Partial short-read resequencing of a highly inbred Iberian pig

A. Esteve (1), R. Kofler(2,5), A.P.Vivancos(2), H. Himmelbauer(2), MAM Groenen(3), JM. Folch(1), MC. Rodríguez(4), M. Pérez-Enciso(1,6)

(1)Departament de Ciència dels Animals i dels Aliments, Facultat de Veterinària, Universitat Autònoma de Barcelona, Spain
(2)Ultrasequencing Unit, Centre de Regulació Genòmica, Barcelona, Spain
(3)Wageningen University, Animal Breeding and Genomics Centre, Wageningen, The Netherlands
(4)Mejora Genética Animal, Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria, Madrid, Spain
(5)Institut für Populationsgenetik, Veterinärmedizinische Universität Wien, Vienna, Austria
(6) Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain

Abstract About 1% of the genome of a highly inbred Iberian pig (Guadyerbas) was sequenced using Illumina's GAII. The ~ 25 million reads were filtered and aligned against the reference genome (Duroc) using 3 alignment tools (MAQ, Mosaik and Gem).

Almost 80.000 SNPs were identified with at least two of these softwares. In each chromosome, nucleotide variability was greater in the telomeres than in the centromeres as predicted by a background selection model, where variability is proportional to recombination rate.

Further, we tested the expected vs. observed number of fixed differences between Duroc and Iberian, and Iberian segregating sites using a multilocus HKA. No strong departures from the neutral model were observed.

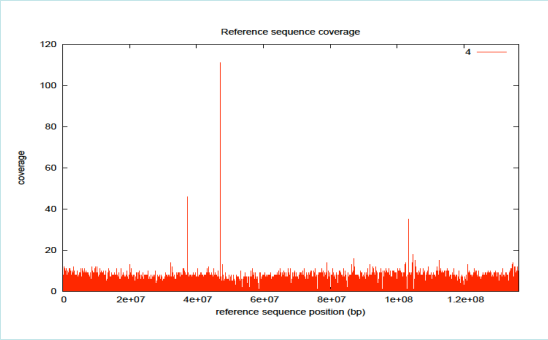
The Iberian Guadyerbas line:
- ✓ One of the survivors of old Iberian lines, maintained in a closed herd (inbreeding F = 40%)
- ✓ Known pedigree since its founding in 1950s.
- ✓ Obese, slow growth, black and hairless
- ✓ We have used it in many QTL experiments.

Methods Reduced Representation Library

- ✓ Genomic DNA with HaeIII digestion
- ✓ 160-200 bp fragment range selection
- ✓ Solexa fragment libraries generation omitting pre-amplification
- ✓ Three lanes were run in GAII
- ✓ Reads filtering
- ✓ Alignment against the reference genome (Duroc)

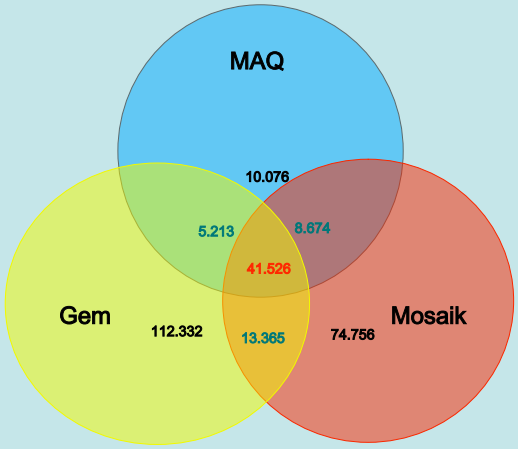
Bioinformatic Results

A total of 5 million of reads matched the assembly unambiguously, spanning 83.1 Mb with at least one read, and 25.1 Mb with at least 3 reads and a maximum coverage of 20. The average coverage was 4x and was rather uniform across all chromosomes.



In this graphic, coverage of chromosome 4 is represented

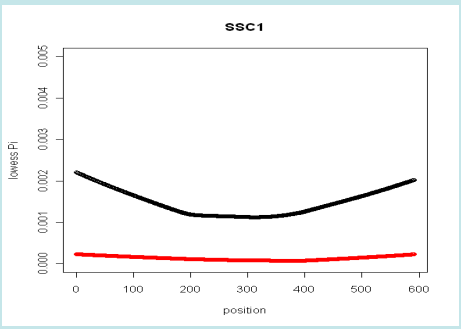
Gem, Mosaik and MAQ detected 172.436, 138.321 and 65.489 SNPs respectively. Of these, 41,526 SNPs were confirmed by the three softwares, while 78,854 SNPs were confirmed by at least two softwares.



MAQ was the most conservative software, while Gem was the less stringent, probably resulting in the largest amount of false positive SNPs. This is probably because Gem does not consider sequence quality in the SNP detection algorithm, in contrast to MAQ or GigaBayes.

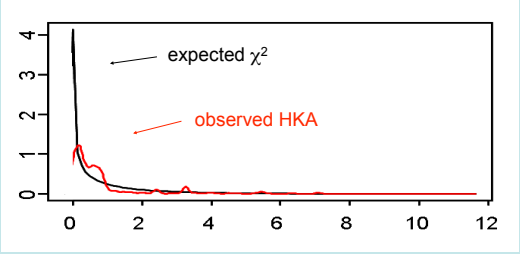
Population Genetics Results

Distribution of variability:A lowess adjusted curve was used to represent the genetic variability per chromosome. In central positions (centromeres) the diversity is markedly lower than distant positions (telomeres).



Nucleotide variability is depicted for chromosome 1
In black fixed differences with the reference genome
In red segregating SNPs in the Iberian sample

Hudson-Kreitman-Aguadé test: We computed observed and expected number of fixed and segregating variants between Duroc and Iberian according to theory in HKA. No strong departures from the neutral model is observed.



Ongoing work

θ_{HKA} was calculated per 500 kB windows in each chromosome and the more extreme windows were identified. The next step we are interested in is to identify which genes are in these regions.

AE is recipient of FPI PhD fellowship from *Ministerio de Ciencia e Innovación* (MICINN), Spain. Work funded by grant AGL2007-65563-C02-01/GAN (MICINN).