**, 1687-96.

Pailhoux E., Vigier B., Chaffaux S., Servel N., Taourit S., Furet J.P., Fellous M., Grosclaude F., Cribiu E.P., Cotinot C., *et al.* (2001) A 11.7-kb deletion triggers intersexuality and polledness in goats. *Nat Genet* **29**, 453-8.

Parker H.G., Kukekova A.V., Akey D.T., Goldstein O., Kirkness E.F., Baysac K.C., Mosher D.S., Aguirre G.D., Acland G.M. & Ostrander E.A. (2007) Breed relationships facilitate fine-mapping studies: a 7.8-kb deletion cosegregates with Collie eye anomaly across multiple dog breeds. *Genome Res* **17**, 1562-71.

Pasaniuc B., Zaitlen N. & Halperin E. (2011) Accurate estimation of expression levels of homologous genes in RNA-seq experiments. *J Comput Biol* **18**, 459-68.

Perez-Enciso M. & Ferretti L. (2010) Massive parallel sequencing in animal genetics: wherefroms and wheretos. *Anim Genet* **41**, 561-9.

Porter V. (1993) *Pigs: A Handbook to the Breeds of the World*. Mountfiled: Helm Information Ltd.

Price E.O. (1999) Behavioral development in animals undergoing domestication. *Applied Animal Behaviour Science* **65**, 245-71.

Pritchard J.K., Seielstad M.T., Perez-Lezaun A. & Feldman M.W. (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol* **16**, 1791-8.

Quilter C.R., Bagga M., Moinie A., Junaid F. & Sargent C.A. (2012) Gene structure and expression of serotonin receptor HTR2C in hypothalamic samples from infanticidal and control sows. *BMC Neurosci* **13**, 37.

Raineri E., Ferretti, L., Esteve-Codina A., Nevado B., Heath S., Pérez-Enciso M. (2012) SNP calling and computing allele frequency by sequencing pooled samples. *BMC Bioinformatics (submitted)*.

Ramayo-Caldas Y., Castello A., Pena R.N., Alves E., Mercade A., Souza C.A., Fernandez A.I., Perez-Enciso M. & Folch J.M. (2010) Copy number variation in the porcine genome inferred from a 60 k SNP BeadChip. *BMC Genomics* **11**, 593.

Ramirez O., Ojeda A., Tomas A., Gallardo D., Huang L.S., Folch J.M., Clop A., Sanchez A., Badaoui B., Hanotte O., *et al.* (2009) Integrating Y-chromosome, mitochondrial, and autosomal data to analyze the origin of pig breeds. *Mol Biol Evol* **26**, 2061-72.

Rinn J.L., Kertesz M., Wang J.K., Squazzo S.L., Xu X., Brugmann S.A., Goodnough L.H., Helms J.A., Farnham P.J., Segal E., *et al.* (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **129**, 1311-23.

Rosengren Pielberg G., Golovko A., Sundstrom E., Curik I., Lennartsson J., Seltenhammer M.H., Druml T., Binns M., Fitzsimmons C., Lindgren G., *et al.* (2008) A cis-acting regulatory mutation causes premature hair graying and susceptibility to melanoma in the horse. *Nat Genet* **40**, 1004-9.

Rubin C.J., Zody M.C., Eriksson J., Meadows J.R., Sherwood E., Webster M.T., Jiang L., Ingman M., Sharpe T., Ka S., *et al.* (2010) Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature* **464**, 587-91.

Salmon Hillbertz N.H., Isaksson M., Karlsson E.K., Hellmen E., Pielberg G.R., Savolainen P., Wade C.M., von Euler H., Gustafson U., Hedhammar A., *et al.* (2007) Duplication of FGF3, FGF4, FGF19 and ORAOV1 causes hair ridge and predisposition to dermoid sinus in Ridgeback dogs. *Nat Genet* **39**, 1318-20.

Salzberg S.L. & Yorke J.A. (2005) Beware of mis-assembled genomes. *Bioinformatics* **21**, 4320-1.

Scandura M., Iacolina L., Crestanello B., Pecchioli E., Di Benedetto M.F., Russo V., Davoli R., Apollonio M. & Bertorelle G. (2008) Ancient vs. recent processes as factors shaping the genetic variation of the European wild boar: are the effects of the last glaciation still detectable? *Mol Ecol* **17**, 1745-62.

Sebat J., Lakshmi B., Malhotra D., Troge J., Lese-Martin C., Walsh T., Yamrom B., Yoon S., Krasnitz A., Kendall J., *et al.* (2007) Strong association of de novo copy number mutations with autism. *Science* **316**, 445-9.

Sella G., Petrov D.A., Przeworski M. & Andolfatto P. (2009) Pervasive natural selection in the Drosophila genome? *PLoS Genet* **5**, e1000495.

Sidjanin D.J., Lowe J.K., McElwee J.L., Milne B.S., Phippen T.M., Sargan D.R., Aguirre G.D., Acland G.M. & Ostrander E.A. (2002) Canine CNGB3 mutations establish cone degeneration as orthologous to the human achromatopsia locus ACHM3. *Hum Mol Genet* **11**, 1823-33.

Skelly D.A., Johansson M., Madeoy J., Wakefield J. & Akey J.M. (2011) A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Res* **21**, 1728-37.

Souza C.A., Paiva S.R., Pereira R.W., Guimaraes S.E., Dutra W.M., Jr., Murata L.S. & Mariante A.S. (2009) Iberian origin of Brazilian local pig breeds based on Cytochrome b (MT-CYB) sequence. *Anim Genet* **40**, 759-62.

Spencer C.C., Deloukas P., Hunt S., Mullikin J., Myers S., Silverman B., Donnelly P., Bentley D. & McVean G. (2006) The influence of recombination on human genetic diversity. *PLoS Genet* **2**, e148.

Stanke M., Diekhans M., Baertsch R. & Haussler D. (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637-44.

Stanke M., Keller O., Gunduz I., Hayes A., Waack S. & Morgenstern B. (2006a) AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* **34**, W435-9.

Stanke M. & Morgenstern B. (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res* **33**, W465-7.

Stanke M., Schoffmann O., Morgenstern B. & Waack S. (2006b) Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**, 62.

Stanke M., Steinkamp R., Waack S. & Morgenstern B. (2004) AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res* **32**, W309-12.

Stanke M., Tzvetkova A. & Morgenstern B. (2006c) AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biol* **7 Suppl 1**, S11 1-8.

Stanke M. & Waack S. (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19 Suppl 2**, ii215-25.

Tajima F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585-95.

Tavare S., Balding D.J., Griffiths R.C. & Donnelly P. (1997) Inferring coalescence times from DNA sequence data. *Genetics* **145**, 505-18.

Tishkoff S.A. & Kidd K.K. (2004) Implications of biogeography of human populations for 'race' and medicine. *Nat Genet* **36**, S21-7.

Tishkoff S.A. & Verrelli B.C. (2003) Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annu Rev Genomics Hum Genet* **4**, 293-340.

Tong P., Prendergast J.G., Lohan A.J., Farrington S.M., Cronin S., Friel N., Bradley D.G., Hardiman O., Evans A., Wilson J.F., *et al.* (2010) Sequencing and analysis of an Irish human genome. *Genome Biol* **11**, R91.

Toro M.A. (2008) Genealogical Analysis of a Closed Herd of Black Hairless Iberian Pigs. *Conservation biology*.

Toro M.A., Rodrigañez J., Silio L. & Rodriguez C. (2000) Genealogical Analysis of a Closed Herd of Black Hairless Iberian Pigs

Análisis Genealógico de una Manada Aislada del Cerdo Ibérico Lampiño Negro. *Conservation biology* **14**, 1843-51.

Trut L., Oskina I. & Kharlamova A. (2009) Animal evolution during domestication: the domesticated fox as a model. *Bioessays* **31**, 349-60.

USDA (2010) Livestock and Poultry: World Markets and Trade. (ed. by Service FA).

Vacic V., McCarthy S., Malhotra D., Murray F., Chou H.H., Peoples A., Makarov V., Yoon S., Bhandari A., Corominas R., *et al.* (2011) Duplications of the neuropeptide receptor gene VIPR2 confer significant risk for schizophrenia. *Nature* **471**, 499-503.

Vaknin K., Goren A. & Ast G. (2009) TEs or not TEs? That is the evolutionary question. *Journal of Biology* **8**, 83.

van De Sluis B., Rothuizen J., Pearson P.L., van Oost B.A. & Wijmenga C. (2002) Identification of a new copper metabolism gene by positional cloning in a purebred dog population. *Hum Mol Genet* **11**, 165-73.

Van Laere A.S., Nguyen M., Braunschweig M., Nezer C., Collette C., Moreau L., Archibald A.L., Haley C.S., Buys N., Tally M., *et al.* (2003) A regulatory mutation in IGF2 causes a major QTL effect on muscle growth in the pig. *Nature* **425**, 832-6.

Van Tassell C.P., Smith T.P., Matukumalli L.K., Taylor J.F., Schnabel R.D., Lawley C.T., Haudenschild C.D., Moore S.S., Warren W.C. & Sonstegard T.S. (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat Methods* **5**, 247-52.

Vicoso B. & Charlesworth B. (2006) Evolution on the X chromosome: unusual patterns and processes. *Nat Rev Genet* **7**, 645-53.

Wang Z., Gerstein M. & Snyder M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57-63.

Watts J.M., Dang K.K., Gorelick R.J., Leonard C.W., Bess J.W., Jr., Swanstrom R., Burch C.L. & Weeks K.M. (2009) Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* **460**, 711-6.

Wheeler D.A., Srinivasan M., Egholm M., Shen Y., Chen L., McGuire A., He W., Chen Y.J., Makhijani V., Roth G.T., *et al.* (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872-6.

Wiedmann R.T., Smith T.P. & Nonneman D.J. (2008) SNP discovery in swine by reduced representation and high throughput pyrosequencing. *BMC Genet* **9**, 81.

Wiener P. & Wilkinson S. (2011) Deciphering the genetic basis of animal domestication. *Proc Biol Sci* **278**, 3161-70.

Wright D., Boije H., Meadows J.R., Bed'hom B., Gourichon D., Vieaud A., Tixier-Boichard M., Rubin C.J., Imsland F., Hallbook F., *et al.* (2009) Copy number variation in intron 1 of SOX5 causes the Pea-comb phenotype in chickens. *PLoS Genet* **5**, e1000512.

Wright S.I., Bi I.V., Schroeder S.G., Yamasaki M., Doebley J.F., McMullen M.D. & Gaut B.S. (2005) The effects of artificial selection on the maize genome. *Science* **308**, 1310-4.

Wu G.S., Yao Y.G., Qu K.X., Ding Z.L., Li H., Palanichamy M.G., Duan Z.Y., Li N., Chen Y.S. & Zhang Y.P. (2007) Population phylogenomic analysis of mitochondrial DNA in wild boars and domestic pigs revealed multiple domestication events in East Asia. *Genome Biol* **8**, R245.

Xie S.S., Li X.Y., Liu T., Cao J.H., Zhong Q. & Zhao S.H. (2011) Discovery of porcine microRNAs in multiple tissues by a Solexa deep sequencing approach. *PLoS One* **6**, e16235.

Xu X., Pan S., Cheng S., Zhang B., Mu D., Ni P., Zhang G., Yang S., Li R., Wang J., *et al.* (2011) Genome sequence and analysis of the tuber crop potato. *Nature* **475**, 189-95.

Zaboli G., Ameur A., Igl W., Johansson A., Hayward C., Vitart V., Campbell S., Zgaga L., Polasek O., Schmitz G., *et al.* (2011) Sequencing of high-complexity DNA pools for identification of nucleotide and structural variants in regions associated with complex traits. *Eur J Hum Genet* **20**, 77-83.

Zeng K., Fu Y.X., Shi S. & Wu C.I. (2006) Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* **174**, 1431-9.

Zhang C. & Plastow G. Genomic Diversity in Pig (Sus scrofa) and its Comparison with Human and other Livestock. *Curr Genomics* **12**, 138-46.

Zhang F., Gu W., Hurles M.E. & Lupski J.R. (2009) Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet* **10**, 451-81.

Zhao X., Mo D., Li A., Gong W., Xiao S., Zhang Y., Qin L., Niu Y., Guo Y., Liu X., *et al.* (2011) Comparative analyses by sequencing of transcriptomes during skeletal muscle development between pig breeds differing in muscle growth rate and fatness. *PLoS One* **6**, e19774.

AKNOWLEGMENTS

I would like to thank my thesis director Miguel Pérez-Enciso to give me the opportunity to develop this exciting research with the most cutting-edge technologies in the area of bioinformatics and genomics, for pushing me to assist and present in many international meetings, to send me abroad every year to gain more experience in new techniques and methodologies and to provide me with a solid background to keep on with my future research. To my co-director, Josep Maria Folch, better known as 'JM', for giving me support at the very beginning of my thesis with your immense experience in the molecular biology field and for being a very good teacher during my Master's.

Thanks Yuliaxis for your inconditional support, your honestity and the nice times we spent together at work and travelling around (Toulouse, Carmona, París and Amsterdam). Thanks William, for your friendly personality, the funny times you gave and your help with statistics. Thanks Sebas, for being such a noble person, your jokes, for bothering me all the time ;), and for your help in population genetics! Thanks Elisa for being my friend and confidant. Thanks Sandra for your hospitality, the dinner's at your place, your infectious laugh and for being such a nice person. Thanks Irene, for your friendliness and optimism, a girl who is always smiling.

Thanks Maria Ballester for encouraging me during my last period. Thanks Ali for our nice conversations during the 'sandwich time'. Thanks Cabrels for being such a crazy and funny girl and always with a smile, Merkels for your sweetness and Barrels for your friendliness. Thanks Castelló, a just and persevering woman always ready to help people at work. Thanks Corominas and Puig, for the nice time we spent at the mud beach in Menorca.

Thanks Bruno, for your help at work and for the nice portuguese words coming out from your mouth, Erica for being such an easy-going person, Sarai for your nice drawings and roomate company, Ariana for your tenderness, Oriol for your happiness and motivation, Xavi, somebody to talk about everything, Ingrid for

your positivity and self-confidence. Thanks Natalia, for your way to happy-live the life without too much worries. Thanks John, for the intellectual and philosophical conversations. Thanks Carola for your delicious typical Peruvian dishes, Yang Bin for your kindness. Thanks Boa and Estellé, two such nice guys whom I appreciate a lot.

Thanks for the friendship and joy of the SGM service: las 'Pin y Pon', Carme, Eli and Lorena. Thanks to the people I shared a nice working environment at the Veterinary Faculty when I started: Aïnhoa, Maria, David, Olga and Laura. Thanks to other group leaders Marcel and Armand.

Thanks Ojeda, for the friendship we had during the year and a half we spent together at the Veterinarian Faculty. Although we do not see each other often, I always think about you and our funny times we spent together. Thanks Maribel to baptize me with the name of 'Queen'.

Thanks to the wise people from the VetmedUni, where I spent 4 months and learnt a lot about RNA-seq. Thanks to the pleasant people in Wageningen, where I spent 3 months and trained me in structural variants detection. Thanks to the lovely people from CIPF, where I spent 2 months and gained some experience in gene annotation.

Thanks to my best friend Clara who always gives me her support in my worst moments and joy in my best moments. Thanks to my best friend Vane for your loyalty and lovely moments we spent since we were teenagers. Thanks Núria for praising my work, for your strength and fortitude.

Thanks to my best group of friends Nené, Aida, Meri, Paula & Gemma for our crazy nights, parties, excursions, beers and more.

Thanks to my family to believe in me and for all the love I received.

Thanks to my love, my company.

APPENDIX

CURRICULUM VITAE

SHORT BIOGRAPHY

Anna Esteve Codina was born Barcelona in 1981, she graduated in Chemistry in 2003 and Biochemistry in 2005 at the Barcelona University. Then she worked as a technician at the Molecular Biology Unit of the Tissue and Blood Bank located at the Vall Hebrón Hospital in Barcelona. In 2007, she moved for half a year in Austria to carry out research in forensic genetics at the Legal Medicine Institute in Innbruck. In 2008, she obtained a FPI fellowship from the Spain Science Ministry to perform a Master's degree in Genetics and a PhD at the Animal Genetics department of the Autonoma University in Barcelona. During the last four years, she also stayed abroad for a short-term period at the Population Genetics Group of the Veterinary Medicine University in Vienna, the Animal Breeding and Genomics Department of the Wageningen University in Holland and the Genomics Expression Group of the Príncipe Felipe Investigation Center in Valencia. Since September 2012, she has a postdoctoral position at one of the biggest genome sequencing centers in Europe, the Genomic Analysis National Center (CNAG) in Barcelona.

PUBLICATIONS

Esteve-Codina A., Paudel, Y., Ferretti L., Rainieri E., Ramos-Onsins S., Megens H.J, Silió L., Pérez-Enciso M. *Dissecting structural and nucleotide genomewide variation in inbred Iberian pigs*. (under development).

Raineri E., Ferretti L., Esteve-Codina A., Nevado B., Heath S., Pérez-Enciso M. *SNP calling and Computing Allele Frequency by sequencing pooled samples*. (submitted to BMC Bioinformatics)

W. Burgos-Paz, C.A. Souza, H.J. Megens, Y. Ramayo-Caldas, M. Melo, C. Lemús-Flores, A. Loarca, H.W. Soto, R. Martínez, L.A. Álvarez, L. Aguirre, V. Iñiguez, M.A. Revidatti, O.R. Martínez-López, S. Llambi, A. Esteve-Codina, R.P.M.A. Crooijmans, S.R. Paiva, L.B. Schook, M.A.M. Groenen, M. Pérez-Enciso. *The porcine colonization of the Americas: A 60k SNP story* (submitted to Heredity).

Yuliaxis Ramayo-Caldas, Nuria Mach, Anna Esteve-Codina, Jordi Corominas, Anna Castelló, Maria Ballester, Jordi Estellé, N. Ibáñez-Escriche, Ana I. Fernández, Miguel Pérez-Enciso, Josep M. Folch. *Liver transcriptome profile in pigs with extreme phenotypes of intramuscular fatty acid composition* (submitted to BMC Genomics)

Esteve-Codina A, Kofler R, Palmieri N, Bussotti G, Notredame C, Pérez-Enciso M. *Exploring the gonad transcriptome of two extreme male pigs with RNA-seq. BMC Genomics*. 2011 Nov 8;12:552.

Esteve-Codina A, Kofler R, Himmelbauer H, Ferretti L, Vivancos AP, Groenen MA, Folch JM, Rodríguez MC, Pérez-Enciso M. *Partial short-read sequencing of a highly inbred Iberian pig and genomics inference thereof. Heredity* (Edinb). 2011 Sep;107(3):256-64. doi: 10.1038/hdy.2011.13. Epub 2011 Mar 16.

Esteve A, Ojeda A, Huang LS, Folch JM, Pérez-Enciso M. *Nucleotide variability of the porcine SERPINA6 gene and the origin of a putative causal mutation associated* *with meat quality*. *Animal Genetics*. 2011 Jun;42(3):235-41. doi: 10.1111/j.1365-2052.2010.02138.x. Epub 2010 Nov 4.

A. Ferrer-Admetlla, M. Sikora, H. Laayouni, A. Esteve, F. Roubinet, A. Blancher, F. Calafell, J. Bertranpetit, and F. Casals. *A Natural History of FUT2 Polymorphism in Humans*. *Mol. Biol. Evol.*, September 1, 2009; 26(9): 1993 - 2003.

Esteve Codina A, Niederstätter H, and Parson W. *"GenderPlex" a PCR multiplex for reliable gender determination of degraded human DNA samples and complex gender constellations*. *International Journal of Legal Medicine*. 123(6):459-64, 2009 Nov.

PRESENTATIONS & POSTERS

Esteve-Codina A., Paudel, Y., Ferretti L., Rainieri E., Ramos-Onsins S., Megens H.J, Silió L., Pérez-Enciso M. *Genome-wide nucleotide diversity of Iberian pigs* (oral presentation). Pig diversity and evolution, May 2012, Menorca.

Anna Esteve, Robert Kofler, Miguel Pérez Enciso. *Measuring gene expression in gonads of two extreme pig breeds with RNAseq* (oral presentation). 3rd COST ACTION Statseq Workshop, April 2011, Toulouse.

A.Esteve, R.Kofler, A.P.Vivancos, H.Himmelbauer, MAM Groenen, JM.Folch, MC.Rodríguez, M.Pérez-Enciso. *Identification of polymorphisms from ultrasequencing data: Comparison of tools* (oral presentation). 1st COST Action StatSeq Workshop, October 2009, Barcelona

A.Esteve, R.Kofler, A.P.Vivancos, H.Himmelbauer, MAM Groenen, JM.Folch, MC.Rodríguez, M.Pérez-Enciso. *Partial short-read resequencing of a highly inbred Iberian pig* (poster). Next generation sequencing: Challenges and opportunities, October 2009, Barcelona

A.Esteve, R.Kofler, A.P.Vivancos, H.Himmelbauer, MAM Groenen, JM.Folch, MC.Rodríguez, M.Pérez-Enciso. *Partial short-read resequencing of a highly inbred Iberian pig*. (oral presentation).Jornades de Biologia Evolutiva, July 2009, Barcelona

Esteve, A., Ojeda, A., Folch, J.M., Pérez-Enciso, M. *Variabilidad nucleotídica del gen cortisol binding globulin (CBG)* (oral presentation). XIII Jornadas de Producción Animal , May 2009, Zaragoza.

Esteve Codina A, Niederstätter H, and Parson W. *Genderplex: a new PCR multiplex for human gender determination* (oral presentation). DNA in Forensics, May 2008, Ancona. 6th International forensic Y-user workshop and the 3rd EMPOP Meeting.

COLOPHON

Anna Esteve Codina is recipient of a FPI PhD fellowship from Ministry of Education, Spain. Work funded by AGL2010-14822 grant to Miguel Pérez Enciso and by a Consolider grant from Spanish Ministry of Research, CSD2007-00036 "Centre for Research in Agrigenomics".