Selection, validation, and utilization of mitogenome SNP array information in cattle breeding

V. Brajkovic[1*], L. Bradic[1], K. Turkalj[1], D. Novosel[2], S. Ristov[3], P. Ajmone-Marsan[4], L. Colli[4], V. Cubric-Curik[1], J. Sölkner[5] and I. Curik[1]

[1] University of Zagreb, Faculty of Agriculture, Svetosimunska 25, 10000, Zagreb, Croatia; [2] Croatian Veterinary Institute, Savska cesta 143, 10000, Zagreb, Croatia; [3] Ruđer Bošković Institute, Bijenička 54, 10000, Zagreb, Croatia; [4]Università Cattolica del Sacro Cuore, Faculty of Agriculture, food and environmental sciences, via Emilia Parmense 84, 29122, Piacenza, Italy; [5] University of Natural Resources & Life Sciences Vienna, Gregor-Mendel-Strasse 33, 1180, Vienna, Austria; [*]vbrajkovic@agr.hr;

Abstract

High-density SNP arrays are available and routinely used for genetic improvement of livestock populations. At the same time, mitogenome SNPs have rarely been used. In this study, we empirically evaluated 331 mitogenome SNPs included in the GGP Bovine 100K SNP array developed in collaboration with NEOGEN Genomics (Lincoln, NE, USA). Here we present a pipeline for the potential use of mitogenome SNPs with some examples: i) classification of haplogroups with emphasis on rare ones (I, P, Q, R and T5), ii) pedigree verification iii) pedigree and haplotype imputation, iv) deleterious mutation detection v) mitogenomic and nuclear diversity, and vi) standard mitogenome classification and phylogenetics. In summary, 317 mitogenome SNPs have been successfully validated and contribute as additional information to autosomal markers in population genomic analyses. We believe that our study will help to promote the use of mitogenome information in cattle breeding and biodiversity management.

Introduction

Mitochondria play a central role in cellular energy production. These organelles contain their own maternally inherited mitogenome, which in most vertebrates including mammalian livestock encodes 37 genes and has a non-coding region that has been widely used in studies of genetic diversity and evolution. While high-density SNP arrays are available and routinely used for genetic improvement of livestock populations, mitogenome SNPs have rarely been used. Therefore, this study aims to increase the use of mitogenome information in animal breeding by selecting and verifying 331 mitochondrial SNPs for the new NEOGEN GGP Bovine 100K SNP Chip and developing a pipeline for practical analysis.

Materials & Methods

Reference Dataset. The idea to include SNP mitogenomes in the new GGP100K SNP chip arose from a meta-analysis of sequence variability of the bovine mitogenome (Cubric-Curik et al., 2021), which in the meantime, since publication, already includes a newly expanded set of 1068 sequences and over 150 breeds.

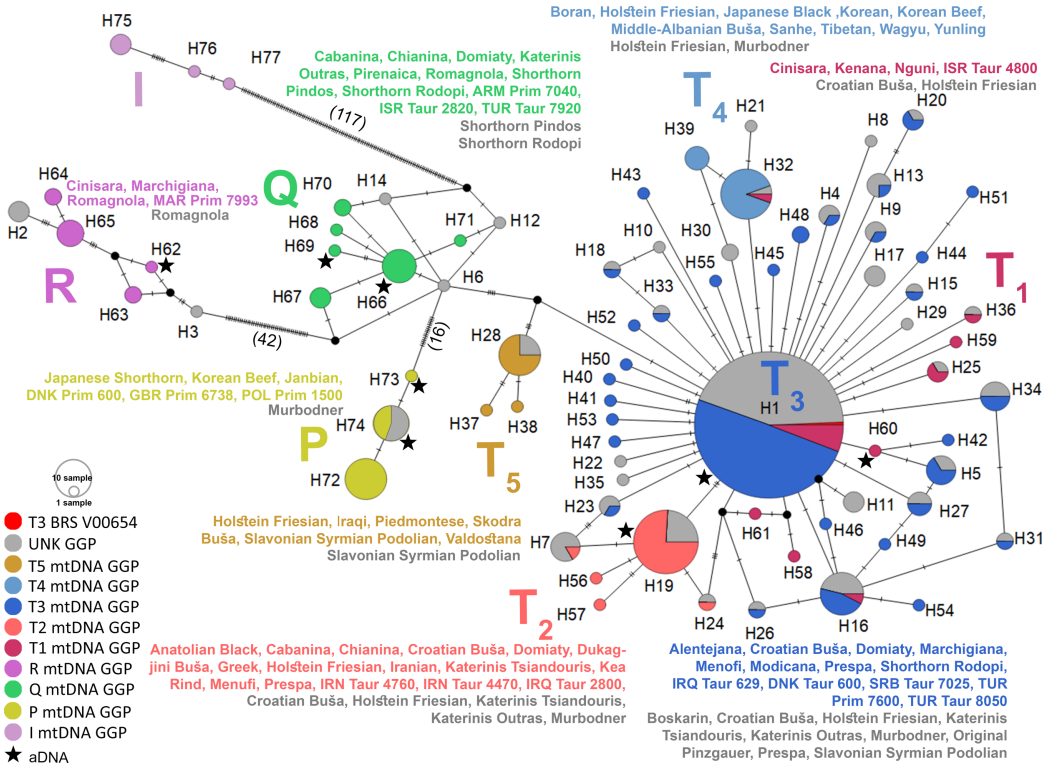
SNP selection. The discriminatory set of mitogenome SNPs was selected based on the pairwise FST values (equal to 1) between all haplogroups among them (I, P, Q, R, T1, T2, T3, T5) (Cubric-Curik et al., 2021). F_{ST} values were calculated using DnaSP v6 software (Rozas

et al., 2017) and Arlequin v3.5 (Exoffier et al., 2015), while SNP primers 150 bp in length were prepared for the GGP Bovine 100K SNP chip. Core SNP selection included 100 SNPs: i) 70 SNPs for haplogroup classification, ii) 20 SNPs for more detailed haplogroup classification within the T3 haplogroup, as this is the most common one in commercial cattle breeds, and iii) 10 SNPs for LHON[1] disease. Additional SNP selection included 231 SNPs: i) 199 SNPs for haplogroup classification and ii) 32 SNPs for MELAS[2] disease.

SNP array. From the study by Cubric-Curik et al. (2021), 129 samples (breeds: Austrian n=9, Greek n=5, Croatian n=4, Slovenian n=1) previously sequenced using NGS technology and an additional 135 new samples (breeds: Greek n=5, Croatian n=4, Austrian n=1) were selected for genotyping.

Results and discussion

Validation. 129 samples (sequenced, genotyped) were for "full" validation, while the new 135 samples (genotyped) were for "blind" functionality of mitogenome SNPs. 317 of 331 SNPs passed the sequence and genotype validation comparison. This is confirmed by the haplogroup classification of the new samples, which can be seen as grey-coloured haplotypes within the known haplogroups in the median-joining (MJ) network constructed with PopART (Leigh and Bryant) (Figure 2). It is important to note that all haplogroups except "I" were successfully validated because we had no samples of that haplogroup. Due to the simplicity of the MJ network representation, SNP positions with gaps in ancient DNA (aDNA) samples were not considered in the construction of the haplotypes. It is worth noting that all aDNA samples were in or close to the core of the haplogroups.



[1] Leber Hereditary Optic Neuropathy
[2] Mitochondrial Encephalopathy, Lactic Acidosis, and Stroke-like episodes

Figure 1. The Median-Joining network for the 181 mitogenome SNPs shows the phylogenetic positions of 77 haplotypes found in 383 samples of *Bos taurus*, *Bos primigenius*, and *Bos indicus* (129 samples sequenced and genotyped; 135 samples genotyped; additional 119 samples from the Cubric-Curik et al. (2021) study for better classification). For aDNA samples the number within the name indicates the year before present.

Mitogenome utilisation pipeline. Here we present a pipeline of potential use of mitogenome SNPs (Figure 2.)

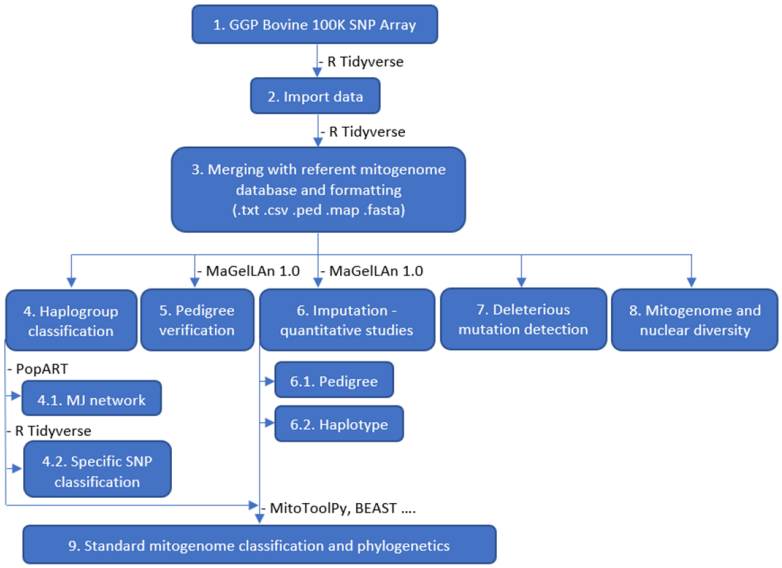


Figure 2. Utilization pipeline steps.

Step 1: Receipt of final report raw data of genotyped samples from NEOGEN Genomics.
Step 2: Import and visualise data with Rstudio (R Core Team, 2019; RStudio Team, 2020).
Step 3: Process the data using the tidyverse package collection (Wickham et al., 2019), merge with the mitogenome reference database, and format according to the type of analysis.
Step 4: 1) Possible classification of haplogroups as shown in Figure 1. 2) specific SNP classification using the tidyverse package collection (Wickham et al., 2019).
Step 5: Verification of complex pedigree errors using mitogenome haplotypes (maternally inherited), which has been well validated on a long and complex Lipizzaner horse pedigree using the computational approach of MaGelLAn 1.0 software (Čačić et al., 2014; Ristov et al., 2016).
Step 6: 1) Since the mitogenome is maternally inherited, MaGelLAn 1.0 makes it possible to quickly identify maternal lineages and assign the corresponding mitogenome sequences to all individuals in the pedigree, so that the phenotypic information can be used as input to any standard software for quantitative genetic (association) analyses (Ristov et al., 2016). Such examples of estimating proportions of phenotypic variances of different traits explained by mitogenome variation can be found in rabbits (Nguyen et al., 2018) and Holstein cattle (Brajkovic et al., 2018). 2) Possible imputation of SNPs to construct the whole mitogenome consensus sequence considering SNPs that were not genotyped. For more information on mitogenome imputation in humans, see Ishiya et al. (2019) and McInerney et al. (2021).
Step 7: Ability to identify mitochondrial genetic disorders associated with LHON and MELAS disease by screening 42 potentially deleterious mtDNA mutations in cattle. While mitochondrial diseases are well documented in humans, to prove the concept, the first

deleterious mutation, T10432C (T10663C in humans) in the ND4L subunit, was found in the Slovenian breed Cika with exophthalmos of the right eye (Novosel et al., 2019), with aetiology and pathogenesis confirmed based on the tissue examined (unpublished data).
Step 8: Various diversity analyses can be performed using autosomal SNP array data. Here, we propose mitogenome-level analysis to enrich for maternal diversity. For example, the F_{ST} value between breeds can be calculated and potentially used to detect upgrading of breeds.
Step 9: Based on the imputed sequences, we can proceed with standard mitogenome classification and phylogenetic analysis, for example, using BEAST and Mitotoolpy.

In summary, 317 mitogenome SNPs have been successfully validated in the new GGP100K SNP chip and contribute as additional information to autosomal markers in population genomic analyses. We believe that our study will help to promote the use of mitogenome information in cattle breeding and biodiversity management.

Acknowledgements
This study was supported by the ANAGRAMS-IP-2018-01-8708, ALGSEQ18-IP-2018-01-7317 and Sir je IN-K.K.01.1.1.04.0058 projects. We also acknowledge financial support from NEOGEN Genomics (Lincoln, NE, USA) and H. Hofeneder-Barclay, J. R. Ferretti, R. K. Nelson, and R. G. Tait Jr. for their efforts related to the inclusion of mitochondrial SNPs in the GGP Bovine 100K.

References
Brajkovic V., Ferenčaković M., Špehar M., Novosel D. *et al.* (2018) Proc. of the 69th EAAP, Dubrovnik, Croatia.
Cubric-Curik V., Novosel D., Brajkovic V., Rota Stabelli O. *et al.* (2021) Evol. Appl. 00:1-16. https://doi.org/10.1111/eva.13315
Čačić M., Cubric-Curik V., Ristov S. and Curik I. (2014) Livest. Sci. 169: 42-47. https://doi.org/10.1016/j.livsci.2014.09.009
Excoffier L. and Lischer, H. E. (2005) Evol. Bioinform 1:117693430500100003. https://doi.org/10.1111/j.1755-0998.2010.02847.x
Ishiya K., Mizuno F., Wang L. and Ueda S. (2019) Bioinform. Biol. Insights. 13:1177932219873884. https://doi.org/10.1177/1177932219873884
Leigh, J. W., and Bryant, D. (2015) Methods Ecol. Evol. 6(9):1110-1116. https://doi.org/10.1111/2041-210X.12410
McInerney T. W., Fulton-Howard B., Patterson C., Paliwal D. *et al.* (2021) BMC Bioinform. 22(1): 1-15. https://doi.org/10.1186/s12859-021-04337-8
Novosel D., Cubric-Curik V., Simcic M., Dovc P. and Curik I. (2019) J. Cent. Eur. Agric. 20(1):19-24. https://doi.org/10.5513/JCEA01/20.1.2413
Nguyen N. T., Brajkovic V., Cubric-Curik V., Ristov S., Veir Z. *et al.* (2018) World Rabbit Sci. *26*(4): 287-298. https://doi.org/10.4995/wrs.2018.7644
Ristov S., Brajkovic V., Cubric-Curik V., Michieli I. and Curik I. (2016) Genet. Sel. Evol. 48(1): 1-10. https://doi.org/10.1186/s12711-016-0242-9
Rozas J., Ferrer-Mata A., Sánchez-DelBarrio J. C., Guirao-Rico S. *et al.* (2017) Mol. Biol. Evol. 34(12):3299-3302. https://doi.org/10.1093/molbev/msx248
R Core Team (2019) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
RStudio Team (2020) RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Wickham H., Averick M., Bryan J., Chang W. *et al.* (2019) J. Open Source Softw. 4(43):1686. https://doi.org/10.21105/joss.01686