GENE-SWitCH

The regulatory GENomE of SWine and CHicken: functional annotation during development

Deliverable D4.3
Validation of developed models with simulated or publicly available data

Deliverable leader: WU

Authors: Mario Calus (WU)**,** Bruno Perez (HG), Marco Bink (HG)

Version: 1.0

Due date of deliverable (as in DoA):	M30
Actual submission date:	20/12/2021 M30

Dissemination level:

PU Public	**X**
CO Confidential, only for members of the consortium (including the Commission Services)	

Research and Innovation Action, SFS-30-2018-2019-2020 Agri-Aqua Labs
Duration of the project: 01 July 2019 – 30 June 2023, 48 months

Table of contents

1 Summary

- **Objectives** Currently used genomic prediction models rely on a reference population including animals with both known genotypes and measured phenotypes. Our objective was to stepwise develop a genomic prediction model that better captures the genetic architecture of traits, and makes use of functional annotations in addition to the common genotype and phenotype data, aiming to improve accuracy of the genomic prediction model. In this report, we describe a validation of the different stages of the developed models, and benchmark those against commonly used genomic prediction models. We used a publicly available mouse data set for benchmarking. This data was particularly suitable for benchmarking given the relatively low level of relatedness in this population compared to livestock, and because of the further advanced annotation maps available for mice.

- **Method:** Commonly used genomic prediction models are linear models that only model additive gene action. Here, as a first step, we compare the performance of the ML method Gradient Boosting Machine (GBM) to linear genomic prediction models, since GBM due to its flexibility is expected to be able to fit non-additive gene action. The second step involved including individual level gene expression as explanatory variable in the model next to SNP genotypes. Since gene expression mediates the impact of the genome on the phenotype, the expectation was that including this information would increase the explained phenotypic variance as well as the model's accuracy to predict phenotypes. The third step involved exploring the benefit of using functional annotation information in the model to up- or down-weight contributions of SNPs to the phenotypic variance. Two types of functional annotation were considered, being either variance in observed expression of different genes, or GERP scores that are a measure of the conservation of loci across evolution, and thereby provide a measure of the potential physiological relevance of loci.

- **Main Results:**

  Our results show that the machine learning model GBM can outperform linear models to predict complex traits in an outbred mice population for traits known to be affected by epistatic effects. Conversely, linear models performed similarly to or better than GBM for more polygenic traits like body weight.

  For most traits, individual gene transcript levels explained more phenotypic variance, and yielded greater prediction accuracy, than SNP genotypes. Adding individual gene transcripts in the model next to SNP genotypes considerably increased the proportion of phenotypic variance explained for all traits, and slightly increased the accuracy of predicting phenotypes for most traits. Adding individual gene transcripts in the model may also improve the accuracy of estimated breeding values, provided that the gene transcripts are conditioned on the SNP genotypes, in which case the gene transcripts do not explain phenotypic variance that is also associated with variation in SNP genotypes.

- Weighting of SNPs based on functional information, being either variation in individual levels of gene transcripts or GERP scores, generally did not affect the estimated heritability. Increases in estimated heritability were only observed when weighting SNPs with positive GERP scores, or when SNPs enriched for functionality were included as a separate model component, next to a model component including all remaining SNPs. Using SNPs located within genes yielded on average a somewhat higher accuracy than using random SNPs. Weighing SNPs by function annotation did

not consistently improve prediction accuracy, possibly due to the medium SNP density used. The observed limited impact of using functional annotation on the accuracy of predicting phenotypes, may in part be due to the density of the SNPs used. In the large scale validations on commercial data (Tasks 4.5 and 4.6), imputed sequence data will be available to overcome this limitation.

- **Teams involved:**

Wageningen University

Hendrix Genetics Research Technology & Services B.V.

The Jackson Laboratory, Bar Harbor, Maine, United States of America

2 Introduction

2.1 Improving genomic prediction accuracy

Genomic prediction is one of the key techniques used in modern livestock breeding programs. Most applications rely on using genotypes of 50k SNPs more or less evenly spread across the genome that are used in a so-called GBLUP model to calculate genomic relationships between animals. In the computations of these relationships, it is assumed that each SNP explains the same amount of genetic variance (VanRaden, 2008). Since the introduction of genomic prediction (Meuwissen et al., 2001), several methods have been proposed that differentiate variances across SNPs (Gianola et al., 2009), which actually estimate SNP-specific variance as an integral part of the genomic prediction model from the phenotypic data. In some cases, these models have outperformed GBLUP kind of models, but these improvements are not consistent across studies, and are generally small (de los Campos et al., 2013).

In the last decade, in practical implementations of genomic prediction, the most common strategy used to increase the accuracy of genomic prediction is increasing the size of the reference population. Although largely successful, eventually there is a limit to further improvements. This is, firstly, because at some stage there simply are not more animals available to be added to the reference population, or because of test capacity or budget constraints to do so. Secondly, the improvement in accuracy due to increasing the size of the reference population is subject to diminishing returns, such as can be seen from e.g. the equation to predict accuracy of genomic prediction developed by Daetwyler et al (2008).

The lack of consistent improvement when actually aiming to model the genetic architecture of traits more closely, suggests that in many cases the models are not sufficiently able to dissect effects of individual loci from effects of other loci, the environment, and all sorts of other possibly confounding factors contribution to variation in phenotypes. To overcome this limitation, it has been suggested to gear prediction models more based on known biologically functional detailed information at the genome level (Fang et al., 2017; Ramstein et al., 2020), in addition to the genotype and phenotype data of a reference population that is typically used in genomic prediction. Another suggested direction to improve genomic prediction models, is to properly account for non-additive gene action (Duenk et al., 2021). In particular, Machine Learning (ML) models have been proposed to use for this (Azodi et al., 2019; Montesinos-López et al., 2021), due to their flexible nature.

2.2 Objectives

We aimed to stepwise expand genomic prediction models, by adding additional sources of functional information, and by considering more flexible models that are able to consider any non-additive gene action. As such, we addressed the following more detailed objectives:

1. Compare the performance of the ML method Gradient Boosting Machine to linear genomic prediction models including GBLUP, Elastic Net and BayesB.
2. Investigate the contribution of using individual level gene expression to the dissection of phenotypic variance and the accuracy of predicting phenotypes.
3. Investigate the added value in terms of increased accuracy of predicting phenotypes, when using functional annotation information in the model to up- or down-weight contributions of SNPs to the phenotypic variance.

Going from 1 to 2, gene expression data is used as a set of additional explanatory variables in the model, next to SNP genotypes. This implies that gene expression data should be available from the same animals for which genotype and phenotype data is available. Going from 1 to 3, the model is expanded by using population-level functional information that is used to derive the relative importance of different SNPs towards the prediction of phenotypes. Two types of functional information were considered here. Firstly, the variance of gene expression across all animals in the population is used as an indicator for variability of gene action for specific genomic regions. Secondly, so-called GERP (Genomic Evolutionary Rate Profiling) scores (Cooper et al., 2005), which are a measure of conservation of loci across evolution, are used to derive weights for the SNPs used. All these analyses were undertaken using a publicly available dataset of Diversity Outbred mice, which is described in more detail in section 2.3.5. We used here publicly available data instead of simulations as being used in Task 4.1, to enable empirical validation of the methods. Results for the first objective have been presented at the 72th Annual Meeting of the European Federation of Animal Science (Annex 1), and are described in a paper that is currently under revision at the journal *G3:Genes|Genomes|Genetics* (Annex 2). A paper about the second objective is in preparation.

2.3 Methods

2.3.1 Gradient Boosting Machine

Gradient boosting machine (GBM) is an ensemble learning technique that applies an iterative process of assembling “weak learners” into a stronger learner, being largely used for both classification and regression problems (Friedman, 2002). It relies on fitting decision trees as the base learner (Hastie et al., 2009a). The first tree is fitted on the errors of an initialized prediction based on the distribution of the response variable and from this point, the algorithm fits sequential trees, in which every subsequent tree aims to minimize the prediction error from the previous one until no further improvement can be achieved. Many different parameters can be used to measure that “improvement”, in the present study we used the relative root-mean-squared-error (RRMSE). GBM does automatic feature selection, prioritizing important variables and discarding ones containing irrelevant or redundant information. We implemented the GBM model using the “h2o.ai” R-package (Click et al., 2016).

The performance of machine learning methods can be sensitive to hyper-parameters (Azodi et al., 2019). To obtain the best possible results from the GBM algorithm, a grid search approach was used to determine the combination of hyperparameters that maximized prediction performance for each trait. Hyperparameters (and range of values) included were number of trees (ntree = 100, 150, 200, 300, 500, 1000, 2000 and 5000), learning rate (lrn_rate = 0.01; 0.05 and 0.10) and maximum tree depth (max_depth = 2, 3, 5 and 10). For each trait analyzed, the hyperparameter tuning scheme was performed inside the reference subset (Figure 1). The best set of hyperparameters was chosen based on the lowest mean squared error obtained from the grid-search. Results reported in the present study for GBM model refer to the best performing model out of the grid search for each trait.

For GBM, the importance of a feature is determined by assessing whether that feature was selected to split on during the tree building process, and the contribution of that to decrease the squared error (averaged over all trees) as a result (Friedman and Meulman, 2003; Hastie et al., 2009b). The feature importance is expressed in a percentage scale that can be ranked to assess the magnitude of importance of each feature.

Considering GBM using all available SNPs as a base model, we additionally investigated if the feature importance performed by the GBM model can be used to improve performance by fitting only extracted relevant SNPs in GBM or any of the other considered models. We considered the top 100, 250, 500 and 1000 SNPs from the base GBM model as input for GBM or the other models. The important features were obtained using the same strategy described for the hyperparameter tuning previously explained, using a random split (80-20) within the reference subset (Figure 1).

Finally, instead of, or in addition to the SNP genotype data, individual gene expression was used as explanatory variables in the prediction model (Table 1).

Table 1. Overview of models applied to SNP genotypes and/or individual levels of gene transcripts.

Model acronym		Explanatory variables		
GBLUP	GBM	SNP genotypes	Gene transcripts	Interaction explicitly modelled
GBLUP	SNP-GBM	Yes	No	No
TBLUP	TGBM	No	Yes	No
GTBLUP	GTGBM	Yes	Yes	No
GTcBLUP		Yes	Yes	No
GTIBLUP		Yes	Yes	Yes

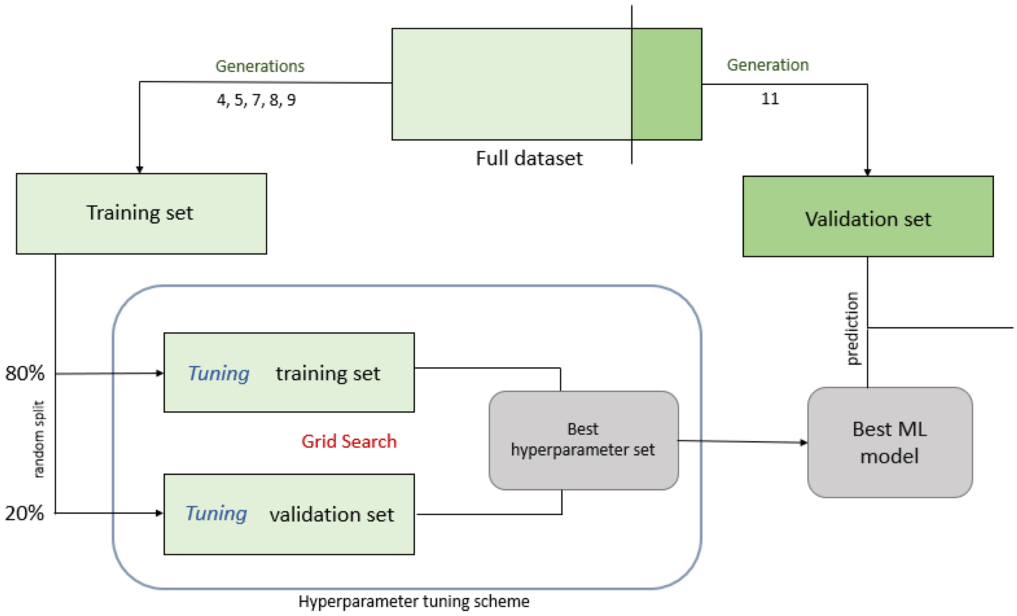

Figure 1. Graphical representation of the hyperparameter tuning grid-search scheme implemented to obtain the best GBM and ENET models

2.3.2 Linear models

For comparison, the analyses based on SNP genotypes only aiming to predict complex phenotypes were also performed using the well-known linear models GBLUP (VanRaden, 2008), BayesB (Meuwissen et al., 2001), and the Elastic Net (ENET) . Note that GBLUP assumes that all SNPs contribute equally to the genetic variance, BayesB performs fairly stringent variable selection in the model by setting per iteration effects of the majority of SNPs to zero, while ENET has two tuning parameters that makes it a mixture of ridge regression (which is equivalent to GBLUP) and the Lasso (Zou and Hastie, 2005), allowing it to be very similar to GBLUP, or the Lasso, or anything in-between (de los Campos et al., 2013). Tuning of these parameters was done using the same procedure as for GBM (as explained in the previous section). GBLUP, any other BLUP models described hereafter, and BayesB were all implemented using the “BGLR” R-package (Pérez and de los Campos, 2014). ENET was implemented using the “h2o.ai” R-package (Click et al., 2016).

2.3.3 Including gene transcripts as explanatory variable in GBLUP and GBM

Both GBLUP and GBM were applied to SNP genotypes, gene transcripts, or both as explanatory variables (Table 1). For the non-parametric GBM, this effectively means that for each individual the vector of individual level gene transcripts is appended to the vector of individual SNP genotypes. This extended vector is then used as explanatory variables, and the GBM model considers all of those, as well as any interactions between them.

For parametric GBLUP, the inclusion of additional explanatory variables requires some choices in terms of parametrization when implementing the model. In total five different BLUP models were used, each including one or two relationship matrices:

GBLUP: $\mathbf{y}^* = \mathbf{1}\mu + \mathbf{g} + \mathbf{e}$

TBLUP: $\mathbf{y}^* = \mathbf{1}\mu + \mathbf{t} + \mathbf{e}$

GTBLUP: $\mathbf{y}^* = \mathbf{1}\mu + \mathbf{g} + \mathbf{t} + \mathbf{e}$

GTIBLUP: $\mathbf{y}^* = \mathbf{1}\mu + \mathbf{g} + \mathbf{t} + \mathbf{g} \times \mathbf{t} + \mathbf{e}$

GTcBLUP: $\mathbf{y}^* = \mathbf{1}\mu + \mathbf{g} + \mathbf{t}_c + \mathbf{e}$

Where $\mathbf{y}^*$ is a vector of pre-corrected phenotypes, $\mathbf{1}$ is a vector of ones, μ is a population mean, and $\mathbf{e}$ is a vector of residuals. The elements associated with $\mathbf{g}$ and $\mathbf{t}$ in each of the models, as well as how each associated relationship matrix is computed, are described in Table 2. The rationale behind the models is that results of GBLUP and TBLUP will show the phenotypic variance and accuracy of phenotypic prediction associated with only SNP genotypes or only gene transcripts. GTBLUP, GTIBLUP and GTcBLUP will show the same quantities, when both are modelled together. The difference is that in GTBLUP and GTIBLUP, any phenotypic variance associated with both variation in SNP genotypes and gene transcripts, can be assigned to either of those, and this will be done on whichever outcome will maximize the likelihood of the model, given the data. With GTcBLUP, effectively any covariance between the gene transcripts and the SNP genotypes, is removed from the gene transcripts. Therefore, this model will assign any phenotypic variance associated with both variation in SNP genotypes and gene transcripts only to the SNP genotypes. Finally, GTIBLUP also explicitly models the interaction between SNP genotypes and gene transcripts.

Table 2. Description of the components of the GBLUP, TBLUP, GTBLUP and GTcBLUP models.

Model component (description)	**Associated relationship matrix**	**Description incidence matrix**
$\mathbf{g}$ (Genomic breeding values)	$\mathbf{G} = \mathbf{XX}'/c$	$\mathbf{X}$ is a matrix containing centred SNP genotypes. $c = \sum_{j=1}^{m} 2p_j(1-p_j)$ and m is the number of SNPs (i.e. the number of columns of $\mathbf{X}$).
$\mathbf{t}$ (Transcriptomic values)	$\mathbf{T} = \mathbf{UU}'/n$	$\mathbf{U}$ is a matrix containing centred and scaled individual gene transcripts. n is the number of genes (i.e. the number of columns of $\mathbf{U}$).
$\mathbf{g} \times \mathbf{t}$ (Interaction between genomic breeding values and gene transcripts)	$\mathbf{G\#T}$	$\mathbf{G\#T}$ is the Hadamard (i.e. the element-wise) product of the matrices $\mathbf{G}$ and $\mathbf{T}$.
$\mathbf{t}_c$ (Transcriptomic values conditioned on $\mathbf{g}$)	$\mathbf{T}_c = \mathbf{VV}'/n$	$\mathbf{V}$ is a matrix containing individual gene transcripts conditioned on SNP genotypes, computed as: $\mathbf{V} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X} + \mathbf{I}\lambda)^{-1}\mathbf{X}')\mathbf{U}.$, where $\mathbf{X}(\mathbf{X}'\mathbf{X} + \mathbf{I}\lambda)^{-1}\mathbf{X}'$ is the so-called "smoother matrix" (Hastie et al., 2009a), $\mathbf{I}$ is an identity matrix, and $\lambda = \frac{m * \sigma_e^2}{\sigma_a^2}$, σ_e^2 is the residual variance, and σ_a^2 is the additive genetic variance. Both variances are estimated with the regular GBLUP model (including only $\mathbf{g}$).

2.3.4 Incorporating functional annotation

The typical computation of a genomic relationship matrix **G**, as described in Table 2, implies that all SNPs have an equal weight of 1, since: $\mathbf{XX}'/c = \mathbf{XIX}'/c$, where $\mathbf{I}$ is an identity matrix (values of 1 on the diagonal, and 0 otherwise). Thus, any weights of SNPs in the **G** matrix can be applied using (VanRaden, 2008): $\mathbf{G}_w = \mathbf{XDX}'/c$, where $\mathbf{D}$ is a diagonal matrix, with weights for each SNP on the diagonal. These weights should represent the variance associated with a particular SNP, and have an average value (across all SNPs) of 1. So, a two times higher weight, means that a SNP explains twice as much variance.

Given that we can write $\mathbf{G}_w = \mathbf{XDX}'/c = \mathbf{XSS}'\mathbf{X}'/c$, where $\mathbf{S}$ is a diagonal matrix for which diagonal elements are the square root of the corresponding elements in $\mathbf{D}$ (i.e. $S_{ii} = \sqrt{D_{ii}}$), we could also first absorb the squares of the weights in the genotype matrix, using e.g.: $\mathbf{W} = \mathbf{XS}$, and then compute $\mathbf{G}_w$ as: $\mathbf{G}_w = \mathbf{WW}'/c$. Equivalent models using regression on SNP genotypes, rather than genomic relationship matrices, simply use $\mathbf{W}$ instead of $\mathbf{X}$.

2.3.5 Data

The DO mice dataset comprising 835 animals was obtained from The Jackson Laboratory (Bar Harbor, ME, US). The animals originated from 6 non-overlapping generations (4, 5, 7, 8, 9 and 11) in which males and females were represented equally. The total number of animals per generation was 97, 48, 200, 184, 99 and 197 for generations 4, 5, 7, 8, 9, and 11, respectively, but numbers of missing records varied across traits. For in total 477 mouse from generations 4, 5, 7 and 11, gene transcript levels data was available for 11,770 genes (Tyler et al., 2017). The proportion of males and females within each diet category was close to 50-50 for all generations. The same was observed for the frequency of males and females within each litter-generation combination (two litters per generation). A detailed description of husbandry and phenotyping methods can be found in Svenson et al. (2012).

Table 3. Description of the traits considered.

Acronym	Description	Age of measurement (weeks)
BMD1	Bone mineral density	12
BMD2	Bone mineral density	21
BW10	Body weight	10
BW15	Body weight	15
BW20	Body weight	20
CHOL1	Circulating cholesterol	8
CHOL2	Circulating cholesterol	19
FATP1	Adjusted body fat percentage	12
FATP2	Adjusted body fat percentage	21
GLUC1	Circulating glucose	8
GLUC2	Circulating glucose	19
TRGL1	Circulating triglycerides	8
TRGL2	Circulating triglycerides	19
INSUL	Circulating insulin	8
UCRT	Urine creatinine	20

Among all phenotypes available we chose 15 traits based on their distinct assumed genetic architectures from previous results with the same dataset (Churchill et al., 2012; Zhang et al., 2012; Tyler et al., 2016; Tyler et al., 2017; Keller et al., 2019; Keenan et al., 2021). The traits considered (see Table 3 for an overview) were bone mineral density at 12 (BMD1) and 21 weeks (BMD2), body weight at 10, 15 and 20 weeks (BW10, BW15 and BW20); circulating cholesterol at 8 (CHOL1) and 19 weeks

(CHOL2), adjusted body fat percentage at 12 (FATP1) and 21 weeks (FATP2), circulating glucose at 8 (GLUC1) and 19 weeks (GLUC2), circulating triglycerides at 8 (TRGL1) and 19 weeks (TRGL2), circulating insulin at 8 weeks (INSUL) and urine creatinine at 20 weeks (UCRT). These traits can be categorized into measurements of body composition (weights and fat percentage), clinical plasma chemistries (triglycerides, glucose, insulin) and urine chemistry (urine creatinine). Across traits, the number of animals with phenotypes ranged from 799 to 834, of which roughly half were male and half were female. The heritabilities of these traits ranged from 0.12 to 0.44. All phenotypes were pre-corrected for fixed effects of diet, generation, litter and sex.

Mice from 8 distinct founder strains were genotyped using either the MUGA or MegaMUGA SNP arrays (Morgan et al., 2016). The variant calls from the arrays in the animals contained in the current dataset were converted to founder haplotypes using a hidden Markov model (Gatti et al., 2014), which uses the order of SNPs in an individual mouse to infer transition points between different DO founder haplotypes. After that, the probability of each parental haplotype at each SNP position in the genome (Gatti et al., 2014) was used to derive SNP genotype probabilities. The complete genotype file used for the analyses was composed of ~64,000 markers reconstructed from the diplotype probabilities from the MUGA and MegaMUGA on an evenly spaced grid, and the average distance between markers was 0.0238 cM. The full genotype data (64K markers) was cleaned based on the following criteria: variants with minor allele frequency < 0.05, call rates < 0.90 and linear correlation between subsequent SNPs > 0.98 were removed. After quality control, a total of 52,840 SNP markers were available for the mice with both phenotypic and genotypic records, as well as for the subset with available gene transcript data.

2.3.6 Considered functional annotations

The first source of functional annotation that we used, is the variance of rank-z transformed (https://rdrr.io/bioc/DOQTL/man/rankZ.html) gene expression across all animals in the population that is used as an indicator for variability of gene action for specific genomic regions. The distribution of those variances is shown in Figure 2. Variability in gene expression were assigned to 11,278 SNPs that were actually located within those genes. These variances, after being scaled to have a mean value of 1, were included in the matrix $\mathbf{D}$ and then used as a weight in GBLUP, or integrated in the weighted genotype matrix $\mathbf{W}$ (as described in section 2.3.4) to be used in GBM. For comparison, unweighted analyses were performed using the same 11,278 SNPs, or using all SNPs as a benchmark.

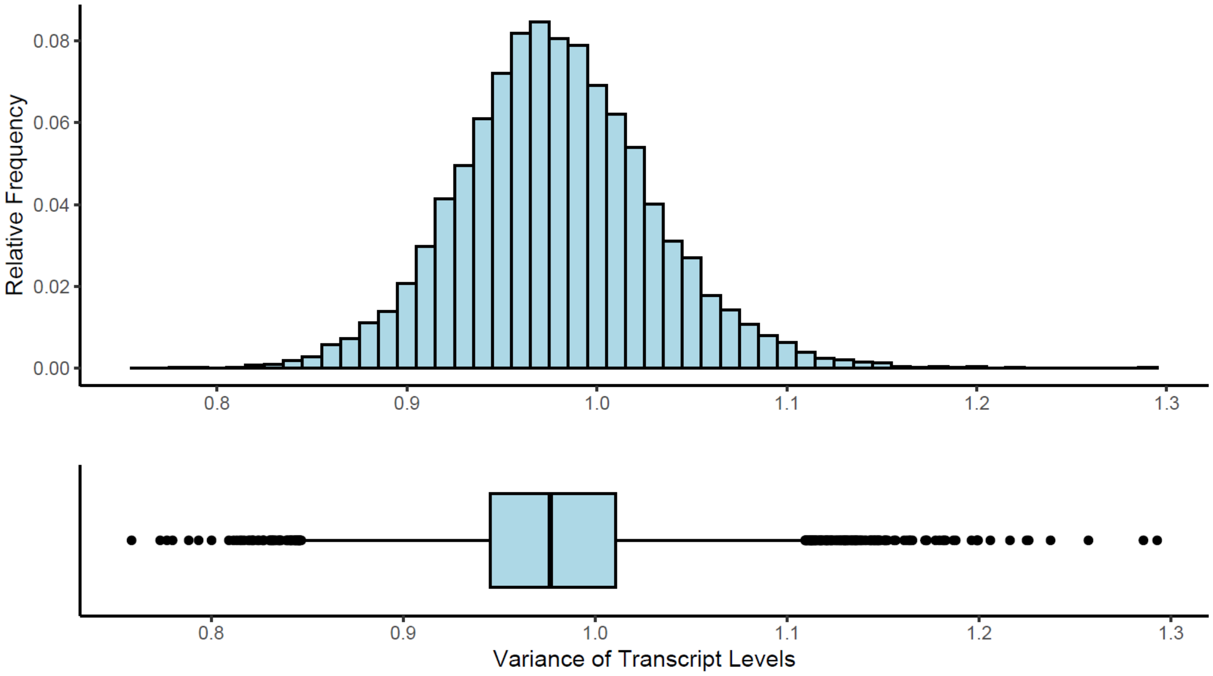

Figure 2. Distribution of variances in transcript levels in the mouse data, depicted as histogram and boxplot in upper and lower panels respectively.

The second source of functional annotation that we used, are GERP scores (Cooper et al., 2005), which are a measure of conservation of loci across evolution. The GERP scores were retrieved via http://ftp.ensembl.org/pub/release-102/compara/conservation_scores/111_mammals.gerp_conservation_score/gerp_conservation_scores.mus_musculus.GRCm38.bw, and were available for in total 53,332 of the 60,883 SNPs in the original map file. In Figure 3, the distribution of the GERP scores is shown. In total, of all 52,840 SNPs that survived the earlier quality control, 14,231 SNPs were associated with a positive GERP score, and 33,068 SNPs were associated with a negative GERP score. For the analyses, we either used only the 14,231 SNPs associated with a positive GERP score, only the 33,068 SNPs associated with a negative GERP score, or both. For comparison, different analyses were run with and without weighting based on GERP scores. See Table 4 for a full overview of all models considered. The weights were obtained by first taking the absolute values of the GERP scores, and then centring them to a value of 1.

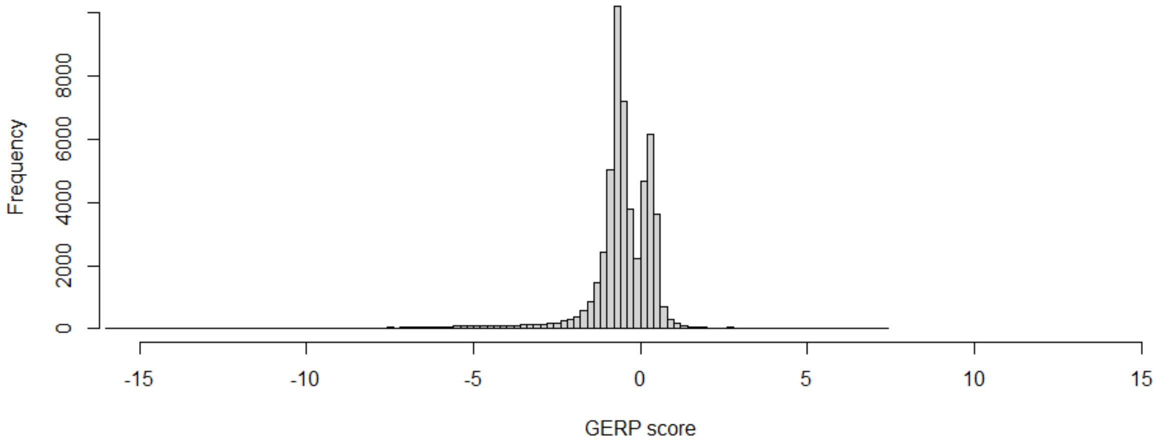

Figure 3. Distribution of GERP scores for 53,332 SNPs in the mouse data.

Table 4. Description of the different models considered to evaluate the impact of using GERP scores

Abbreviation[1]	**SNPs included**[2]	**Nr of G matrices**	**Weights used for:**
GBLUP_pos_unwt	pos	1	-
GBLUP_neg_unwt	neg	1	-
GBLUP_pos_wt	pos	1	pos
GBLUP_neg_wt	neg	1	neg
GBLUP_pos+neg_unwt	pos + neg	1	-
GBLUP_pos+neg_wt	pos + neg	1	pos + neg
GBLUP_pos+remain_unwt	pos + remain	2	-
GBLUP_neg+remain_unwt	neg + remain	2	-
GBLUP_pos+remain_wt	pos + remain	2	pos
GBLUP_neg+remain_wt	neg + remain	2	neg
GBLUP_rnd_Npos.14k	rnd (14k)	1	-
GBLUP_rnd_Nneg.33k	rnd (33k)	1	-
GBLUP_all_SNP	all 50k	1	-

[1] unwt: all included SNPs receive equal weights; wt: sets of SNPs listed in the rightmost column of the Table are weighted by their absolute GERP scores.

[2] pos: 14,231 SNPs with a positive GERP score; neg: 33,068 SNPs with a negative GERP score; remain: all other remaining SNP from the entire 50k panel; rnd: random subset of 14,231 (14k) or 33,068 (33k); all 50k: the entire panel.

2.3.7 Evaluation of model performance

Performance of predictions from the models was measured by the Pearson correlation and the relative root mean squared error of prediction (RRMSE) between predicted and pre-corrected phenotypes. In all analyses, we used a forward prediction validation scheme in which animals from older generations were used as the reference and animals from the younger generation as the validation subset. For the analyses using the full SNP genotype data, the reference included generations 4, 5, 7, 8 and 9, and generation 11 formed the validation subset. For the analyses using the full SNP genotype data, the reference included generations 4, 5, and 7, and generation 11 formed the validation subset.

3 Results

3.1 Genomic prediction based on SNP genotypes

The GBM model showed the highest prediction accuracy for BMD1, CHOL2 and GLUC2 (Figure 4), and the lowest RRMSE for the same traits (Table 5). For other traits, prediction accuracy from GBM varied from being competitive to the linear models for BW10, BW15 and TRGL2, to a poorer performance observed for UCRT. It only showed the worst predictive ability among all models for FATP1, but with a small difference from the next performing model (- 1.76% absolute difference). Interestingly, for the three traits of which GBM showed greater prediction accuracy than the linear models (BMD1, CHOL2 and GLUC2) there is strong evidence in literature of a relevant portion of phenotypic variance being explained by epistatic effects (e.g. Tyler et al., 2016; Tyler et al., 2017).

A more extensive report of the comparison of the different models is provided in Annex 2, which is a full manuscript currently under revision at the journal *G3: Genes, Genomes, Genetics*. This includes deeper analyses of the results presented here, to further understand and clarify the similarities and differences between the models, as well as genomic prediction results when there is a greater distance between the training and the validation data, or when a subset of informative SNPs is used that is pre-selected using the GBM model.

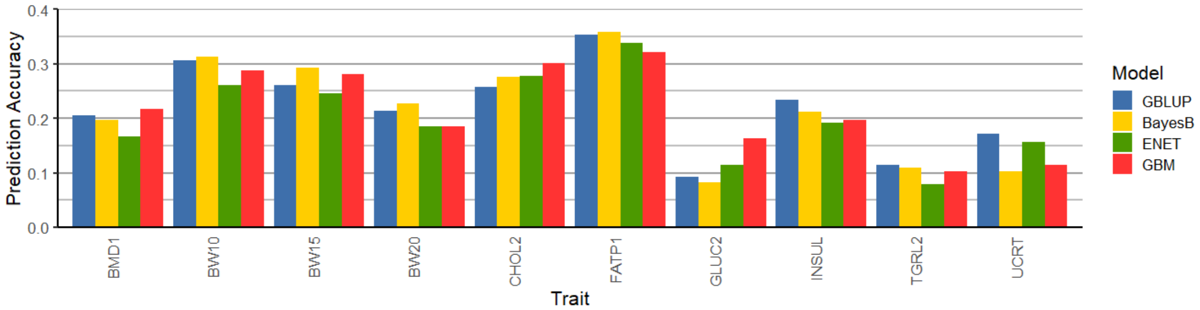

Figure 4. Prediction accuracy obtained from GBLUP, BayesB, elastic net (ENET) and gradient boosting machine (GBM). For a description of the traits, see Table 3.

Table 5. Relative root-mean-square error obtained from GBLUP, BayesB, ENET and GBM for 10 phenotypes analyzed in the diversity outbred mouse population. For each trait, the lowest value is indicated in bold.

Trait[1]	GBLUP	BayesB	ENET	GBM
BMD1	0.94	0.97	0.95	**0.93**
BW10	**0.72**	0.81	0.76	0.75
BW15	**0.71**	0.76	0.73	0.72
BW20	**0.71**	0.75	0.72	0.74
CHOL2	0.80	0.94	0.86	**0.78**
FATP1	**0.71**	0.74	0.72	0.72
GLUC2	0.94	0.99	0.95	**0.91**
TRGL2	**1.02**	1.18	1.09	1.06
INSUL	**0.83**	0.88	0.86	0.84
UCRT	**0.88**	0.95	0.91	0.90

[1] For a description of the traits, see Table 3.

3.2 Partitioning of phenotypic variance based on SNP genotypes and gene transcripts

In Table 6, the percentages of the variance explained by g (the SNP genotypes) and t (the transcript levels) are presented for the different models applied. The expectation was that g would explain most variance when included alone in the model (GBLUP), but actually for most traits most variance was explained by g if t_c (the gene transcripts conditioned on the SNP genotypes) was included (GTcBLUP). The component t explained most variance when included alone in the model, as expected. When g and t were included together in the model (GTBLUP), the variance explained by both g and t reduced compared to the counterpart models that only included either of those (GBLUP and TBLUP)

Table 6. Percentage of the phenotypic variance explained by SNP genotypes (g), transcript levels (t), the interaction between SNP genotypes and transcript levels (gt), and the residual (e), using a GBLUP (G), TBLUP (T), GTBLUP (GT), GTIBLUP (GTI), or GTcBLUP model[1]. For each model component, the highest percentage (across models) is indicated in bold.

	Model component													
	g (SNPs)				t (gene transcripts)				gt	e (residual)				
Trait	G	GT	GTI	GTc	T	GT	GTI	GTc	GTI	G	T	GT	GTI	GTc
BMD1	**38**	28	25	37	**50**	23	21	2	5	**62**	50	49	49	61
BMD2	40	32	29	**50**	**58**	28	27	7	4	**60**	42	40	40	43
BW10	42	7	1	**44**	**72**	66	60	4	10	**58**	28	27	29	52
BW15	35	3	0	**45**	**77**	74	46	22	21	**65**	23	23	33	33
BW20	37	1	0	**54**	**81**	80	52	27	15	**63**	19	19	33	19
CHOL1	31	30	28	**36**	**28**	4	3	1	5	**69**	72	66	64	63
CHOL2	**44**	39	24	43	**58**	10	8	1	17	**56**	42	51	51	**56**
FATP1	33	11	10	**44**	**74**	63	60	22	2	**67**	26	26	28	34
FATP2	25	4	4	**40**	**83**	79	78	26	3	**75**	17	17	15	34
GLUC1	**21**	19	17	19	**13**	7	6	1	1	79	87	74	76	**80**
GLUC2	8	3	0	**13**	**23**	21	15	12	6	**92**	77	76	79	75
TRGL1	**28**	19	10	27	**24**	15	1	1	16	**72**	76	66	73	**72**
TRGL2	20	10	9	**23**	**36**	26	25	4	1	**80**	64	64	65	73

[1] For a description of the models, see Table 1.

3.3 Genomic prediction based on SNP genotypes and gene transcripts

When only considering SNP genotypes in the model GBLUP outperformed GBM for 8 out of the 13 traits, but differences were generally small (Table 7). The TGBM model yielded higher accuracies than TBLUP for 9 out of 14 traits, but also here differences were generally small. For most of the traits, with the exception of BMD1, CHOL1, CHOL2, GLUC1, TRGL1, TBLUP outperformed GBLUP, likely because individual gene transcripts explain more phenotypic variances than SNPs do (Table 6). This trend was largely confirmed by the corresponding GBM models, SNP-GBM and TGBM. Modelling both SNPs and gene transcripts with BLUP (i.e. GTBLUP), hardly led to any further improvement of prediction accuracy compared to only modelling gene transcript (TBLUP). A similar trend was observed for GBM, where actually modelling both (i.e. GTGBM) in several cases resulted in somewhat lower prediction accuracies compared to modelling gene transcripts only (TGBM). Finally, the prediction based on the SNP component of GTcBLUP, i.e. the breeding estimated with this model, generally had a similar accuracy as GBLUP, and considerably outperformed GBLUP for BW20, FATP1 and FATP2 (results not shown). For the same three traits, the variance associated with the SNPs was considerably higher for GTcBLUP than for GBLUP. This suggests that conditioning SNP genotypes on gene transcripts in some cases results in model that estimates breeding values with a higher accuracy.

Table 7. Accuracy of predicting phenotypes using different models. For a description of the models, see Table 1. For each group of models, the one with the highest accuracy is indicated in bold.

Trait[1]	GBLUP	SNPGBM	TBLUP	TGBM	GTBLUP	GTIBLUP	GTcBLUP	GTGBM
BMD1	**0.200**	0.189	0.117	**0.139**	**0.203**	0.192	0.199	0.183
BMD2	**0.286**	0.280	**0.376**	0.370	**0.406**	0.404	0.287	0.375
BW10	0.216	**0.230**	**0.482**	0.380	0.482	**0.485**	0.204	0.356
BW15	**0.186**	0.144	0.519	**0.525**	**0.523**	**0.523**	0.217	0.494
BW20	**0.245**	0.224	**0.609**	0.548	**0.611**	0.584	0.299	0.545
CHOL1	0.159	**0.182**	0.131	**0.148**	**0.185**	0.136	0.168	0.180
CHOL2	**0.285**	0.204	0.071	**0.147**	-0.107	-0.111	0.140	**0.197**
FATP1	0.153	**0.214**	0.446	**0.458**	**0.451**	0.439	0.281	0.447
FATP2	**0.234**	0.218	**0.540**	0.489	**0.543**	**0.543**	0.313	0.483
GLUC1	**0.123**	0.119	0.026	**0.040**	0.131	0.086	0.114	**0.146**
GLUC2	0.013	**0.032**	0.032	**0.046**	0.045	**0.070**	-0.052	0.059
TRGL1	0.101	**0.108**	0.059	**0.076**	**0.067**	0.061	0.025	0.060
TRGL2	**0.148**	0.139	0.153	**0.186**	0.164	0.167	0.125	**0.192**

[1] For a description of the traits, see Table 3.

3.1 Genomic prediction using variance in gene transcripts as functional annotation

Using variances in gene transcripts to weight the SNPs in the G matrix ("wtGRM"), yielded the same heritabilities compared to using an unweighted G matrix ("unwtGRM") based on the same SNPs (Figure 6). Both approaches tended to yield a slightly lower heritability than using the same number of SNPs (11,278) randomly selected from the entire panel ("rndGRM"), most likely because the random selection resulted in a more evenly distribution of the selected SNPs across the genome.

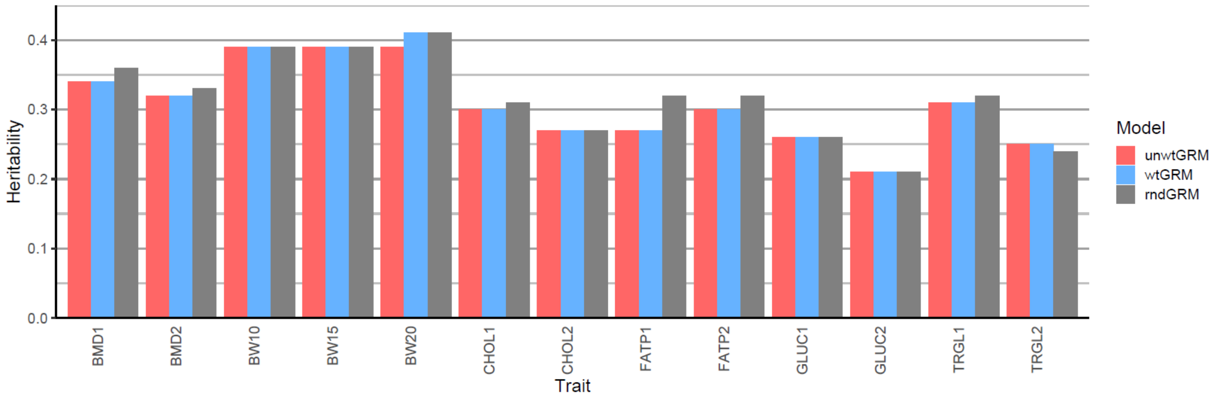

Figure 6. Heritabilities estimated using 11,278 SNP residing in genic regions using a G matrix weighted by variance in gene transcript levels ("wtGRM") or not ("unwtGRM"), or using 11,278 SNPs randomly selected from the entire panel (in this case the result is the average of 10 replicates).

For all traits, prediction accuracies were very similar when the SNPs in genic regions were weighted or not (Figure 7). For 8 out of the 13 traits, using SNPs in genic regions yielded a higher accuracy than using the same number of SNPs being selected random across the genome. These variable differences across traits, are in line with previously reported results on including gene annotation in genomic prediction (Morota et al., 2014; Gao et al., 2017).

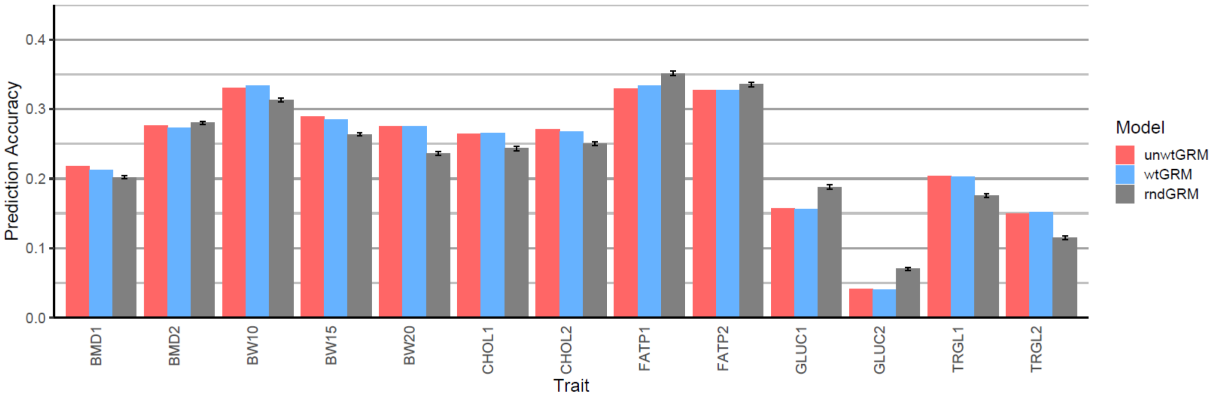

Figure 7. Prediction accuracies obtained when using 11,278 SNP residing in genic regions in a G matrix weighted by variance in gene transcript levels (“wtGRM”) or not (“unwtGRM”), or using 11,278 SNPs randomly selected from the entire panel (in this case the result is the average of 10 replicates). For a description of the traits, see Table 3.

3.2 Genomic prediction using GERP scores as functional annotation

Estimated heritabilities, when not applying any weights, and using either the subsets of SNPs with positive or negative GERP scores, the random subsets, or all SNPs, were very close to each other (Figure 8). Estimated heritabilities increased somewhat if two instead of one G matrices were used in the model. Finally, whenever the SNPs with a positive GERP score were weighted, the estimated heritability considerably increased.

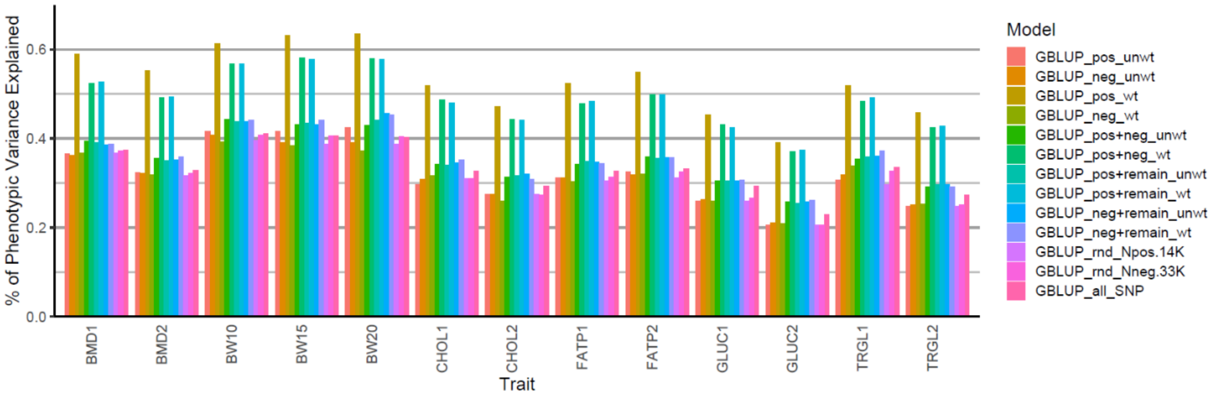

Figure 8. Heritabilities estimated with models using different sets of SNPs, using differential weighing, or not (see Table 4 for a full description of the models used). For a description of the traits, see Table 3.

The observed increases in estimated heritability when either using two G matrices or when weighing the SNPs with a positive GERP score (Figure 8), did not translate into a noticeable increase in prediction accuracy (Figure 9). The prediction accuracies show some fluctuations across models within traits, but there is no general pattern recurring for all traits (Figure 9).

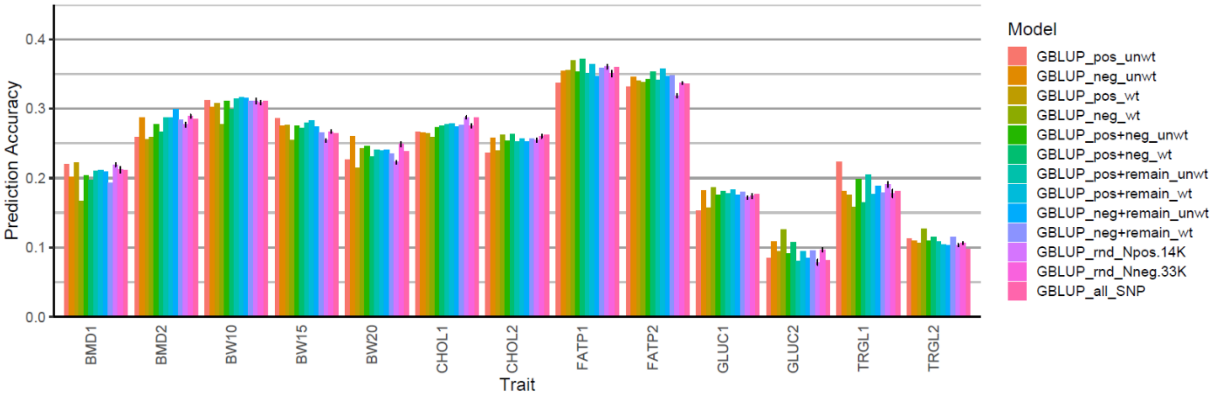

Figure 9. Prediction accuracy estimated with models using different sets of SNPs, using differential weighing, or not (see Table 4 for a full description of the models used). [1]For a description of the traits, see Table 3.

4 Conclusion

Our results show that the machine learning model GBM is a competitive method to predict complex traits in an outbred mice population for traits known to be affected by epistatic effects. For these traits, GBM outperformed BLUP, both when using SNP genotype or gene transcript data. Conversely, linear models performed similarly to or better than GBM for more polygenic traits like body weight.

Adding individual gene transcripts in the model considerably increased the proportion of phenotypic variance explained for all traits, and slightly increased the accuracy of predicting phenotypes for most traits. Using only individual gene transcript levels in the model resulted for most traits in explaining a larger proportion of the phenotypic variance, and in greater prediction accuracy, than using only SNP genotypes. Adding individual gene transcripts in the model may actually improve the accuracy of estimated breeding values, provided that the gene transcripts are conditioned on the SNP genotypes as described here, in which case the gene transcripts do not explain phenotypic variance that is also associated with variation in SNP genotypes.

Weighting of SNPs based on functional information, being either variation in individual levels of gene transcripts or GERP scores, generally did not affect the estimated heritability. Increases in estimated heritability were only observed when weighting SNPs with positive GERP scores, or when SNPs enriched for functionality were included as a separate model component, next to a model component including all remaining SNPs. For most of the traits, using SNPs in genic regions yielded a higher accuracy than using the same number of SNPs being selected random across the genome, but weighing SNPs by variance in gene transcript levels did not affect the accuracy. Using GERP scores to differentiate between groups of SNPs, or to weigh the SNPs, did not lead to a consistent change in prediction accuracy across traits.

The observed limited impact of using functional annotation on the accuracy of predicting phenotypes, may in part be due to the density of the SNPs used. In the large scale validations on commercial data (Tasks 4.5 and 4.6), imputed sequence data will be available to overcome this limitation. Conversely, these larger datasets with higher genotype resolution combined with several annotations maps, may also lead to some computational limitations. In particular, the extensive hyperparameter tuning required to optimize GBM may hamper a full exploitation of putative models and combinations of all annotation maps.

In summary:

- The benefit of using ML for genomic prediction depends on the architecture of the trait.
- Using individual level gene transcript data increases the amount of phenotypic variance explained, the accuracy of predicting phenotypes, and in some cases the accuracy of predicting breeding values, if gene transcript levels are conditioned on SNP genotypes.
- Using SNPs in genic regions yielded on average a somewhat higher accuracy than using random SNPs. Weighing SNPs by function annotation did not consistently improve prediction accuracy, possibly due to the medium SNP density used.

5 Deviations or delays

None.

6 Acknowledgements

We acknowledge interactions with others partners within the GENE-SWitCH consortium, in particular Andrea Rau and Fanny Mollandin from INRAE, about the research done. We acknowledge Gary Churchill from the Jackson laboratory (Bar Harbor, Maine, US) for helping us with answering specific questions regarding the data, the different traits, and their specific genetic architecture. Finally, we acknowledge Martijn Derks from Animal Breeding and Genomics, Wageningen University, for his help in subtracting and interpreting the GERP scores.

7 References

Azodi, C. B., E. Bolger, A. McCarren, M. Roantree, G. de los Campos, and S.-H. Shiu. 2019. Benchmarking Parametric and Machine Learning Models for Genomic Prediction of Complex Traits. G3 (Bethesda). 9:3691-3702. https://doi.org/10.1534/g3.119.400498

Churchill, G. A., D. M. Gatti, S. C. Munger, and K. L. Svenson. 2012. The diversity outbred mouse population. Mamm. Genome. 23:713-718. https://doi.org/10.1007/s00335-012-9414-2

Click, C., M. Malohlava, A. Candel, H. Roark, and V. Parmar. 2016. Gradient Boosted Models with H2O. http://h2o-release.s3.amazonaws.com/h2o/master/3568/docs-website/h2o-docs/booklets/GBMBooklet.pdf. http://h2o-release.s3.amazonaws.com/h2o/master/3568/docs-website/h2o-docs/booklets/GBMBooklet.pdf

Cooper, G. M., E. A. Stone, G. Asimenos, E. D. Green, S. Batzoglou, and A. Sidow. 2005. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 15:901-913. http://www.genome.org/cgi/doi/10.1101/gr.3577405

Daetwyler, H. D., B. Villanueva, and J. A. Woolliams. 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS ONE. 3:e3395. https://doi.org/10.1371/journal.pone.0003395

de los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, and M. P. L. Calus. 2013. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics. 193:327-345. https://doi.org/10.1534/genetics.112.143313

Duenk, P., P. Bijma, Y. C. J. Wientjes, and M. P. L. Calus. 2021. Review: Optimizing genomic selection for crossbred performance by model improvement and data collection. J. Anim. Sci. 99:1–24. https://doi.org/10.1093/jas/skab205

Fang, L., G. Sahana, P. Ma, G. Su, Y. Yu, S. Zhang, M. S. Lund, and P. Sørensen. 2017. Use of biological priors enhances understanding of genetic architecture and genomic prediction of complex traits within and between dairy cattle breeds. 18:604. https://doi.org/10.1186/s12864-017-4004-z

Friedman, J. H. 2002. Stochastic gradient boosting. Comp. Stat. Data Anal. 38:367-378. https://doi.org/10.1016/S0167-9473(01)00065-2

Friedman, J. H. and J. J. Meulman. 2003. Multiple additive regression trees with application in epidemiology. Stat. Med. 22:1365-1381. https://doi.org/10.1002/sim.1501

Gao, N., J. W. R. Martini, Z. Zhang, X. Yuan, H. Zhang, H. Simianer, and J. Li. 2017. Incorporating gene annotation into genomic prediction of complex phenotypes. Genetics. 207:489-501. https://doi.org/10.1534/genetics.117.300198

Gatti, D. M., K. L. Svenson, A. Shabalin, L.-Y. Wu, W. Valdar, P. Simecek, N. Goodwin, R. Cheng, D. Pomp, A. Palmer, E. J. Chesler, K. W. Broman, and G. A. Churchill. 2014. Quantitative trait locus mapping methods for diversity outbred mice. G3 (Bethesda). 4:1623-1633. https://doi.org/10.1534/g3.114.013748

Gianola, D., G. de los Campos, W. G. Hill, E. Manfredi, and R. Fernando. 2009. Additive genetic variability and the Bayesian alphabet. 183:347-363. https://doi.org/10.1534/genetics.109.103952

Hastie, T., R. Tibshirani, and J. Friedman. 2009a. The Elements of Statistical Learning. Springer, New York.

Hastie, T., R. Tibshirani, and J. Friedman. 2009b. Support vector machines and flexible discriminants. *in* The Elements of Statistical Learning. Springer, New York.

Keenan, B. T., J. C. Webster, A. S. Wiemken, N. Lavi-Romer, T. Nguyen, K. L. Svenson, R. J. Galante, G. A. Churchill, S. Pickup, A. I. Pack, and R. J. Schwab. 2021. Heritability of fat distributions in male mice from the founder strains of the Diversity Outbred mouse population. G3 (Bethesda). 11. https://doi.org/10.1093/g3journal/jkab079

Keller, M. P., M. E. Rabaglia, K. L. Schueler, D. S. Stapleton, D. M. Gatti, M. Vincent, K. A. Mitok, Z. Wang, T. Ishimura, S. P. Simonett, C. H. Emfinger, R. Das, T. Beck, C. Kendziorski, K. W. Broman, B. S. Yandell, G. A. Churchill, and A. D. Attie. 2019. Gene loci associated with insulin secretion in islets from nondiabetic mice. J. Clin. Investig. 129:4419-4432. https://doi.org/10.1172/JCI129143

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 157:1819-1829. https://doi.org/10.1093/genetics/157.4.1819

Montesinos-López, O. A., A. Montesinos-López, P. Pérez-Rodríguez, J. A. Barrón-López, J. W. R. Martini, S. B. Fajardo-Flores, L. S. Gaytan-Lugo, P. C. Santana-Mancilla, and J. Crossa. 2021. A review of deep learning applications for genomic selection. BMC Genomics. 22:19. https://doi.org/10.1186/s12864-020-07319-x

Morgan, A. P., et al. 2016. The mouse universal genotyping array: from substrains to subspecies. G3 (Bethesda). 6:263-279. https://doi.org/10.1534/g3.115.022087

Morota, G., R. Abdollahi-Arpanahi, A. Kranis, and D. Gianola. 2014. Genome-enabled prediction of quantitative traits in chickens using genomic annotation. BMC Genom. 15:109. https://doi.org/10.1186/1471-2164-15-109

Pérez, P. and G. de los Campos. 2014. Genome-wide regression and prediction with the BGLR statistical ppsoackage. Genetics. 198:483-495. https://doi.org/10.1534/genetics.114.164442

Ramstein, G. P., S. J. Larsson, J. P. Cook, J. W. Edwards, E. S. Ersoz, S. Flint-Garcia, C. A. Gardner, J. B. Holland, A. J. Lorenz, M. D. McMullen, M. J. Millard, T. R. Rocheford, M. R. Tuinstra, P. J. Bradbury, E. S. Buckler, and M. C. Romay. 2020. Dominance effects and functional enrichments improve prediction of agronomic traits in hybrid maize. Genetics. 215:215-230. https://doi.org/10.1534/genetics.120.303025

Svenson, K. L., D. M. Gatti, W. Valdar, C. E. Welsh, R. Cheng, E. J. Chesler, A. A. Palmer, L. McMillan, and G. A. Churchill. 2012. High-resolution genetic mapping using the mouse diversity outbred population. Genetics. 190:437-447. https://doi.org/10.1534/genetics.111.132597

Tyler, A. L., L. R. Donahue, G. A. Churchill, and G. W. Carter. 2016. Weak epistasis generally stabilizes phenotypes in a mouse intercross. PLOS Genet. 12:e1005805. https://doi.org/10.1371/journal.pgen.1005805

Tyler, A. L., B. Ji, D. M. Gatti, S. C. Munger, G. A. Churchill, K. L. Svenson, and G. W. Carter. 2017. Epistatic networks jointly influence phenotypes related to metabolic disease and gene expression in diversity outbred mice. Genetics. 206:621-639. https://doi-org.ezproxy.library.wur.nl/10.1534/genetics.116.198051

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414-4423. https://doi.org/10.3168/jds.2007-0980

Zhang, W., R. Korstanje, J. Thaisz, F. Staedtler, N. Harttman, L. Xu, M. Feng, L. Yanas, H. Yang, W. Valdar, G. A. Churchill, and K. DiPetrillo. 2012. Genome-wide association mapping of quantitative traits in outbred mice. G3 (Bethesda). 2:167-174. https://doi.org/10.1534/g3.111.001792

Zou, H. and T. Hastie. 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. B-Stat. Meth. 67:301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

8 Annexes

Annex 1:

Perez, B. C., M. C. A. M. Bink, G. A. Churchill, and M. P. L. Calus. 2021. Gradient boosting is a competitive method for genomic prediction of complex traits in outbred mice. Page 456 in Book of Abstracts of the 72nd Annual Meeting of the European Federation of Animal Science (Book of abstracts No. 27 (2021)), Davos, Switzerland.

Annex 2:

Perez, B.C, M.C.A.M. Bink, K.L. Svenson, G.A. Churchill, M. P. L. Calus. Prediction performance of linear models and gradient boosting machine on complex phenotypes in outbred mice. *Under review at G3:Genes|Genomes|Genetics.*

9 Glossary

Model terms:

BLUP: Best Linear Unbiased Prediction

ENET: Elastic Net

GBLUP: BLUP based on SNP genotypes

GBM: Gradient Boosting Machine

GRM: Genomic Relationship Matrix

TBLUP: BLUP based on gene transcripts

GTBLUP: BLUP based on SNP genotypes and gene transcripts

GTcBLUP: BLUP based on SNP genotypes and gene transcripts conditioned on SNP genotypes

GERP: Genomic Evolutionary Rate Profiling

ML: Machine Learning

RRMSE: relative root-mean-squared-error

rnd: analysis using a subset of SNPs chosen at random

unwt: analysis in which SNPs are not weighted (and effectively all have an equal weight of 1)

wt: analysis in which SNPs are weighted

Trait abbreviations: see Table 3.