Informational steps and tools in Genomic Selection of livestock. A review.

Mihail Alexandru Gras*, Horia Grosu
*corresponding author: gras_mihai@yahoo.com

National Research-Development Institute for Animal Biology and Nutrition, No.1, Calea Bucuresti, 077015, Balotesti, Romania, (0040)213512081, (0040)213512080,

ABSTRACT

An important approach in genomic selection of livestock is the use of Single Nucleotide Polymorphism (SNP) array technology containing a large quantity of genetic data. Between livestock species, dairy cattle were the most intense studied from genomic perspective, with the largest variety of SNP chips. Genomic data formats depend of SNP chip customization and producers, and this affect the capacity of data integration and exchange. Due to the huge amount of genomic data, are required specialized software tools for analysis. In this review, we presented the steps involved in Genome - Wide Association Studies (GWAS) – SNP array data management, imputation, GWAS and their association with software solutions used on those steps.

Keywords: data management, imputation, livestock species, single nucleotide polymorphism, imputation.

INTRODUCTION

In the last decades there is an increasing requirement for deciphering genetic architecture of the traits who may be used in the livestock genetic improvement programs (Nejati-Javaremi et al., 1997; Weller, 2001). Following those efforts, some methods are developed, offering possibility to estimate a genomic value for individuals at the beginning of their life (Meuwissen et al., 2001).

From year 2000 until now there where published several livestock genomes, like gallinacean genome (published by International Chicken Genome Sequencing Consortium in 2004), bovine genome (Bovine Genome Sequencing & Analysis Consortium in 2009), swine (by Groenen et al. in 2012), caprine (Dong et al. in 2013), and later, Jiang et al. (2014) published ovine genome. Starting with 2008, Illumina company developed the first commercial high-density SNP-chip for bovine (Illumina BovineSNP50 BeadChip). From this point forward genomic breeding value estimation became more popular. The major success of these commercial SNP chips

leaded to the use of this technology in genomic evaluation for other species like swine (Ramos et al. 2009), ovine (Kijas et al. 2009), chicken (Kranis et al. 2013), etc.

Excepting chickens and salmon, where first commercial chips were 600k and 6k respectively, the SNP number in the first generation of chips were 50-60k (Lien et al. 2011). Further developments and market requests leaded to the development of other SNP-arrays with other densities. For example, are available 22 commercial chips for bovine genotyping (3k, LD v1, LD v1.1, LD v2.0, GGPLD v1, GGPLD v2, GGPLD v3, SNP50 v1, SNP50 v2, SNP50 v3, GGPHD, GGP 150k, GGP Bos Indicus & HD, Axiom Bos & HD, INFINIUM XT, INFINIUM iSelect HD, INFINIUM iSelect HTS, Axiom Buffalo), 10 chips for swine (GGPLD v1, SNP60 v1, SNP60 v2, SNP80, INFINIUM XT, INFINIUM iSelect HD, INFINIUM iSelect HTS, GGP Porcine HD, GGP Porcine LD, Axiom Porcine), 7 for equine (SNP50 v1, SNP70 v1, INFINIUM XT, INFINIUM iSelect HD, INFINIUM iSelect HTS, GGP Equine, Axiom Equine), 6 for sheep and goats (SNP50 v1, HD, INFINIUM XT, Axiom Ovine, SNP50 v1, INFINIUM XT) and 5 for chickens (INFINIUM XT, INFINIUM iSelect HD, INFINIUM iSelect HTS, Axiom Chicken). All of the chips are developed based on two genotyping technologies: Illumina and Affymetrix. In addition, a lot of non-commercial chips are made it for research purposes. Increased number of available SNP-chips were not accompanied by a common effort for data standardization. Allelic coding, SNP names and genomic coordinates are difficult to compare and update, especially for older SNP chips.

Consistence and standardization of SNP are of overwhelming importance. However, excepting Affymetrix, commercial companies don't offer "reference SNP Ids" (RefSeq SNP or rs IDs) for SNP names. This thing makes practically impossible for developers to connect a SNP chip information with a public database. This situation precludes data utilization from individuals genotyped with a single SNP chip or data integration of different SNP chips. Also, protocols integration and standardization became more important when is necessary to combine two different genotypes obtained using different genotyping technologies.

Beyond SNP name standardization is necessary to standardize position in genome (for example chromosome and base-pair position). The coordinates must refer always to reference genome assembly (RGA) who can be adapted or replaced completely in time. RGA quality and stability depends upon specie. For some species, RGA where recently configurated (ex. goat, by Dong et al., 2013), and in the others suffer multiple revisions and improvements. RGA change frequency depends by sequence quality and magnitude of scientific investments on that specie. Additionally, can exist more "official" RGAs, like the case of bovine with Maryland State University (UMD 3.1.1) and Bovine Genome International Consortium (Btau 4.6.1). The first attempt to surpass those difficulties was an on-line application named SNAT (Jiang et al., 2011).

This tool, not available anymore, address of few SNP chips and who's regarding more annotation aspects then SNP chip data integration. Some years later, Nicollazi et al., (2015) introduced SNPchiMp v.1 application, who actually is at version 3 (http://bioinformatics.tecnoparco.org/SNPchimp/), first SNP chip integration and standardization targeted application. From the emergence application develop considerably his bovine SNP database and was extended for all six major domestic species. Beyond SNP data standardization and integration (including SNP notation for different RGA), application can help for finding particular genes using data mining application Ensembl BioMart. Behind SNP chip data integration and standardization difficulties, data analysis is influenced also by multiple input/output data format used by increasing number of analisys softwares.

Beside the most important tools which could affect the process of genomic selection, we underline here the tools used for SNP array data management, missing genotypes imputation process and genome-wide association studies (GWAS).

SNP array data management

Large datasets generated in the SNP array genotyping process made very difficult even the simple task like data manipulation and storage. From this reason were developed informational tools, particularly designed for those datasets.

In time where developed many management tools initially created just for data analysis. For example, PLINK (Purcell et al., 2007) was created initially for SNP array data analysis in the context of tenths of thousands of SNPs genotyped on multiple individuals. The speed and stability of PLINK recommended this tool like standard software for SNP data management. On the other hand, PLINK can register data sets in diverse formats. Current version of PLINK is 1.07 with 1.9 beta and 2.0 alfa versions with improvements at speed, memory use and new function like genomic kinship matrix calculation.

Recently, statistical analysis software R (https://www.r-project.org/) became a powerful SNP data management resource. It contains developed specific libraries and scripts for management and storage of SNP data from Illumina snpQC.

SNP data management using relational databases is a good solution when datasets are small. When complexity and dimensions are increased, the use of dedicated SNP data management software is recommended. The R package used for data management on Illumina datasets is *gdmp* package. This package can be used to manage and analyze high-dimensional SNP data from Illumina chips with multiple densities. Input data is Illumina Genome Studio file who is turned into an array. For Affymetrix datasets exist R packages designed to work under Bioconductor platform. Those are open sources solutions.

Commercial solutions on the other hand are the most completed, like JMP GENOMICS, Progeny lab and BCPlatforms which are capable to use multiple data formats and have superior functionalities by compared to free solutions.

For big datasets exists also open access alternatives, like SNPpy (Mitha et al., 2011), a Phyton package for sequencing data analysis, genotypic browser GBrowse (Donlin 2009) and Chado database (Mungall et al., 2007), Pearl based packages. A friendly Java and Web-based application was developed by Broad Institute and the Regents of the University from California, named INTEGRATIVE GENOMIC VIEWER (IGV) (Robinson et al., 2011). Regarding SNP array data format conversion, the speed of changes from analysis software publishing, send to minimize the development of a comprehensive application capable to use all the input/output formats.

Some software packages cover the main data formats (PGDSpider and fcGENE for example). PGDSpider is a conversion tool designed for population genetics and genomics applications, capable to manage many genotypic data (from PCR-RFLP to whole genome sequences) (Lischer & Excoffier, 2012). PGDSpider can manage a large number of input/output formats but cannot handle with large datasets because of usual PC`s memory limitation.

fcGENE package is more focused on imputing and GWAS (Roshyara & Scholz, 2014). Designed at the beginning for human genetics the package is able to translate PLINK files in inputs used in animal genetics.

Imputation process

Imputation is the process of filling the missing genotypes resulted from SNP-array technology or when are used SNP-chips of different densities. Missing allele/genotype imputation process represent a preliminary step for a large panel of genetic analysis because most of the programs and models from genomics cannot handle missing values. Imputing process is very efficient working with 99% accuracy (Weigel et al., 2010). Imputation methods are basically divided in two classes: methods based exclusively on linkage disequilibrium and allele frequency calculation and those who include pedigree information.

The most popular software used in missing genotypes imputation are BEAGLE (Browning & Browning, 2007) and IMPUTE2 (Howie et al., 2009). Both of them are designed for human genetics and can impute very efficient simple populational structures but cannot use pedigree information.

In animal genetics and genetic improvement are used software who use pedigree information. This kind of software are MaCH (Li et al., 2010), findhap.f90 (VanRaden et al., 2011), PedImpute (Nicolazzi et al., 2013), Fimpute (Sargolzaei et al., 2014) and AlphaPhase (Hickey 2011). findhap.f90, PedImpute, Fimpute and AlphaPhase software were created for use in domestic animal breeding and genetics, especially cattle population. In terms of computational speed, deterministic models tend to be faster than numerical

ones. Imputation accuracy depend on imputation method, specie, SNP-array chip density, sample size, genetic structure and history of population, etc. All those programs use own input file format and is necessary to use a conversion program.

Genome-wide association studies

Genome-wide association studies are hypothesis free methods of findings associations between genetic regions (SNP) and phenotypic variations (traits).

High density SNP-array technology makes GWAS standard method for disease or important traits affecting loci identification on domestic animals.

GWAS follows some steps:

- SNP genotype determination and data control quality;
- Detection and correction for population stratification;
- Missing genotypes imputation;
- Testing association of SNP genotypes with categorical or discontinuous phenotypes;
- Global significance analysis and multiple testing correction;
- Data presentation (Manhattan plots)
- Cross replication and meta-analysis of data.

This generalization not account some factors like statistical model used in analysis for example. Bayesian statistic models do not need statistical significance testing so programming and analysis are modified in conformity with that.

From many years a panel of software are at researcher disposition (Table 1). The most used in GWAS analysis was PLINK. But PLINK is not the best option for domestic animals (GWAS options are not effective if are not accounted effective population size, stratification or chromosome number of the species). R library GenABEL cover all that deficiencies, offer a better data control and management, and estimating kinship for dense panel markers (Aulschenko et al., 2007). Estimated kinship can be used further in linear mixed models used in quantitative traits analysis. Also, GenABEL have a multitude of data visualization options.

Other open-source software is GENOME-WIDE COMPLEX TRAIT ANALYSIS (GCTA). Build initially for SNP explained phenotypic variance proportion estimation for complex traits, the software includes more GWAS and mixed linear model analysis options (mlma). Data management and input are MaCH and PLINK specific.

Table 1. Software for SNP management and analysis (Nicollazi et al. 2015, updated)

Utility	Name	OS*	License **	Input	Link
SNP management	PLINK 1.90	W/L/M	F	Own	https://www.cog-genomics.org/plink2
	SNPQC***	W/L/M	OS	Illumina raw	http://www-personal.une.edu.au/~cgondro2/snpQC.htm
	GDMP***	W/L/M	OS	Illumina	https://cran.r-project.org/web/packages/gdmp/gdmp.pdf
	JMP GENOMICS	W/L/M	L	Many	https://www.jmp.com/en_us/software/genomics-data-analysis - software.html
	GOLDEN HELIX SNP & VARIATION SUITE	W/L/M	L	Many	http://www.goldenhelix.com/SNP_Variation/index.html
	PROGENY LAB	W/L	L	Many	http://www.progenygenetics.com/
	BCPLATFORMS	L	L	Many	https://www.bcplatforms.com/product/
	SNPPY	L	OS	Affymetrix/ Illumina raw	https://bitbucket.org/faheem/snppy
	GBROWSE	L/M	OS	GFF3	http://gmod.org/wiki/GBrowse
	IGV	W/L/M	OS	Many	http://www.broadinstitute.org/igv/
	PGDSPIDER	W/L/M	OS	Many	http://www.cmpg.unibe.ch/software/PGDSpider/
	FCGENE	W/L	OS	Many	http://sourceforge.net/projects/fcgene/
Imputation	FIMPUTE	L/M	F	Own	http://www.aps.uoguelph.ca/~msargol/fimpute/
	BEAGLE	L/W/M	F	Own, PLINK, VCF	http://faculty.washington.edu/browning/beagle/beagle.html
	IMPUTE2	L/W/M	F	Own	https://mathgen.stats.ox.ac.uk/impute/impute_v2.html

Imputation	MACH	L/W/M	F	Own ~PLINK	http://www.sph.umich.edu/csg/abecasis/MACH/tour/imputation.html
	PEDIMPUTE	L/M	OS	Own	https://bio.tools/pedimpute
	ALPHAPHASE	W/L/M	F	Own	https://sites.google.com/site/hickeyjohn/alphaphase
	FINDHAP	L	OS	Own	http://aipl.arsusda.gov/software/findhap/
GWAS	GENABEL (R)***	W/L/M	OS	MACH, PLINK, Illumina/ Affymetrix raw	https://www.rdocumentation.org/packages/GenABEL/versions/1.8-0
	GCTA	L	OS	PLINK, MACH	http://ctgg.qbi.uq.edu.au/software/gcta/index.html
	GEMMA	W/L	OS	PLINK, BIMBAM	https://www.xzlab.org/software/GEMMAmanual.pdf
	SSGBLUP	W/L/M	OS	Own	http://nce.ads.uga.edu/wiki/doku.php?id=readme.pregsf90
Population genomics and signatures of selection	SWEED	L	OS	VCF	http://sco.h-its.org/exelixis/web/software/sweed/index.html
	ARLEQUIN	W/L/M	F	Own	http://cmpg.unibe.ch/software/arlequin35/
	SELSCAN	L/W/M	F	Own (~PLINK)	https://github.com/szpiech/selscan
	VCFTOOLS	L/M	F	VCF	http://vcftools.sourceforge.net/
	BAYESCAN	L/M	F	Own	http://cmpg.unibe.ch/software/BayeScan/index.html
	ADMIXTURE	L/M	F	PLINK	http://software.genetics.ucla.edu/admixture/
	FASTSTRUCTURE	W/L/M	F	Own, PLINK	https://web.stanford.edu/group/pritchardlab/structure.html
	BAPS	W/L/M	F	Own, GENEPOP	http://www.helsinki.fi/bsg/software/BAPS/

Population genomics and signatures of selection	LFMM	W/L/M	F	Own	http://membres-timc.imag.fr/Olivier.Francois/lfmm/index.htm
	MATSAM	W	F	Own	http://www.bioinfoindia.org/MatSAM/
	DIY-ABC	W/L/M	F	Own	http://www1.montpellier.inra.fr/CBGP/diyabc/
	POPABC	W/M/L	F	Own	https://code.google.com/p/popabc/
	ABCTOOLBOX	W/L	F		http://cmpg.unibe.ch/software/ABCtoolbox/
Genomic predictions	GS3	W/L/M	OS/LA	Own	http://snp.toulouse.inra.fr/~alegarra/
	ASREML	W/L/M	L	Own	http://www.vsni.co.uk/software/asreml
	GENSEL	L/M	L	Own	
	BGLR (R)***	W/L/M	OS	Own, PLINK	http://www.soph.uab.edu/ssg/software/bglr
	RRBLUP (R)***	W/L/M	OS	Own	http://cran.r-project.org/web/packages/rrBLUP/
	BLUPF90 SUITE	W/L/M	F/LA	Own	http://nce.ads.uga.edu/wiki/doku.php

Note: *OS - W, Windows; L, Linux; M, MacOS; **License - OS, open source (source code available); F, free (source code not provided); LA, licensed but free for academia; L, licensed; *** (R) - R packages

GEMMA is also an open-source software who implement Genome-wide Efficient Mixed Model Association Algorithm developed by Zhou and Stephens (2012). GEMMA use Bayesian models and use PLINK input files.

A different approach is adopted by PREGSF90–POSTGSF90 software (Aguilar et al., 2014), interfaces for pre- and post-processing genomic information obtained from BLUPF90 family software. BLUPF90 is a Fortran90/95 software collection used in mixed model animal breeding value estimation. PREGSF90 can be used in data pre-processing before single-step methodology (Legarra et al., 2009, Misztal et al., 2009) implementation and POSTGSF90 estimate SNP effects using methodology described by Wang et al., in 2012(a, b). SNP effects are plotted using GNUPLOT, R graphic package (Zhang et al., 2015) or SNPEVG (Wang et al., 2012).

Genomic selection

SNP-array genotyping technology and GWAS develop the concept of genomic prediction and genomic selection on livestock (Meuwissen et al., 2001). Those things are possible due to large data quantity generated from

SNP genotyping and sequencing, and the accelerated development of data processing capacity in bioinformatics. Now are available many methodologies and models used in genomic prediction. Most popular methods are genomic BLUP (GBLUP) and Bayesian methods (Bayes A/B/C/Cπ etc.).

GBLUP can be applied in any statistical analysis software considering genomic kinship matrix a (co)variance matrix. Some software is from Restricted Maximum Likelihood (REML) family, like DMU (Madsen et al., 2006), WOMBAT (Meyer 2007) and ASREML (Gilmour et al., 2009). Other software family come from quantitative genetics era (BLUPF90), being used for fitting the models used in (co)variance components and breeding values estimation. This software family require advanced knowledge in statistics and working file parametrization, but offer a good fitting for a large number of models. In BLUPF90 was created the first extension GBLUP used in single-step methodology (Misztal at al., 2009).

Another software is the open-source software GS3, used by INRA for genomic predictions applying GBLUP methodologies with Bayes Cp and Bayesian Lasso Models. GS3 bear generally models with additive, dominant and infinitesimal effect adding permanent environmental effects. A disadvantage for GS3 is the complicated formatting of input and output data.

The same disadvantage and inability to manage missing genotypes showed GENSEL software (Fernando & Garrick 2013), a suite who use Bayes A/B/C/Cπ models from Marcov Chain Monte Carlo (MCMC) methodology.

For genomic breeding value estimation are available packages in R software. The package BGLR (Bayesian Generalized Linear Regression; Perez & los Campos 2014) implement Gibbs algorithm from MCMC methodology on Bayesian Regression Models. Also, this package solves the problem p>>n (number of variables - SNP's - is much larger than sample size - genotyped individuals -) using parametric adjustment of GBLUP with Bayes A, B, C; Bayesian Lasso and Bayesian Ridge Regression models.

BGLR is well suited for continue, categorical and censured traits analysis. An alternative could be rrBLUP package, an R library for a fast implementation of Ridge Regression or GBLUP standard.

CONCLUSION

SNP array management, standardization and integration difficulties are, in most of the cases, underrated. For a better use of genomic information in livestock improvement are necessary some improvements: 1. SNP ID and SNP array information must be standardized and internationally recognized. 2. Because now is used Human Reference Genome, are necessary to develop more allelic standardization and coding software, designed for livestock species due to different numbers of chromosome and Reference Genome

Assembly. 3. For an easier use, input files for SNP array management and analysis are better to be standardized.

ACKNOWLEDGEMENTS

This research was supported by funds from the National Project ADER 8.1.6./2019-2022, granted by the Romanian Ministry of Agriculture and Rural Development - MADR, Sectoral Plan ADER. The publication was supported by funds from the National Research Development Project Projects to finance excellence (PFE) - 17/2018-2020 granted by the Romanian Ministry of Research and Innovation.

REFERENCES

Aguilar I., Misztal I., Tsuruta S., Legarra A. & Wang H. (2014) PREGSF90 – POSTGSF90: Computational Tools for the Implementation of Single-step Genomic Selection and Genome-wide Association with Ungenotyped Individuals in BLUPF90 Programs. In: Proceedings, 10th World Congress of Genetics Applied to Livestock Production, Vancouver, Canada, August 2014.

Aulschenko Y.S., Ripke S., Isaacs A. & van Duijn C.M. (2007) GenABEL: An R library for genome-wide association analysis. Bioinformatics 23, 1294–6.

Browning S.R. & Browning B.L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. American Journal of Human Genetics 81, 1084–97.

Dong Y., Xie M., Jiang Y. et al. (2013) A reference genome of the domestic goat (Capra hircus) generated by Illumina sequencing and whole genome mapping. Nature Biotechnology 31, 135–41.

Donlin M.J. (2009) Using the generic genome browser (GBrowse). Current Protocols in Bioinformatics 9, 9.

Fernando R.L. & Garrick D.J. (2013). Bayesian methods applied to GWAS. In Genome-Wide Association Studies and Genomic Prediction (Ed. by C. Gondro, J.H.J. van der Werf & B. Hayes), pp. 237–74. Springer Series: Methods in Molecular Biology. Humana Press, Berlin.

Gilmour A.R., Gogel B., Cullis B. & Thompson R. (2009) ASREML User Guide Release 3.0. Hemel Hempstead, VSN International Ltd, Hemel Hempstead, HP11ES, UK.

Groenen M.A., Archibald A.L., Uenishi H. et al. (2012) Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491(7424), 393–8.

Hickey J.M. (2013) Sequencing millions of animals for genomic selection 2.0. Journal of Animal Breeding and Genetics 130, 331–2.

Howie B.N., Donnelly P. & Marchini J. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genetics 5, e1000529.

International Chicken Genome Sequencing Consortium (2004) A genetic variation map for chicken with 2.8 million single nucleotide polymorphisms. Nature 432, 717–22.

Jiang J., Jiang L., Zhou B., Fu W., Liu J.F. & Zhang Q. (2011) SNAT: a SNP annotation tool for bovine by integrating various sources of genomic information. BMC Genetics 12, 85.

Jiang Y., Xie M., Chen W. et al. (2014) The sheep genome illuminates biology of the rumen and lipid metabolism. Science, 344, 1168–1173.

Kijas J.W., Townley D., Dalrymple B.P. et al. & the International Sheep Genomics Consortium (2009) A genome wide survey of SNP variation reveals the genetic structure of sheep breeds. PLoS ONE 4, e4668.

Kranis A., Gheyas A.A., Boschiero C. et al. (2013) Development of a high density 600K SNP genotyping array for chicken. BMC Genomics 14, 59.

Legarra A., Aguilar I. & Misztal I. (2009) A relationship matrix including full pedigree and genomic information. Journal of Dairy Science 92, 4656–63.

Li Y., Willer C.J., Ding J., Scheet P. & Abecasis G.R. (2010) MACH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genetic Epidemiology 34, 816–34.

Lien S., Gidskehaug L., Moen T., Hayes B.J., Berg P.R., Davidson W.S., Omholt S.W. & Kent M.P. (2011) A dense SNP-based linkage map for Atlantic salmon (Salmo salar) reveals extended chromosome homologies and striking differences in sex-specific recombination patterns. BMC Genomics 12, 615.

Lischer H.E.L. & Excoffier L. (2012) PGDSPIDER: an automated data conversion tool for connecting population genetics and genomics programs. Bioinformatics 28, 298–9.

Madsen P., Sorensen P., Su G., Damgaard L.H., Thomsen H. & Labouriau R. (2006) DMU – a package for analysing multivariate mixed models. In: Proceedings of the 8th World Congress on Genetics Applied to Livestock Production, Belo Horizonte, Minas Gerais, Brazil, 2006.

Meuwissen T.H.E., Hayes B.J. & Goddard M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819–29.

Meyer K. (2007) WOMBAT – A tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). Journal of Zhejiang University Science B 8, 815–21.

Misztal I., Legarra A. & Aguilar I. (2009) Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. Journal of Dairy Science 92, 4648–55.

Mitha F., Herodotou H., Borisov N., Jiang C., Yoder J. & Owzar K. (2011) SNPpy – database management for SNP data from genome wide association studies. PLoS ONE 6, e24982.

Mungall C.J., Emmert D.B. and The FlyBase Consortium (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. Bioinformatics 23, i337–46.

Nejati-Javaremi A., Smith C. & Gibson J.P. (1997) Effect of total allelic relationship on accuracy of evaluation and response to selection. Journal of Animal Science 75, 1738–45.

Nicolazzi E.L., Biffani S. & Jansen G. (2013) Short communication: imputing genotypes using PEDIMPUTE fast algorithm combining pedigree and population information. Journal of Dairy Science 96, 2649–53.

Nicolazzi E.L., Caprera A., Nazzicari N., Cozzi P., Strozzi F., Lawley C., Pirani A., Soans C., Brew F., Jorjani H., Evans G., Simpson B., Tosser-Klopp G., Brauning R., Williams J.L., Stella A. (2015). SNPchiMp v.3: integrating and standardizing single nucleotide polymorphism data for livestock species. BMC genomics, 16:283

Pérez P. & de los Campos G. (2014) Genome-wide regression & prediction with the BGLR statistical package. Genetics 114, 483–95.

Purcell S., Neale B., Todd-Brown K. et al. (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics 81, 559–71.

Ramos A.M., Crooijmans R.P.M.A., Affara N.A. et al. (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.

Robinson J.T., Thorvaldsdottir H., Winckler W., Guttman M., Lander E.S., Getz G. & Mesirov J.P. (2011) Integrative genomics viewer. Nature Biotechnology 29, 24–6.

Roshyara N.R. & Scholz M. (2014) FCGENE: a versatile tool for processing and transforming SNP datasets. PLoS ONE 9, e97589.

Sargolzaei M., Chesnais J.P. & Schenkel F.S. (2014) A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15, 478.

VanRaden P.M., O'Connell J.R., Wiggans G.R. & Weigel K.A. (2011) Genomic evaluations with many more genotypes. Genetics Selection Evolution 43, 10.

Wang H., Misztal I., Aguilar I., Legarra A. & Muir W.M. (2012a) Genome-wide association mapping including phenotypes from relatives without genotypes. Genetic Research 94, 73–83.

Wang S., Dvorkin D. & Da Y. (2012b) SNPEVG: a graphical tool for GWAS graphing with mouse clicks. BMC Bioinformatics 13, 319.

Weigel K.A., Van Tassell C.P., O'Connell J.R.O., VanRaden P.M. & Wiggans G.R. (2010) Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. Journal of Dairy Science 93, 2229–38.

Weller J.I. (2001) Quantitative Trait Loci Analysis in Animals. CABI Publishing, UK.

Zhou X. & Stephens M. (2012) Genome-wide efficient mixed-model analysis for association studies. Nature Genetics 44, 821–4.