WIAS PhD Proposal Graduate programme 2016

General information

Chair Group (s): Animal Breeding and Genomics Centre (ABGC)

Project title (English): Genomic prediction to infer function from conservation

Start date – End date: 01-10-2017 to 30-09-2021

Start date must be within 8 months after the grant is awarded.

Composition of the project group and scheduled time for project

Name	Role	Funded by	hours per week
Fernando Bueno Gutiérrez	PhD candidate	WIAS	40
Dr. John Bastiaansen	Daily supervisor	WUR	2.5
Dr. Hendrik-Jan Megens	Daily supervisor	WUR	2.5
Prof. Dr. Roel Veerkamb	Supervisor	WUR	pm
Prof. Dr. Martien Groenen	Promotor	WUR	pm

Cooperation with organisations outside WIAS

Wageningen UR	Other Graduate Schools:
	Research Institutes:
The Netherlands	Universities:
	Research Institutes:
	Industry and organisations:
International	Universities:
	Research Institutes:
	Industry and organisations: Breed4foidentifiedod partners companies
	Consortium: FAANG (Functional Annotation of Animal Genomes)

Where will the project be carried out:
Wageningen University and Research, the Netherlands.

Will vertebrate animals be used:	NO
Does the project involve biotechnological research:	NO

If one or both answers are 'yes', please, take care yourself of appropriate submission to the relevant committee and other legal aspects.

Summary (max. 300 words)

Summary of objectives and hypotheses

The main objective of this project is to develop methodologies that can optimally utilize evolutionary conservation (EC) data and validations from genomic prediction (GP) to gain understanding of the function that the single nucleotide polymorphisms (SNPs) have on the phenotypic traits. The efficient use of EC data is expected to increase the accuracy of GP. Moreover, the inference of SNPs' functionality at the phenotypic level is expected to have wide applications in genetic improvement of plants and livestock species, medicine and biodiversity. The main research question is: How can we utilize EC data to increase the accuracy of GP and generate functional annotations at the phenotypic level? This research question is further divided into four objectives:

1. How can we use EC data to identify putative functional SNPs?
2. How can we estimate probabilities of functional importance for the SNPs identified in objective -1-?
3. How can we improve the accuracy of GP using the estimates from objective -2-?
4. How can we include the results from objective -3- to improve the estimates from objective -2- and generate functional annotations at the phenotypic level?

Relevance for the WIAS mission

The aim of this project is to develop methods that utilize EC data and GP to infer functionality of SNPs at the level of phenotypic traits. The efficient use of EC data is expected to improve GP. Thus, the project will contribute to more efficient and sustainable breeding programs. Moreover, the project lies within the framework set out by the FAANG consortium (Functional Annotation of Animal Genomes) which aims to produce comprehensive maps of functional elements in the genomes of domesticated animal species.

Data management (do you follow the data management policy of the chair group; are there any additional issues)

The data for this project is available due to the collaboration between ABGC with Topigs-Norsvin and other Breed4food partner companies, and with the FAANG consortium. The data will concern traits that are not economically important, avoiding publications using these data being commercially sensitive. Data will be handled carefully and in compliance with the data management rules of the chair group.

Feasibility

How is adequate supervision guaranteed?

For the period of the project, I will have regular meetings with the supervisors. The supervisors of the project come from a wide area of expertise. Prof. Dr. Martien Groenen is an expert in animal genomics. Dr. Hendrik-Jan Megens is an expert in genetic architecture. Prof. Dr. Roel Veerkamb and

Dr. John Bastiaansen are experts in quantitative genetics and animal breeding. Dr. John Bastiaansen and Dr. Hendrik-Jan Megens are the daily supervisors. Weekly, a meeting will be held with the daily supervisors. Every two months, there will be a meeting with the whole supervising group. Moreover, there is an industry user group meeting every six months. In addition to the normal supervisory group, the user group includes industry representatives from the Breed4food partner companies. The supervisory team will be available in case of emergency questions.

How is the execution of the research guaranteed? (facilities, technical assistance)

The group ABGC at Wageningen is the main group involved in the project. The group has the required expertise and knowledge on statistical genetics, theory, software and computer facilities. State of the art computing facilities are available via access to High Performance Cluster (HPC), a versatile platform that can handle jobs having raw computing power and heavy memory demands. Pig data will be made available by Topigs-Norsvin and the FAANG consortium.

Which agreements have been made regarding cooperation with other groups/universities/institutes?

At Wageningen UR, I will be physically working at the Animal Breeding and Genomics Centre (ABGC) five days a week.
Currently there are three projects from STW-Breed4Food Partnership Programme, that are doing related research in ABGC. These projects are partly supervised by the same supervisors as this project:

- STW-Breed4Food Partnership Programme. Project 14283: From sequence to phenotype: detecting deleterious variation by prediction of functionality.
- Method of association study for rare variants using NGS.
- Towards Precision Breeding using genomic prediction.

We intend to collaborate with these projects. At the moment, there are no agreements made with other institutes or universities.

Content (max. 2500 words)

Review of literature

Maps of functional elements

Understanding the function that the DNA elements have at the molecular and phenotypical levels are major goals in biology. Combined efforts such as the recently initiated FAANG consortium (Functional Annotation of Animal Genomes) ([1], see also www.faang.org) aim for the construction of comprehensive maps of functional elements in the genome. Currently, these maps contain only information at the molecular level. Biochemical tools such as chromatin state and DNA methylation allow inferring the function of DNA elements at the level of individual molecules. The applications of this information are limited, though. Disciplines like genetic improvement of plants and livestock species, medicine and biodiversity science could benefit more if the maps of functional elements also contained information at the level of phenotypic traits.

Genomic prediction

Nowadays, genomic prediction (GP) is the technique preferred by breeding organizations to estimate the breeding value of animals. The development of dense panels of single nucleotide polymorphism (SNP) markers provides precious information about the genetic variation of animals. Subsequently, GP has made the prediction of breeding values more accurate [2-5]. Currently, GP does not seem to improve with more dense panels of markers or whole-genome sequence data, as our proficiency to generate data from the genome has exceeded our ability to process it [6]. The number of potential explanatory variables (SNPs) and their possible interactions is now much larger than the number of animals (i>>n) [7]. To overcome this problem, the development of strategies that minimize the number of SNPs has been proposed [8]. The Bayesian methods, for instance allow discarding the less important SNPs; hence the number of explanatory variables can be minimized without losing information. To identify the less important SNPs, the Bayesian methods account for information external to the data, such as derived from maps of functional elements. This information can be used as an indicator of the weights that the SNPs should have in the prediction model. Since these weights are based on information external to the data, they receive the name of "priors". In Bayesian GP, the priors are defined based on how much the SNPs are expected to affect the phenotypic traits.

GP improvement

The use of priors may enable to benefit from whole-genome sequence data because the SNPs with low priors can be discarded. Furthermore, the use of priors can improve GP because the SNPs will receive adequate weights in the prediction models. Currently, the accuracy of prediction is ~60-70% depending on the trait. It is expected that the use of priors could increase the accuracy by ~40% [9]. Besides, the use of priors is expected to improve other aspects of GP, such as higher efficiency in small populations (local breeds) [10], and less dependency on phenotypic recording [11].

Recent research in GP has estimated priors based on the molecular-level functional annotations that are available. The increase in accuracy using these priors, however, has not been substantial [12-14], which could be explained by the connection between molecular function and phenotypic traits being unclear. It is expected that a large portion of the functional elements are not functionally important when it comes to the phenotypic level [15]. Evolutionary conservation can be used to derive accurate estimates of functional importance.

Evolutionary conservation

In evolutionary conservation (EC), the genome of different species is compared, and important SNPs (or sets of SNPs) can be identified using the principle of conservation genetics [16]. According to this principle, the genomic segments that remain constant through evolution (conserved elements) presumably result from purifying selection and are thereby indicative of functional importance [16-18]. The EC approaches consist of estimating probabilities of functional importance for the SNPs based on the degree of conservation. For instance, nucleotides that are polymorphic (SNPs) in all mammals will receive low conservation scores, whereas nucleotides that only show variation in *Sus scrofa* will receive high scores.

The efficiency of EC approaches to estimate functional importance has been previously tested. Kircher and co-workers [19] showed that SNPs associated with human disease in Genome-wide association studies (GWAS) are 1.37-fold enriched in conserved elements relative to all SNPs. Alföldi et al. [20] showed that the combination of GWAS and EC can be very powerful for identifying important SNPs. For instance, the alleles of the SNPs within conserved regions are likely to be present

at frequencies close to fixation and GWAS cannot identify these. EC approaches have proved to be very efficient in identifying deeply conserved elements, some of them turning out to have high significance for a breeding trait. (i.e. DGAT gene in dairy cattle [21], IGF2 in pigs [22], IGF1R in dogs [23]).
Despite their success, EC approaches have one important limitation. Although the majority of functional elements are conserved, some of them are only weakly conserved and cannot be identified with simple EC analyses [24,25]. One possible solution to this problem is to perform EC analyses at a more shallow phylogenetic scope. By comparing the genome of different breeds within a species, it is possible to identify weakly conserved SNPs [22,26,27]. Moreover, these analyses allow the identification of SNPs from putative transcription factors (TFs) and transcription factors binding sites (TFBS).

An important aspect to consider is how the principle of conservation genetics applies to the different DNA classes. For instance, it is known that the evolution of gene promoters can be better explained by this principle than the evolution of enhancers [28]. An important question for further research would therefore be: How do conservation and function overlap for the different DNA classes?

From EC to functional annotation
Nowadays there are biochemical tools that allow inferring function at the molecular level. However, these tools are complex to use and information is generated at a low rate. Information is however, rapidly increasing for other less precise annotations. EC, for instance, can make use of the accumulating sequence data to generate valuable information. The main advantage of EC is that it allows estimating the functional importance of the SNPs rather than the function itself. The potential of EC to derive genome annotation information has been previously noted [29-31]. Information derived from EC can be used to improve the accuracy of GP. Furthermore, through subsequent GP validations, it should be feasible to increase the reliability of the EC information and generate functional annotations at the level of phenotypic traits.

Formulation of the problem

Current functional annotations concern the molecular level and are limited in their applications. The attempts to improve GP using priors derived from these annotations have not been successful. This may be explained by a large portion of functional elements not being functionally important when it comes to the phenotypic level. It may thus be more efficient to derive priors from estimates of functional importance than from molecular-level annotations. EC offers an opportunity to derive estimates of functional importance. These estimates can, in turn, be used to estimate priors and improve GP. Furthermore, since the increments in GP accuracy are measurable, it should be feasible to improve the estimates through GP validations, and generate functional annotations at the phenotypic level. These annotations are expected to have wide applications, from identifying the SNPs involved in specific traits to further improving GP.
The proposed project uses a wealth of genome data available and state-of-the-art genomics technology to increase the accuracy of GP in the absence methods capable of processing all genome data. Moreover, in an era in which the generation of functional annotations is a scientific priority, the

proposed project provides methodologies to extract functional information from the accumulating sequence data. The main research question is: How can we utilize EC data to increase the accuracy of GP and generate functional annotations at the phenotypic level? This research question is further divided into four objectives:

1. How can we use EC data to identify putative functional SNPs?
2. How can we estimate probabilities of functional importance for the SNPs identified in objective -1-?
3. How can we improve the accuracy of GP using the estimates from objective -2-?
4. How can we include the results from objective -3- to improve the estimates from objective -2- and generate functional annotations at the phenotypic level?

Methodology

The methodologies can be applied to any livestock species. We will use data from pigs (*Sus scrofa*) because a wealth of data is available for this project. Figure 1 explains how the genomic information is classified and shows the flux of activities of this project.

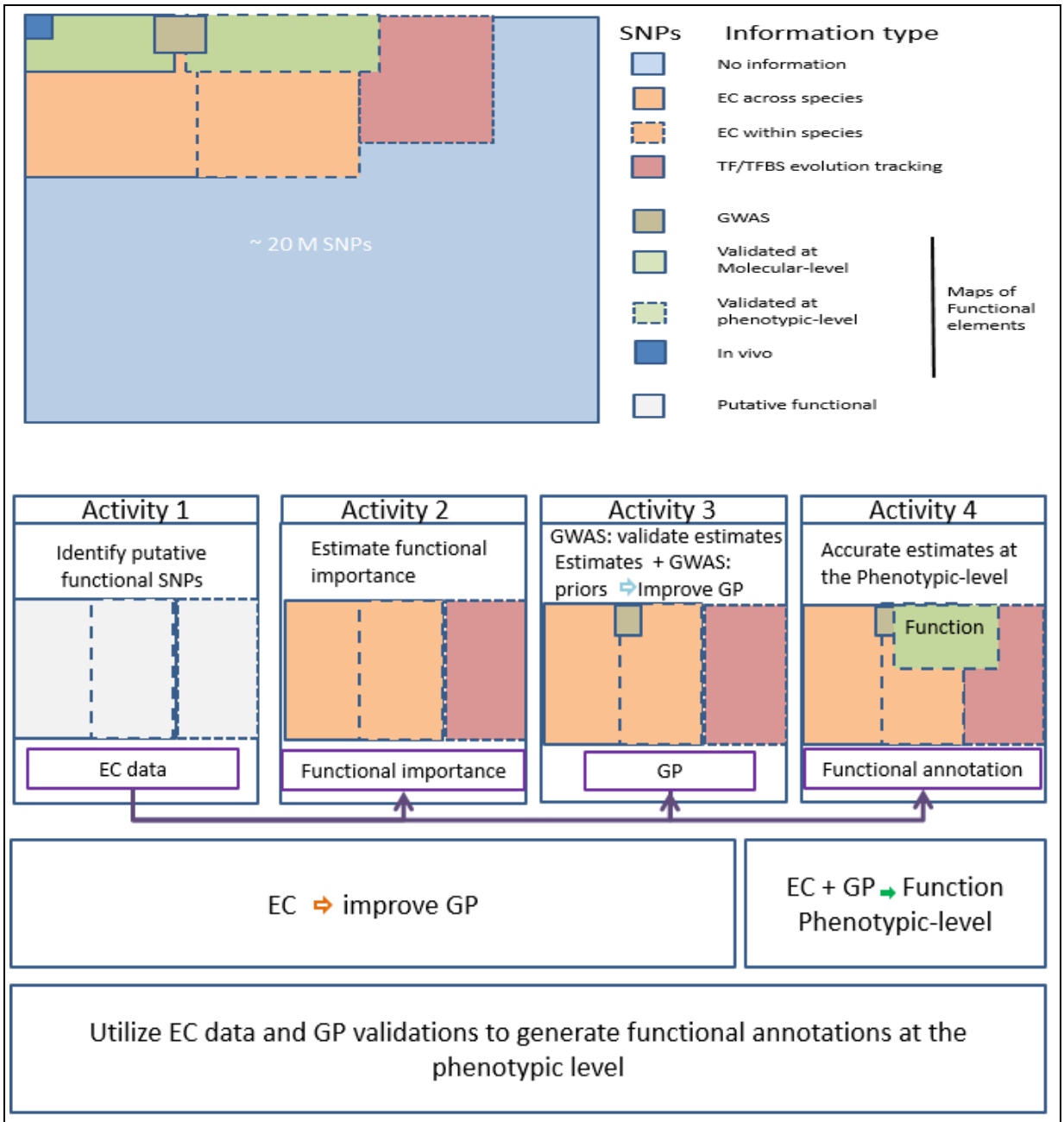

Figure 1. Genome information and flux of activities.

(Above): Hypothetical portions of SNPs (areas) with respect to the total SNPs in the genome, with different information types (colors). (Below): Diagram of activities. In Activity 1 we identify putative functional SNPs; In Activity 2 we compute estimates of functional importance; In Activity 3 we combine the estimates with GWAS; In Activity 4 we infer function at the phenotypic level.

EC: Evolutionary Conservation, GP: Genomic prediction, GWAS: Genome-wide association studies, TF: Transcription factors, TFBS: Transcription factor binding sites.

Objective 1:
How can we use EC data to identify putative functional SNPs?

Activity 1:
The University of California Santa Cruz (USCS) (http://genome.ucsc.edu) possesses information on putative functional SNPs for *Sus scrofa* that were identified with EC analyses across species. We will perform pairwise comparisons between the genomes of different breeds within *Sus scrofa* and related outgroups by mapping SNPs and constrained elements in a reciprocal manner, as explained in

Diego Villar et al. [28]. From combining the public data with results from our own analysis, we will identify putative functional SNPs and derive information such as the degree of sequence divergence, and the length of the conserve regions.

In order to enlarge the collection of putative SNPs, we will identify potential (TFs/TFBS) using information from the EC analysis. Lastly, we will define catalogues of variation for the different breeds and outgroups and discard all conserved elements that do not correspond to genomic variation (i.e. SNPs).
The main deliverable from this activity is an analysis of conservation within *Sus scrofa* and with other species. For the other objectives, this will result in a list of putative SNPs with diverse information for *Sus scrofa* and catalogues of variation.

Objective 2:
How can we estimate probabilities of functional importance for the SNPs identified in objective -1-?

Activity 2:
We will compute conservation scores for the SNPs identified in activity -1- using the most appropriate method available. Currently, there are diverse methods that compute rates of nucleotides substitutions for the SNPs based on information such as provided in activity -1- [32-36]. Then, we will use a birth-death model to track the evolution of individual (TFs/TFBS) and compute conservation scores [37].
The probabilities of functional importance will be estimated by integrating the conservation scores available at USCS, with the conservation computed in this Activity. For this purpose, we will use one of the methods that allow integrating annotations [19,38-40]. We will investigate which method is most appropriate for integrating different EC information types by using crossvalidation.

Deliverable from this activity is a methodology to predict functional importance of SNPs. Lists of probability estimates of functional importance of SNPs will become available as well as, and a pipeline to compute them. These estimates will be independent of the breed and trait but specific for *Sus scrofa*.

Objective 3:
How can we improve the accuracy of GP using the estimates from objective -2-?

Activity 3:
We will estimate priors for GP by integrating the estimates from objective -2- with GWAS, as explained by Alföldi et al. [20]. This will allow us to validate the estimates from objective -2- and improve their reliability.

We will pool all the genotypic data available and we will perform GWAS for four traits. The difference between the traits will be considered. For instance, fertility traits may be subject to natural selection. Then, we will perform multi-SNP Bayesian GP in 5-10 scenarios, based on different statistical models (i.e. BAYES_R [41], BayesBπ [42]…), the criteria used for subsetting SNPs, and other aspects. For this, we will work in collaboration with ongoing research in ABGC that aims the development of priors derived from molecular-level annotations. We will test the increments in GP accuracy in each commercial line. The use of breeding data allows inferring function in measurable traits. Thus, further research could use this information to infer function at the molecular level.

Deliverables from this activity include methodologies to estimate trait-specific quantitative priors for the SNPs, and values of accuracy of prediction under different scenarios. This activity should be reproducible in case another type of scores became available, such as scores derived from phenotypic-level and molecular-level annotations.

Objective 4:
How can we include the results from objective -3- to improve the estimates from objective -2- and generate functional annotations at the phenotypic level?

Activity 4:

The aim of this activity is to identify the SNPs that have the highest effect on the increase of GP accuracy for a particular trait. By leveraging the results from GP and crossing the information with the catalogues of variation of the commercial lines (Activity 1), we will identify the sets of SNPs that are responsible for the increase in accuracy of each commercial line in a particular trait. Then, we will estimate the effect of smaller sets of SNPs or individual SNPs. Machine learning based methods and subsequent GP validation are possible methods to investigate which schemes of assignment of priors leads to better results.

Dealing with the interactions between SNPs within clusters of functional elements will be particularly challenging. To minimize this problem, we will try to identify these and we will characterize them as individual units. Information from the maps of functional elements, as well as DNA class and conservation of the SNPs, can be used to identify these clusters [35]. The deliverables from this activity include a methodology to identify SNPs associated with phenotypic traits, as well as lists of SNPs (or clusters of SNPs) and their estimated effect on four traits considered.

Ongoing research in the ABGC aims to integrate all functional annotations into a single C-score per SNP. If time allows, we will estimate the correlation between the C-scores and the conservation scores obtained in this project. More importantly, we will investigate how the principle of conservation genetics (conservation implies function) applies to the different DNA classes.

Data
The project will take place in the ABGC. The group is part of the FAANG consortium and has access to wide genome annotation resources and genotypic data. This includes sequence data from outgroups closely related to *Sus scrofa*, from ~15M years ago until ~2M years ago, and from multiple breeds within *Sus scrofa* since ~1.5 M years ago, including breeds from Europe and Asia [26,27,43,44]. Genotypic data from ~22206 animals from 4 commercial lines will be provided by Topigs-Norsvin (Table 1).

Table 1. Number of genotyped animals in the commercial lines.

Breed	Number of individuals	SNP chip
Large White	~12000	60-80-660K-imputed to sequence
Dutch Landrace	~5000	80-660K-imputed to sequence
Duroc	~5000	80-660K-imputed to sequence
Large White	32	80K-sequence
Dutch Landrace	12	80K-sequence
Duroc	162	80K-sequence

Work Plan (for the entire project, including writing of the thesis)

Activity	Year 1				Year 2				Year 3				Year 4			
Activity 1																
Activity 2																
Activity 3																
Activity 4																
Finalize thesis																

Requested budget

		year 1	year 2	year 3	year 4
Personnel (mm)		35720	42738	45983	49487
Research costs (k€)					
	Equipment	2			
	Consumables*	8	9	9	9
	Fieldwork				

Explanation and/or remarks to the proposed budget:
Equipment is a computer and other office expenditures. Consumables include HPC costs, scientific training and travelling. HPC costs are based on the following pricing (depending on the use of the HPC a discount is given of up to 40%). Total disk space of the sequence data used within the project is around 10 Tb.
Computing: €0.08 /core/hour
Storage (no backup) €204 /Tb/year
Storage (backup) €360 /Tb/year

Are companies involved in this project?
Data will be made available in agreement with the Breed4food partner companies. The companies do not have any compromise with the project.

Signatures

Chair holder WIAS group*
Name: Martien Groenen

Signature 

Applicant
Name: Fernando Bueno Gutiérrez

Signature 

* With this signature the Chair holder guarantees the additional funding needed for the PhD project (research costs)

Literature

1. Consortium TF, Andersson L, Archibald AL, et al. Coordinated international action to accelerate genome-to-phenome with FAANG , the Functional Annotation of Animal Genomes project. Genome Biol. 2015:4-9. doi:10.1186/s13059-015-0622-4.
2. Dalton R. No bull : genes for better milk. Nature. 2009:149
3. Heffner EL, Jannink J, Sorrells ME. Genomic Selection Accuracy using Multifamily Prediction Models in a Wheat Breeding Program. Plant Genome. 2011;4:65-75. doi:10.3835/plantgenome2010.12.0029.
4. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. Genomic selection in dairy cattle: progress and challenges. J Dairy Sci. 2009;92(2):433-443. doi:92/2/433 [pii];10.3168/jds.2008-1646 [doi].
5. Schaeffer LR. Strategy for applying genome-wide selection in dairy cattle. J Anim Breed Genet. 2006;123(2001):218-223.
6. VanRaden PM, Van Tassell CP, Wiggans GR, et al. Invited Review: Reliability of genomic predictions for North American Holstein bulls. J Dairy Sci. 2009;92(1):16-24. doi:10.3168/jds.2008-1514.
7. Deriziotis P, Fisher SE. Neurogenomics of speech and language disorders : the road ahead. Genome Biol. 2013:1-12.
8. Gianola D. Priors in Whole-Genome Regression : The Bayesian Alphabet Returns. Genetics. 2013;194(July):573-596. doi:10.1534/genetics.113.151753.
9. Pérez-Enciso M, Rincón JC, Legarra A. Sequence- vs. chip-assisted genomic selection: accurate biological information is advised. Genet Sel Evol. 2015;47(1):43. doi:10.1186/s12711-015-0117-5.
10. Oscar O M I, Woolliams JA, Yu X, Wellmann R, Meuwissen THE. Within - and across - breed genomic prediction using whole - genome sequence and single nucleotide polymorphism panels. Genet Sel Evol. 2016:1-15. doi:10.1186/s12711-016-0193-1.
11. Meuwissen T, Goddard M. Accurate Prediction of Genetic Values for Complex Traits by Whole-Genome Resequencing. Genetics. 2010;185. doi:10.1534/genetics.110.116590.
12. Binsbergen Van R, Calus MPL, Bink MCAM, van Eeuwijk FA, Schrooten C, Veerkamp RF. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. Genet Sel Evol. 2015;47(1):71. doi:10.1186/s12711-015-0149-x.
13. Ober U, Ayroles JF, Stone EA, et al. Using whole-genome sequence data to predict quantitative trait phenotypes in Drosophila melanogaster. PLoS Genet. 2012;8(5). doi:10.1371/journal.pgen.1002685.
14. Brøndum RF, Su G, Janss L, et al. Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. J Dairy Sci. 2015;98(6):4107-4116. doi:10.3168/jds.2014-9005.
15. Cooper GM, Brown CD. Qualifying the relationship between sequence conservation and molecular function. Genome Res. 2008:201-205. doi:10.1101/gr.7205808.4.
16. Kimura, M. (1983). The neutral theory of molecular evolution (Cambridge University Press, Cambridge/New York)
17. Boffelli D, Nobrega MA, Rubin EM. Comparative genomics at the vertebrates extremes. Nature. 2004;5(June):456-465. doi:10.1038/nrg1350.

18. Dermitzakis ET, Reymond A AS. Conserved non-genic sequences — an unexpected feature of mammalian genomes. Nat Genet. 2005;6(February):1-7. doi:10.1038/nrg1526.
19. Kircher M. A general framework for estimating the relative pathogenicity of human genetic variants. Nat g. 2014;46(3):310-315. doi:10.1038/ng.2892.A.
20. Alföldi J, Lindblad-toh K. Comparative genomics as a tool to understand evolution and disease. Genome Res. 2013;23:1063-1068. doi:10.1101/gr.157503.113.Freely.
21. Rosse C, Steinberg S, Santos R, et al. Novel SNPs and INDEL polymorphisms in the 3 0 UTR of DGAT1 gene : in silico analyses and a possible association. 2014:4555-4563. doi:10.1007/s11033-014-3326-z.
22. Rubin C, Megens H, Martinez A, Maqbool K, Sayyab S. Strong signatures of selection in the domestic pig genome. PNAS. 2012;109(48):19529-19536. doi:10.1073/pnas.1217149109.
23. Li C, Sun D, Zhang S, et al. Genome wide association study identifies 20 novel promising genes associated with milk fatty acid traits in Chinese Holstein. PLoS One. 2014;9(5). doi:10.1371/journal.pone.0096186.
24. Torkamani A, Schork NJ. Predicting functional regulatory polymorphisms. Bioinformatics. 2008;24(16):1787-1792. doi:10.1093/bioinformatics/btn311.
25. Meader S, Ponting CP, Lunter G. Massive turnover of functional sequence in human and other mammalian genomes. Genome Res. 2010;20:1335-1343. doi:10.1101/gr.108795.110.
26. Herrero-medrano JM, Megens H, Groenen MAM, Bosse M, Pérez-enciso M, Crooijmans RPMA. Whole-genome sequence analysis reveals differences in population management and selection of European low-input pig breeds. BMC Genomics. 2014:1-12.
27. Groenen M. et al. Analyses of pig genomes provide insight into porcine demography and evolution. Nature. 2012;491(7424):393-398. doi:10.1038/nature11622.
28. Villar D, Berthelot C, Flicek P, et al. Enhancer Evolution across 20 Mammalian Species Article Enhancer Evolution across 20 Mammalian Species. Cell. 2015;160(3):554-566. doi:10.1016/j.cell.2015.01.006
29. Koonin EV, Galperin MY. Sequence—evolution—function: computational approaches in comparative genomics. Boston: Kluwer Academic; 2003. p. xiii. p., 411 plates
30. Schultz RG, Copley RR, Andrade MA, Bork P. The use of sequence information to frame structural , functional , and evolutionary hypotheses represents a major challenge for the postgeno- mic era . Central to an understanding of the evolution of sequence families is the concept of the domain : a stru. Adv Protein Chem. 2000;54.
31. Gaucher EA, Gu X, Miyamoto MM, Benner SA, Benner SA. Predicting functional divergence in protein evolution by site-specific rate shifts. Trends Biochem Sci. 2002;27(6):315-321.
32. Siepel A, Bejerano G, Pedersen JS, et al. Evolutionarily conserved elements in vertebrate , insect , worm , and yeast genomes. Genome Res. 2005:1034-1050. doi:10.1101/gr.3715005.
33. Cooper GM, Stone EA, Asimenos G, et al. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005:901-913. doi:10.1101/gr.3577405.
34. Margulies EH, Blanchette M, Comparative N, Program S, Haussler D, Green ED. Identification and Characterization of Multi-Species Conserved Sequences. Genome Res. 2003:2507-2518. doi:10.1101/gr.1602203.emerged.

35. Asthana S, Roytberg M, Stamatoyannopoulos J, Sunyaev S. Analysis of Sequence Conservation at Nucleotide Resolution. PLOS Comput Biol. 2007;3(12). doi:10.1371/journal.pcbi.0030254.
36. Davydov E V, Goode DL, Sirota M, Cooper GM, Sidow A. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP ++. PLOS Comput Biol. 2010;6(12). doi:10.1371/journal.pcbi.1001025.
37. Yokoyama KD, Zhang Y, Ma J. Tracing the Evolution of Lineage-Specific Transcription Factor Binding Sites in a Birth-Death Framework. PLOS Comput Biol. 2014;10(8). doi:10.1371/journal.pcbi.1003771
38. Ryan NM, Morris SW, Porteous DJ, Taylor MS, Evans KL. SuRFing the genomics wave : an R package for prioritising SNPs by functionality. Genome Med. 2014:1-13. doi:10.1186/s13073-014-0079-1.
39. Ritchie GRS, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. Nat Methods. 2014;11(3). doi:10.1038/nmeth.2832.
40. Lappalainen T, Sboner A, Lochovsky L, Chen J. Integrative Annotation of Variants from 1092 Humans: Application to Cancer Genomics. *Science (80-)*. 2014;342(6154):1-21. doi:10.1126/science.1235587.Integrative.
41. MACLEOD, I. M., HAYES, B. J. & GODDARD, M. E. The Effects of Demography and Long-Term Selection on the Accuracy of Genomic. Genetics. 2014;198(December):1671-1684. doi:10.1534/genetics.114.168344.
42. Gao N, Li J, He J, et al. Improving accuracy of genomic prediction by genetic architecture based priors in a Bayesian model. BMC Genet. 2015:1-11. doi:10.1186/s12863-015-0278-9.
43. Frantz LAF, Schraiber JG, Madsen O, et al. Genome sequencing reveals fine scale diversification and reticulation history during speciation in Sus. Genome Biol. 2013;14(9):R107. doi:10.1186/gb-2013-14-9-r107.
44. Bosse M, Lopes MS, Madsen O, et al. Artificial selection on introduced Asian haplotypes shaped the genetic architecture in European commercial pigs. Proc R Soc B. 2015;282.

Abbreviations and acronyms

ABGC	Animal Breeding and Genomics Centre
EC	Evolutionary Conservation
FAANG	Functional Annotation of Animal Genomes
GP	Genomic Prediction
GWAS	Genome-Wide Association Studies
HPC	High Performance Cluster
SNP	Single Nucleotide Polymorphisms
TF	Transcription Factor
TFBS	Transcription Factors Biding Sites
USCS	University of California Santa Cruz
WUR	Wageningen University and Research

Suggestion for referees

Please name three referees who are not involved in the project or in the participating groups. These referees should be able to give an independent judgement on the scientific quality and feasibility of the project. At least two of the referees must be from abroad.

Referee 1
name: Chris Maliepaard
affiliation: Wageningen University & Research
area of expertise: Breeding programmes, Bioinformatics, Biometrics, Genetics, Genomic
full address: Droevendaalsesteeg 1 6708PB Wageningen. The Netherlands
phone: +31317480855
fax:
e-mail: chris.maliepaard@wur.nl

Referee 2
name: Dirk-Jan de Koning
affiliation: Swedish University of Agricultural Sciences
area of expertise: Animal Genetics, Animal Breeding, Quantitative Genetics, Genomic selection
full address: Ulls väg 26, Uppsala. Sweden
phone: 018-672039
fax:
e-mail: dj.de-koning@slu.se

Referee 3
name: Miguel Ángel Toro Ibañez
affiliation: Compluetense University of Madrid
area of expertise: Evolutionary and applied genetics
full address: Etsi Agronómica, Aliment. y Biosistemas (Producción Agraria (N)), Madrid. Spain
phone: 914524900
fax:
e-mail: miguel.toro@upm.es