Design of a medium density chip (microarray) for the genetic management of Pura Raza Española horses and related breeds

A. Encina[1]; M. Valera[2], N. Laseca[3], A. Rodríguez[1]; S. Demyda[3]; D.I. Perdomo-González[2], P.J. Azor[1], A. Gil[1], M. Ripollés[2], G. Anaya[3], I. González[1], C. Medina[3], M.J. Sánchez-Guerrero[2], J. Poyato[1], A. Molina[3]

[1]Real Asociación Nacional de Criadores de Caballos de Pura Raza Española (ANCCE), Sevilla, Spain.
[2]Department of Agronomy, ETSIA, Universitu of Seville, Sevilla, Spain.
[3]Deparment of Genetics, University of Cordoba, Córdoba, Spain.
Corresponding Author: aencina@lgancce.com

Abstract

The GO EQUIGENOM is a project funded by Spanish Ministry of Agriculture, Fisheries and Food funds, which involves the development of an optimized medium density array for parentage control, disease diagnosis, detection of traits of economic importance and dedicated to research and development in equine genomics, specifically focused on the Pura Raza Española Horse (PRE) and other genetically related breeds. The design of this chip has been structured in several phases, currently in the last phase, validation in an independent PRE population. Firstly, various GWAS analyses were carried out to search for molecular markers associated with diseases of great importance for the PRE. In this sense, using information from animals diagnosed and genotyped with a HD array (867 animals) or MD array (738 animals) and sequenced (284 animals), 257 markers were selected. Subsequently, an exhaustive search for all genetic markers associated with equine diseases described in the bibliography, and in international databases such as OMIA, or Ensembl genome browser was carried out. To this search, markers related to other traits of economic importance, 1,240 markers have been found.

Second phase, a reference population was selected for imputation analysis to the currently available HD and MD array. For this purpose, a selection of 4,490 representative animals (1,907 males and 2,583 females) belonging to 1,718 studs has been made, based on the maximum variability available in this breed and the maximum effective contribution to increase the reliability of the assessments in the evaluations (greater number of offspring in performance control of the different selection objectives of this breed). Genotyping of these animals in HD (2,359) and MD (1,781) allowed us to select 26,017 markers for imputation to HD (99.97 % imputation) and 15,705 markers for imputation to MD (99.96 % imputation). Based on this information, a selection of markers with MAF $\geq$ 0.4 was carried out to fine-tune the genomic selection. Our results have demonstrated the effectiveness of this methodology for improving the accuracy of genetic assessments, especially in traits such as morphology, and to a lesser extent in others such as reproductive efficiency.

Finally, a selection of markers has been made to distribute them homogeneously over all the chromosomes of the equine genome. To this end, priority was given to the quality of genotyping of these markers (based on the information generated with previously genotyped animals) and their informative power in genomic assessment (maximum MAF). In the final design, there are 90,938 markers with an average distance of 26 k. Furthermore, 1,165 ECAY and 1,000 mitochondrial markers selected thanks to the 300 sequenced PRE animals have been included.

Once developed, the chip must be validated to ensure its efficacy and accuracy in the identification and evaluation of genetic markers associated with phenotypic traits of interest, such as morphology, sport performance and health. Final objective of the project is to integrate the information obtained through this MD array into existing selection programs for the PRE using an ssBLUP genomic assessment strategy, to improve the efficiency and accuracy of the selection of breeding stock.

Keywords: SNP, horses, genetic markers, genomic selection, equine.
Presented at the ICAR Anual Conference 2024 in Bled at the Session 9: Genomic's impact on Livestock Sustainability

Introduction

The Pura Raza Española (PRE) horse is a native Spanish equine breed that has been officially recognized since the 15th century. Actually, the PRE is the most popular equine breed in Spain, representing the 70% of all registered equids. The total PRE population, 282,066 horses, are mainly located in Spain but also in other 67 countries (ANCCE, 2024). Recently, the Ministry of Agriculture, Fisheries and Food has awarded the innovative project GO EQUIGENOM to a consortium led by ANCCE and the Royal Spanish Federation of Selected Cattle (RFEAGAS), involving researchers from the PAIDI Group AGR-273 of the University of Seville, and AGR-158 of the University of Cordoba.

The aim of the project is to develop an economic medium density (MD) genotyping chip for the equine species, which simultaneously enables parentage control, the early diagnosis of hereditary and chromosomal diseases, the detection of economically important traits (coat colour markers, or those related to sporting ability, among others) and the development of genomic selection. Additionally, it will be complemented with the development of a robust and easy to use digital tool, which integrates the genetic information (from the breeding programme) and/or genomic information (generated by the chip) of each animal, allowing its use in a quick and easy to interpret way for the technician and for the breeder, in order to make breeding decisions, as well as the detection and control of hereditary diseases, avoiding a considerable economic expense and anticipating future problems and economic losses; with the consequent economic benefit for the farm.

The results of the project will have a direct impact, firstly, on the PRE breed, but also on the whole equine sector, both nationally and internationally.

Material and methods

Animals genotyped and sequenced

In total, 4,490 representative animals (1,907 males and 2,583 females) belonging to 1,718 studs were selected for genotyping, based on retained the maximum variability available in this breed and the maximum effective contribution to increase the reliability of the assessments in the evaluations (greater number of offspring in performance control of the different selection objectives of this breed).

Genomic DNA was isolated from blood or hair samples using a DNeasy Blood and Tissue extraction kit (Qiagen, Germantown, MD, USA). 2,359 horses were genotyped with the high density (HD) Affymetrix AxiomTM Equine 670K SNP Genotyping Array (Thermofisher, Spain), including 670,804 markers uniformly distributed across the entire genome (Schaefer *et al*., 2017). The raw genotype data were processed following the "Best Practices Workflow" procedure in the Axiom Analysis Suite package v5.0 with default parameter (DishQC $\geq$ 0.82). In the same way, 1,781 horses were genotyping with medium density (MD) GGP Equine Array (NEOGEN), including 70K SNPs

distributed across the entire genome. The raw genotypic data were filtered using PLINK v1.9 software (Purcell *et al.*, 2007). Only SNP markers showing a high-quality genotyping rate (call-rate >0.95), with a known genomic position located on the autosomes, mitochondrial and sexual chromosomes (XY) were kept.

Finally, 284 individuals were completely sequenced with a minimum depth of 4X. 144 of the samples were sent to NEOGEN Genomics (Lincoln, NE, USA) and 140 were sent to Psomagen, Inc. (Rockville, MD, USA) to the sequencing on the Illumina NovaSeq 6000 platform (Illumina Inc., San Diego, CA, USA). After filtering by quality, the adapters were removed with the fastp software. The remaining high-quality sequences were aligned with the Equus Caballus v.3.0 (to obtain the ECA1 to ECA31, ECAX and mitochondrial sequences) and MH341179 (to obtain the ECAY sequence) using the Burrows-Wheeler Aligner (BWA) software. Variants from the ECA1 to ECA31, ECAX and mitochondrial chromosomes were called from the Equus Caballus v.3.0 reference genome while the variants from the ECAY chromosome were called from the MH341179 reference sequence. After variant calling (27.76M), SNPs were filtered out with a MAF < 0.2 and individuals which missingness was greater than 30% using PLINK v1.9. 1.6M variants were kept for whole-genome association analysis and population-based linkage analysis.

Imputation analysis

A reference population of animals genotyped at high density (2,359) and medium density (1,781) were used to select the markers that best imputed at high and medium density. Imputation and phasing were carried out using the Beagle 5.4 software (Browning *et al.*, 2018; Browning *et al.*, 2021). All SNPs (default) and genotype probabilities > 0.85, 0.90, and 0.95 were considered for imputed genotype calls. The accuracy of imputation was calculated by the ratio of true and imputed genotypes and the presence of missing and non-missing SNPs in the imputed population.

Bibliographic search for genomic markers associated with traits of economic importance

An exhaustive search for all genetic markers (SNPs, CNV, Indel...) associated with traits of economic importance such as diseases, coat colour and control parentage described in the bibliography, and in international databases such as OMIA, Ensembl genome browser, ISAG etc. was carried out.

Final chip design

For the final selection of the markers to be included in the array, a chromosome-to-chromosome simulation was performed trying to provide an average distance of 26k between markers and had a maximum genotyping quality. The genotyping quality was determined by Thermofisher according to internal parameters and empirical results, as well as by the type of allele detected.

Results and discussion

SNPs and traits of economic importance in the horse

A total of 1,240 markers were selected from the literature and published international databases for different traits of economic importance for the horse such as diseases, behaviour, sport performance, coat colour…

SNPs selected to improve imputation

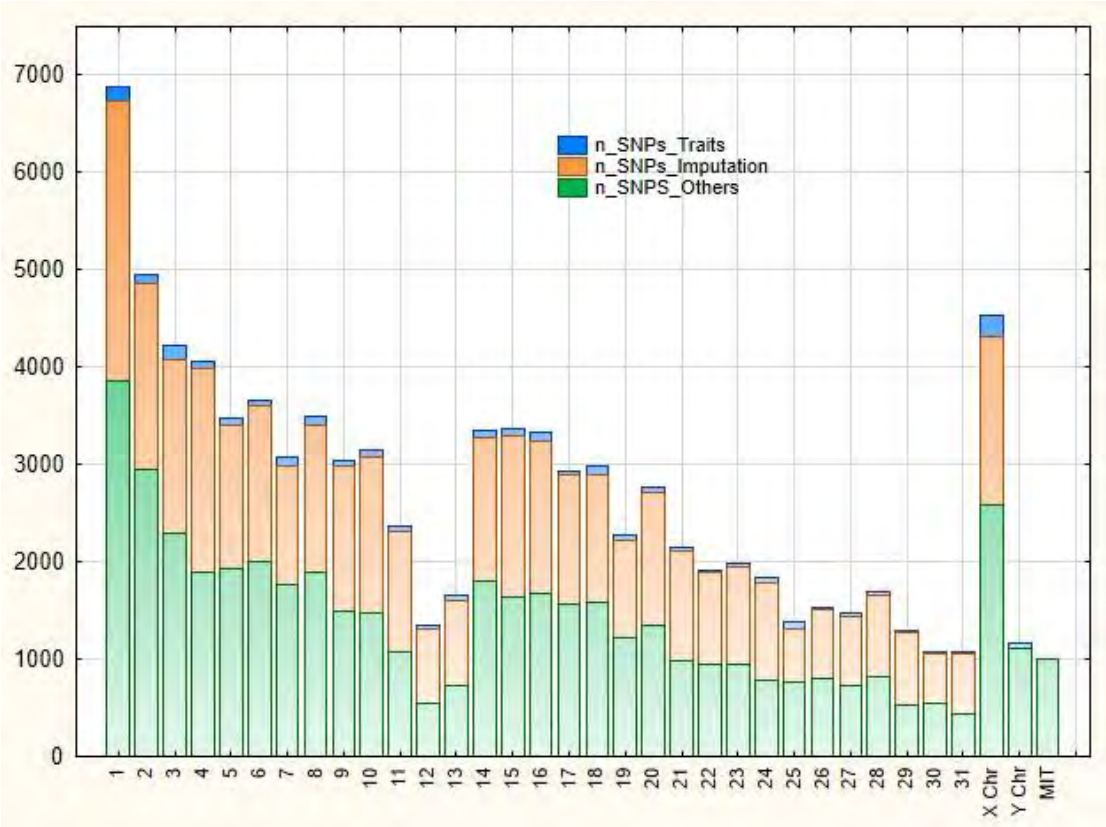
Genotyping HD (2,359) and MD (1,781) animals allowed us to select 26,017 SNPs markers for HD imputation (99.97% imputation) and 15,705 markers for MD imputation (99.96% imputation)..

Final chip design

The number of filler markers was 45,263 SNPs. The total number of markers included in the final design was 90,938 SNP.

Acknowledgments

The authors would like to thank the Royal National Association of Spanish Horse Breeders (ANCCE) for providing the data used in this study. This work has been financed with FEADER funds by the EQUIGENOM Operational Group (Ministry of Agriculture, Fisheries and Food, through the Spanish Agrarian Guarantee Fund, FEGA).



Picure 1. Total number of SNPs found on each chromosome.

List of references

Browning, B.L., Tian, X., Zhou, Y., Browning, S.R., 2021. Fast two-stage phasing of large-scale sequence data. The American Journal of Human Genetics 108, 1880-1890. doi:https://doi.org/10.1016/j.ajhg.2021.08.005.

Browning, B.L., Zhou, Y., Browning, S.R., 2018. A One-Penny Imputed Genome from Next-Generation Reference Panels. The American Journal of Human Genetics 103, 338-348. doi:https://doi.org/10.1016/j.ajhg.2018.07.015.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. American journal of human genetics 81, 559-575. doi:10.1086/519795.

Schaefer, R.J., Schubert, M., Bailey, E., Bannasch, D.L., Barrey, E., Bar-Gal, G.K., Brem, G., Brooks, S.A., Distl, O., Fries, R., Finno, C.J., Gerber, V., Haase, B., Jagannathan, V., Kalbfleisch, T., Leeb, T., Lindgren, G., Lopes, M.S., Mach, N., da Câmara Machado, A., MacLeod, J.N., McCoy, A., Metzger, J., Penedo, C., Polani, S., Rieder, S., Tammen, I., Tetens, J., Thaller, G., Verini-Supplizi, A., Wade, C.M., Wallner, B., Orlando, L., Mickelson, J.R., McCue, M.E., 2017. Developing a 670k genotyping array to tag ~2M SNPs across 24 horse breeds. BMC Genomics 18, 565. doi:10.1186/s12864-017-3943-8.