Efficient Use of Marker Profiles in Genomic Selection

Marco Scutari[1], Ian Mackay[2] and David Balding[1]
m.scutari@ucl.ac.uk, ian.mackay@niab.com, d.balding@ucl.ac.uk

[1] Genetics Institute, University College London (UCL)
[2] National Institute of Agricultural Botany (NIAB)

September 5, 2012

Efficient Use of Marker Data

The ever-increasing amount of genetic information available in plant and animal genetics requires sophisticated computational approaches to perform GS and GWAS efficiently. In this talk we will try to address two broad issues.

1. The number of genotyped markers has been increasing for many years. Do we really need such dense, genome-wide profiles, or is focusing on a smaller set of suitably chosen markers just as effective? In other words, is it possible to perform feature selection without losing relevant information?
2. Many GS models explicitly use of a kinship matrix in the estimation of genetic effects, e.g. GBLUP, RR-BLUP. Which marker-based approach to compute such a matrix makes the best use of the profiles?

Feature Selection

It is not possible for all markers in a profile to be relevant for a trait (and we don't expect them to), both because they usually outnumber the varieties under study ($n \ll p$) and because some markers provide essentially the same information due to LD.

Therefore, both GS and GWAS can be cast as a feature selection problems. We aim to find the subset of markers $\mathbf{S} \subset \mathbf{X}$ such that

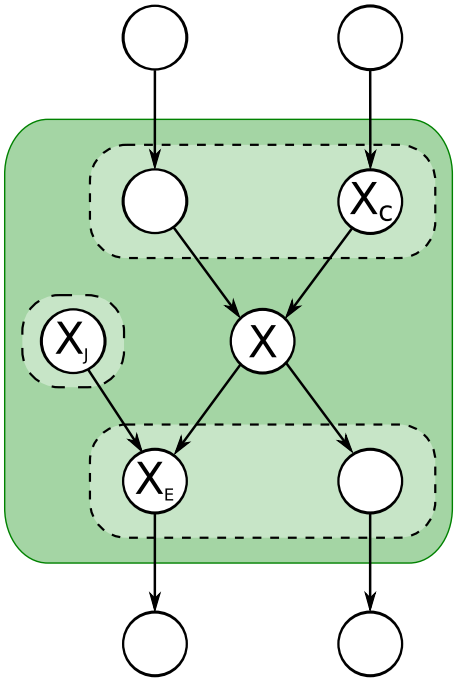
$$\mathrm{P}(\mathbf{y} \,|\, \mathbf{X}) = \mathrm{P}(\mathbf{y} \,|\, \mathbf{S}, \mathbf{X} \setminus \mathbf{S}) \approx \mathrm{P}(\mathbf{y} \,|\, \mathbf{S}),$$

that is, the subset of markers $(\mathbf{S})$ that makes all other markers $(\mathbf{X} \setminus \mathbf{S})$ redundant as far as the trait $\mathbf{y}$ we are studying is concerned.

Markov Blankets

There are several ways to identify $\mathbf{S}$; some models above do that implicitly (*e.g.* LASSO). A probabilistic approach that does that explicitly is Markov blanket learning [9, 13], which originates in graphical modelling (Bayesian and Markov networks). A Markov blanket (MB) is a minimal set $\mathcal{B}(\mathbf{y})$ that satisfies

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



and is unique under very mild conditions. It can be learned from the data with one of several algorithms (*e.g.* Incremental Association Markov Blanket, IAMB) in polynomial time using a sequence of conditional independence tests involving small subsets of markers.

Kinship Estimation

Three kinship matrix estimators have been considered:

- Habier *et al.* [5]

$$\mathbf{K} = \frac{(\mathbf{X} - \mathbf{P})(\mathbf{X} - \mathbf{P})^T}{2\sum_i p_i(1 - p_i)}$$

  where $\mathbf{P} = [2p_1 \cdots 2p_m]$ and p_i is the allele frequency of the ith marker;

- Astle & Balding [1],

$$\mathbf{K} = \overline{\mathbf{X}}\,\overline{\mathbf{X}}^T$$

  where $\overline{\mathbf{X}}$ is the standardised $\mathbf{X}$.

- Speed et al. [12] LD-adjusted kinship matrix, which adjusts for over-estimation of causal variants in high-LD regions and under-estimation in low-LD regions.

Data Sets

Data sets used as benchmarks are:

- the barley marker profiles from the AGOUEB project [2, 15] (227 profiles with 810 SNPs), with yield as the trait;
- the WTCCC [11, 14] mice heterogeneous population (2K profiles with 12K SNPs) with growth rate as the trait;
- the *Oryza sativa* rice [17] (414 profiles with 74K SNPs), with the number of seeds per panicle as the trait.

All the data sets were pre-processed by removing highly-correlated markers ($r > 90\%$), those with $> 20\%$ missing values and those with $\texttt{MAF} < 0.01$.

GS Models & Software

We considered 4 GS models which do not account explicitly for kinship:

- Partial Least Squares (`R` package `pls`);
- Ridge Regression (`R` packages `penalized` and `glmnet`);
- LASSO (`R` packages `penalized` and `glmnet`);
- Elastic Net (`R` packages `penalized` and `glmnet`);

and 2 models which do:

- GBLUP (`R` package `synbreed`);
- RR-BLUP (`R` package `synbreed`).

The kinship matrices from Habier *et al.* [5] and Astle & Balding [1] have been estimated with the `synbreed` `R` package, and the one from Speed *et al.* has been estimated with `ldak` (`http://www.ldak.org/`).

Markov blanket feature selection was performed with the IAMB algorithm as implemented in the `bnlearn` `R` package.

Predictive Power: Markov Blankets

Model	COR	COR_{MB}	Δ	CV	CV_{MB}	Δ
AGOUEB, YIELD (184.9 SNPs out of 810, 22.82%)						
PLS	0.812	0.805	−0.007	0.495	**0.495**	+0.000
Ridge	0.817	0.765	−0.051	**0.501**	0.489	−0.012
LASSO	**0.829**	**0.811**	−0.018	0.400	0.399	−0.001
Elastic Net	0.806	0.752	−0.054	**0.500**	0.489	−0.011
MICE, GROWTH RATE (543.1 SNPs out of 12K, 4.32%)						
PLS	0.716	0.882	+0.166	0.344	0.388	+0.044
Ridge	**0.841**	0.889	+0.047	0.366	0.394	+0.028
LASSO	0.717	0.881	+0.164	0.390	0.394	+0.004
Elastic Net	0.751	**0.893**	+0.142	**0.403**	**0.401**	−0.001
RICE, SEEDS PER PANICLE (293 SNPs out of 74K, 0.39%)						
PLS	0.853	0.923	+0.070	0.583	0.601	+0.018
Ridge	0.950	0.921	−0.029	0.601	**0.612**	+0.011
LASSO	0.885	**0.939**	+0.054	0.516	0.580	+0.064
Elastic Net	**0.958**	0.917	+0.040	**0.602**	**0.612**	+0.010

Predictive Power: Kinship

Model	GBLUP COR	GBLUP CV	RR-BLUP COR	RR-BLUP CV
AGOUEB, YIELD (810 SNPs)				
Habier *et al.*	0.847	0.512	0.846	**0.459**
Astle & Balding	**0.848**	0.513	0.845	**0.460**
Speed *et al.*	0.832	**0.521**	**0.847**	**0.460**
MICE, GROWTH RATE (12K SNPs)				
Habier *et al.*	0.656	0.366	0.654	0.306
Astle & Balding	0.688	0.388	0.656	0.308
Speed *et al.*	**0.695**	**0.400**	**0.666**	**0.310**
RICE, SEEDS PER PANICLE (74K SNPs)				
Habier *et al.*	**0.933**	0.590	0.932	**0.595**
Astle & Balding	**0.933**	**0.598**	0.933	**0.596**
Speed *et al.*	0.918	0.594	**0.935**	**0.595**

Markov Blankets and Kinship Estimation (GBLUP)

Model	GBLUP					
	COR_{MB}	Δ	CV_{MB}	Δ	CV_{MB}^{KIN}	Δ
AGOUEB, YIELD (810 SNPs)						
Habier *et al.*	**0.881**	+0.033	0.412	−0.100	0.482	−0.030
Astle & Balding	**0.881**	+0.033	0.414	−0.099	**0.491**	−0.022
Speed *et al.*	**0.882**	+0.049	**0.415**	−0.105	0.475	−0.045
MICE, GROWTH RATE (12K SNPs)						
Habier *et al.*	0.858	+0.201	0.118	−0.248	0.357	−0.008
Astle & Balding	0.870	+0.182	0.176	−0.211	0.363	−0.025
Speed *et al.*	**0.876**	+0.181	**0.195**	−0.204	**0.379**	−0.021
RICE, SEEDS PER PANICLE (74K SNPs)						
Habier *et al.*	**0.950**	+0.017	**0.428**	−0.161	**0.592**	+0.002
Astle & Balding	0.941	+0.008	**0.429**	−0.168	0.589	−0.008
Speed *et al.*	0.949	+0.031	0.425	−0.169	**0.591**	−0.003

Conclusions

- Among the models considered, the Elastic Net and GBLUP consistently outperformed the other models in terms of predictive ability.
- Speed *et al.* LD-adjusted kinship matrix usually provides better predictive power than other kinship estimators, often outperforming them for GBLUP.
- Performing feature selection by learning the Markov blanket of a trait can reduce the size of the marker profile severalfold with no significant loss (or with a small increase) in predictive power.
- Computing kinship after feature selection results in a substantial loss of predictive power for GBLUP; fitting the models after feature selection but with the kinship matrix computed from the full marker profiles works fine.

Acknowledgements

Thanks:

Anne-Marie Bochard
Zivan Karaman
the biostatistic team at Limagrain
all the people involved in the MIDRIB project

This work has been supported through the MIDRIB consortium, funded by the UK Technology Strategy Board and the BBSRC.

References I

W. Astle and D. J. Balding.

Population Structure and Cryptic Relatedness in Genetic Association Studies.
Statistical Science, 24(4):451–471, 2009.

J. Cockram, J. White, D. L. Zuluaga, D. Smith, J. Comadran, M. Macaulay, Z. Luo, M. J. Kearsey, P. Werner, D. Harrap, C. Tapsell, H. Liu, P. E. Hedley, N. Stein, D. Schulte, B. Steuernagel, D. F. Marshall, W. T. Thomas, L. Ramsay, I. Mackay, D. J. Balding, AGOUEB Consortium, R. Waugh, and D. M. O'Sullivan.

Genome-Wide Association Mapping to Candidate Polymorphism Resolution in the Unsequenced Barley Genome.
PNAS, 107(50):21611–21616, 2010.

J. H. Friedman, T. Hastie, and R. Tibshirani.

Regularization Paths for Generalized Linear Models via Coordinate Descent.
Journal of Statistical Software, 33(1):1–22, 2010.

J. J. Goeman.

penalized R package, 2012.
R package version 0.9-41.

D. Habier, R. L. Fernando, and J. C. M. Dekkers.

The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values.
Genetics, 177:2389–2397, 2007.

T. Hastie, R. Tibshirani, and J. Friedman.

The Elements of Statistical Learning: Data Mining, Inference, and Prediction.
Springer, 2nd edition, 2009.

References II

T. H. E. Meuwissen, B. J. Hayes, and M. E. Goddard.
Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps.
Genetics, 157:1819–1829, 2001.

B.-H. Mevik, R. Wehrens, and K. H. Liland.
pls: Partial Least Squares and Principal Component Regression, 2011.
R package version 2.3-0.

J. Pearl.
Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.
Morgan Kaufmann, 1988.

M. Scutari.
Learning Bayesian Networks with the bnlearn R Package.
Journal of Statistical Software, 35(3):1–22, 2010.

L. C. Solberg, W. Valdar, D. Gauguier, G. Nunez, A. Taylor, S. Burnett, C. Arboledas-Hita,
P. Hernandez-Pliego, S. Davidson, P. Burns, S. Bhattacharya, T. Hough, D. Higgs, P. Klenerman W. O. Cookson, Y. Zhang, R. M. Deacon, J. N. Rawlins, R. Mott, and J. Flint.
A protocol for high-throughput phenotyping, suitable for quantitative trait analysis in mice.
Mammalian Genome, 17:129–146, 2006.

D. Speed, G. Hermani, M. R. Johnson, and D. J. Balding.
Improved Heritability Estimation from Genome-Wide SNPs.
American Journal of Human Genetics. Submitted.

References III

I. Tsamardinos, C. F. Aliferis, and A. Statnikov.

Algorithms for Large Scale Markov Blanket Discovery.

In *Proceedings of the 16th International Florida Artificial Intelligence Research Society Conference*, pages 376–381, 2003.

W. Valdar, L. C. Solberg, D. Gauguier, S. Burnett, P. Klenerman, W. O. Cookson, M. S. Taylor, J. N. Rawlins, R. Mott, and J. Flint.

Genome-Wide Genetic Association of Complex Traits in Heterogeneous Stock Mice.

Nature Genetics, 8:879–887, 2006.

R. Waugh, D. Marshall, B. Thomas, J. Comadran, J. Russell, T. Close, N. Stein, P. Hayes, G. Muehlbauer, J. Cockram, D. O'Sullivan, I. Mackay, A. Flavell, AGOUEB, BarleyCAP, and L. Ramsay.

Whole-Genome Association Mapping in Elite Inbred Crop Varieties.

Genome, 53(11):967–972, 2010.

V. Wimmer, T. Albrecht, H.-J. Auinger, and C.-C. Schoen.

synbreed: Framework for the Analysis of Genomic Prediction Data Using R, 2012.

R package version 0.9-3.

K. Zhao, C. Tung, G. C. Eizenga, M. H. Wright, M. L. Ali, A. H. Price, G. J. Norton, M. R. Islam, A. Reynolds, J. Mezey, A. M. McClung, C. D. Bustamante, and S. R. McCouch.

Genome-Wide Association Mapping Reveals a Rich Genetic Architecture of Complex Traits in Oryza Sativa.

Nature communications, 2:467, 2011.