STW-Breed4Food Partnership Programme

Project 14283
From sequence to phenotype: detecting deleterious variation by prediction of functionality

---

Short Title: Predicting Impact of Variation in Livestock

Contact details

1. Applicants

Main applicant

Name initials and title : Prof. dr. Martien Groenen

Research institute : Animal Breeding and Genetics Group, Wageningen University
Droevendaalsesteeg 1, 6708PB Wageningen, The Netherlands

Phone number : +31 317 483747
E-mail : martien.groenen@wur.nl

Part-time percentage : 100%
Permanent position : yes

Co-applicant(s)

Name initials and title : Prof. dr. ir. Dick de Ridder

Research institute : Bioinformatics Group, Wageningen University
Droevendaalsesteeg 1, 6708PB Wageningen, The Netherlands

Phone number : +31 317 484074
E-mail : dick.deridder@wur.nl

Part-time percentage : 100%
Permanent position : yes

Co-applicant(s)

Name initials and title : Prof. dr. ir. Marcel Reinders

Research institute : Delft Bioinformatics Lab, Delft University of Technology
Mekelweg 4, 2628 CD Delft, The Netherlands

Phone number : +31 15 2786024
E-mail : m.j.t.reinders@tudelft.nl

Part-time percentage : 100%
Permanent position : yes

Keywords

Pig, chicken, machine learning, next generation sequencing, deleterious, genome variation

Project description

2. Summaries

Research summary.

The overall goal of the project is to develop procedures in livestock species to utilise the wealth of sequence and functional information currently accumulating. The proposed research project is based on data available at Wageningen University and at the Breed4Food partners for chicken, turkey, pig and cattle. The data includes whole-genome sequences of hundreds of individuals for each of these four species and genotypes and phenotypes for a large number of individuals from breeding lines of the Breed4Food partners. This sequence data will be used for the identification of genetic variants (SNPs) predicted to affect the phenotype. Computational methods based on machine learning will be developed to predict the likelihood of a given variant in the data being deleterious or pathogenic, by integrating genome annotation, evolutionary conservation and functional genomics data, and transferring knowledge obtained from other species (human) as far as possible. These predictors will be tested and validated in chicken, as it allows for relatively fast, inexpensive experimentation, but will be designed to be generally applicable in livestock breeding. Within the available genotyped pedigrees (pigs, chicken) as well as in specific experimental crosses between carriers of a selection of the identified variants (chicken), we will verify whether offspring homozygous for a variant are indeed affected or non-viable.

The main objectives of the project are to:

1. Develop an approach combining whole-genome sequencing, bioinformatics and machine learning to identify potentially deleterious variants with particular emphasis on lethal or pathogenic variants;
2. Explore the frequency of such variants in commercial breeding lines of the four livestock species;
3. Demonstrate the utility of specific marker-assisted breeding to remove deleterious and negative mutations from the population, resulting in improved performance of the commercial breeding lines.
4. Predict the likelihood of any mutation to have an effect on the phenotype

Utilisation summary

The project makes optimal use of state of the art genomics technology to improve genetic breeding stock by identifying likely deleterious and/or pathogenic variants. This will allow subsequent removal of these variants from the population. The approach will be complementary to other molecular breeding tools such as genomic selection, which are restricted to dominant and additive variants and to variants found at relatively high frequencies in the population. The procedure that will be developed will initially focus mainly on protein altering variants in these species. However, as additional functional information becomes available, the proposed methods will become more effective for predicting additional deleterious mutations affecting the expression and regulation of genes in the near future.

The project is expected to contribute to a reduction in inbreeding depression and in increased performance of the lines analysed. It will define an improved strategy for molecular breeding combining the selection against negative variants with on-going genomic selection in breeding. The main deliverables directly applicable towards this utilisation are:

1. A list of C-scores, predictions of the deleteriousness of variants, for every position in the genomes of the livestock species considered in this project, that can be used to generate priors for specific SNPs used in genomic selection
2. An analysis pipeline for regular updates of these C-scores
3. A list of potentially deleterious variants in the livestock species used within the project

3. Composition of the group

The current group

The proposed project brings together a group of leading experts in the area of animal genomics (Prof Groenen) and bioinformatics (Prof de Ridder and Prof Reinders). Increasingly, experiments in genetics and genomics produce large, complex data sets. To fully utilize the wealth of information in this data requires insight in the biology of the species as well as expertise in state-of-the-art analysis techniques (bioinformatics). **Prof Groenen** is head of genomics research within the Animal Breeding and Genomics Centre (ABGC) at Wageningen University (WU) with a focus on the characterization of molecular variation underlying phenotypic variation in livestock species. He has a strong track record in genome sequencing and variant discovery in a variety of species (chicken, pig, duck, turkey, the great tit) [1-5]. He is also a member of the steering committees of international consortia to further functionally characterize these genomes (FAANG, Functional Annotation of Animal genomes; the International Swine Methylome Consortium) and has direct access to data relevant for the project. **Prof de Ridder** heads the Bioinformatics Group at WU, which focuses on fundamental and applied bioinformatics research in the green life sciences, developing and applying novel computational methods for the analysis and integration of -omics data [6-9]. In particular, he studies how to combine such data with prior knowledge about a problem to be solved in machine learning-driven models and algorithms. **Prof Reinders** heads the Pattern Recognition and Bioinformatics group that conducts research into pattern recognition, computer vision and bioinformatics at Delft University of Technology (DUT) [7-11]. His group is recognized as international experts on machine learning and in his bioinformatics research he applies the cutting-edge machine learning expertise to develop data-driven analysis methodologies to progress molecular biology insights. He initiated work on molecular classification and genetic network modelling, and has a strong track record on next-generation sequencing analysis, network-based analysis, and integration of genomic data. He is scientific co-director of the Netherlands Bioinformatics Centre, and member of the scientific advisory board of the Dutch Techcentre for Life Sciences.

The ABGC has a strong collaboration with the industrial partners within Breed4Food. These partners participate in the project by providing the following data: **Topigs Norsvin** and **Hendrix Genetics** will provide pedigree information with genotypes for the validation experiments; **Topigs Norsvin** and **Cobb-Vantress** provide whole genome sequence data for additional pigs and chickens, respectively. **Hendrix Genetics** will perform specific chicken crosses for the validation experiment.

Available infrastructure

The ABGC has well equipped laboratories for state of the art molecular analyses (DNA isolation, sequencing, PCR, SNP analyses). For large scale SNP typing we routinely use external providers and the necessary procedures required are well established. The majority of the analyses within the project are computational analyses of large whole-genome data sets. Both the groups at WU and DUT have extensive expertise in the field of computational genomics. WU in collaboration with the Breed4Food partners in 2013-2014 has established a high performance computing (HPC) infrastructure in Wageningen. This HPC consists of 896 cores, and includes two fat nodes (64 cores each) with 1 Tb of memory for each of the 2 nodes. Furthermore, both DUT and WU are connected to the Dutch Life Science Grid with access to an additional high memory machine (2 Tb). All sequence data is available on a large parallel file system on the HPC (600 Tb) directly accessible for our analyses. Pipelines for whole genome sequence alignments and variant calling for next generation sequence data are operational.

4. Scientific description

Background and state-of-the-art

The Holy Grail in functional genomics is the ability to understand, at the molecular level, the function of all elements in the genome. Key questions to be answered are how their genome sequences eventually contribute to the morphology and functioning of an organism, and how variants in these sequences affect a specific phenotype. Most of the traits important in livestock breeding are complex in nature, with a large number of genes affecting the trait. One of the most compelling challenges of modern genetics has therefore been the unravelling of the genetic factors underlying these complex multifactorial traits. An important approach extensively applied in livestock species, has been the identification of quantitative trait loci (QTL). Although numerous studies have, by now, reported thousands of QTL [12], identification of the actual genes underlying these QTL has proven to be extremely difficult with only a few variants with large effect actually being used in animal breeding. Implementation of information on functional relevance of variants in breeding, as a result, is rare.

The recent development of medium and high-density SNP chips, now available for most livestock species, has opened up an alternative approach of using molecular tools in breeding, generally referred to as Genomic Selection (GS). Although very successful, GS treats the genome as a black box and is not very effective at addressing non-additive effects and variants that occur at low allele frequencies in the population. Furthermore, GS is limited to using traits measured during the training phase, where both SNPs and phenotypes are being collected on a large number of individuals. Another disadvantage of GS is that, whilst effective within lines, it is not very powerful in crosses between lines. This is of particular importance in poultry and pig breeding, where most commercial products are based on 3- or 4-way crosses between different lines.

The rapid improvements in genome sequencing have opened up new possibilities to explore and use the genetic variation seen in these animals. We have sequenced over 270 individual pigs, 290 individual chicken (layers and broilers), 65 bulls and 160 individual turkeys from a variety of breeds, and identified tens of millions of genetic variants (single nucleotide polymorphisms or SNPs). In this project, we will combine our joint expertise in bioinformatics, machine learning and animal genomics to use this wealth of information in future breeding programs.

Genome variation

The amount of variation seen in a genome depends on past and present population demography and can vary substantially in different species and/or populations. Our sequencing results as well as those of others have demonstrated a 2-4 fold higher variation in livestock species than in humans [1, 3, 13-15]. Accordingly, the number of non-synonymous mutations seen in these animals is higher than that seen in individual humans, on average 10,000 [16]. As in human 25-50% of the rare non-synonymous variants are estimated to be deleterious [17], this suggests a large reservoir of potential deleterious variants in livestock as well. Although many of the variants are likely only mildly deleterious and contribute to the phenotypic variation within a population, each individual is expected to harbour a considerable number of lethal and/or pathogenic variants as well. In a systematic survey of loss of function (LoF) variants in human, it was observed that on average an individual carries around 100 LoF variants, many of which are found at low frequency in the population [18]. The higher observed number of genetic variants in livestock species makes it likely that these species likewise carry a large number of potential lethal or pathogenic alleles, which recently has been confirmed in several studies in cattle [14, 15, 19]. Therefore, the ability to be able to predict such variants in genomes is expected to become highly relevant, facilitating the promise of precision breeding.

Distinguishing benign from potentially deleterious mutations

Our hypothesis is that, as shown for human, any individual animal genome contains hundreds of potential harmful mutations. It is to be expected that the majority of these variants occur at a low allele

frequency in the population. Although many variants most likely are recessive, it is to be expected that heterozygote carriers have a slightly lower fitness than individuals that are homozygous for the wild type. Therefore, despite their low frequency, collectively, these variants will have an overall negative effect on the fitness of the population. However, they are difficult to detect or breed against using classic or genomic selection because of the low allele frequency and relatively small (non-additive) effect of the individual variants in the heterozygote state.

Predicting phenotype from sequence has seen rapid progress in human genetics. Large-scale efforts such as the 1000 Genomes and ENCODE projects have generated the data required for this development. As a result, a wide range of tools has become available to predict functions from sequence information. The majority of these tools – such as SIFT, SNPS&GO, PolyPhen-2 and VEP – focus on the prediction of deleterious variants that affect the protein sequence (for recent reviews, see [20, 30]). Although these tools are effective in identifying variants likely to affect the function of a protein, they only address the variation in the coding fraction of the genome. This fraction has been estimated to only represent around 20% of the genetic variation underlying the observed phenotypic variation [21]. Most current tools focus on conservation-based features of variants, and some tools are "trained" (i.e. parameters are optimized) to distinguish between variants known to be related to a phenotype (e.g. disease, as found in OMIM [22]) and all other variants. In an attempt to combine all available information, both coding and non-coding, Kircher and co-workers [23] recently developed a method (combined annotation-dependent depletion, CADD) that integrates all relevant information available into a single score (C-score), without resorting a priori to such training. Based on machine learning, this C-score was then shown to be able to distinguish benign from potentially deleterious mutations. A number of recent publications improve on this approach by integrating more data [31] and using deep learning approaches [32].

In this project, we will follow an approach similar to the one taken by CADD and similar tools. A predictor for the likelihood for coding and non-coding variants being deleterious will be developed, exploiting the large number of whole-genome sequence datasets available for chicken (290 individuals), turkey (160 individuals) pig (270 individuals) and cattle (65 bulls and access to the 1000 bull genomes data set). The variants identified in these individuals will be complemented with a set of non-variants and a classification algorithm (e.g. support vector machine or random forest) will be trained to distinguish deleterious from non-deleterious variants based on the available functional information (annotation).

Activity 1a: Effect prediction for mammals (pig, cow)

The incomplete annotation and draft status of livestock genomes will be particularly challenging, and will require innovative approaches to obtain reliable predictions of functional effects (C-scores) in livestock species. An interesting question in that respect is to what extent the rich data available for the human and mouse genomes can be used to improve predictions. While it is not likely that knowledge on individual genes or annotations can be transferred, it is to be expected that similar features hold predictive value. A recent study on 20 mammalian genomes [24] indicates that while enhancers evolve very rapidly, for promoter regions function is generally conserved. Another avenue of research is to improve the predictor, leveraging the results of subsequent experimentation and the increasing availability of functional genomics data in livestock. Once properly trained, the classifier will provide a likelihood for any given variant being deleterious or pathogenic. Although, the information available for livestock species is currently biased to coding sequences, extensive RNAseq resources are available for these species, as well as a small number of ENCODE type data sets [24]. With the recently initiated FAANG consortium [25], these resources are expected to increase substantially over the next few years. Likewise, comparative genomics information has shown to be a powerful

component for computing the C-scores in human [23]. Briefly, this activity consist of the following steps:

- Identify lineage specific derived alleles present at high allele frequency (90-95%)
- Create random variant set (using mutation model)
- Construct matrix with functional information (RNAseq, methylation data, histone modifications, comparative genomics data such as GERP scores, SIFT/PolyPhen2 scores optimized for pig/cow)
- Use machine learning based methods to compute C-score (several iterations to improve the algorithm)

Activity 1b: Effect prediction for poultry (chicken, turkey)

A similar approach to Activity 1a will be applied to calculate C-scores for chicken and turkey. However, because of the higher evolutionary distance to mammals it is expected that the proposed methods require optimization for use in birds. The recent completion of the genome sequences of a large number of birds [33] and the establishment of Avianbase [34] will aid in the use of bird specific conservations scores.

Activity 2: Identification of potential deleterious alleles from whole genome sequence data using SIFT/Polyphen2 scores

Although coding variants are estimated to represent only 20% of the genetic variation underlying phenotypic variation, they are much easier to identify and validate. In addition, having a catalogue of such variants would provide the necessary information to benchmark other activities proposed. The emphasis, therefore, will initially be on non-synonymous and LoF coding variants in the individuals for which genome sequences are available. Variants have already been identified using our NGS analysis pipelines, but these need additional stringent filtering to remove false positives/artefacts that result from the draft nature of current reference genomes and their incomplete annotation. We will randomly select and experimentally validate variants representing different classes. Validation will be done by PCR/Sanger sequencing for a representative number of variants.

Activity 3: Missing homozygotes: analysis of population/SNP genotyping data

For both chicken and pigs, through our collaboration with the breeding companies Topigs-Norsvin, Hendrix Genetics and Cobb-Vantress, we have access to a large data set of pedigreed and genotyped individuals. Because of genomic selection, genotypes are available for tens of thousands of individuals. We will use this data to test for statistical depletion or absence of specific homozygous haplotypes in these pedigrees [26], i.e. perform homozygosity mapping. Significant depletion of such haplotypes is an indication of decreased viability. Intersecting results from Activity 3 with results of Activities 1 and 2 will allow for a direct assessment of decrease in viability of specific variants, particularly those variants that occur at a somewhat higher frequency. Due to the very large number of pedigreed animals, this approach is expected to be powerful even for rare haplotypes/alleles (~10%-5% occurrence and less).

Activity 4: Validation 1: Specific SNP assays for predicted lethal variants

Most of the sequenced individuals are derived from the same populations as used under activity 3, allowing comparison with our predictions under activities 1 and 2, and enabling the identification of the potential underlying causative mutations for some of the regions identified. During the first year of the project the emphasis will be on activity 2, but as our C-score predictions improve, additional non-coding sequences will be taken into account as well. We will design SNP assays for the most promising potentially causative variants in the regions identified and use these to genotype representative individuals, i.e. from pedigrees depleted for homozygotes at these specific chromosomal locations (for pigs, turkey and chicken).

Activity 5: Validation 2: Specific matings between carriers of predicted lethal alleles in chicken

A more direct approach is to set up specific crosses between two carriers of the same predicted lethal mutation to test whether indeed homozygous individuals are affected or non-viable. We will do this in chicken, since experimental crosses can be established relatively cheaply and quickly, generating a large number of full sibs. Furthermore, the possibility to examine eggs that do not hatch, will allow further genetic and phenotypic verification of non-viable homozygotes. Of the 290 chickens for which genome sequences are available, 240 are part of three large populations of pedigreed individuals that are genotyped with a 60K SNP chip. These 240 chickens represent the majority of haplotypes segregating in the 3 populations. We will develop SNP assays for the most likely lethal alleles and identify carriers of potential lethal alleles in the current populations. We plan to use the "Illumina add on" option, where additional SNPs can be added to the 60K chip used for genomic selection by the Breed4Food partners. Given the budget, this option would allow us to genotype around 3000 individuals for up to 1000 SNPs. Heterozygous carriers will be used for experimental crosses for the 20 most promising mutations. For each of the crosses, we will monitor whether homozygous offspring are viable or not, and whether homozygous offspring display any visible negative phenotypic characteristics. The short generation interval in chicken, and the tight collaboration with the breeding companies makes this achievable within the time frame of the project.

Activity 6: Haplotype analysis and population dynamics of potential deleterious alleles

We will explore the possibility to use the sequences of the 240 animals in combination with the genotypes obtained from activity 5 to impute the deleterious alleles into additional (phenotyped) animals of the three layer lines. The research question to be addressed within this activity is to explore to what extent such variants underlie phenomena like inbreeding depression and heterosis. The use of crosses between divergent inbred populations, as is done in modern animal breeding, provides a powerful case for investigating cumulative effects of mildly deleterious mutations, particularly those that occur at a high frequency in the inbred population. Specifically, parents that have the highest predicted cumulative load of homozygous, mildly deleterious alleles should produce offspring that display the most pronounced heterosis effects. The large numbers of genotyped pure line animals and crossbreds will allow statistical detection of interactions between such variants and phenotype.

Activity 7: Provide predictor results in format to be easily used by breeders

While in this project the C-scores will be used to identify potentially lethal variants, they will be more widely useful to the community, for example to develop priors in genomic selection. To this end, we will disseminate the C-scores both in tabular formats, as well as in the form of custom tracks (i.e. bigwig files) to be used in genome browsers such as Ensembl or UCSC. To support decision making based on the C-scores provided, we will investigate the possibility of adding information on the most relevant features underlying the prediction at each position, i.e. which type(s) of annotation at what value(s) most influence the C-score. C-scores will be provided for all livestock species relevant for the breed4food partners (pig, chicken, turkey, cattle)

Activity 8: Provide optimized analysis pipeline for regular updates of C-scores

While the architecture of the predictor developed in activity 1 will at some point be fixed, regular re-training is essential to exploit updates to the data sources used and to allow new functional data to be included. To allow for this, we will develop a fully autonomous pipeline that allows the addition of new annotations to the existing ones, followed by re-training of the predictor and regeneration of the C-scores, tables and genome browser tracks. This pipeline will also provide information on the utility of the novel data sources.

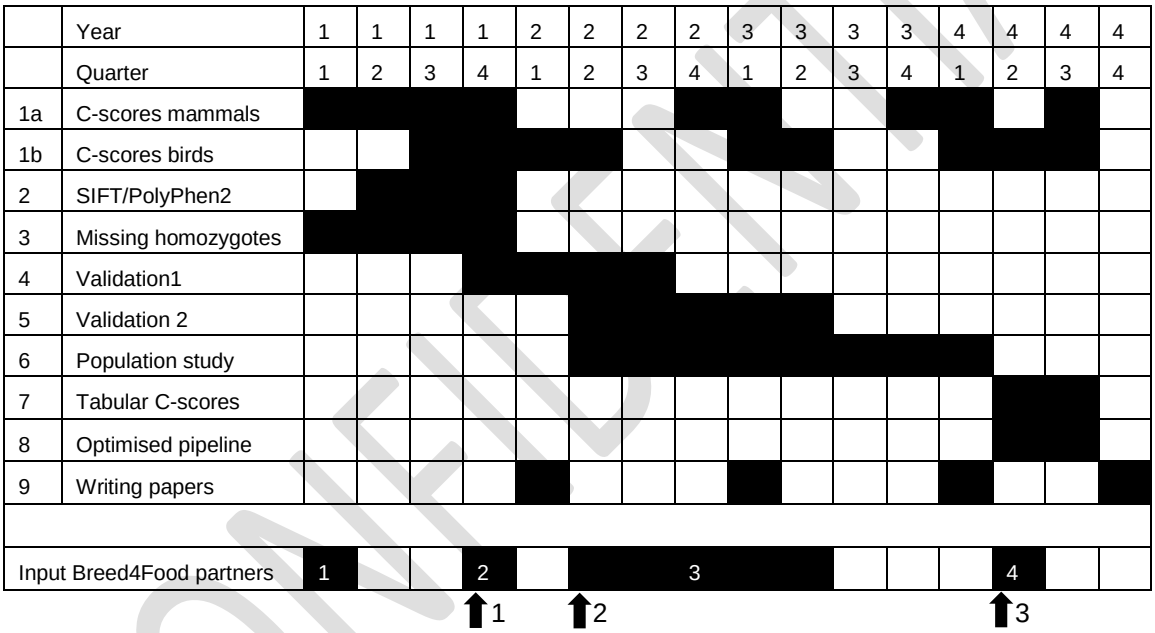
Required personnel and equipment

The proposed project is multi-disciplinary and key to its success will be the integration of the bioinformatics and biological data within the project. We therefore ask for funding of two PhD students with a background in bioinformatics and molecular genetics respectively. Successful communication and integration is achieved by monthly project meetings with the PI's and direct supervisors of the PhDs. Crucial for the successful integration and implementation of the approaches developed within the project are the validation assays in year 2 and 3. This will be further strengthened by the appointment of a post-doc with expertise in both molecular genetics and bioinformatics during this period of the project. The post-doc will also explore the use of the imputed variants in the chicken crosses between the lines analysed in this project (activity 6).

All the necessary infrastructure is available at WU and DUT and no additional equipment is required. Specific crosses for activity 5 will be provided in kind by one of the Breed4Food partners (Hendrix Genetics).

.

Time schedule and allocation of tasks



The major decision points (1-3, indicated by an arrow) are at the end of year 1/beginning of year 2 and the middle of year 4.

1. List of potential deleterious coding variants available. Selection of SNP assays to be designed.
2. First list of C-scores. Selection of SNPs to be used for validation experiment 2.
3. Evaluation of results and decision on optimal way to provide C-scores to be used for breeding.

Input Breed4Food partners

1. Provide pedigree data, genotypes (pig, poultry) and additional sequences (broilers, Cobb).
2. Selection of assays and individuals to genotype (provide DNA).
3. Selection of assays and individuals to genotype (provide DNA) and make crosses between carriers.
4. Discuss optimal form of data to be used within breeding programs.

Division of activities

1. Activities 1, 2, 7, 8: Bioinformatics PhD student
2. Activities 2, 3, 4, 5: ABGC PhD student
3. Activities 5, 6, 7: Post-doc
4. Activity 9: All.

5. Fit within the research topics of the programme

The overall goal of the project is to develop and experimentally validate novel computational methods to identify potentially deleterious variants in livestock species, with particular emphasis on pig and poultry. Such methods will not only provide insight into the genetic makeup of populations of these species, but will also provide new genomics tools to breeders to improve genetic breeding stock, as detection of such deleterious variations will allow their subsequent removal from the population. The approach will be complementary to other methods such as genomic selection, which are restricted to dominant and additive variants found at relatively high frequencies in the population. Genomic selection provides a framework in which additional genomic information, such as envisioned to be discovered in the current application, can be leveraged. Combined, these molecular breeding methods are expected to contribute to a reduction in inbreeding depression and to increased performance, by incorporating selection against negative variants into on-going genomic selection in breeding.

The development of such tools is at the core of the STW Breed4Food programme, which calls (among others) for new genomics and bioinformatics methods for accurate prediction of phenotypes, to enable genetic improvement in new and complex traits. It specifically addresses the first research area in the programme, i.e. prediction of phenotypes by using all genetic information available for a given species and within a specific population.

Connections with other research

The current proposal builds upon the vastly increasing amount of genomic data and on computational developments in human. An important aspect of the C-score analysis is the availability of functional "ENCODE" type of genome data for the livestock species. While a few data sets in pigs are already available (WU, [24]), it is expected that this number will greatly increase in the coming years. Most relevant in that respect is the establishment of the FAANG consortium (Functional Annotation of Animal Genomes) that recently has presented its plans in a commentary in Genome Biology ([25], see also www.faang.org). As a member of the steering committee of FAANG, Prof Groenen has early access to relevant data to be used in the current proposal. With respect to C-score analysis in cattle, WU will closely collaborate with Prof Michel Georges of the University of Liege. Prof Georges was awarded an ERC Advanced grant that, among other things, focuses on deleterious variants in cattle.
The work outlined in this proposal is focused on a prediction how likely a specific variant is expected to affect the phenotype based on the wealth of functional genome information. How best to implement that information as a prior in genomic selection is extremely relevant but outside the scope of the current proposal. Much of the research within the ABGC (Dr. Mario Calus, Prof Roel Veerkamp) is focused specifically on this aspect however, and we will ensure close collaboration and discussions between the proposed project and ongoing projects on genomic selection.

6. Utilisation plan

The challenge from the practice and the proposed solution

This project will deliver novel, experimentally validated computational tools to identify potentially deleterious variants in any livestock species. Identification and removal of such deleterious variants from the population will allow an overall increase of the fitness and performance. The innovative aspect of this approach is that, rather than trying to increase the frequency of the positive alleles as done in the more classical approaches, it allows breeders to decrease the frequency of a large number of rare potential negative variants. By doing so, the net result will be the same: increased performance of the population and increased animal welfare. From within the genomics selection community there is direct interest in implementing the information obtained through the proposed project. Since genomic selection in modern animal breeding has laid the groundwork for genomics-

based management of populations, both the theoretical and practical basis for implementation is ensured. For instance, weighting the negative effects of alleles can be implemented in GS models to avoid such variants from establishing a high frequency in the population..

Some deliverables of the project can be directly applied on the short term in breeding programs, while others are more directed towards the medium and longer term.
Short term: Identification of deleterious coding variants (non-synonymous and LoF variants. Critical for identifying these variants is the adaptation of existing tools towards livestock species and the additional filtering required due to the draft nature and incomplete annotation of the livestock genomes. The validation experiments within the project will be done in close collaboration with the Breed4Food partners, further ensuring optimal implementation of the results.
Medium term: The resulting C-scores will be made available to the breeding industry such that they can be directly used as a prior for estimating genomic breeding values in genomic selection. Because our results (predictor scores and potential lethal alleles) are identified using actual breeding lines of Topigs Norsvin, Hendrix Genetics, CR-Delta and Cobb-Vantress, the results can be directly applied and tested in current breeding programs of these companies.
Long term: While it is to be expected that in particular information on coding variants will prove valuable as a prior in genomic selection in the medium term, we expect that over time we increasingly will be able to identify potential deleterious alleles in the non-coding regulatory part of the genome. This will be on ongoing process, with predictions becoming more powerful as more data will accumulate. In this respect initiatives like FAANG are particularly relevant. One of the deliverables of the project will be a bioinformatics pipeline that enables to update the C-scores making sure that the priors to be used within genomic selection are updated and improved continuously.

Cooperation with Breed4Food

WU already has an intensive collaboration with the individual Breed4Food partners for many years, and currently Prof Groenen has bi-monthly meetings with Topigs Norsvin and Hendrix Genetics to discuss ongoing research collaboration. This collaboration has already resulted in the rich data sets (genome sequences) to be used within the current project. Within the current project this will continue and with the additional user meetings for the current project, this collaboration will be further extended to the bioinformatics groups of WU and DUT. The breeding companies realize that bioinformatics expertise increasingly is becoming important in genetics, and some of the companies have already taken this a step further in attracting bioinformaticians to work within the R&D of the company (i.e. Hendrix Genetics). Within this project we will actively stimulate the exchange of expertise between universities and breeding companies. Also, the PhD students trained in this project may prove valuable as future employees at these companies.
Another important aspect is how best to implement the results of the project (potential deleterious alleles, C-scores) within current breeding programs. While this aspect is not directly the focus of the current proposal, we are aware of other proposals specifically addressing this. In case these proposals are successful, we will exchange information and work closely together with these other projects to facilitate the incorporation within breeding programs. More specifically, we will explore joint project and user meetings to optimise exchange between the different projects.

Past performance in utilisation

As indicated in the previous paragraph, the ABGC has a long-lasting collaboration with the different Breed4Food partners. This has already resulted in a variety of activities and tools that are used by these breeding companies:

- We have successfully implemented genotyping protocols for specific qualitative traits (e.g. the *FMO3* gene and genes affecting feather colours) in poultry and up to now have performed over 50,000 assays for Hendrix Genetics.
- We have developed marker panels (initially microsatellites, in the last decade SNPs) for parentage control in livestock pedigrees (in particular chicken and turkey).
- We have led a consortium that developed the Illumina 60K bead chip used worldwide by the pig breeding industry for genomic selection. We also developed an updated version of the chip that is used exclusively by the genotyping company GeneSeek in the US.
- More recently, we have developed a 665K Affymetrix Axiom high density SNP chip for pigs, which has been announced at the Plant and Animal genome Meeting in San Diego, January 2015 (*M. Groenen, "Development of a high-density Axiom® porcine genotyping array to meet research and commercial needs*").
- In collaboration with the USDA, we developed a 650 K Axiom SNP chip for turkey.
- The ABGC is involved in a number of collaborative projects with the breeding industry to improve Genomic Selection approaches.
- We have identified a major locus with an effect on "Boar taint" and SNPs at this locus are used by Topigs Norsvin to select against this trait [27-29].

Profs de Ridder and Reinders have extensive experience in public-private partnerships, mostly in biotechnology. They are faculty members of the Kluyver Centre for Genomics of Industrial Fermentation, now part of BE-Basic, and have worked in several projects with industrial partners (DSM, Amyris), among others the Platform Green Synthetic Biology. Some of this work led to patent applications, e.g. on the use of machine learning in protein redesign (WO 2010102982 A1, WO 2013160316 A1).

Prof Reinders (DUT) developed a within sample copy number detection tool for next generation sequencing data that is being used for prenatal screening in clinical genetics departments of Dutch medical centres. He also has ample experience in prioritizing genomic variants, amongst others using CADD, in a number of projects involving human data (longevity, cancer, and Alzheimer).

7. Contracts and patents

In an initial search, we did not find any patents directly related to the ideas put forward in this project.

8. Budget

	Year1	Year2	Year3	Year4	Total
PhD1	39,072	46,224	49,152	52,296	186,744
PhD2	39,072	46,224	49,152	52,296	186,744
Post-doc		30,564	61,596	31,032	123,192
Technician		3,794	7,588		11,382
Consumables	26,000	77,000	77,000	30,000	210,000
Travel costs	5,000	7,500	7,500	5,000	25,000
Total					743,062

Consumables include the following costs:

- DNA isolation, PCR and Sanger sequencing € 20,000
- Open access publications (2 PhD, Post-doc) € 10,000
- SNP genotyping (validation experiments) €100,000
- High Performance Computing (HPC) € 80,000

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 40 Tb.

Computing:	€0.083 /core/hour
Storage (no backup)	€304 /Tb/year
Storage (backup)	€468 /Tb/year

In addition to the requested contribution from STW as outlined in the table, the Breed4Food partners will contribute in kind by providing 60K SNP genotype data for over 10,000 individuals (Hendrix Genetics, Topigs Norsvin), sequencing data for at least 20 individual broilers sequenced at 20x (Cobb Vantress) and by setting up the experimental crosses as described for activity 5 (Hendrix Genetics).

Literature

1. Groenen MA, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, Rogel-Gaillard C, Park C, Milan D, Megens HJ et al. Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491:393-398, 2012
2. Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, Bork P, Burt DW, Groenen MA, Delany ME, Dodgson JB,et al. International Chicken Genome Sequencing Consortium. (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432:695-716
3. Bosse M, Megens HJ, Frantz LA, Madsen O, Larson G, Paudel Y, Duijvesteijn N, Harlizius B, Hagemeijer Y, Crooijmans RP, Groenen MA. (2014) Genomic analysis reveals selection for Asian genes in European pigs following human-mediated introgression. Nat Commun. 5:4392
4. Bosse M, Megens HJ, Madsen O, Paudel Y, Frantz LA, Schook LB, Crooijmans RP, Groenen MA. (2012) Regions of homozygosity in the porcine genome: consequence of demography and the recombination landscape. PLoS Genet. 8:e1003100.
5. Frantz LA, Madsen O, Megens HJ, Groenen MA, Lohse K. (2014) Testing models of speciation from genome sequences: divergence and asymmetric admixture in Island South-East Asian Sus species during the Plio-Pleistocene climatic fluctuations. Mol Ecol.23:5566-5574
6. Aflitos S, Schijlen E, de Jong H, de Ridder D, Smit S et al. (2014) Exploring genetic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing. Plant Journal 80(1):136–48
7. van den Berg BA, Reinders MJ, de Ridder D, de Beer TA. Insight into neutral and disease-associated human genetic variants through interpretable predictors. PLoS ONE, in press.
8. de Ridder D, de Ridder J and Reinders MJ. Pattern recognition in bioinformatics. Briefings in Bioinformatics 14(5):633-647, 2013.
9. Nijkamp JF, Pop M, Reinders MJ and de Ridder D. Exploring variation-aware contig graphs for (comparative) metagenomics using MaryGold. Bioinformatics 29(22):2826-34, 2013.
10. Straver R, Sistermans EA, Holstege H, Visser A, Oudejans CB, Reinders MJ. (2014) WISECONDOR: detection of fetal aberrations from shallow sequencing maternal plasma based on a within-sample comparison scheme. Nucleic Acids Research 42(5):e31,
11. van den Akker EB, Verbruggen B, Heijmans BT, Beekman M, Kok JN, Slagboom PE, Reinders MJ. (2011) Integrating protein-protein interaction networks with gene-gene co-expression networks improves gene signatures for classifying breast cancer metastasis. J Integr Bioinform. 8:188.
12. Hu Z, Park CA, Wu XL and Reecy JM. Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. *Nucleic Acids Research* 41 (D1):D871-D879, 2013.
13. Wong GK, Liu B, Wang J, Zhang Y, Yang X, Zhang Z, Meng Q, Zhou J, Li D, Zhang J, Ni P, Li S, et al. International Chicken Polymorphism Map Consortium. (2004) A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. Nature 432:717-722
14. Daetwyler HD, Capitan A, Pausch H, Stothard P, van Binsbergen R, Brøndum RF,Liao X, Djari A, Rodriguez SC, Grohs C, Esquerré D, Bouchez O, Rossignol MN, Klopp C, Rocha D, Fritz S, Eggen A, Bowman PJ, Coote D, Chamberlain AJ, Anderson C, VanTassell CP, Hulsegge I, Goddard ME, Guldbrandtsen B, Lund MS, Veerkamp RF, Boichard DA, Fries R, Hayes BJ. (2014) Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet. 46:858-65.

15. Jansen S, Aigner B, Pausch H, Wysocki M, Eck S, Benet-Pagès A, Graf E, Wieland T, Strom TM, Meitinger T, Fries R. (2013) Assessment of the genomic variation in a cattle population by re-sequencing of key animals at low to medium coverage. BMC Genomics. 14:446.
16. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME and McVean GA. A map of human genome variation from population-scale sequencing. *Nature* 467(7319):1061-73, 2010.
17. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422):56-65, 2012.
18. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science 335(6070):823-8, 2012.
19. Pausch H, Kölle S, Wurmser C, Schwarzenbacher H, Emmerling R, Jansen S, Trottmann M, Fuerst C, Götz KU, Fries R. (2014) A nonsense mutation in TMEM95 encoding a nondescript transmembrane protein causes idiopathic male subfertility in cattle. PLoS Genet. 2014 10:e1004044
20. Peterson TA, Doughty E and Kann MG. Towards precision medicine: advances in computational approaches for the analysis of human variants. *Journal of Molecular Biology* 425(21):4047-63, 2013.
21. Ponting CP and Hardison RC (2011) What fraction of the human genome is functional? Genome Res. 21:1769-1776
22. Hamosh A, Scott AF, Amberger JS, Bocchini CA and McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Research 33 (S1):D514-7, 2005.
23. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM and Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics* 46(3):310-5, 2014.
24. Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, Park TJ, Deaville R et al. (2015) Enhancer evolution across 20 mammalian species. Cell 160: 554-566
25. The FAANG consortium. Coordinated international action to accelerate genome to phenome with FAANG, the Functional Annotation of Animal Genomes project (2015) Genome Biology *in press*
26. VanRaden PM, Olson KM, Null DJ, and Hutchison JL (2011) Harmful recessive effects on fertility detected by absence of homozygous haplotypes. J. Dairy Sci. 94 :6153–6161
27. Ramos AM, Duijvesteijn N, Knol EF, Merks JW, Bovenhuis H, Crooijmans RP, Groenen MA, Harlizius B. (2011) The distal end of porcine chromosome 6p is involved in the regulation of skatole levels in boars. BMC Genet. 12:35
28. Duijvesteijn N, Knol EF, Merks JW, Crooijmans RP, Groenen MA, Bovenhuis H, Harlizius B. (2010) A genome-wide association study on androstenone levels in pigs reveals a cluster of candidate genes on chromosome 6. BMC Genet. 11:42
29. Duijvesteijn N, Knol EF, Bijma P. (2014) Boar taint in entire male pigs: a genomewide association study for direct and indirect genetic effects on androstenone. J Anim Sci. 92:4319-28.
30. Ritchie, GR and Flicek, P. Computational approaches to interpreting genomic sequence variation. Genome Medicine 6:87, 2014.
31. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day IN, Gaunt TR and Campbell C. (2015.) An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics, advance access publication online
32. Quang D, Chen Y and Xie X. (2014) DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics, advance access publication online
33. Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, et al. (2014) Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346:1320–1331
34. Eöry L, Gilbert MT, Li C, Li B, Archibald A, Aken BL, Zhang G, Jarvis E, Flicek P and Burt DW (2015) Genome Biology 16:21

9. Key words, abbreviations and acronyms

Pig, chicken, machine learning, next generation sequencing, deleterious, genome variation

ABGC	Animal Breeding and Genomics Centre
DUT	Delft University of Technology
CADD	Combined Annotation-Dependent Depletion
ENCODE	Encyclopedia of DNA Elements
FAANG	Functional Annotation of Animal genomes
GERP	Genomic Evolutionary Rate Profiling
GS	Genomic Selection
HPC	High Performance Computing
LoF	Loss of Function
NGS	Next Generation Sequencing
OMIM	Online Mendelian Inheritance in Man
PCR	Polymerase Chain Reaction
PolyPhen2	Polymorphism Phenotyping v2
QTL	Quantitative Trait Loci
SIFT	Sorting Intolerant From Tolerant
SNP	Single Nucleotide Polymorphism
Tb	Terabytes
UCSC	University of California at Santa Cruz
VEP	Variant Effect Predictor
WU	Wageningen University