Review history

First round of review

Reviewer 1

In this work, the authors proposed a multi-kernel learning method GMKL for genomic prediction/interpretation, where multiple kernel matrices are defined based on distinct inheritance mechanisms and heuristic methods are used to combine these kernels for use in kernel SVA. Simulations with the known genetic architecture are used to demonstrate the improved predictive performance of GMKL when the inheritance mode deviates significantly from the traditional additive assumption. Moreover, by comparing the "correlation" of each kernel/kinship matrix with the ideal one defined based on the observed phenotypes, the major inheritance mechanism can be speculated. Finally, GMKL is applied to two independent IBD datasets. Although the prediction performance of GMKL is similar to traditional methods, the authors made interpretations about the major inheritance mechanism based on the weights on the respective kernel/kinship matrices. Overall, the manuscript is well written, the idea makes sense, the evaluation plan is well designed, and the presentation of the results is clear. However, the methodological innovation is moderate since all tools used are well established. The performance on the IBD datasets is a bit weak. Finally, the interpretation based on the weights on different kernel/kinship matrices is rather exploratory. Following are my major comments:
(1) The simulation settings are well designed. However, it seems only one simulation run is performed (if my understanding is correct). Thus, it is not adequate to assess the run-to-run variability.
(2) The authors compared different versions of GMKL. But it offers no practical guidance as which one to use in real applications. Cherry-picking the best GMKL is subject to over-fitting.
(3) The results on the IBD datasets are weak. It seems the GBLUP based on the additive model is very comparable to the GMKL methods, casting doubt on the extent of epistasis in IBD. Moreover, the rankings based on the FH and CKA are very different, and in both datasets, CKA places the highest rank on the Dominance inheritance mechanism, further clouding the conclusions made.

Reviewer 2

- line numbers are repeated for every page, which is inconvenient, it would be better to have one single continuous numbering of lines
- line numbers are not well aligned with the text, sometimes it is difficult to identify the correct line for revision
- there is an imperfect understanding of genomic predictions, genomic selection and their applications in plant and animal breeding, and in human medicine (but also veterinary medicine); this makes the Introduction and Discussion filled with imprecise and incorrect statements (see specific comments below)
- in your discussion of kernel learning methods, you fail to mention the broad family of RKHS (reproducing kernel hilbert space) regression methods, which are extensively used in the field of genomic predictions (see for example "Reproducing Kernel Hilbert Spaces Regression Methods for Genomic Assisted Prediction of Quantitative Traits" by Gianola and van Kaam 2008; or the R package BGLR that you yourselves used in your work; and related works by Gustavo de los Campos & colleagues). I think this must be acknowledged and discussed in somewhere your article (introduction, methods, discussion)

Introduction

P1

L43: genomic predictions and genomic selection are not the same thing: whole-genome predictions (a.k.a. genomic or genome-enabled predictions) are a mathematical/statistical tool that allow scientists and professionals to predict the unknown genetic value of future phenotypic performance for a specific trait of interest. Genomic selection is the term used for the artificial selection programmes revolutionised by the use of genomics in plant & animal breeding
L43: I don't think that whole-genome predictions (WGP) and genomic selection (GS) are a branch of genome interpretation: GI is a relatively new/unusual term which is not used very often in literature (I see that there's a couple of citations from the same authors of this manuscript). I suggest that you give your definition of genome interpretation and present/discuss preexisting concepts like WGP and GS in the context of GI
L44: WGP and GS are used not only in animal breeding but also in plant breeding (e.g. Heffner et al. 2010 "Plant Breeding with Genomic Selection: Gain per Unit Time and Cost")
L44-45: WGP is used also in human medicine for the prediction, for example, of genetic predisposition to developing complex diseases (but laso for other traits: behavior etc.); see for instance the following publications (but there are many more):
- Leet et al. 2008 "Predicting Unobserved Phenotypes for Complex Traits from Whole-Genome SNP Data"
- Daetwyler et al. 2008 "Accuracy of predicting the genetic risk of disease using a genome-wide approach"
- de los Campos et al. 2010 "Predicting genetic predisposition in humans: the promise of whole-genome markers"
GS is obvioulsy not used in humans ;-)
L51: in WGP you don't "identify specific markers (SNP) associated with the individuals' estimated breeding values (EBV)" (this is a form of GWAS that uses EBVs as phenotypes), rather you combine the information from all markers into a single GEBV (genomic estimated breeding value) (or polygenic risk score / predicted genomic score) for each individual for the studied phenotype
L53-54: again, not only livestock and plants, but also humans (and insects etc.)
L58: replace "agricultural technology" with agriculture (or, better, plant & animal breeding)
L58: BLUP/GBLUP (and the equivalent parameterizations known as RR-BLUP, SNP-BLUP, Bayesian alphabet etc.) is also used for genetic/genomic predictions in human medicine/biology, though more limited compared to plant & animal breeding

P2
L8-10: there is a little bit of confusion: WGP models, like GBLUP models, can be used mainly for two purposes (besides estimating variance components): 1) to predict the unobserved genetic value (genetic propensity) with respect to the phenotype of interest: this is used essentially in agriculture (plant & animal breeding) to guide delection decisions aimed at transmitting genetic merit from parents to their offspring, with the objective of improving the genetic make-up of the population (i.e. move the average towards the desired direction); in pricniple, thi can be applied also to huaman and veterinary medicine to provide information and counselling on matings and on the risk of giving birth to offspring that can develop / have specific diseases. 2) to predict the unknown future phenotypic performance of an individual: this is of course of interest in human (and veterinary) medicine with respect to the probability of developing diseases, but is also in interest in agriculture, i.e. when the phenotypic traits are measured over multiple time points (e.g. years): you can find an example in peach trees (Biscarini et al. 2017 "Genome-enabled predictions for fruit weight and quality from repeated records in European peach progenies"), but there are many more (for example when working with multiple lactations in dairy cows), including in the applications of neural networks to genomi predictions, that you mentioned earlier (see a review by Montesinos-Lopez et al. 2021 "A review of deep learning applications for genomic selection")
L13: though I agree with the fact that the term polygenic risk score (PRS) is most often used for the combination of SNP effects from GWAS studies, keep in mind that WGP models have been (and are) used in humans to predict the genetic predisposition to diseases (genomic prediction of

disease risk): see for example the above-mentioned works by Daetwyler et al. or by de los Campos et al.
L15: I would specify that they MOSTLY model only additive effects, since there are several works on how to incorporate dominance and epistasis into GWAS and WGP models (e.g. RKHS regression models)
L19-21: I am not totally sure about this: RKHS (reproducing kernel hilbert space) regression models are used in plant and animal breeding to incorporate multiple kernels (e.g. additive, dominance and epsistasis kinship matrices: see for example de los Campos et al. 2009 "Reproducing kernel Hilbert spaces regression: A general framework for genetic evaluation"), and RKHS and SVM are somehow mathematically related. However, a more formal or specific illustration of the connections between SVM and RKHS models for WGP from your work would certainly be a very welcome addition to the scientific literature!
L32-33: please evaluate your proposed GMKL framework with the RKHS framework already used for WGP models
L45: why only MKL here?

Results

P3
[up to L17]

Methods

P12
L11: be aware that these SNP data are already pre-filtered for quality parameters (call-rate, MAF etc.)
L14: citation/link to the Simphe R package
L54: "25000 SNPs from other diseases" I think that there is a bit of ambiguity here, since SNP chips may be designed to target specific objectives, but it can be misleading to imply that ~200k SNPs are associated with CD and UO, and 25k SNPs are associated with other diseases (which?): the actual number of truly associated SNPs with any given disease is bound to be much smaller
L55: "The build of the SNPs is given in GRCh37/hg19." I guess you mean that SNPs in the Immunochip are mapped on the GRCh37/hg19 build of the human genome

P13
L15: "value of 0,1,2 for each sample representing the count of that variant": this implies that these variants are biallelic, right? Please be clear
L16-18: this is not clear: you filtered out SNPs with MAF > 0.2? Usually SNPs with low MAF (e.g. < 0.05) are filtered out; what you did seem to be that you removed the more variable/diverse SNPs in your dataset, which would be a poor choice to compute a kinship matrix (artificially higher homozygosity). Besides, you would be retaining monomrphic SNPs, non informative for this type of analysis
L18: it is also weird that you had RAM problems with a non-too large dataset, as in genomics we often deal with millions of variants on a similar number of samples. If really needed, a better way to reduce the data size, after QC filtering, would be to randomly pick 50%/75% of the SNPs, or to use some sort of LD-pruning. This should not affect too much the resulting kinship matrix, alleviating the memory problems

equation (2): it is not clear what the I to which the i's belong represent. This equation is a bit confusing: I guess you want to say that you may have a model with only additive genetic effects (y = Xb + Za + e), or a model with additive and dominant effects (e.g. y = Xb + Za + Wd + e) and so on.

Make an effort to convey clearly this message to the readers.

equations (1) and (2): what did you do for the binary phenotypes? Did you use the logistic/probit version of GBLUP? Or just used the linear GBLUP (no transformation of the y, no link function), relying on the fact that the prediction accuracy (the rank) is expected to not be majorly affected by this simplification?

P14
L5: the kinship matrix is the covariance matrix in the model
L5: something is missing between computed and different
L21-25: I believe that GBLUP is a special case (a subset) of the more general family of kernel methods, where the kinship matrix acts as the kernel (see for instance Morota & Gianola 2014 "Kernel-based whole-genome prediction of complex traits: a review"). This is also reflected in your sentence "The prediction of unseen samples x_j is computed as the weighted sum of a similarity function k(x_j, x_i) between the target sample x_j and the training samples x_i.": this is also what happens in GBLUP when you want to make predictions on unobserved "new" samples.

P16
L48-52: as you correctly acknowledge, in genomic predictions the rank of of the samples is relevant, not the absolute value of predictions (which is on an arbitrary scale of unobserved genetic values, usually centered at 0 with std = square root of the genetic variance). This is why I wonder why you did not calculate also the Spearman rank correlation coefficient as a metric to evaluate the performance of your models.

Figure 1: please add to the caption the explanation of pheno:0 to pheno:8

Authors' response to reviewers

Reviewer #1: ===
In this work, the authors proposed a multi-kernel learning method GMKL for genomic prediction/interpretation, where multiple kernel matrices are defined based on distinct inheritance mechanisms and heuristic methods are used to combine these kernels for use in kernel SVA. Simulations with the known genetic architecture are used to demonstrate the improved predictive performance of GMKL when the inheritance mode deviates significantly from the traditional additive assumption. Moreover, by comparing the "correlation" of each kernel/kinship matrix with the ideal one defined based on the observed phenotypes, the major inheritance mechanism can be speculated. Finally, GMKL is applied to two independent IBD datasets. Although the prediction performance of GMKL is similar to traditional methods, the authors made interpretations about the major inheritance mechanism based on the weights on the respective kernel/kinship matrices. Overall, the manuscript is well written, the idea makes sense, the evaluation plan is well designed, and the presentation of the results is clear. However, the methodological innovation is moderate since all tools used are well established. The performance on the IBD datasets is a bit weak. Finally, the interpretation based on the weights on different kernel/kinship matrices is rather exploratory.

Thanks for your comments.

Following are my major comments:

(1) The simulation settings are well designed. However, it seems only one simulation run is performed (if my understanding is correct). Thus, it is not adequate to assess the run-to-run

variability.

We have now repeated the computation on the synthetic phenotypes 4 more times, and we provide an estimate of the standard deviation of the predictions in Suppl. Fig. S1 . The results indicate that the standard deviation of the Pearson correlations is always lower than 0.1 for all the phenotypes. The simulation of the phenotypes with Simphe is a time consuming procedure and it is therefore not feasible to repeat it several times. We refer to Suppl. Fig. S1 in Section 2.2.

(2) The authors compared different versions of GMKL. But it offers no practical guidance as which one to use in real applications. Cherry-picking the best GMKL is subject to over-fitting.

We agree with your comment. We now discuss this aspect in the new Discussion Section 3.3, explaining that we suggest to use CKACLOSED, since it is the most mathematically sound method among the ones found in literature. We also mention it in Section 2.5.1.

We also elaborate further on this point in the answer to your following comment (number 3).

(3) The results on the IBD datasets are weak. It seems the GBLUP based on the additive model is very comparable to the GMKL methods, casting doubt on the extent of epistasis in

IBD. Moreover, the rankings based on the FH and CKA are very different, and in both datasets, CKA places the highest rank on the Dominance inheritance mechanism, further clouding the conclusions made.

For what concerns the detection of epistasis, we used the synthetic data to test GMKL on controlled experimental settings, and we successfully showcased its ability in terms of prediction and interpretation of the underlying inheritance mechanisms.

The two IBD datasets provide a more complex scenario in which other effects come into play. IBD is a complex disease, consisting of a heterogenous group of patients, exhibiting a broad spectrum of symptoms with variations in disease severity, location and extra-intestinal manifestations. The underlying mechanisms of this wide range of clinical presentations involve a variable set of multiple genetic factors that interact both amongst themselves and with environmental factors.

Concerning the role of epistasis in IBD, we consider the fact that adding kinship matrices to our GMKL models always increases the prediction performances shown in Figure 2 and 3 an indication that including forms of epistasis in the model is indeed beneficial and therefore likely aligned with the underlying disease mechanism.
This increase is particularly striking on the IBDSNP dataset (see Fig. 2), where, for example, FH goes from 0.5711 for FH:1 to 0.6954 for FH:5. This is also in line with our previous findings [10.1186/s13059-023-03064-y], and it has been hypothesized in other studies [10.1073/pnas.1119675109]. We have strengthened this point in Discussion.

This is further illustrated by the fact that, if we collapse the 3 types of epistasis into a single category E∗, we see that **every method** assigns a non negligible relevance to it in Tables 2 and 3.
Finally, if we build GMKL models considering only {A, D} or {A, D, E*} effects, on both datasets **the addition of the E* kinship matrices provide a significant increase in AUC**, indicating once again the relevance of this Inheritance Mechanism for the predictions (see DeLong test p-values at the beginning of page 10, Section 3.1).

For what concerns the predictions, the additive GBLUP:1 model indeed has high performance on both IBDSNP and IBDWES, and shows very little benefit from the introduction of additional kinship matrices.
This can be explained by the generally noisy and heterogeneous conditions of these real-life datasets, which may confound the inheritance mechanisms signals. In these settings the low complexity (high bias) of the linear models is beneficial as it translates into robustness.
Moreover, the high performances of additive models like GBLUP:1 can be further explained by the possibility to **approximate epistatic effects by their additive component**. Studies like [DOI:10.1038/nrg3627] indeed show that **even purely epistatic effects might result in a strong additive component** [see Fig. 2 in DOI:10.1038/nrg3627, DOI:10.1038/nrg1407, DOI:10.1038/nrg3382, 10.1073/pnas.1119675109], which could explain why non-epistatic mechanisms are still likely to be picked up as relevant, in particular at relatively low sample size. We now discuss this aspect in the new Section 2.6.1.

For what concerns the discrepancy between FH and CKA methods, the differences are due to the different choices of kernel alignment functions used in the two papers proposing these methods.
While they all build on the idea of kernel alignment, in CKA and CKACLOSED, the alignment is computed after centering the kernel matrices, therefore turning the alignment function into something closer to an actual correlation metric. This only difference explains the disagreements between CKA/CKACLOSED and FH/GD. The centering of the kernels is a theoretical improvement of CKA/CKACLOSED over FH/GD, as explained in the section in which we describe the CKA method (Methods Section 5.4.3) and in the original paper. Our goal here was to show how different multiple kernel learning methods performed, and we therefore compared few approaches from the literature and of our own design (i.e. GD).

While CKA indeed ranks Dominance higher in both datasets, the weights of the 5 different IMs are very similar, showing CKA's inability to model the correlations between IMs. Moreover, CKACLOSED performs a whitening transformation on the kernel alignments to remove their linear correlations. Thanks to this improvement, D-D epistasis ranked second in IBDSNP, and first in IBDWES.

Due to the fact that CKACLOSED algorithm is the most mathematically sound of the ones proposed in literature so far, we believe that its results should be considered to be the closest to the underlying disease mechanisms of IBD. We clarified this point in the text, following your previous comment number 2. We now discuss these aspects in detail in Section 2.5.1 and in the new Discussion Section 3.3.

Reviewer #2: Title : "Genomic Prediction with kinship-based Multiple Kernel Learning produces hypothesis on the underlying inheritance mechanisms of phenotypic traits"

- line numbers are repeated for every page, which is inconvenient, it would be better to have one single continuous numbering of lines

We agree that it is not ideal, but the line numbers have been added by the BMC platform after submission. We originally uploaded a PDF without numbers.

- line numbers are not well aligned with the text, sometimes it is difficult to identify the correct line for revision

Again, this is an issue introduced by the BMC platform after we uploaded the PDF, which does not contain numbers.

- there is an imperfect understanding of genomic predictions, genomic selection and their applications in plant and animal breeding, and in human medicine (but also veterinary medicine); this makes the Introduction and Discussion filled with imprecise and incorrect statements (see specific comments below)

Thanks for your insight, we changed the text according to your corrections.

We believe that many of these points are due to the difference in the vocabularies that are used by researchers converging to the “Genome Interpretation problem” (i.e. predicting the relationship between genotype and phenotype) from different disciplines. We are happy that clarifying these points will broaden the audience that we will be able to reach effectively.

We hope we were able to address all your comments. As you noticed, it was not always trivial to correctly identify which part of the text should be amended due to the misalignment of the line numbers. We are happy to further change the text if necessary.

- in your discussion of kernel learning methods, you fail to mention the broad family of RKHS (reproducing kernel hilbert space) regression methods, which are extensively used in the field of genomic predictions (see for example "Reproducing Kernel Hilbert Spaces Regression Methods for Genomic Assisted Prediction of Quantitative Traits" by Gianola and van Kaam 2008; or the R package BGLR that you yourselves used in your work; and related works by Gustavo de los Campos & colleagues). I think this must be acknowledged and discussed in somewhere your article (introduction, methods, discussion)

We believe that this point is indeed due to one of these differences in vocabulary between researchers approaching the same problem from different fields.
The Reproducing Kernel Hilbert Space regression techniques are powerful and flexible approaches in Machine Learning and Statistics for nonparametric regression. They are based on the fundamental concept that the modeling of nonlinear relationships in a computationally tractable way is achieved through the embedding of the input data into a high-dimensional (possibly infinite-dimensional) space with the use of kernel functions.
The Support Vector Machines we base our paper on belong to the RKHS methods family, since they are based on the Representer Theorem and the properties of the RKHS. Other examples of RKHS methods are the Kernel Ridge Regression and Gaussian Processes. The term “RKHS methods” is not commonly used in Machine Learning literature.

We agree that our writing should be more inclusive and we therefore rephrase this part of the introduction, explicitly mentioning the RKHS family of methods and citing the papers you recommended.

Introduction

P1
L43: genomic predictions and genomic selection are not the same thing: whole-genome predictions (a.k.a. genomic or genome-enabled predictions) are a mathematical/statistical tool that allow scientists and professionals to predict the unknown genetic value of future phenotypic performance for a specific trait of interest. Genomic selection is the term used for the artificial selection programmes revolutionised by the use of genomics in plant & animal breeding

Thanks for the suggestion, we changed the text.

L43: I don't think that whole-genome predictions (WGP) and genomic selection (GS) are a branch of genome interpretation: GI is a relatively new/unusual term which is not used very often in literature (I see that there's a couple of citations from the same authors of this manuscript). I suggest that you give your definition of genome interpretation and present/discuss pre existing concepts like WGP and GS in the context of GI

Following your suggestion, we rephrased the incipit of the introduction, clarifying what we mean by Genome Interpretation.

Genome Interpretation (GI) is a relatively recent term, but it is used and well understood in Bioinformatics. For example, every two years the Critical Assessment of Genome Interpretation (CAGI) is held, which is a world-wide challenge with the goal of assessing the current state of GI methods on several different tasks.

L44: WGP and GS are used not only in animal breeding but also in plant breeding (e.g. Heffner et al. 2010 "Plant Breeding with Genomic Selection: Gain per Unit Time and Cost")

Thanks for pointing this out, we added the reference.

L44-45: WGP is used also in human medicine for the prediction, for example, of genetic predisposition to developing complex diseases (but laso for other traits: behavior etc.); see for instance the following publications (but there are many more):

- Leet et al. 2008 "Predicting Unobserved Phenotypes for Complex Traits from Whole-Genome SNP Data"
- Daetwyler et al. 2008 "Accuracy of predicting the genetic risk of disease using a genome-wide approach"
- de los Campos et al. 2010 "Predicting genetic predisposition in humans: the promise of whole-genome markers"

GS is obvioulsy not used in humans ;-)

Thanks for the references, we added them and we clarified the text as well.

L51: in WGP you don't "identify specific markers (SNP) associated with the individuals' estimated breeding values (EBV)" (this is a form of GWAS that uses EBVs as phenotypes), rather you combine the information from all markers into a single GEBV (genomic estimated breeding value) (or polygenic risk score / predicted genomic score) for each individual for the studied phenotype

We hope the new version of the text is now clearer.

L53-54: again, not only livestock and plants, but also humans (and insects etc.)

We changed the text to make it clearer.

L58: replace "agricultural technology" with agriculture (or, better, plant & animal breeding)

We changed the text following your suggestion.

L58: BLUP/GBLUP (and the equivalent parameterizations known as RR-BLUP, SNP-BLUP, Bayesian alphabet etc.) is also used for genetic/genomic predictions in human medicine/biology, though more limited compared to plant & animal breeding

Probably due to historical reasons, the use of Linear Mixed Models in clinical genetics is, from our experience, very limited, and methods like Polygenic Risk Scores are preferred instead. We mention this peculiarity in the introduction.

P2
L8-10: there is a little bit of confusion: WGP models, like GBLUP models, can be used mainly for two purposes (besides estimating variance components): 1) to predict the unobserved genetic value (genetic propensity) with respect to the phenotype of interest: this is used essentially in agriculture (plant & animal breeding) to guide delection decisions aimed at transmitting genetic merit from parents to their offspring, with the objective of improving the genetic make-up of the population (i.e. move the average towards the desired direction); in pricniple, thi can be applied also to huaman and veterinary medicine to provide information and counselling on matings and on the risk of giving birth to offspring that can develop / have specific diseases. 2) to predict the unknown future phenotypic performance of an individual: this is of course of interest in human (and veterinary) medicine with respect to the probability of developing diseases, but is also in interest in
agriculture, i.e. when the phenotypic traits are measured over multiple time points (e.g. years): you can find an example in peach trees (Biscarini et al. 2017 "Genome-enabled predictions for fruit weight and quality from repeated records in European peach progenies"), but there are many more (for example when working with multiple lactations in dairy cows), including in the applications of neural networks to genomi predictions, that you mentioned earlier (see a review by Montesinos-Lopez et al. 2021 "A review of deep learning applications for genomic selection")

We rephrased the entire paragraph in an attempt to make it more clear. We also cited the papers you mentioned.
We based the original text on the comparative reasoning proposed in “Naomi R Wray, Kathryn E Kemper, Benjamin J Hayes, Michael E Goddard, and Peter M Visscher. *Complex trait prediction from genome data: contrasting ebv in livestock to prs in humans: genomic prediction*. Genetics, 211(4):1131–1141, 2019.

L13: though I agree with the fact that the term polygenic risk score (PRS) is most often used for the combination of SNP effects from GWAS studies, keep in mind that WGP models have been (and are) used in humans to predict the genetic predisposition to diseases (genomic prediction of disease risk): see for example the above-mentioned works by Daetwyler et al. or by de los Campos et al.

We agree that there are cases of use of linear mixed models in genetics as well, but from our personal experience and literature search, they are vastly outnumbered by the widespread combination of GWAS+PRS. We have acknowledged in the text the two papers you

suggested above, but we think that mentioning explicitly that these are two exceptions to the otherwise striking popularity of PRS would make the text harder to understand and it would go outside the scope of our paper.

The goal of our introduction is to give a reasonably high-level introduction to the field to readers with different backgrounds, for example including bioinformaticians and machine learning researchers and the clinical geneticists with which we collaborate. We understand that the way in which we phrase things is not completely orthodox for what concerns the field of plant quantitative genetics, but we are trying to propose a synthesis between different ways to look at similar data and scientific goals. Our approach indeed touches different disciplines and we are trying to walk the narrow edge between them, talking to researchers with different backgrounds.

L15: I would specify that they MOSTLY model only additive effects, since there are several works on how to incorporate dominance and epistasis into GWAS and WGP models (e.g. RKHS regression models)

We changed the text following your suggestion.

L19-21: I am not totally sure about this: RKHS (reproducing kernel hilbert space) regression models are used in plant and animal breeding to incorporate multiple kernels (e.g. additive, dominance and epsistasis kinship matrices: see for example de los Campos et al. 2009 "Reproducing kernel Hilbert spaces regression: A general framework for genetic evaluation"), and RKHS and SVM are somehow mathematically related. However, a more formal or specific illustration of the connections between SVM and RKHS models for WGP from your work would certainly be a very welcome addition to the scientific literature!

RKHS regression is a term that indicates a broad class of models, such as Gaussian processes, Kernel Ridge regressions and SVMs. It basically identifies Machine Learning approaches that rely on the Representer Theorem to achieve the modeling of non-linear relations within the data by transforming the input data into a possibly infinitely dimensional feature space constructed by the kernel function, and then performing their inference in this enhanced space. The use of kernels allows this transformation to be tractable.

SVMs are a specific class of RKHS methods that look for *the optimal hyperplane* modeling or separating the data in the constructed Hilbert space, using constrained optimization. SVMs are therefore RKHS methods which use a more computational expensive optimization procedure (quadratic programming) to ensure that they will not just find **a hyperplane**, but **the optimal hyperplane** (i.e. the one with the largest margin) in kernels space.

Following your suggestion, we added the formal illustration of the connection between SVM and KRHS models for WPG in the new Supplementary Section S1.

L32-33: please evaluate your proposed GMKL framework with the RKHS framework already used for WGP models

The same reasoning applies. We changed the text accordingly.

L45: why only MKL here?

We changed the text.

Results

P3
[up to L17]

Methods

P12
L11: be aware that these SNP data are already pre-filtered for quality parameters (call-rate, MAF etc.)

We clarified this point and referenced the papers in which this information can be retrieved.

L14: citation/link to the Simphe R package

We added the reference to the package.

L54: "25000 SNPs from other diseases" I think that there is a bit of ambiguity here, since SNP chips may be designed to target specific objectives, but it can be misleading to imply that ~200k SNPs are associated with CD and UO, and 25k SNPs are associated with other diseases (which?): the actual number of truly associated SNPs with any given disease is bound to be much smaller

We clarified the text in the manuscript.

L55: "The build of the SNPs is given in GRCh37/hg19." I guess you mean that SNPs in the Immunochip are mapped on the GRCh37/hg19 build of the human genome

Thanks for noticing, we changed the text.

P13
L15: "value of 0,1,2 for each sample representing the count of that variant": this implies that these variants are biallelic, right? Please be clear

Thanks for noticing, we changed the text.

L16-18: this is not clear: you filtered out SNPs with MAF > 0.2? Usually SNPs with low MAF (e.g. < 0.05) are filtered out; what you did seem to be that you removed the more variable/diverse SNPs in your dataset, which would be a poor choice to compute a kinship

matrix (artificially higher homozygosity). Besides, you would be retaining monomrphic SNPs, non informative for this type of analysis

Monomorphic SNPs are always filtered out with an additional filtering step, since they are not informative for our purposes. We clarified this in the text.

Unfortunately, we were forced to introduce this MAF filtering in the IBDWES dataset due to limitations of the Sommer package, which uses excessive RAM (>32GB) to compute kinship matrices if we allow too many variants (SNPs) in it. In future works we will address this problem by streamlining the entire pipeline and removing the R parts, but for now it is outside the scope of this paper.

To address your concerns, we ran additional analyses, testing the alternative filtering suggestions you made:

- We tried a random selection of 50000 SNPs.
- We selected all the variants sharing less than 0.5 of Pearson correlation between all the others, thereby reducing the redundancy among them. This filtering yields 89881 SNPs.
- We selected the variants using the 0.2<MAF<0.5, to check whether we were previously removing relevant variants, as you suggested.

The comparison between our initial approach and the variations you suggested is now available as Suppl. File S1. In 10 cases of the 17 tested, the original MAF filtering we proposed is the best performing in terms of AUC. The second best method is the Pearson-based redundancy reduction, which has the highest AUC 5/17 times. We discuss these additional tests in Sections 2.6 and 5.1.3.

L18: it is also weird that you had RAM problems with a non-too large dataset, as in genomics we often deal with millions of variants on a similar number of samples. If really needed, a better way to reduce the data size, after QC filtering, would be to randomly pick 50%/75% of the SNPs, or to use some sort of LD-pruning. This should not affect too much the resulting kinship matrix, alleviating the memory problems

The scalability issues are due to the R package we used in the pipeline.

We followed your suggestions and we tried different ways to reduce the number of variants (SNPs), to make the computation of the kinship matrices feasible (requiring less than 32Gb of RAM) for Sommer. As explained in the answer to your previous comment, we tried to randomly sample 50k variants, to sample ~90k non-redundant variants (reduced LD between them), and to consider only the most variable variants (0.2<MAF<0.5). In the majority of the cases the original 0.01 < MAF < 0.2 filtering we originally proposed in the paper resulted to be the one yielding the highest AUC (59% of the times). The second best approach (29% of the times) was the non-redundant set of SNPs.

We show these results in Suppl. File S1 and we discuss them in the new text in Sections 2.6 and 5.1.3.

equation (2): it is not clear what the I to which the i's belong represent. This equation is a bit confusing: I guess you want to say that you may have a model with only additive genetic effects (y = Xb + Za + e), or a model with additive and dominant effects (e.g. y = Xb + Za + Wd + e) and so on. Make an effort to convey clearly this message to the readers.

This equation is the same used in the DOI:10.1038/s41598-022-24405-0 paper. The i iterates over different random variables $u_i = N(0, G_i \sigma_i)$ that are associated to kinship matrices G_i representing different Inheritance Mechanisms (Additive, Dominance and the three types of epistasis). All these effects are **summed** and therefore combined in the same model, as you mention.

We agree that we did not explain the meaning of the I set. We have now improved the explanation by calling the I variable “IM” (Inheritance Mechanisms) instead, and explaining better what it represents.

equations (1) and (2): what did you do for the binary phenotypes? Did you use the logistic/probit version of GBLUP? Or just used the linear GBLUP (no transformation of the y, no link function), relying on the fact that the prediction accuracy (the rank) is expected to not be majorly affected by this simplification?

We tried both approaches, but we ended up using the linear GBLUP (response=”gaussian” in the BGLR package) because the computation of the AUC gives higher scores when provided with continuous values (i.e. probability like predictions). The AUC indeed marginalizes over all the possible thresholds that can be used for binarizing real valued predictions, and the most relevant aspect is just the rank of the samples in the predictions. This configuration was the best performing for the BGLR models.

P14
L5: the kinship matrix is the covariance matrix in the model

We changed the text, and we hope it is now clearer.

L5: something is missing between computed and different

Thanks for noticing this issue, we fixed that.

L21-25: I believe that GBLUP is a special case (a subset) of the more general family of kernel methods, where the kinship matrix acts as the kernel (see for instance Morota & Gianola 2014 "Kernel-based whole-genome prediction of complex traits: a review"). This is also reflected in your sentence "The prediction of unseen samples x_j is computed as the weighted sum of a similarity function k(x_j, x_i) between the target sample x_j and the training samples x_i.": this is also what happens in GBLUP when you want to make predictions on unobserved "new" samples.

We indeed agree with your statement. We added a new Suppl. Section S1 in which we discuss the similarity between SVMs and GBLUP. We did not change the text in Section 5.4 because our goal there is to introduce the SVM methods to the reader.

P16
L48-52: as you correctly acknowledge, in genomic predictions the rank of of the samples is relevant, not the absolute value of predictions (which is on an arbitrary scale of unobserved genetic values, usually centered at 0 with std = square root of the genetic variance). This is why I wonder why you did not calculate also the Spearman rank correlation coefficient as a metric to evaluate the performance of your models.

Following your suggestion, we changed Fig. 1, including replacing the NDCG metric with the Spearman correlation, since it is indeed a more natural metric for regression.

Figure 1: please add to the caption the explanation of pheno:0 to pheno:8

We added the caption.

Second round of review

Reviewer 1

The authors have addressed my comments very well.

Reviewer 2

Dear authors,

first of all I apologise for the time taken to go through your answers to my previous comments. I must say that you answered satisfactorily to all comments, and that you did a lot of work of high technical value.
Related to you difficulties with the sommer package and the kinship matrix, I just want to point out that the kinship matrix can be calculated outside of R (e.g. with Plink) and then read by sommer for the GP models. I am not suggesting that you do this now, but be aware of these possibilities to make your workflow more efficient.
Before finalising this review task and recommend publication of your article, I however need you to help me with checking how you incorporated the changes in the text of the manuscript: the pdf file I have access to has two major problems:

1. the bibtex references have not been resolved and are therefore absent in the file
2. you need to somehow add correct/aligned line numbers or point me to the page and lines where you made changes in response to my comments

Without the above, it is way too time consuming and difficult for me to correctly evaluate the revised version of your article.

Thank you.

Authors' response to reviewers

Reviewer #2: Dear authors,

First of all I apologise for the time taken to go through your answers to my previous comments. I must say that you answered satisfactorily to all comments, and that you did a lot of work of high technical value.

Thanks a lot for your positive comment, we appreciate it very much.

Related to you difficulties with the sommer package and the kinship matrix, I just want to point out that the kinship matrix can be calculated outside of R (e.g. with Plink) and then read by sommer for the GP models. I am not suggesting that you do this now, but be aware of these possibilities to make your workflow more efficient.

Thanks, we will try your suggestion!

Before finalising this review task and recommend publication of your article, I however need you to help me with checking how you incorporated the changes in the text of the manuscript: the pdf file I have access to has two major problems:

1. the bibtex references have not been resolved and are therefore absent in the file
2. you need to somehow add correct/aligned line numbers or point me to the page and lines where you made changes in response to my comments

Without the above, it is way too time consuming and difficult for me to correctly evaluate the revised version of your article.

Thank you.

We are sorry for the issue. The journal's web platform seems to be bugged, and we could not get our latex to compile (see also comment from the Editor at the beginning of this document). To avoid problems, we also uploaded a correct pre-compiled PDF version, but we're not sure it reached the reviewers in the end.

Anyway, to help you with the review, we are now providing a new compiled PDF version in which we show in red only the changes related to your previous comments (the changes due to Reviewer 1 are in black). We also amended our previous answers to your comments (see text below) to mention in green the current location of the corresponding changes in the text, using page numbers and section numbers. We hope it is now more understandable.

First round of revision by Reviewer 2:

Reviewer #2: Title : "Genomic Prediction with kinship-based Multiple Kernel Learning produces hypothesis on the underlying inheritance mechanisms of phenotypic traits"

- line numbers are repeated for every page, which is inconvenient, it would be better to have one single continuous numbering of lines

We agree that it is not ideal, but the line numbers have been added by the BMC platform after submission. We originally uploaded a PDF without numbers.

- line numbers are not well aligned with the text, sometimes it is difficult to identify the correct line for revision

Again, this is an issue introduced by the BMC platform after we uploaded the PDF, which does not contain numbers.

- there is an imperfect understanding of genomic predictions, genomic selection and their applications in plant and animal breeding, and in human medicine (but also veterinary medicine); this makes the Introduction and Discussion filled with imprecise and incorrect statements (see specific comments below)

Thanks for your insight, we changed the text according to your corrections.

We believe that many of these points are due to the difference in the vocabularies that are used by researchers converging to the "Genome Interpretation problem" (i.e. predicting the relationship between genotype and phenotype) from different disciplines. We are happy that clarifying these points will broaden the audience that we will be able to reach effectively.

We hope we were able to address all your comments. As you noticed, it was not always trivial to correctly identify which part of the text should be amended due to the misalignment of the line numbers. We are happy to further change the text if necessary.

We addressed your comment by rewriting the first half of Introduction (end of page 1, first half of page 2).

- in your discussion of kernel learning methods, you fail to mention the broad family of RKHS (reproducing kernel hilbert space) regression methods, which are extensively used in the field of genomic predictions (see for example "Reproducing Kernel Hilbert Spaces Regression Methods for Genomic Assisted Prediction of Quantitative Traits" by Gianola and van Kaam 2008; or the R package BGLR that you yourselves used in your work; and related works by Gustavo de los Campos & colleagues). I think this must be acknowledged and discussed in somewhere your article (introduction, methods, discussion)

We believe that this point is indeed due to one of these differences in vocabulary between researchers approaching the same problem from different fields.
The Reproducing Kernel Hilbert Space regression techniques are powerful and flexible approaches in Machine Learning and Statistics for nonparametric regression. They are based on the fundamental concept that the modeling of nonlinear relationships in a computationally tractable way is achieved through the embedding of the input data into a high-dimensional (possibly infinite-dimensional) space with the use of kernel functions.
The Support Vector Machines we base our paper on belong to the RKHS methods family, since they are based on the Representer Theorem and the properties of the RKHS. Other examples of RKHS methods are the Kernel Ridge Regression and Gaussian Processes. The term "RKHS methods" is not commonly used in Machine Learning literature.

We agree that our writing should be more inclusive and we therefore rephrase this part of the introduction, explicitly mentioning the RKHS family of methods and citing the papers you recommended.

We addressed your comment by rewriting part of the introduction (explicitly mentioning RKHS) in page 2. We also did the same in Methods section 5.3. We also added Suppl.

Section S1 to discuss the connection between SVMs and the RKHS regression methods used in GP.

Introduction

P1

L43: genomic predictions and genomic selection are not the same thing: whole-genome predictions (a.k.a. genomic or genome-enabled predictions) are a mathematical/statistical tool that allow scientists and professionals to predict the unknown genetic value of future phenotypic performance for a specific trait of interest. Genomic selection is the term used for the artificial selection programmes revolutionised by the use of genomics in plant & animal breeding

Thanks for the suggestion, we changed the text.
We addressed your comment by rewriting the first part of the introduction (page 1).

L43: I don't think that whole-genome predictions (WGP) and genomic selection (GS) are a branch of genome interpretation: GI is a relatively new/unusual term which is not used very often in literature (I see that there's a couple of citations from the same authors of this manuscript). I suggest that you give your definition of genome interpretation and present/discuss pre existing concepts like WGP and GS in the context of GI

Following your suggestion, we rephrased the incipit of the introduction, clarifying what we mean by Genome Interpretation.

Genome Interpretation (GI) is a relatively recent term, but it is used and well understood in Bioinformatics. For example, every two years the Critical Assessment of Genome Interpretation (CAGI) is held, which is a world-wide challenge with the goal of assessing the current state of GI methods on several different tasks.

We addressed your comment by rewriting the first part of the introduction (page 1).

L44: WGP and GS are used not only in animal breeding but also in plant breeding (e.g. Heffner et al. 2010 "Plant Breeding with Genomic Selection: Gain per Unit Time and Cost")

Thanks for pointing this out, we added the reference.
We addressed your comment by rewriting the first part of the introduction (page 1).

L44-45: WGP is used also in human medicine for the prediction, for example, of genetic predisposition to developing complex diseases (but laso for other traits: behavior etc.); see for instance the following publications (but there are many more):

- Leet et al. 2008 "Predicting Unobserved Phenotypes for Complex Traits from Whole-Genome SNP Data"
- Daetwyler et al. 2008 "Accuracy of predicting the genetic risk of disease using a genome-wide approach"

- de los Campos et al. 2010 "Predicting genetic predisposition in humans: the promise of whole-genome markers"
GS is obvioulsy not used in humans ;-)

Thanks for the references, we added them and we clarified the text as well.
We added these references to the first half of the introduction.

L51: in WGP you don't "identify specific markers (SNP) associated with the individuals' estimated breeding values (EBV)" (this is a form of GWAS that uses EBVs as phenotypes), rather you combine the information from all markers into a single GEBV (genomic estimated breeding value) (or polygenic risk score / predicted genomic score) for each individual for the studied phenotype

We hope the new version of the text is now clearer.
We addressed your comment by rewriting the first part of the introduction.

L53-54: again, not only livestock and plants, but also humans (and insects etc.)

We changed the text to make it clearer. We addressed your comment by rewriting the first part of the introduction.

L58: replace "agricultural technology" with agriculture (or, better, plant & animal breeding)

We changed the text following your suggestion. We addressed your comment by rewriting the first part of the introduction.

L58: BLUP/GBLUP (and the equivalent parameterizations known as RR-BLUP, SNP-BLUP, Bayesian alphabet etc.) is also used for genetic/genomic predictions in human medicine/biology, though more limited compared to plant & animal breeding

Probably due to historical reasons, the use of Linear Mixed Models in clinical genetics is, from our experience, very limited, and methods like Polygenic Risk Scores are preferred instead. We mention this peculiarity in the introduction. We explain this in the red text at the beginning of page 2.

P2
L8-10: there is a little bit of confusion: WGP models, like GBLUP models, can be used mainly for two purposes (besides estimating variance components): 1) to predict the unobserved genetic value (genetic propensity) with respect to the phenotype of interest: this is used essentially in agriculture (plant & animal breeding) to guide delection decisions aimed at transmitting genetic merit from parents to their offspring, with the objective of improving the genetic make-up of the population (i.e. move the average towards the desired direction); in pricniple, thi can be applied also to huaman and veterinary medicine to provide information and counselling on matings and on the risk of giving birth to offspring that can develop / have specific diseases. 2) to predict the unknown future phenotypic performance of an

individual: this is of course of interest in human (and veterinary) medicine with respect to the probability of developing diseases, but is also in interest in
agriculture, i.e. when the phenotypic traits are measured over multiple time points (e.g. years): you can find an example in peach trees (Biscarini et al. 2017 "Genome-enabled predictions for fruit weight and quality from repeated records in European peach progenies"), but there are many more (for example when working with multiple lactations in dairy cows), including in the applications of neural networks to genomi predictions, that you mentioned earlier (see a review by Montesinos-Lopez et al. 2021 "A review of deep learning applications for genomic selection")

We rephrased the entire paragraph in an attempt to make it more clear. We also cited the papers you mentioned.
We based the original text on the comparative reasoning proposed in “Naomi R Wray, Kathryn E Kemper, Benjamin J Hayes, Michael E Goddard, and Peter M Visscher. *Complex trait prediction from genome data: contrasting ebv in livestock to prs in humans: genomic prediction*. Genetics, 211(4):1131–1141, 2019.

The changes are shown in red in page 2, we hope we addressed your concerns.

L13: though I agree with the fact that the term polygenic risk score (PRS) is most often used for the combination of SNP effects from GWAS studies, keep in mind that WGP models have been (and are) used in humans to predict the genetic predisposition to diseases (genomic prediction of disease risk): see for example the above-mentioned works by Daetwyler et al. or by de los Campos et al.

We agree that there are cases of use of linear mixed models in genetics as well, but from our personal experience and literature search, they are vastly outnumbered by the widespread combination of GWAS+PRS. We have acknowledged in the text the two papers you suggested above, but we think that mentioning explicitly that these are two exceptions to the otherwise striking popularity of PRS would make the text harder to understand and it would go outside the scope of our paper.

The goal of our introduction is to give a reasonably high-level introduction to the field to readers with different backgrounds, for example including bioinformaticians and machine learning researchers and the clinical geneticists with which we collaborate. We understand that the way in which we phrase things is not completely orthodox for what concerns the field of plant quantitative genetics, but we are trying to propose a synthesis between different ways to look at similar data and scientific goals. Our approach indeed touches different disciplines and we are trying to walk the narrow edge between them, talking to researchers with different backgrounds.

The changes are shown in red in the first half of the introduction.

L15: I would specify that they MOSTLY model only additive effects, since there are several works on how to incorporate dominance and epistasis into GWAS and WGP models (e.g. RKHS regression models)

We changed the text following your suggestion.

We did exactly what you suggested, by saying “Most GBLUP methods use kinship matrices computed solely from…” in page 2.

L19-21: I am not totally sure about this: RKHS (reproducing kernel hilbert space) regression models are used in plant and animal breeding to incorporate multiple kernels (e.g. additive, dominance and epsistasis kinship matrices: see for example de los Campos et al. 2009 "Reproducing kernel Hilbert spaces regression: A general framework for genetic evaluation"), and RKHS and SVM are somehow mathematically related. However, a more formal or specific illustration of the connections between SVM and RKHS models for WGP from your work would certainly be a very welcome addition to the scientific literature!

RKHS regression is a term that indicates a broad class of models, such as Gaussian processes, Kernel Ridge regressions and SVMs. It basically identifies Machine Learning approaches that rely on the Representer Theorem to achieve the modeling of non-linear relations within the data by transforming the input data into a possibly infinitely dimensional feature space constructed by the kernel function, and then performing their inference in this enhanced space. The use of kernels allows this transformation to be tractable.

SVMs are a specific class of RKHS methods that look for *the optimal hyperplane* modeling or separating the data in the constructed Hilbert space, using constrained optimization. SVMs are therefore RKHS methods which use a more computational expensive optimization procedure (quadratic programming) to ensure that they will not just find **a hyperplane**, but **the optimal hyperplane** (i.e. the one with the largest margin) in kernels space.

Following your suggestion, we added the formal illustration of the connection between SVM and KRHS models for WPG in the new Supplementary Section S1.

We changed the text in page 2, in section 5.3 (just after Eq 1), and we added a new Supplementary Section 1 to discuss this.

L32-33: please evaluate your proposed GMKL framework with the RKHS framework already used for WGP models

The same reasoning applies. We changed the text accordingly in the introduction.

L45: why only MKL here?

We changed the text in the introduction.

Results

P3
[up to L17]

Methods

P12
L11: be aware that these SNP data are already pre-filtered for quality parameters (call-rate, MAF etc.)

We clarified this point and referenced the papers in which this information can be retrieved.

We address this comment in the new red text in Section 5.1.1.

L14: citation/link to the Simphe R package

We added the reference to the package.

We address this comment in the new red text in Section 5.1.1.

L54: "25000 SNPs from other diseases" I think that there is a bit of ambiguity here, since SNP chips may be designed to target specific objectives, but it can be misleading to imply that ~200k SNPs are associated with CD and UO, and 25k SNPs are associated with other diseases (which?): the actual number of truly associated SNPs with any given disease is bound to be much smaller

We clarified the text in the manuscript. We changed the text in Section 2.5.1.

L55: "The build of the SNPs is given in GRCh37/hg19." I guess you mean that SNPs in the Immunochip are mapped on the GRCh37/hg19 build of the human genome

Thanks for noticing, we changed the text. We changed the text in Section 2.5.1.

P13
L15: "value of 0,1,2 for each sample representing the count of that variant": this implies that these variants are biallelic, right? Please be clear

Thanks for noticing, we changed the text.
We address this comment in the new red text in Section 5.1.3.

L16-18: this is not clear: you filtered out SNPs with MAF > 0.2? Usually SNPs with low MAF (e.g. < 0.05) are filtered out; what you did seem to be that you removed the more variable/diverse SNPs in your dataset, which would be a poor choice to compute a kinship matrix (artificially higher homozygosity). Besides, you would be retaining monomrphic SNPs, non informative for this type of analysis

Monomorphic SNPs are always filtered out with an additional filtering step, since they are not informative for our purposes. We clarified this in the text.

Unfortunately, we were forced to introduce this MAF filtering in the IBDWES dataset due to limitations of the Sommer package, which uses excessive RAM (>32GB) to compute kinship matrices if we allow too many variants (SNPs) in it. In future works we will address this problem by streamlining the entire pipeline and removing the R parts, but for now it is outside the scope of this paper.

To address your concerns, we ran additional analyses, testing the alternative filtering suggestions you made:

- We tried a random selection of 50000 SNPs.
- We selected all the variants sharing less than 0.5 of Pearson correlation between all the others, thereby reducing the redundancy among them. This filtering yields 89881 SNPs.
- We selected the variants using the 0.2<MAF<0.5, to check whether we were previously removing relevant variants, as you suggested.

The comparison between our initial approach and the variations you suggested is now available as Suppl. File S1. In 10 cases of the 17 tested, the original MAF filtering we proposed is the best performing in terms of AUC. The second best method is the Pearson-based redundancy reduction, which has the highest AUC 5/17 times. We discuss these additional tests in Sections 2.6 and 5.1.3.

We changed the text in Section 2.6 and 5.1.3.

L18: it is also weird that you had RAM problems with a non-too large dataset, as in genomics we often deal with millions of variants on a similar number of samples. If really needed, a better way to reduce the data size, after QC filtering, would be to randomly pick 50%/75% of the SNPs, or to use some sort of LD-pruning. This should not affect too much the resulting kinship matrix, alleviating the memory problems

The scalability issues are due to the R package we used in the pipeline.

We followed your suggestions and we tried different ways to reduce the number of variants (SNPs), to make the computation of the kinship matrices feasible (requiring less than 32Gb of RAM) for Sommer. As explained in the answer to your previous comment, we tried to randomly sample 50k variants, to sample ~90k non-redundant variants (reduced LD between them), and to consider only the most variable variants (0.2<MAF<0.5). In the majority of the cases the original 0.01 < MAF < 0.2 filtering we originally proposed in the paper resulted to be the one yielding the highest AUC (59% of the times). The second best approach (29% of the times) was the non-redundant set of SNPs.

We show these results in Suppl. File S1 and we discuss them in the new text in Sections 2.6 and 5.1.3. We changed the text in Section 2.6 and 5.1.3. In the future we will follow your suggestion concerning the use of PLINK.

equation (2): it is not clear what the I to which the i's belong represent. This equation is a bit confusing: I guess you want to say that you may have a model with only additive genetic effects (y = Xb + Za + e), or a model with additive and dominant effects (e.g. y = Xb + Za + Wd + e) and so on. Make an effort to convey clearly this message to the readers.

This equation is the same used in the DOI:10.1038/s41598-022-24405-0 paper. The i iterates over different random variables $u_i = N(0, G_i \sigma_i)$ that are associated to kinship matrices G_i representing different Inheritance Mechanisms (Additive, Dominance and the three types of epistasis). All these effects are **summed** and therefore combined in the same model, as you mention.

We agree that we did not explain the meaning of the I set. We have now improved the explanation by calling the I variable "IM" (Inheritance Mechanisms) instead, and explaining better what it represents.

We changed the text below Eq. 2.

equations (1) and (2): what did you do for the binary phenotypes? Did you use the logistic/probit version of GBLUP? Or just used the linear GBLUP (no transformation of the y, no link function), relying on the fact that the prediction accuracy (the rank) is expected to not be majorly affected by this simplification?

We tried both approaches, but we ended up using the linear GBLUP (response="gaussian" in the BGLR package) because the computation of the AUC gives higher scores when provided with continuous values (i.e. probability like predictions). The AUC indeed marginalizes over all the possible thresholds that can be used for binarizing real valued predictions, and the most relevant aspect is just the rank of the samples in the predictions. This configuration was the best performing for the BGLR models.

P14
L5: the kinship matrix is the covariance matrix in the model

We changed the text, and we hope it is now clearer. We fixed the text below Eq. 2, thanks for noticing.

L5: something is missing between computed and different

Thanks for noticing this issue, we fixed that. We fixed the text below Eq. 2, thanks for noticing.

L21-25: I believe that GBLUP is a special case (a subset) of the more general family of kernel methods, where the kinship matrix acts as the kernel (see for instance Morota & Gianola 2014 "Kernel-based whole-genome prediction of complex traits: a review"). This is also reflected in your sentence "The prediction of unseen samples x_j is computed as the

weighted sum of a similarity function k(x_j, x_i) between the target sample x_j and the training samples x_i.": this is also what happens in GBLUP when you want to make predictions on unobserved "new" samples.

We indeed agree with your statement. We added a new Suppl. Section S1 in which we discuss the similarity between SVMs and GBLUP. We did not change the text in Section 5.4 because our goal there is to introduce the SVM methods to the reader.

P16
L48-52: as you correctly acknowledge, in genomic predictions the rank of of the samples is relevant, not the absolute value of predictions (which is on an arbitrary scale of unobserved genetic values, usually centered at 0 with std = square root of the genetic variance). This is why I wonder why you did not calculate also the Spearman rank correlation coefficient as a metric to evaluate the performance of your models.

Following your suggestion, we changed Fig. 1, including replacing the NDCG metric with the Spearman correlation, since it is indeed a more natural metric for regression. The change is in Fig. 1 and in the corresponding text in Section 2.3.1.

Figure 1: please add to the caption the explanation of pheno:0 to pheno:8

We added the caption. We added the caption in Fig. 1.

Third round of review

Reviewer 2

I carefully read the response of the authors to my previous comments and I am happy with their answers.
My current recommendation is to accept this article by Raimondi et al. for publication in Genome Biology.