Incorporating high-dimensional omics phenotypes into models for predicting breeding values

Ole F. Christensen

Aarhus University, Center for Quantitative Genetics and Genomics

BovReg course 2023

Omics data for breeding value estimation

- High-dimensional (omics) phenotypes are becoming increasingly abundant in breeding and genetics: metabolomics, microbiota, NIR, trancriptomics, etc
- Aims of this presentation:
  - Present previously developed model for incorporating such data into genetic evaluation.
  - Our experiences with this model on real data sets.

Our point of departure: Whole genomic and omics prediction

- Two components: omics + genomics (e.g. Morgante et al 2020).
- Model

$$\mathbf{y} = \mathbf{Xb} + \mathbf{M}\boldsymbol{\alpha} + \mathbf{a} + \mathbf{e}$$

  - Omics intensities $\mathbf{M}$, $\boldsymbol{\alpha}$ effects of omics.
  - Genomic effect $\mathbf{a}$ (e.g. genomic relationship)

Our point of departure: Whole genomic and omics prediction

- Two components: omics + genomics (e.g. Morgante et al 2020).
- Model

$$\mathbf{y} = \mathbf{Xb} + \mathbf{M}\boldsymbol{\alpha} + \mathbf{a} + \mathbf{e}$$

  - Omics intensities $\mathbf{M}$, $\boldsymbol{\alpha}$ effects of omics.
  - Genomic effect $\mathbf{a}$ (e.g. genomic relationship)

- Majority of cases, including both omics and genomics had highest predictive performance.

Whole genomic and omics prediction

- $\mathbf{y} = \mathbf{Xb} + \mathbf{M}\boldsymbol{\alpha} + \mathbf{a} + \mathbf{e}$
- What is breeding value?
- An approach for EBV:
  - Construct omics-derived phenotype
  - Include omics-derived phenotype as correlated trait (Hayes et al, 2017, metabolomics in barley).
  - A method, not a model!
- Our motivation: genetic model for omics and phenotypes, with breeding value defined within model.

Omics-genetic model

(Christensen et al. 2021)

- Phenotypes (given omics)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{M}\boldsymbol{\alpha} + \mathbf{Z}\mathbf{a}_d + \boldsymbol{\epsilon}$$

- Omics intensities of features

$$\mathbf{m}_i = \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}}_i + \tilde{\mathbf{Z}}\mathbf{g}_i + \mathbf{e}_i, \quad i = 1, \ldots, k$$

Omics-genetic model

(Christensen et al. 2021)

- Phenotypes (given omics)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{M}\boldsymbol{\alpha} + \mathbf{Z}\mathbf{a}_d + \boldsymbol{\epsilon}$$

- Omics intensities of features

$$\mathbf{m}_i = \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}}_i + \tilde{\mathbf{Z}}\mathbf{g}_i + \mathbf{e}_i, \quad i = 1, \ldots, k$$

- All effects are Gaussian, independent, and

$$\text{Var}(\boldsymbol{\alpha}) = \sigma^2_\alpha \mathbf{I} \quad \text{Var}(\mathbf{a}_d) = \sigma^2_{a,d}\mathbf{G} \quad \text{Var}(\boldsymbol{\epsilon}) = \sigma^2_\epsilon \mathbf{I}$$

$$\text{Var}(\mathbf{g}_i) = \sigma^2_{g,i}\mathbf{G} \quad \text{Var}(\mathbf{e}_i) = \sigma^2_{e,i}\mathbf{I}$$

Omics-genetic model. heritability and BV

- Heritability: $h^2 = c_m^2 * h_m^2 + h_d^2$
- Direct heritability: h_d^2
- Indirect heritability: $c_m^2 * h_m^2$
  - Proportion of variance explained by omics: c_m^2
  - Heritability (common) of omics features: h_m^2
- Ideally, both c_m^2 and h_m^2 should be large.

Omics-genetic model. heritability and BV

- Heritability: $h^2 = c_m^2 * h_m^2 + h_d^2$
- Direct heritability: h_d^2
- Indirect heritability: $c_m^2 * h_m^2$
  - Proportion of variance explained by omics: c_m^2
  - Heritability (common) of omics features: h_m^2
- Ideally, both c_m^2 and h_m^2 should be large.
- Breeding value: $BV = \sum_{i=1}^{k} \mathbf{g}_i \alpha_i + \mathbf{a}_d$
  - Direct BV: $\mathbf{a}_d$.
  - Indirect BV: $\sum_i \mathbf{g}_i \alpha_i$.

Estimated breeding values (BLUP method 1)

- Complete data (omics on all individuals)
- Estimate effect of omics features on phenotype, $\hat{\boldsymbol{\alpha}}$, and direct BV, $\hat{\mathbf{a}}_r$, by solving MME for phenotypes.
- Estimate genetic effects on omics $\hat{\mathbf{g}}_i$ by solving MME for omics feature $i = 1, \ldots, k$.
- EBV $= \sum_i \hat{\mathbf{g}}_i \hat{\alpha}_i + \hat{\mathbf{a}}_d$
- Implemented in standard software for genetic evaluation

Estimated breeding values (BLUP method 1)

- Complete data (omics on all individuals)
- Estimate effect of omics features on phenotype, $\hat{\boldsymbol{\alpha}}$, and direct BV, $\hat{\mathbf{a}}_r$, by solving MME for phenotypes.
- Estimate genetic effects on omics $\hat{\mathbf{g}}_i$ by solving MME for omics feature $i = 1, \ldots, k$.
- EBV $= \sum_i \hat{\mathbf{g}}_i \hat{\alpha}_i + \hat{\mathbf{a}}_d$
- Implemented in standard software for genetic evaluation
- Need to solve MME for all k omics features!.

Estimated breeding values (BLUP method 2)

- Two steps:
  - Estimate omics effect on phenotype, $\widehat{\mathbf{M}\boldsymbol{\alpha}}$, and direct BV, $\hat{\mathbf{a}}_r$ by solving a MME
  - Estimate indirect BV, $\hat{\mathbf{a}}_m$, by solving another MME with $\widehat{\mathbf{M}\boldsymbol{\alpha}}$ as response

- EBV $= \hat{\mathbf{a}}_m + \hat{\mathbf{a}}_d$

- Both steps: Implemented in standard software for genetic evaluation

Estimated breeding values (BLUP method 2)

- Two steps:
  - Estimate omics effect on phenotype, $\widehat{\mathbf{M}\boldsymbol{\alpha}}$, and direct BV, $\hat{\mathbf{a}}_r$ by solving a MME
  - Estimate indirect BV, $\hat{\mathbf{a}}_m$, by solving another MME with $\widehat{\mathbf{M}\boldsymbol{\alpha}}$ as response
- EBV $= \hat{\mathbf{a}}_m + \hat{\mathbf{a}}_d$
- Both steps: Implemented in standard software for genetic evaluation
- Incomplete omics data: method not presented here.
- Alternative: Bayesian inference + McMC ; JWAS package (Zhao et al 2022).

Experiences with Omics-genetic model

- Milk traits in dairy sheep with microbiota data
  - Guillermo Martinez Boggio (visiting PhD student)
  - Abundances of Operational Taxonic Units (heritable!)
  - EBV's similar to GBLUP - makes sense since microbiota did not explain much variance.

- Malting quality in barley with metabolomic data (next part)

Metabolomic-genomic prediction of malting quality in barley

Guo X., P. Sarup, A. Jahoor, J. Jensen and O.F. Christensen (2023).
Metabolomic-genomic prediction can improve prediction accuracy for malting quality traits in barley. Genetics Selection Evolution. (*Accepted for publication*)

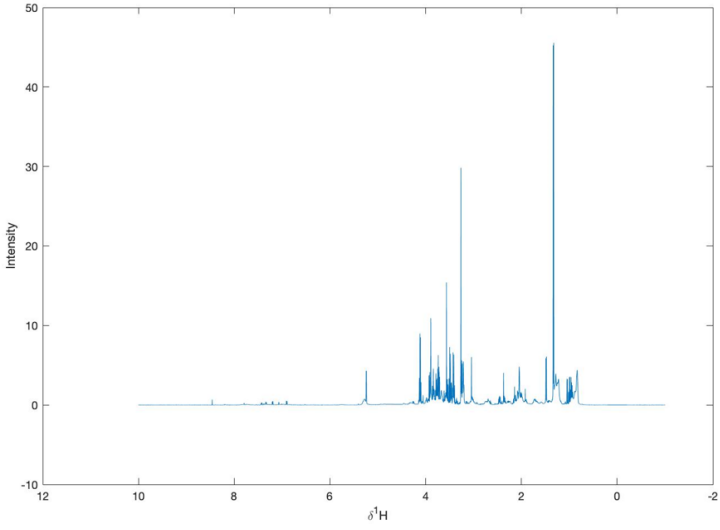
Metabolomic-genomic prediction of malting quality in barley

- Based on a published data set
- Previous analysis (my coauthors):
  - GBLUP developed
  - NMR features are heritable.
  - Cross-validation: high predictive performance when including metabolomic data.
- Our objectives:
  - Estimate model parameters
  - Investigate accuracies of EBV

Data

- 2,430 plots from 562 lines
- Two locations and three years
- A sample from each plot was malted, metabolomics on malt
- Five traits: filtering speed (FS), extract yield (EY), wort color (WC), beta-glucan content (BG), and wort viscosity (WV).

Example: metabolomics (NMR spektra)



NMR - steps

- Several bioinformatics steps (remove water peak, clustering, alignment steps, etc)
- Intensities of 24,018 metabolomic features.
- Rows of $\mathbf{M}$ normalised (to remove effect of dilution)
- Columns of $\mathbf{M}$ centered to zero and standardised to unit variance.
- Similarity matrix $\mathbf{Q} = \mathbf{M}\mathbf{M}^T/q$

GBLUP

- Model

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}_g\mathbf{g} + \mathbf{Z}_g\mathbf{l} + \mathbf{i}_g + \mathbf{i}_l + \mathbf{Z}_t\mathbf{t} + \mathbf{e}$$

  - $\mathbf{b}$: fixed (location x year x trial) effect
  - $\mathbf{g}$: additive genomic effect
  - $\mathbf{l}$: line effect
  - $\mathbf{i}_g$: Genomic by E (six location x year environments) effect
  - $\mathbf{i}_l$: line by E effect
  - $\mathbf{t}$: batch effect

Metabolomic-genomic model

- Phenotypes (given metabolomics)

$$\mathbf{y} = \mathbf{X}\mathbf{b}_1 + \mathbf{M}\boldsymbol{\alpha} + \mathbf{Z}_g\mathbf{g}_1 + \mathbf{Z}_g\mathbf{l}_1 + \mathbf{Z}_g\mathbf{i}_{g_1} + \mathbf{i}_{l_1} + \mathbf{Z}_t\mathbf{t}_1 + \mathbf{e}_1$$

- Metabolomic intensities

$$\mathbf{m}_j = \mathbf{X}\mathbf{b}_j + \mathbf{Z}_g\mathbf{g}_{j,2} + \mathbf{Z}_g\mathbf{l}_{j,2} + \mathbf{Z}_g\mathbf{i}_{g_{j,2}} + \mathbf{i}_{l_{j,2}} + \mathbf{Z}_t\mathbf{t}_{j,2} + \mathbf{e}_{j,2} \quad j = 1, \ldots, k$$

Estimated breeding values

- Two steps:
  - $MGBLUP_1$: Estimate metabolomic effect on phenotype, $\widehat{\mathbf{M}\boldsymbol{\alpha}}$, and direct BV, $\hat{\mathbf{g}}_1$.

$$\mathbf{y} = \mathbf{X}\mathbf{b}_1 + \mathbf{M}\boldsymbol{\alpha} + \mathbf{Z}_g\mathbf{g}_1 + \mathbf{Z}_g\mathbf{l}_1 + \mathbf{Z}_g\mathbf{i}_{g_1} + \mathbf{i}_{l_1} + \mathbf{Z}_t\mathbf{t}_1 + \mathbf{e}_1$$

  - $MGBLUP_2$: Estimate Indirect EBV, $\hat{\mathbf{g}}_2$

$$\widehat{\mathbf{M}\boldsymbol{\alpha}} = \mathbf{X}\mathbf{b}_2 + \mathbf{Z}_g\mathbf{g}_2 + \mathbf{Z}_g\mathbf{l}_2 + \mathbf{i}_{l_2} + \mathbf{Z}_t\mathbf{t}_2 + \mathbf{e}_2$$

  - EBV $= \hat{\mathbf{g}}_1 + \hat{\mathbf{g}}_2$

- Parameters estimated using $MGBLUP_1$ and $MGBLUP_1$ (parameter estimation using $MGBLUP_2$: an approximation)

Results: variance estimates

	GBLUP		*MGBLUP*$_1$		
	genet	total	genet	metabol	total
FS	0.014		0.005		
EY	0.28		0.22		
WC	0.14		0.04		
BG	2630		348		
WV	0.0005		0.0001		

- Direct genetic variance in *MGBLUP*$_1$ reduced compared to *GBLUP*.

Results: variance estimates

	GBLUP		*MGBLUP*$_1$		
	genet	total	genet	metabol	total
FS	0.014		0.005	0.345	
EY	0.28		0.22	1.65	
WC	0.14		0.04	3.15	
BG	2630		348	111922	
WV	0.0005		0.0001	0.0155	

- Direct genetic variance in *MGBLUP*$_1$ reduced compared to *GBLUP*.
- Metabolomic variance is very large

Results: variance estimates

	GBLUP		*MGBLUP*$_1$		
	genet	total	genet	metabol	total
FS	0.014	0.236	0.005	0.345	0.616
EY	0.28	1.68	0.22	1.65	3.17
WC	0.14	0.50	0.04	3.15	3.32
BG	2630	11636	348	111922	116102
WV	0.0005	0.0030	0.0001	0.0155	0.0170

- Direct genetic variance in *MGBLUP*$_1$ reduced compared to *GBLUP*.
- Metabolomic variance is very large
- Total variance in *MGBLUP* much larger than in *GBLUP*

Results: heritabilities and variance ratios

	GBLUP	*MGBLUP*			
	h^2	h_1^2	c_m^2	h_m^2	h^2
FS	0.06	0.01	0.67		
EY	0.17	0.07	0.52		
WC	0.28	0.01	0.95		
BG	0.23	0.00	0.96		
WV	0.15	0.01	0.91		

- h_1^2 (small) and c_m^2 (large) estimated from $MGBLUP_1$;

Results: heritabilities and variance ratios

	GBLUP	*MGBLUP*			
	h^2	h_1^2	c_m^2	h_m^2	h^2
FS	0.06	0.01	0.67	0.17	
EY	0.17	0.07	0.52	0.23	
WC	0.28	0.01	0.95	0.28	
BG	0.23	0.00	0.96	0.27	
WV	0.15	0.01	0.91	0.25	

- h_1^2 (small) and c_m^2 (large) estimated from $MGBLUP_1$;
- h_m^2 (heritability of MF) estimated from $MGBLUP_2$. Depends on the trait!

Results: heritabilities and variance ratios

	GBLUP	*MGBLUP*			
	h^2	h_1^2	c_m^2	h_m^2	h^2
FS	0.06	0.01	0.67	0.17	0.12
EY	0.17	0.07	0.52	0.23	0.19
WC	0.28	0.01	0.95	0.28	0.28
BG	0.23	0.00	0.96	0.27	0.27
WV	0.15	0.01	0.91	0.25	0.24

- h_1^2 (small) and c_m^2 (large) estimated from $MGBLUP_1$;
- h_m^2 (heritability of MF) estimated from $MGBLUP_2$. Depends on the trait!
- $MGBLUP$ $h^2 = c_m^2 * h_2^2 + h_1^2$

Predictive performance

- Leave one year out (LOYO)
- Scenarios:

	Validation
$GBLUP_g$	geno
$GBLUP_{gp}$	geno, pheno
$MGBLUP_g$	geno
$MGBLUP_{gm}$	geno, meta
$MGBLUP_{gmp}$	geno, meta, pheno

- $GBLUP_g$ and $MGBLUP_g$: cross-validation
- Others: method LR

Cross-validation

Cor(EBV, y_c) [measure of accuracy of EBV]

	$GBLUP_g$	$MGBLUP_g$
FS	0.11	0.11
EY	0.26	0.26
WC	0.27*	0.27*
BG	0.19*	0.21*
WV	0.06	0.06

- Prediction accuracies for *GBLUPg* and $MGBLUP_g$ are similar, except BG.

Cross-validation

Cor(EBV, y_c) [measure of accuracy of EBV]

	$GBLUP_g$	$MGBLUP_g$
FS	0.11	0.11
EY	0.26	0.26
WC	0.27*	0.27*
BG	0.19*	0.21*
WV	0.06	0.06

- Prediction accuracies for *GBLUPg* and $MGBLUP_g$ are similar, except BG.
- No gain in accuracy with metabolomic and phenotypes on same plots (in general)

Ratios of accuracies - method LR

- Cor($EBV_{partial}, EBV_{whole}$): measure of $acc_{partial}/acc_{whole}$ in VP,

	GBLUP acc_g/acc_{gp}	MGBLUP acc_g/acc_{gmp}
FS	0.73	0.73
EY	0.78	0.79
WC	0.71	0.71
BG	0.69	0.71
WV	0.59	0.60

- Ratios acc_g/acc_{gp} for GBLUP and acc_g/acc_{gmp} for MGBLUP are similar.

Ratios of accuracies - method LR

- Cor($EBV_{partial}, EBV_{whole}$): measure of $acc_{partial}/acc_{whole}$ in VP,

	GBLUP acc_g/acc_{gp}	MGBLUP acc_g/acc_{gmp}
FS	0.73	0.73
EY	0.78	0.79
WC	0.71	0.71
BG	0.69	0.71
WV	0.59	0.60

- Ratios acc_g/acc_{gp} for GBLUP and acc_g/acc_{gmp} for MGBLUP are similar.

- Since acc_g is similar for GBLUP and MGBLUP, then acc_{gp} for GBLUP and acc_{gmp} for MGBLUP are similar.

Ratios of accuracies - method LR

	MGBLUP		
	acc_g/acc_{gm}	acc_{gm}/acc_{gmp}	acc_g/acc_{gmp}
FS			0.73
EY			0.79
WC			0.71
BG			0.71
WV			0.60

- Low ratio = high increase in accuracy
- Ratios of accuracies: confusing to look at.

Ratios of accuracies - method LR

	MGBLUP		
	acc_g/acc_{gm}	acc_{gm}/acc_{gmp}	acc_g/acc_{gmp}
FS	0.84	0.87	0.73
EY	0.99	0.81	0.79
WC	0.88	0.89	0.71
BG	0.81	0.88	0.71
WV	0.81	0.80	0.60

- Low ratio = high increase in accuracy
- Ratios of accuracies: confusing to look at.

MGBLUP accuracies

(accuracies easier to look at than ratios)

	$MGBLUP_g$	$MGBLUP_{gm}$	$MGBLUP_{gmp}$
FS	0.32		
EY	0.60		
WC	0.51		
BG	0.40		
WV	0.12		

- Accuracies for $MGBLUP_g$ computed from cross-validation results.

MGBLUP accuracies

(accuracies easier to look at than ratios)

	$MGBLUP_g$	$MGBLUP_{gm}$	$MGBLUP_{gmp}$
FS	0.32	0.33	
EY	0.60	**0.60**	
WC	0.51	0.53	
BG	0.40	0.42	
WV	0.12	0.13	

- Accuracies for $MGBLUP_g$ computed from cross-validation results. Other accuracies are computed from ratios.

- Accuracies increase with metabolomics, except EY

MGBLUP accuracies

(accuracies easier to look at than ratios)

	$MGBLUP_g$	$MGBLUP_{gm}$	$MGBLUP_{gmp}$
FS	0.32	0.33	0.34
EY	0.60	**0.60**	0.64
WC	0.51	0.53	0.55
BG	0.40	0.42	0.44
WV	0.12	0.13	0.13

- Accuracies for $MGBLUP_g$ computed from cross-validation results. Other accuracies are computed from ratios.
- Accuracies increase with metabolomics, except EY
- and further increase with phenotypes

Conclusion on study

- Proportion of genetic effect mediated by metabolome is substantial.
- Very large metabolomic variance (sign of model-deficiency?)
- Metabolomics data and phenotypes on same plots: same accuracy as GBLUP, possibly except BG.
- Own metabolomic data, but not own phenotype: increase in accuracy compared to GBLUP, except EY.

Conclusion on study

- Proportion of genetic effect mediated by metabolome is substantial.
- Very large metabolomic variance (sign of model-deficiency?)
- Metabolomics data and phenotypes on same plots: same accuracy as GBLUP, possibly except BG.
- Own metabolomic data, but not own phenotype: increase in accuracy compared to GBLUP, except EY.
- Implications for practical breeding:
  - Reduce the amount of phenotyping
  - NMR was done on malt; so very close to the actual phenotypes.

Work in progress/Future work

- Metabolomics in pigs (NMR on 8500 pigs)
  - Investigating different metabolomic similarity matrices (ADG as a model trait)
  - Meat quality records on 1000 pigs with NMR
- Metabolomics in barley (with NMR on leaves)

Acknowledgements

- Andres Legarra, Vinzent Börner and Luis Varona (coauthors on Omics model paper, 2021)
- Guillermo Martinez Boggio (microbiota paper)
- Xiangyu Guo, Pernille Sarup, Ahmed Jahoor, Just Jensen (Coauthors on metabolomics in barley paper, 2023)
- Project "Metablomic selection in pig and barley"
  - Nordic Seed, Breeding and genetics in pig, DanBred, Aarhus University.
  - Project leader: Tage Ostersen
  - Green Development and Demonstration program (GUDP).

Take-home messages

- You can incorporate high-dimensional (omics) into genetic evaluation.
- Requirements (omics): both heritable and related to the trait of interest.
- Model parameters and similarity matrix: More experience and better understanding needed.
- Accuracy of EBV assessed using method LR