Introduction to Genomic Prediction

David Balding
Professor of Statistical Genetics
University of Melbourne, and
University College London

3/4 Feb 2016

1. Prediction and model selection

2. Measuring success

3. Statistical models for prediction

4. MultiBLUP

1. Prediction and model selection

2. Measuring success

3. Statistical models for prediction

4. MultiBLUP

Prediction of phenotype from GW data

Our goal here is, given a training set of data (Y_i, X_i, Z_i) for $i = 1, \ldots, n$ individuals, where

- Y_i is the phenotype,
- X_i is a vector of (usually genome-wide) genotypes,
- Z_i is a vector of recorded covariates (e.g. age, location, treatment);

to predict the unobserved phenotype Y_* of a future individual given the corresponding X_* and Z_* values.

Why?

1. **Genomic selection:** (plant/animal breeding) select individuals to mate or to be carried forward in a breeding program using estimated breeding values (EBV); BV = random effect in mixed model.
2. **Health care:** identify high-risk individuals in order to plan prophylactic interventions such as a drug treatment or lifestyle change, or to modify treatments to avoid adverse outcomes.

Prediction versus model selection

1 above is currently being successfully implemented, 2 has been less successful so far but we believe that it holds promise for the future:

- a gain of a few percent may be economically important in plant/animal breeding;
- in many settings only very high predictive accuracy is useful for human health interventions, but incremental advances can be useful (e.g. enhancing classical risk scores for drug prescription).

In both human and plant/animal genetic studies, the focus in recent years has shifted from

- model selection – identifying the SNPs that are significantly associated with phenotype, with a view to understanding mechanisms or introducing beneficial alleles;

towards

- prediction of phenotype, aimed at choosing optimal interventions or improved selection.

For prediction, we don't mind so much about including variables (e.g. SNPs) in the model if they turn out to be uninformative:

- a non-informative "predictor" usually does little harm, and by adopting a liberal approach we can glean more information from including many weakly-informative predictors than we lose from including non-predictors;
- however, it can still be beneficial to implement some screening to remove non-predictors.

Therefore, the stringent requirements for genome-wide significance that have emerged in GWAS are not appropriate for selecting predictors.

- This has led to the widespread view that the statistical model for association should be different from that for prediction, with the latter having less stringent criteria for "feature selection".
- An alternative view (that I support) is that there should be just one model that reflects your best understanding of the reality being modelled - the difference comes in how you use that model.

1. Prediction and model selection
2. Measuring success
3. Statistical models for prediction
4. MultiBLUP

How to measure success?[1]

After fitting a prediction model in a training sample, we can measure success using a test sample for which the phenotype is available (but these individuals must not form part of the training population).

- If we don't have a suitable test sample, we can artificially create one by holding back a fraction (say 0.1) from the training population;
  - predictive accuracy then tends to be understated because the full training sample is not used to fit the model;
  - conversely the test individuals are statistically the same as the training individuals, which may lead to an overstatement of predictive accuracy relative to future individuals with slightly different characteristics.
- Repeatedly estimating prediction accuracy by holding back a fraction of the training population is called cross-validation (CV). The held-back individuals may be resampled at random each time, or sampled systematically so that each individual is a member of the test sample a fixed number of times (typically once, called k-fold CV where $1/k$ is the fraction of individuals in each test sample).

[1] For ways to go wrong see: Wray N *et al.* (2013) Pitfalls of predicting complex traits from SNPs, *Nat Rev Genet* 14:507-515.

Metrics of predictive accuracy: continuous traits

Suppose that in a test sample of size k we have predictions $\hat{Y}_1, \ldots, \hat{Y}_k$ an, and the observed values (not used in the prediction process) are then revealed to be $Y_1, \ldots, Y_k$. The closer the $\hat{Y}_i$ to the Y_i the better, but there are many ways to measure closeness. The different metrics are typically highly correlated but they are not equivalent, and so there is no canonical way to measure predictive success.

Some measures of predictive accuracy for continuous Y are:

- The correlation $\mathrm{cor}(\hat{Y}, Y)$ or else the squared correlation (which is related to variance explained in a regression);
- The mean absolute error or the (root) mean square error:

$$\frac{1}{k}\sum_{i=1}^{k}|\hat{Y}-Y| \qquad \text{or} \qquad \frac{1}{k}\sum_{i=1}^{k}(\hat{Y}-Y)^2$$

Metrics of predictive accuracy: binary traits

When predicting a binary outcome, rather than return as prediction one of the two states (e.g. $Y_* \in \{\text{negative,positive}\}$), it is usually more useful to return a continuous value say π_*, which is typically constrained to the interval [0,1] and interpreted as

$$\pi_* = \mathbb{P}(Y_* = \text{positive}).$$

A threshold $\alpha \in (0, 1)$ can be assigned such that

$$\pi_* > \alpha \qquad \Leftrightarrow \qquad Y_* = \text{positive}.$$

By varying α the prediction algorithm can be tuned to optimise desired properties. Two important properties are the

- true positive rate (TPR) (or "sensitivity") which is the proportion of all true positives that are predicted to be positive, and
- false positive rate (FPR) which is the proportion of all true negatives that are predicted to be positive (1-FPR="specificity").

ROC and AUC

A Receiver Operating Characteristic (ROC) curve is a plot of the TPR (y-axis) against the FPR (x-axis) for all $\alpha \in [0,1]$. When $\alpha = 0$ all predictions are positive, so TPR=FPR=1. When $\alpha = 1$ all predictions are negative, so TPR=FPR=0.

The AUC provides a single-number summary of the ROC curve that is often used to measure predictive success:

- AUC can be interpreted as the probability that a random true $+$ve is assigned a higher probability to be a $+$ve than a random $-$ve.[2]
- AUC=0.5 $\Rightarrow$ random guessing, no information in the predictions;

Although the AUC is popular it has limitations, including:

- the AUC value is dominated by parts of the ROC curve for which the false positive rate is too high to be of interest.

[2] Wray N *et al.* (2010) The genetic interpretation of area under the ROC curve in genomic profiling. PLoS Genet 6: e1000864

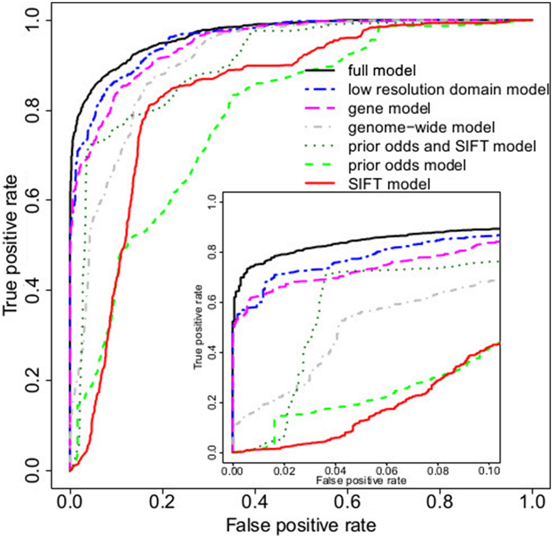
Positive Predictive Value (PPV)

Another important measure of predictive success for a given threshold is

$$\text{PPV} = \frac{\text{Prior odds x True positive rate}}{(\text{Prior odds x True positive rate}) + \text{False positive rate}}$$

The PPV is the probability that a positive call corresponds to a true positive. This differs from TPR and FPR in two important respects:

- it answers a question of direct interest;
- it requires a value for the prior probability that a queried individual will be positive, which can vary according to circumstances even if the prediction algorithm does not change.



From: Ruklisa *et al.*, Bayesian models for syndrome- and gene-specific probabilities of novel variant pathogenicity, *Genome Medicine* (2015) 7:5 doi: 10.1186/s13073-014-0120-4

PPV 0.9	**Sensitivity 0.9**	**PPV 0.95**	**Sensitivity 0.95**	**PPV 0.99**	**Sensitivity 0.99**
1	79	1	60	1	36
0.999	82	1	67	1	42
1	39	1	28	1	4
0.999	76	1	60	1	35

1. Prediction and model selection

2. Measuring success

3. Statistical models for prediction

4. MultiBLUP

The Linear Model

The workhorse of genomic prediction is the multiple linear regression model

$$Y = \mathbf{Z}\theta + \mathbf{X}\beta + \epsilon$$

where

- Y is an n-vector of phenotypes;
- $\mathbf{X}$ is an $n \times p$ matrix of (suitably coded) genotypes;
- β is an p-vector of genetic effect parameters;
- $\mathbf{Z}$ is an $n \times m$ matrix of covariates (first one constant = intercept), including treatments;
- θ is an m-vector of covariate effect parameters;
- ϵ is an n-vector of errors (or "noise"), assumed to be iid and usually assumed to be normally distributed.

Modelling assumptions

- As for model selection, most often only additive genetics effects are modelled, thus ignoring dominance and epistasis).
- Although it is difficult to confidently detect dominance and epistasis, modelling such effects may nevertheless be valuable for prediction.
- Independence of the ϵ implies that all kinship effects are assumed to be accounted for through the markers.

In practice covariates can be very important in prediction, but from now on we ignore them and focus on prediction from genomic data only.

Too few or too many predictors?

Including only genome-wide significant SNPs in a prediction model usually leads to poor prediction:

- the polygenic nature of many complex traits means that many true predictors do not reach GW significance;
- each individually conveys little information, but collectively they can be important.

On the other hand, including many predictors in a model risks over-fitting:

- parameter estimates achieve close matching of fitted values to the observed data, which appears good but ...
- much of this apparent success amounts to “fitting statistical noise”: parameters are tuned to irreproducible features of the data, leading to poor fit to new data (poor generalisability).

Polygenic risk scores[3]

A compromise is to include SNPs that are significant at a weaker threshold, chosen to optimise prediction in a training set:

1. Test each SNP one-at-a-time in the training sample and record those that are significant at level α and their estimated effect sizes.
2. The polygenic risk score for each test individual is the sum over SNPs of the effect size estimate times the individual's genotype.
3. It is common to repeat for different α in order to try to maximise predictive success, but note that the final measure of predictive success is then likely to be upwardly biased.

The effect sizes can be re-estimated in step 2 by fitting a multiple regression model including all the SNPs significant in step 1 - only useful if $\#$ SNPs (p) is $\ll$ size of the training sample (n).

[3] Dudbridge F (2013) Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genet* 9(3): e1003348; Purcell S *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* 460, 748-52 (2009).

Over-fitting and shrinkage methods (penalised regression)

A better solution to the over-fitting problem is offered by penalised (or shrinkage) regression in which a penalty in the residual sum of squares or log-likelihood “shrinks” parameter estimates towards zero.

- The form of the penalty function can be justified empirically, in terms of performance on test datasets (e.g. using CV).
- It can also be motivated in Bayesian terms: the penalty function should reflect available knowledge about the true distribution of effect sizes of marker alleles, i.e. your prior distribution.

Although the models can be implemented in a Bayesian way (integrating out the parameters – more below), for computational reasons it is more common to **maximise** over the parameters – **maximum penalised likelihood (MPL)**.

Ridge Regression (RR)

RR is an MPL method with a independent mean-0 Gaussianpenalty/prior on each genetic effect. This leads to a quadratic term in the log-likelihood:

$$\hat{\beta}_{\mathrm{ridge}} = \underset{\beta}{\text{argmin}} \left\{ \sum_{i=1}^{n} \epsilon_i^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \qquad \epsilon_i = Y_i - \sum_{j=1}^{p} X_{ij}\beta_j$$

RR is equivalent to a best linear unbiased predictor (BLUP) in a mixed model with allelic-correlation kinships computed from the marker genotypes.[4]

- BLUP the shrinkage parameter λ is estimated from the data whereas in RR it is often treated as a tuning parameter to be chosen by the investigator.

[4] Goddard *et al.* (2009) Estimating Effects and Making Predictions from Genome-Wide Marker Data, *Statist Sci* 24(4): 517-29.

GBLUP and mixed-model BLUP

The RR formulation of BLUP is sometimes called Random Regression-BLUP or Ridge Regression-BLUP (RR-BLUP) since, unlike classical regression models, the coefficients are assumed to be random.

- RR-BLUP is equivalent to genomic BLUP or GBLUP.[5]
- Switching between mixed-model (random effect + correlation matrix) BLUP and shrinkage regression (individual predictors + penalty/prior) RR-BLUP can be convenient for describing and implementing models.

[5] Meuwissen *et al.* Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001 157(4):1819-29.

LASSO Regression

LASSO (Tibshirani 2006) is similar to RR, but assumes a Laplace (double exponential) penalty on the genetic effects, equivalent to a linear term in the log-likelihood (L_1 rather than L_2):

$$\hat{\beta}_{\text{lasso}} = \underset{\beta}{\text{argmin}} \left\{ \sum_{i=1}^{n} \epsilon_i^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

- Because the Laplace distribution has a sharp peak at zero, many genetic effects will be estimated at zero, so LASSO combines model selection with prediction.
- In fact the number of non-zero effects is constrained to be $\leq n$, which may be sub-optimal if the true genetic model is highly polygenic.
- In regions of high LD typically only 1 SNP has $\hat{\beta}_j \neq 0$, which can generate sub-optimal tagging of an ungenotyped causal variant.

LASSO extensions

- Bayesian LASSO[6] is the same model but uses Gibbs sampling for integration rather than maximisation over the β_j.
- Extended Bayesian LASSO[7]: some SNPs have an individual variance term allowing bigger effects, while weaker effects are absorbed into the usual polygenic term.
- HyperLASSO[8]: gamma prior on Laplace rate parameter, 2 parameters – shape and scale. Good performance for association analysis[9].
  - LASSO is a special case when the gamma shape parameter is large.
  - Smaller values of the shape parameter give flatter tails and a sharper peak at the origin.

[6] Park T, Casella G (2008) The Bayesian Lasso *J Am Statt Assoc* 103(482): 681-6.

[7] Mutshinda C, Sillanpää M. (2010) Extended Bayesian LASSO for multiple quantitative trait loci mapping and unobserved phenotype prediction. *Genetics*, 186(3):1067-75

[8] Hoggart C *et al.* (2008) Simultaneous Analysis of All SNPs in Genome-Wide and Re-Sequencing Association Studies. *PLoS Genet* 4(7): e1000130.

[9] Ayers K, Cordell H (2010) SNP Selection in genome-wide and candidate gene studies via penalized logistic regression, *Genet. Epi.* 34: 879.

Elastic Net Regression

Elastic Net combines RR and LASSO by weighting their penalties as follows:

$$\hat{\beta}_{\text{enet}} = \underset{\beta}{\text{argmin}} \left\{ \sum_{i=1}^{n} \epsilon_i^2 + \lambda \sum_{j=1}^{p} (\alpha \beta_j^2 + (1-\alpha)|\beta_j|) \right\}.$$

The Elastic Net selects variables like the LASSO, and shrinks together the coefficients of correlated predictors like RR.

Partial Least Squares (PLS) Regression

- PLS identifies orthogonal linear combinations of the genotypes $\mathbf{z}_1, \ldots, \mathbf{z}_k$, like principal components but that maximise the correlation with phenotype rather than the variance.
- These can then be used as regression predictors, with parameters estimated via least squares in the usual way

$$\hat{\mathbf{b}}_{\text{pls}} \approx \underset{\mathbf{b}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (Y_i - \mu - \sum_{j=1}^{k} z_{ij} b_j)^2 \right\}.$$

The dimension of the problem is greatly reduced compared with including all the individual SNPs as predictors

- so no need for penalty term, but
- the model does not provide estimates of the genetic effects β.

Some pros & cons of different shrinkage models

- Finding a near-optimal value for λ in RR and LASSO is straightforward given an optimality criterion, but two major candidates are not equivalent: CV predictive correlation and predictive log-likelihood.
- Tuning the Elastic Net is time consuming if CV is performed over a grid for (α, λ), but simple hill-climbing searches can be effective and much faster than a grid.
  - Once tuned, Elastic Net outperforms both RR and LASSO.[10]
- PLS is the easy to tune, as it has a single parameter *k* and predictive correlation and predictive log-likelihood usually are unimodal in *k*.

[10] Scutari M *et al.* (2013) Improving the Efficiency of Genomic Selection. *Statist Appl Genet Mol Biol* 12 (4), 517-527.

Bayesian methods

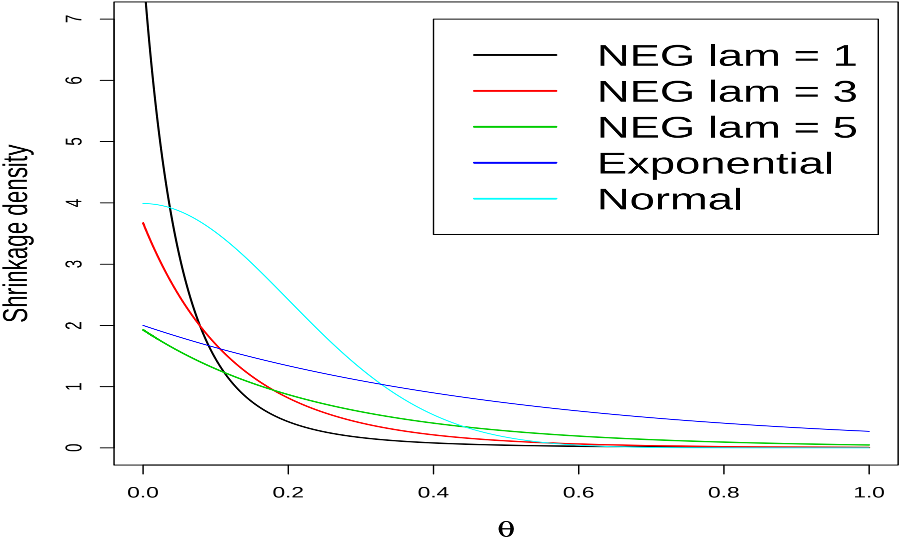
Bayesian models often have the form:

$$\prod_{i=1}^{n} N\left(Y_i | \left(\mu + \sum_{j=1}^{p} X_{ij}\beta_j\right), \sigma^2\right) \quad \times \quad p(\sigma^2)\prod_{j=1}^{p} p(\beta_j|\omega)$$

$$\text{likelihood} \quad \times \quad \text{prior}$$

- ω: vector of hyperparameters used to specify the prior; they can be
  - assumed given;
  - integrated out with respect to a prior (fully Bayesian) or
  - estimated from the data (empirical Bayes).
- σ^2 is commonly assigned a $\chi^{-2}(\nu, S)$ prior distribution.
- Assigning a Gaussian prior to β implies that the posterior means are the GBLUP estimates.
- Assigning a double-exponential or Laplace prior is the density used in the Bayesian LASSO (ref above)

Priors for SNP effect sizes



Prior densities of marker effects (mean=0, variance=1)

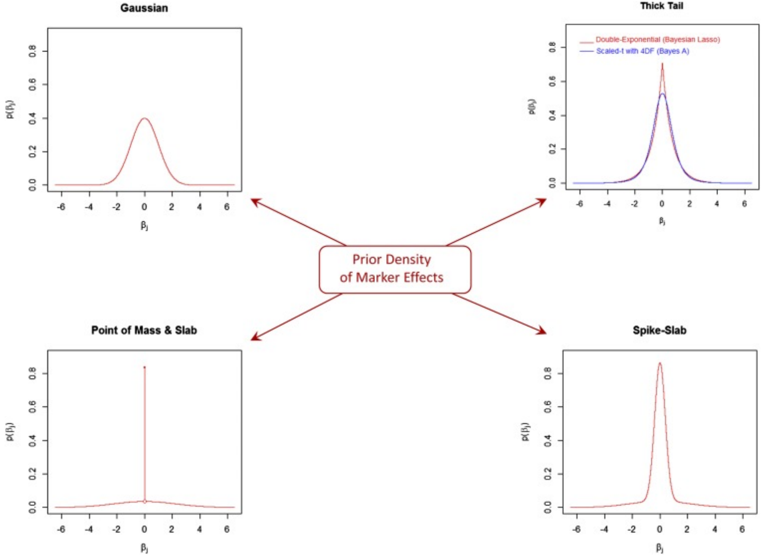


Fig 1 of: De los Campos, G *et al.* (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193(2): 327-45.

Zoo of prediction algorithms for genomic selection

Bayes A: similar to ridge regression but

- t-distribution prior (rather than Gaussian) for the β_j;
- variance comes from an inverse-χ^2 instead of being fixed.

Estimation via Gibbs sampling.

Bayes B: similar to A but with a "spike" (probability mass π) at the origin, forcing the effects of many SNPs to zero to enforce sparsity.

- Estimation via a combination of Metropolis-Hastings and Gibbs sampling: computationally intensive.

Fast Bayes B: similar to Bayes B, but uses a Laplace distribution to obtain a closed form posterior and thus speed up model estimation.[a]

[a] Meuwissen T *et al.* (2009) A fast algorithm for BayesB type of prediction of genome-wide estimates of genetic value", *Genet Sel Evol*, 41(1): 2.

Bayes Cπ: uses a “rounded spike” (low-variance Gaussian) at origin

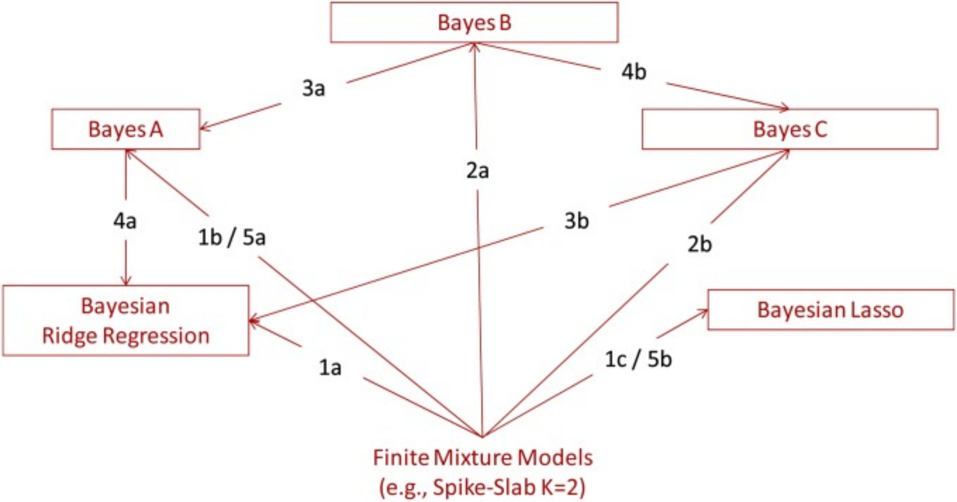
- many small effects can contribute to polygenic component,
- reduces the dimensionality of the model (makes Gibbs sampling feasible).

Bayes Dπ: similar to C, but with a *t*-distribution prior for SNP effects, allowing for different variances.

Bayes R: Hierarchical Bayesian mixture model with 4 Gaussian components, with variances scaled by 0, 10^{-4}, 10^{-3}, and 10^{-2}.[a]

The choice of prior for the β_j should ideally reflect the genetic architecture of the trait, and will vary (perhaps a lot) across traits.

[a] Moser G *et al.* (2015) Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. PLoS Genet 11(4): e1004969.



$$p\left(\beta_j \middle| \pi_1,\ldots\pi_K, \boldsymbol{\omega}_1,\ldots,\boldsymbol{\omega}_K\right)=\sum{}_{k=1}^{K}\pi_j\, p\left(\beta_j \middle| \pi_j, \boldsymbol{\omega}_j\right)$$

Fig 2 of: De los Campos, G *et al.* (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193(2): 327-45.

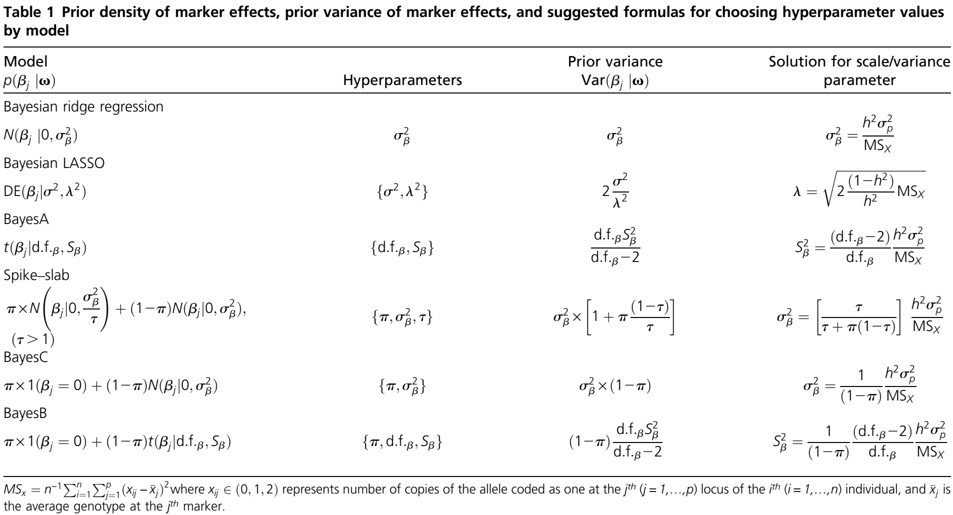


Table 1 Prior density of marker effects, prior variance of marker effects, and suggested formulas for choosing hyperparameter values by model

Model $p(\beta_j \mid \boldsymbol{\omega})$	Hyperparameters	Prior variance $\mathrm{Var}(\beta_j \mid \boldsymbol{\omega})$	Solution for scale/variance parameter
Bayesian ridge regression			
$N(\beta_j \mid 0, \sigma_\beta^2)$	σ_β^2	σ_β^2	$\sigma_\beta^2 = \frac{h^2\sigma_p^2}{\mathrm{MS}_X}$
Bayesian LASSO			
$\mathrm{DE}(\beta_j \mid \sigma^2, \lambda^2)$	$\{\sigma^2, \lambda^2\}$	$2\frac{\sigma^2}{\lambda^2}$	$\lambda = \sqrt{2\frac{(1-h^2)}{h^2}\mathrm{MS}_X}$
BayesA			
$t(\beta_j \mid \mathrm{d.f.}_\beta, S_\beta)$	$\{\mathrm{d.f.}_\beta, S_\beta\}$	$\frac{\mathrm{d.f.}_\beta S_\beta^2}{\mathrm{d.f.}_\beta - 2}$	$S_\beta^2 = \frac{(\mathrm{d.f.}_\beta - 2)}{\mathrm{d.f.}_\beta}\frac{h^2\sigma_p^2}{\mathrm{MS}_X}$
Spike–slab			
$\pi \times N\left(\beta_j \mid 0, \frac{\sigma_\beta^2}{\tau}\right) + (1-\pi)N(\beta_j \mid 0, \sigma_\beta^2),$ $(\tau > 1)$	$\{\pi, \sigma_\beta^2, \tau\}$	$\sigma_\beta^2 \times \left[1 + \pi\frac{(1-\tau)}{\tau}\right]$	$\sigma_\beta^2 = \left[\frac{\tau}{\tau + \pi(1-\tau)}\right]\frac{h^2\sigma_p^2}{\mathrm{MS}_X}$
BayesC			
$\pi \times 1(\beta_j = 0) + (1-\pi)N(\beta_j \mid 0, \sigma_\beta^2)$	$\{\pi, \sigma_\beta^2\}$	$\sigma_\beta^2 \times (1-\pi)$	$\sigma_\beta^2 = \frac{1}{(1-\pi)}\frac{h^2\sigma_p^2}{\mathrm{MS}_X}$
BayesB			
$\pi \times 1(\beta_j = 0) + (1-\pi)t(\beta_j \mid \mathrm{d.f.}_\beta, S_\beta)$	$\{\pi, \mathrm{d.f.}_\beta, S_\beta\}$	$(1-\pi)\frac{\mathrm{d.f.}_\beta S_\beta^2}{\mathrm{d.f.}_\beta - 2}$	$S_\beta^2 = \frac{1}{(1-\pi)}\frac{(\mathrm{d.f.}_\beta - 2)}{\mathrm{d.f.}_\beta}\frac{h^2\sigma_p^2}{\mathrm{MS}_X}$

$MS_x = n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{p}(x_{ij} - \bar{x}_j)^2$where $x_{ij} \in (0, 1, 2)$ represents number of copies of the allele coded as one at the j^{th} ($j = 1,\ldots,p$) locus of the i^{th} ($i = 1,\ldots,n$) individual, and $\bar{x}_j$ is the average genotype at the j^{th} marker.

Table 1 of: De los Campos, G *et al.* (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* 193(2): 327-45.

Many other statistical models

- Random forest
- Neural networks
- Reproducing kernel Hilbert spaces.

The performance of 10 algorithms, with/without bagging/boosting, over 8 crop datasets, is given by: Heslot *et al.* (2012) Genomic Selection in Plant Breeding: A Comparison of Models *Crop Sci* 52 (1): 146-60.

Many other method comparisons reviewed in: De los Campos, G *et al.* (2013) Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. *Genetics* 193(2): 327-45.

In general, prediction accuracy depends on:

- size of the training sample,
- trait heritability;
- the number of loci affecting the trait,
- the genetic relatedness between training and test samples.

Feature Selection

Redundant predictors have a (small) adverse impact on prediction, and imply a cost in additional genotyping, so it is often of interest to incorporate an element of model selection into genomic prediction.

The task of identifying a minimally-nonredundant set of predictors from a large set has come to be known as feature selection. We aim to find a minimal subset of markers $\mathbf{S} \subset \mathbf{X}$ such that

$$\mathrm{P}\left(\mathbf{Y} \,|\, \mathbf{X}\right) \approx \mathrm{P}\left(\mathbf{Y} \,|\, \mathbf{S}\right),$$

Markers in $\mathbf{X} \setminus \mathbf{S}$ are redundant only for Y, and in the case of high LD they may be redundant for many traits.

Markov Blankets & Feature Selection

LASSO and EN perform model selection implicitly. An explicit approach is Markov blanket learning. A Markov blanket (MB) is a minimal set $\mathcal{B}(\mathbf{Y})$ that satisfies

$$(\mathbf{Y} \perp\!\!\!\perp X \setminus \mathcal{B}(\mathbf{Y})) \,|\, \mathcal{B}(\mathbf{Y})$$

and is unique under very mild conditions.

- An algorithm for identifying $\mathcal{B}(\mathbf{Y})$ approximately satisfying this condition can be performed in polynomial time using a sequence of conditional independence tests involving small subsets of markers.
- The markers in $\mathcal{B}(\mathbf{Y})$ can then be used for GS with one of the linear models illustrated above.
- Whether or not feature selection is worth the additional computational effort, or indeed conveys any benefit at all overall, is a matter of debate.

Ranking and GS

- The main goal of GS is to select new varieties with better values for the trait of interest.
- For comparison across methods it can be useful to focus on ranks of the predicted EBVs:

The most common statistic to compare rankings is Kendall's τ:

$$\tau = \frac{(\text{concordant pairs}) - (\text{discordant pairs})}{\frac{1}{2}(n)(n-1)}$$

where concordant pairs are pairs of EBVs such that the highest ranked EBV of the pair is the same in both rankings; otherwise they are discordant.

Rank-based model averaging

Combining predictions of different models using their predicted ranks:

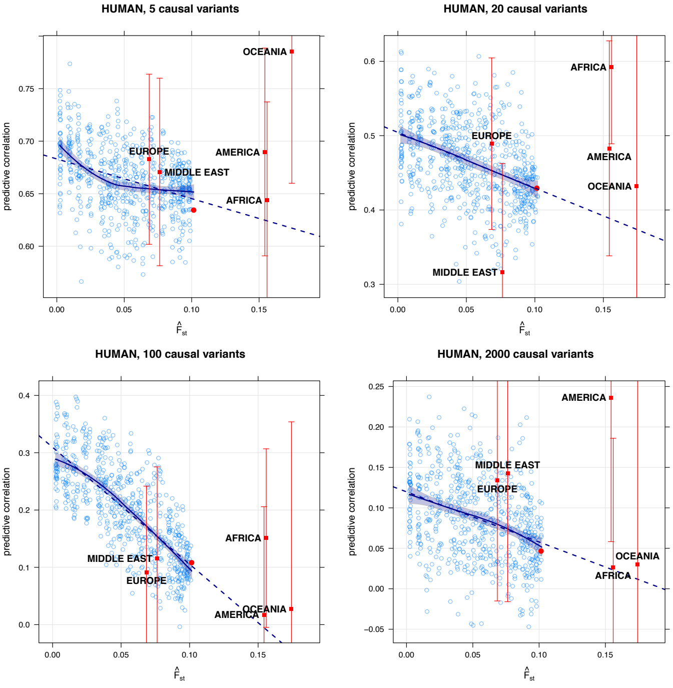
- minimises effects of prediction errors made by just one model;
- allows the combination of the predictions based on different information, because different models are better at capturing different kinds of genetic effects;
- averaged models are "smoother" than the original ones, and have been shown to have better predictive power for many classes of statistical models.

Model averaging based on ranks is also called rank aggregation.

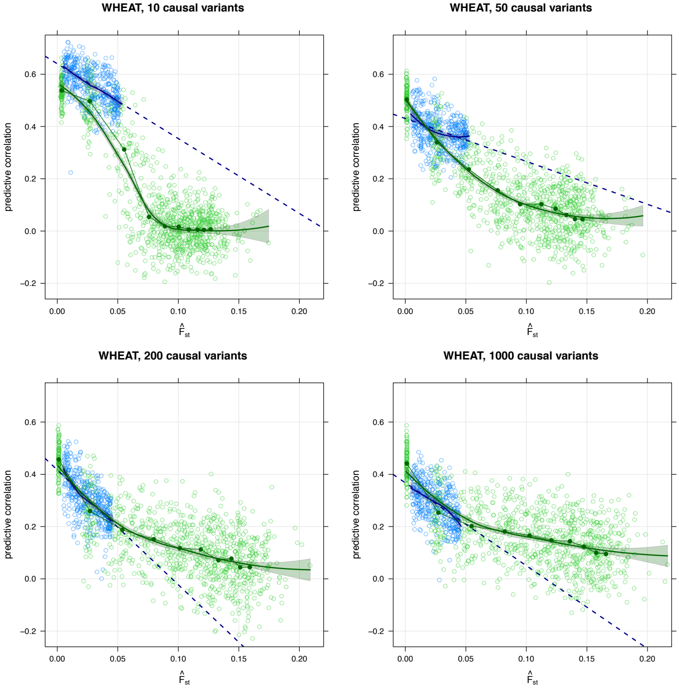
Prediction beyond the training sample

- Under CV, individuals in the test sample are statistically the same as those in the training sample.
- In genomic selection, we want to predict individuals in several future generations; in other settings we may want to predict into populations lacking sufficient training data.
- F_{ST} is a measure of average kinship between two samples and can give a (imperfect) guide to predictive accuracy.

In a paper with Mackay and Scutari, under review at PLoS Genetics, we examine the decline of predictive accuracy with increasing F_{ST} between training and test samples.



Simulated phenotype with # causal variants as indicated. Blue circles are obtained by resampling “maximally distinct” subsets from Asian data. Red vertical lines are 95% CI for prediction in the labelled populations.



Simulation of a 10-generation breeding program using 200 wheat varieties generated from 2002-07 data. Blue circles represent predictions from training data obtained by resampling "maximally distinct" subsets. Green represents predictions from subsequent generations of a simulated breeding program.

1. Prediction and model selection

2. Measuring success

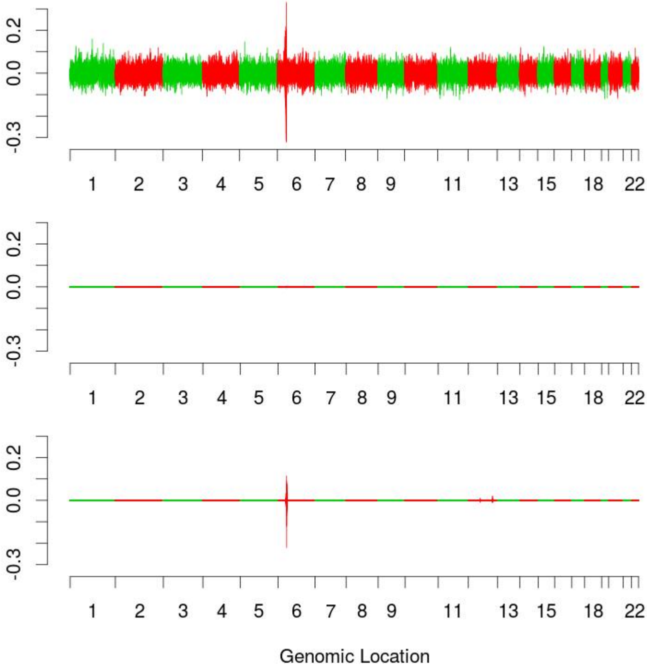
3. Statistical models for prediction

4. MultiBLUP

Generalising BLUP

- Different generalisations of GBLUP considered above have differed mainly in choice of prior/penalty, but with the same choice being applied to all SNPs.
- **MultiBLUP**[11] (incorporated in the LDAK software) extends BLUP by allowing reduced shrinkage for SNPs in promising genomic regions.
- Easiest to describe in mixed-model formulation: $Y = \gamma_1 + \gamma_2 + \gamma_3 + \ldots + \epsilon$ where $\mathrm{Var}[\gamma_m] = \sigma_m^2 K_m$ with K_m computed from SNPs in *m*th region.
- The regions can be pre-specified or chosen by MultiBLUP – adaptive MultiBLUP.

[11] Speed D, Balding D, MultiBLUP: improved SNP-based prediction for complex traits *Genome Res*, 24: 1550-1557 (2014).



Genetic Profile Risk Scores

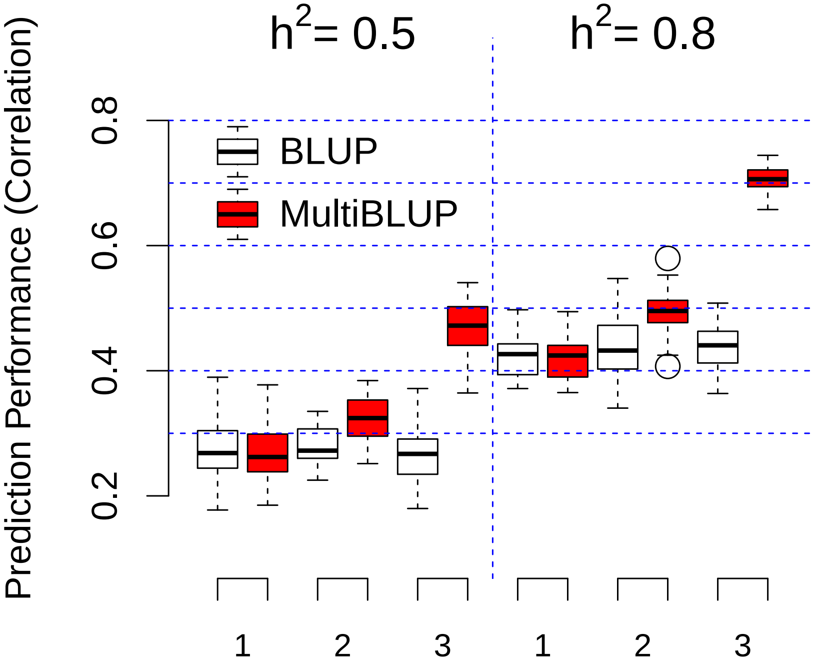
(no shrinkage)

BLUP

(uniform shrinkage)

Adaptive MultiBLUP

(flexible shrinkage, here 3 regions)


h^2= 0.5
h^2= 0.8
Prediction Performance (Correlation)
0.8
0.6
0.4
0.2
BLUP
MultiBLUP
1
2
3
1
2
3

Prediction for Crohn's Disease with 5 *a priori* regions: 3 pathways + 2 genes

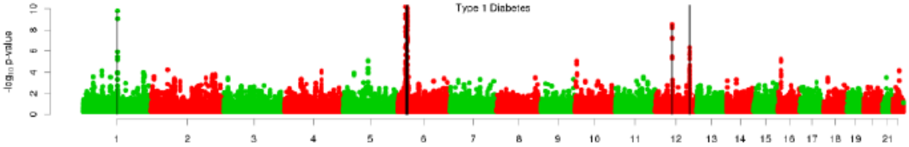
Random Effect	Region h^2	Region r^2
IL-9 Signalling	0.006	0.003
IL-2 Receptor Beta Chain	0.003	0.001
IL12 Pathway	0.019	0.016
Gene NOD2	0.012	0.012
Gene IL23R	0.008	0.007
Background Region	0.96	0.09

Correlation of predicted and true values in CV improves from 0.10 (BLUP) to 0.12 (MultiBLUP with 5 regions).

Adaptive MultiBLUP

Step 1: Divide genome into (say) 75kbp overlapping chunks.

Step 2: Test each chunk for association (using GBAT).



Step 3: Identify all significant chunks (say $P < 10^{-5}$).
(Merge these chunks with neighbouring chunks with $P < 0.01$.)

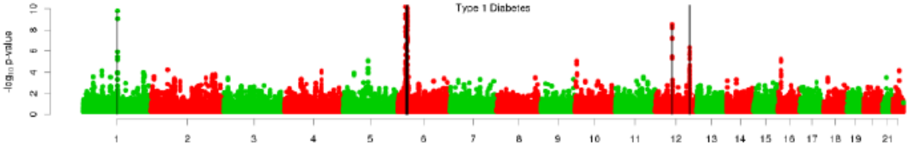
E.g., for Type 1 Diabetes, obtain 4 local regions.

Step 4: Run MultiBLUP with five random effects.

Adaptive MultiBLUP

Step 1: Divide genome into (say) 75kbp overlapping chunks.

Step 2: Test each chunk for association (using GBAT).



Step 3: Identify all significant chunks (say $P < 10^{-5}$).
(Merge these chunks with neighbouring chunks with $P < 0.01$.)

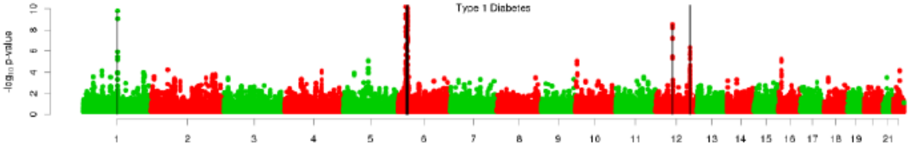
E.g., for Type 1 Diabetes, obtain 4 local regions.

Step 4: Run MultiBLUP with five random effects.

Adaptive MultiBLUP

Step 1: Divide genome into (say) 75kbp overlapping chunks.

Step 2: Test each chunk for association (using GBAT).

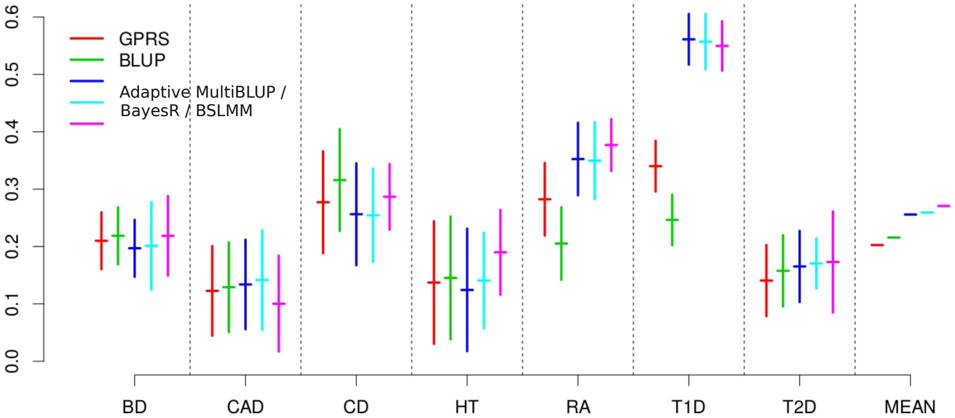


Step 3: Identify all significant chunks (say $P < 10^{-5}$).
(Merge these chunks with neighbouring chunks with $P < 0.01$.)

E.g., for Type 1 Diabetes, obtain 4 local regions.

Step 4: Run MultiBLUP with five random effects.

Adaptive MultiBLUP vs other methods: WTCCC diseases



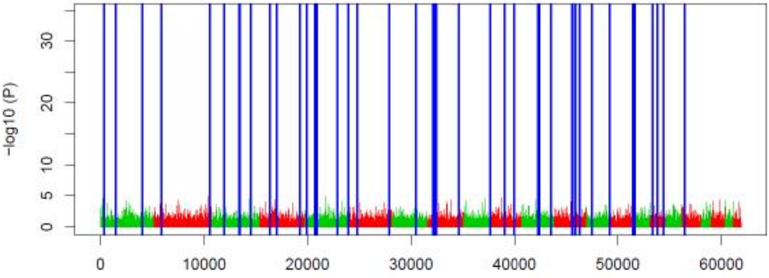
Compute times: Risk score / BLUP: < 1 hr, Stepwise Regression: 2 hrs to 5 days, MultiBLUP: 2-3 hrs, BSLMM: 8-30 hrs.

Adaptive MultiBLUP, BayesR, BSLMM achieve very similar performance, but noticeably better than Polygenic Risk Scores and BLUP for these data.

Adaptive MultiBLUP is Computationally Efficient

WT analyses ($n \approx 5000$, $N \approx 300000$) take $\sim$40 min and $< 1Gb$ memory.

Adaptive MultiBLUP can handle upwards of 50 000 individuals and imputed SNP data.



Inflammatory Bowel Disease ($n = 12,678$ $N \approx 1.5M$); 2 hours, $< 1Gb$.

Some larger datasets

BSLMM[12] and BayesR not feasible.
Performance, measured as correlation (AUC):

Irritable Bowel Disease (12,678 individuals, 1.5M SNPs):

- BLUP: 0.15 (0.58)
- Risk Score: 0.21 (0.63)
- MultiBLUP: 0.34 (0.68)

Celiac Disease (15,283 individuals, 200k SNPs):

- BLUP: 0.40 (0.76)
- Risk Score: 0.44 (0.78)
- MultiBLUP: 0.54 (0.84)

[12] Guan & Stephens (2011) *Ann. Appl. Stat.* 5(3): 1780-1815.

Conclusions

- Genome-wide SNPs allow us to think differently about both heritability and prediction.
- Many different models have been proposed in literature for genomic prediction, with different strengths and weaknesses.
- Different models will suit different trait architectures; e.g. some give more weight to rare alleles than others.
- The key elements of a statistical model include
    - whether to maximise over or integrate out genetic effects;
    - prior/penalty assumed for effect sizes.
- A polygenic term is an important component of many models,
    - its correlation structure can be specified by kinships in a mixed model
    - or it is implicit in the genome-wide distribution of effect sizes.
- Using ranks instead of the predicted EBVs can be more robust for GS.