Sequencing the Iberian pig genomes

A. Esteve*, R. Kofler'***, H. Himmelbauer***, M. Groenen****,
MC. Rodríguez***** and M. Pérez-Enciso***

*ICREA, Universitat Autònoma de Barcelona, Bellaterra (Spain)
**VetMedUni Wien, Vienna (Austria)
***Centre for Genomic Regulation (CRG), Barcelona (Spain)
****Wageningen University and Research Centre, Wageningen (The Netherlands)
*****Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA), Madrid (Spain)

Abstract. The Iberian pig is one of the most important varieties of local pig breeds, both economically and culturally. Contrary to most widespread breeds, like Duroc or Large White, it seems not to have been intercrossed with Chinese pigs. Modern sequencing technologies allow us to sequence and analyze complete genomes for costs in the order of a few thousand euros, democratizing the access to this kind of data. Here we present preliminary results of the partial sequencing a highly inbred animal of the Guadyerbas strain.

Keywords. Iberian pig – Guadyerbas strain – Next Generation Sequencing.

Séquençage des génomes du porc ibérique

Résumé. *Le porc Ibérique est une des plus importantes races locales, économiquement et culturellement. Au contraire d'autres races comme Duroc et Large White, il n'est pas hybridé avec des races asiatiques. Les techniques de séquençage modernes nous permettent de séquencer et d'analyser des génomes complets pour un prix raisonnable, démocratisant les outils génomiques. On présente ici les résultats préliminaires de la séquence d'un porc Ibérique de la lignée Guadyerbas.*

Mots-clés. *Porc Ibérique – Guadyerbas – Séquençage à haute performance.*

I – Introduction

Next generation sequencing (NGS) has revolutionized genomics research, making it difficult to overstate its impact in Biology. NGS will immediately allow researchers working in non mainstream species to obtain complete genomes together with a comprehensive catalogue of variants. A reasonable question to ask, nevertheless, is whether we *really* need so much more sequence. Our answer is yes, we do. There are important advantages on having full sequence rather than genotypes: removing SNP ascertainment bias, uncovering all extant variability, recovering the full unbiased demographic history of populations, or setting the theoretical ground for a unified framework that combines coalescence and association mapping. Additional applications involve RNA-seq, to quantify precisely the transcriptome and study allele specific expression.

We have been studying the Iberian pig from a genetic point of view for many years. The Iberian pig is one of the most important varieties of local pig breeds, both economically and culturally. Contrary to most widespread breeds, like Duroc or Large White, it seems not to have been intercrossed with Chinese pigs. However, a current danger for Iberian germplasm is uncontrolled introgression with Duroc genomes, a fact that may have been exacerbated by recent increasing demand of Iberian pig products. Here, we present ongoing efforts carried out at the Universitat Autonoma de Barcelona and INIA in cooperation with the Centre for Genomics Research in Barcelona, University of Vienna, and Wageningen University in Holland to sequence several Iberian pig genomes and transcriptomes. Thus far, we only have analyzed about 1% of the Guadyerbas strain, and this is what is discussed in this communication.

II – Methods

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 ~ 0.40 and ~ 0.46 for X chromosome.

We prepared a Reduced representation Library (RRL) by digesting with HaeIII enzyme and selecting the band of 200 ± 10 bp. From the three runs, a total of 25.3 million base called reads were obtained. Reads were trimmed to 40 bp due to low 3' end quality. We aligned the quality filtered 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) allowing up to 3 mismatches and we retained for variant calling only those reads that mapped unambiguously. We identified SNPs with GEM, MAQ and GigaBayes (Quinlan *et al*., 2008). When mapping the filtered reads with GEM, we used default options except for the mismatches allowed in each read to the reference genome. For the identification of SNP we used custom Perl scripts. For the assignation of SNP_F, the minimum allele frequency of the alternative allele was set to 0.9. For SNP_H with a coverage of 3×, the reference allele had to be found 1× and the alternative allele 2×; for the rest of the cases ($SNP_H \geq 4\times$), the minimum allele count was 2, the maximum frequency of the reference allele was 0.4 and the maximum frequency of any allele was 0.8%.

After filtering, see methods, and removing ambiguous matching reads, we retained five million reads for further analysis. The total length assembled was 2.3 Mb. 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. The average coverage, 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.

We classified the SNPs into two classes, fixed (SNP_F) when the differences were between the assembly and the Iberian reads, and segregating (SNP_H) when the Iberian pig was heterozygous. We computed the average number of SNPs, fixed differences and heterozygous, along non - overlapping contiguous 500 kb windows. We also obtained Hudson – Kreitman – Aguadé (HKA) diversity (θ) estimates (Hudson *et al*., 1987). Briefly, HKA 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.

III – Results and discussion

We found a global autosomal Iberian heterozygosity rate of $\pi = 0.5\times10^{-3}$. This value should be taken with some caution, though, because it might be inflated by false positives; alternatively, some true SNPs have been certainly discarded because of stringent criteria in identifying SNPs. Nevertheless, assuming a mutation rate (μ) of 10^{-8}, this results in an estimate of effective size $Ne = \pi/4\mu{\sim}10^4$. This value is larger than expected, in particular considering that this is a highly inbred animal and suggests that the actual effective size in the founder herd might be actually double. 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. Therefore, we can hypothesize that loss of genetic variability in the Iberian pig population is very recent and caused by recent inbreeding. Sequencing more animals is needed to investigate these issues further.

To gain further insight into the variability distribution, we plotted the percentage of fixed and segregating SNPs in non-overlapping contiguous windows of 500 kb, Figure 1 shows the results for two chromosomes, SSC1 and SSCX as examples. For clarity, we also present the lowess adjusted curves. A trend of increasing variability toward the telomeres is clearly visible, both in fixed and in heterozygous SNPs. This occurs as well in the sex chromosome, although less marked than in autosomes because the overall level of variability is lower.

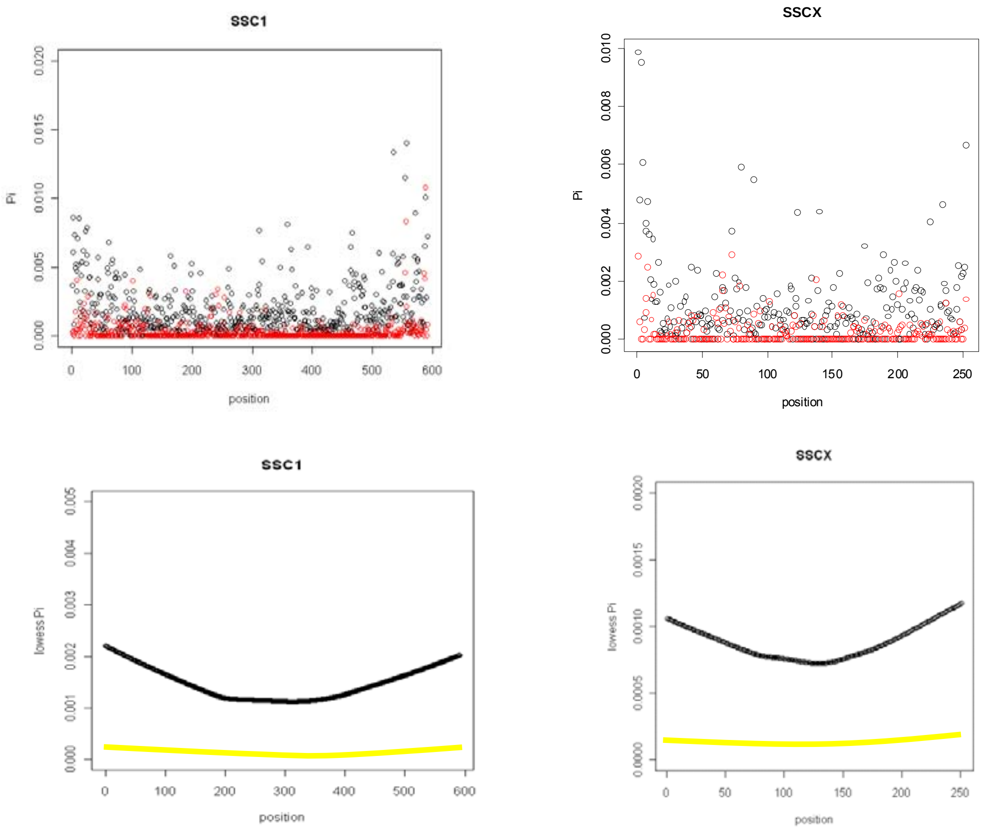

Fig. 1. Average rate of fixed (dark) and heterozygous (light) SNPs per bp in 500 kb windows (Pi) along chromosomes 1 and X. Position refers to window number. The top panels show the observed values whereas the bottom panels are the lowess adjusted curves, shown to underline an increased variability towards the telomeric regions.

Overall, the HKA test showed no strong departures of neutrality. 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.

IV – Conclusion and perspectives

Having complete genomes sequenced is no longer a luxury available only to large sequencing centers. Our challenge now is to convert this huge amount of information into useful knowledge for the conservation and improvement of local breeds, as well as to identify the variants responsible for the distinctiveness of these breeds compared to highly selected, international breeds. On our side, the next immediate steps are: i) to sequence the transcriptome in a few Iberian and Large White samples; ii) to sequence pools of Iberian pigs; iii) to increase coverage of the Guadyerbas animal sequenced; and iv) to sequence partially American pigs of Iberian descent adapted to extreme climate environments, i.e., altitude and heat.

Acknowledgments

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, and by a Consolider grant from Spanish Ministry of Research.

References

Amaral A. *et al*., 2009a. Finding selection footprints in the swine genome using massive parallel sequencing In: *Conference on Next Generation Sequencing: Challenges and Opportunities*, Barcelona

Amaral A. *et al*., 2009b. *Application of massive parallel sequencing to whole genome snp discovery in the porcine genome*. BMC Genomics 10: 374.

Hasin-Brumshtein Y., Lancet D. and Olender T., 2009. Human olfaction: From genomic variation to phenotypic diversity. In: *Trends Genet* 25: 178-184.

Hudson R. R., Kreitman M. and Aguade M., 1987. A test of neutral molecular evolution based on nucleotide data. In: *Genetics* 116: 153-159.

Li H., Ruan J. and Durbin R., 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. In: *Genome Research* 18: 1851-1858.

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

Quinlan A. R., Stewart D. A., Stromberg M. P. and Marth G. T., 2008. Pyrobayes: An improved base caller for snp discovery in pyrosequences. In: *Nat Methods* 5: 179-181.

Sanchez-Gracia A., Vieira F. G. and Rozas J., 2009. Molecular evolution of the major chemosensory gene families in insects. In: *Heredity* 103: 208-216.