Predicting Genetic Predisposition using Prior Biological Information

Peter Sørensen

Center for Quantitative Genetics and Genomics (QGG)

Department of Molecular Biology and Genetics
Aarhus University
Denmark

Genetic Predisposition

Goal: Predict individual genetic predisposition from genotypes

Phenotype = Genotype + Environment

Terminology: Genetic predisposition, polygenic risk scores, genetic risk, genetic value, breeding value

Genetic Predisposition

Animal and Plant Breeding:

- predict genetic value used for selection decisions (e.g. improve feed efficiency and health)
- predict consequences of selection decisions (e.g. if we select for feed efficiency what happens to health?)

Healthcare Systems and Pharma Industry:

- optimize the design of a trial for testing a drug (e.g. which people should be included in trial?)
- optimize the use of a drug (e.g. which people will likely respond positively or negatively to the drug?)

Genetic Predisposition

Genetic predisposition is computed as a weighted sum of centered and scaled genotypes (or allelic counts) given by:

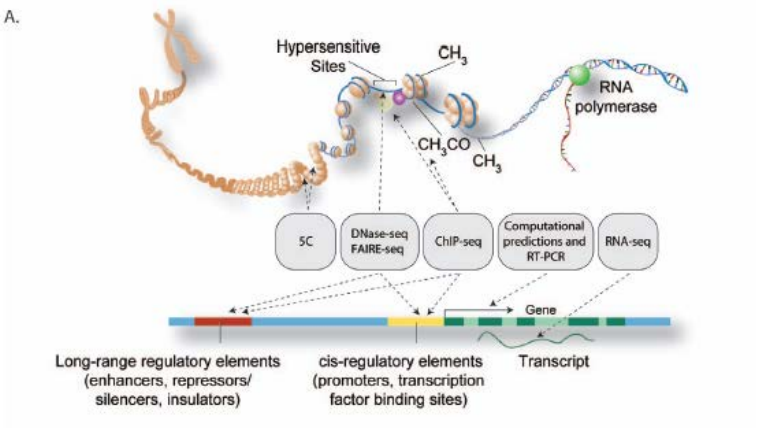
$$\mathbf{g} = \sum_{i=1}^{m} \mathbf{W}_i \, \hat{s}_i$$

where $\mathbf{W}_i$ is the centered and scaled genotypes, and $\hat{s}_i$ the weight of the i'th single genetic marker.

Weights can be log odds ratios or coefficients obtained from single- or multi-marker association analyses.

Prior Biological Information

Whole-genome sequences and **multiple novel trait phenotypes** from large numbers of individuals from **multiple populations**



Encyclopedia of DNA Elements (ENCODE)

- **Molecular phenotypes** (e.g. transcriptome, proteome,metabolome, methylome) associated to the traits/diseases of interest
- **Molecular-interaction maps** that provide insight into the structural and functional organization of the genomes

Genomic Feature Models

Develop statistical models that can **use prior biological information**

- increase prediction accuracy of phenotypes or genetic predisposition
- provide novel insights into the genetic basis of the traits

Phenotype = Genome + Metabolome +Transcriptome+...+ residual

- Phenotyping technologies (many novel traits)
- Sequencing and related technologies (multiple layers of omics data)
- Rapid accumulation of biological information in databases

- Genetic architecture (few large, many small effects, gene by gene,.....)

"Easy" to detect using methods that allow for differential shrinkage (e.g. Bayesian mixture models).

Difficult to detect

Overall Hypotheses

1. **Causal mutations** are clustered in regions on the genome defined genomic features such as:

   - biological pathways
   - gene or sequence ontologies
   - prior QTL regions
   - expression or methylation patterns
   - protein-protein or protein-metabolite interactions .....

2. If we use a **statistical model that quantifies the effect of a set of genetic variants** defined by a genomic feature we can

   - increase detection power for causal variants with small effects
   - increase prediction accuracy of complex trait phenotypes

Statistical Modeling Approaches

Genomic Feature Best Linear Unbiased Prediction Models

- Mapping Variants to Gene Ontology Categories Improves Genomic Prediction for Quantitative Traits in *Drosophila melanogaster*. Edwards SM, Sørensen IF, Sarup P, Mackay TF, Sørensen P. (2016). Genetics 203 (4): 1871-1883.

Bayesian Multiple Feature and Multiple Trait Prediction Models

- Genetic Control of Environmental Variation of Two Quantitative Traits of Drosophila melanogaster Revealed by Whole-Genome Sequencing. Sørensen P, de los Campos G, Morgante F, Mackay TFM, Sorensen DA. (2015). Genetics 201(2):487-497.

Genomic BLUP derived Marker Set Test

- Multiple Trait Covariance Association Test Identifies Gene Ontology Categories Associated with Chill Coma Recovery Time in *Drosophila melanogaster*. Sørensen IF, Edwards SM, Duun Rohde P, Sørensen P. (2017). Scientific Reports 7:2413

Drosophila Genome Reference Population

- **205 inbred lines** from the DGRP population derived from 20 generations of full sib mating

- **Whole Genome Sequence data**
  ~ 1.8M SNPs
  ~ 20 SNP pr Kb

dgrp.gnets.ncsu.edu/data

- A **range of complex trait phenotypes** (e.g. starvation resistance, startle response, chill coma recovery)

- **Access to a wealth of annotation data** that can be used to link the SNPs to different types of genomic features (e.g. genes, pathways, QTL regions, MAFs, gene and sequence ontologies, and so on)

Chill Coma Recovery

Chill coma recovery is a measure of how long time it takes before a fly right itself and stand on its legs (after on ice for three hours).

Precise phenotyping:

n = 32,231 phenotypic observations

- males and females
- 50 observation pr line/sex in replicates of 2
- 159 lines used

Genomic BLUP Model

GBLUP model is based on a **one component** linear mixed model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{g} + \mathbf{e}$$

$$(\mathbf{y} = \mathbf{Xb} + \mathbf{Ws} + \mathbf{e})$$

g is the genomic value based on all genetic markers

$$\mathbf{g} \sim N(\mathbf{0}, \mathbf{G}\sigma_g^2) \qquad \mathbf{G} = (\mathbf{WW}')/m$$

Estimate parameters using REML and predict using BLUP

Cross validation study:

- assess predictive ability of statistical model

- 10% test and 90% train
- predictions across lines
- repeated 50 times

$$PA = \frac{1}{50}\sum_{i=1}^{50} Corr(y_{pred}, y_{obs})$$

Genomic Feature BLUP Models

Predictions in GBLUP

$$\mathbf{g}_{\text{pred}} = \left(\mathbf{G}_{\text{vt}} \cdot \sigma^2_{\text{g}}\right)\left[\left(\mathbf{G}_{\text{tt}} \cdot \sigma^2_{\text{g}}\right) + \mathbf{I} \cdot \sigma^2_{\text{e}}\right]^{-1}\left(\mathbf{y}_{\text{t}} - \mathbf{Xb}\right)$$

Genetic predisposition of individuals without phenotypes

Genetic covariance between individuals with phenotypes (t) or without phenotypes (v)

Inverse of phenotypic covariance among individuals with phenotypes

Observed phenotypes adjusted for non-genetic factors

Genomic BLUP Model

Predictive Ability of GBLUP model

Females: 0.055±0.029
Males: 0.00±0.032

Fitting a Bayesian mixture model with two classes (small and large effects) yielded similar predictive ability as GBLUP model.

- "Black box" models (e.g. GBLUP or Bayesian mixture models) ignoring prior information could not predict phenotypes

- Big data, but little information

Genomic Feature BLUP Model

GFBLUP model is based on a **two component** linear mixed model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{f} + \mathbf{r} + \mathbf{e}$$

f is the genomic values based on markers in the genomic feature
r is the genomic values for the remaining set of genetic markers

$$\mathbf{f} \sim N(\mathbf{0}, \mathbf{G}_f \sigma^2_f) \qquad \mathbf{G}_f = (\mathbf{W}_f \mathbf{W}_f')/m_f$$
$$\mathbf{r} \sim N(\mathbf{0}, \mathbf{G}_r \sigma^2_r) \qquad \mathbf{G}_r = (\mathbf{W}_r \mathbf{W}_r')/m_r$$

Estimate parameters using REML and predict using BLUP

Cross validation study:

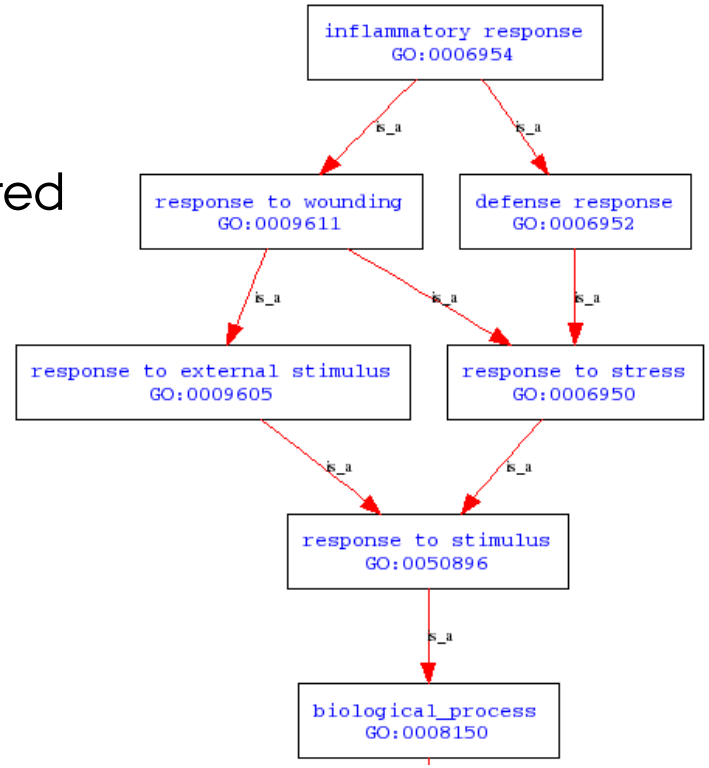
- assess predictive ability of statistical model

- 10% test and 90% train
- predictions across lines
- repeated 50 times

Genomic Feature BLUP Model

Genomic Features defined by **Gene Ontology (GO)**

Gene grouped according to biological processes such as mitosis or immune response, that are accomplished by ordered assemblies of molecular functions



- map SNP to Gene to GO
- within open reading frame of gene
- 1161 SNP sets (genomic feature sets)

Genomic Feature BLUP Model

Table 1 Ten most significant predictions for chill coma recovery with the GFBLUP model.

Sex	GO id[1]	PA[2*] ± SEM[3]	padj[4]	LR[5]	$\hat{h}_f^{2}$[6]	nsets[7]	Gene Ontology term
Female	GO:0007266	0.370±0.022	1.8×10^{-10}	11.39	0.31	3139	Rho protein signal transduction
	GO:0005100	0.365±0.023	4.0×10^{-10}	12.67	0.37	3886	Rho GTPase activator activity
	GO:0007173	0.343±0.026	1.9×10^{-8}	15.96	0.92	9674	Epidermal growth factor receptor signaling pathway
	GO:0030031	0.318±0.027	6.7×10^{-7}	11.49	0.74	2700	Cell projection assembly
	GO:0035160	0.309±0.027	1.5×10^{-6}	9.65	0.39	5011	Maintenance of epithelial integrity. open tracheal system
	GO:0016323	0.299±0.026	2.7×10^{-6}	9.13	0.47	7761	Basolateral plasma membrane
	GO:0035277	0.280±0.027	2.3×10^{-5}	11.12	0.56	7582	Spiracle morphogenesis. open tracheal system
	GO:0007494	0.263±0.025	5.8×10^{-5}	8.86	0.58	9614	Midgut development
	GO:0006406	0.288±0.033	8.3×10^{-5}	12.53	0.52	1530	mRNA export from nucleus
	GO:0005089	0.253±0.026	2.2×10^{-4}	9.54	0.67	11922	Rho guanyl-nucleotide exchange factor activity

- 32 GO terms in females and 16 in males had predictive values that were significantly different from GBLUP model
- GO terms ‘Rho protein signal transduction’ (GO:0007266) and ‘Rho GTPase activator activity’ (GO:0005100) had the highest prediction accuracies for male and female chill coma recovery time.
- Evidence from previous studies suggest several ways in which Rho genes may functionally affect the time to recover from a chill-induced coma.

Genomic Feature BLUP Models

Predictions in GFBLUP

GBLUP same weights on genomic relationships

$$\mathbf{g}_{\text{pred}} = \left(\mathbf{G}_{\text{vt}} \cdot \sigma^2_{\text{g}}\right)\left[\left(\mathbf{G}_{\text{tt}} \cdot \sigma^2_{\text{g}}\right) + \mathbf{I} \cdot \sigma^2_{\text{e}}\right]^{-1}\left(\mathbf{y}_{\text{b}} - \mathbf{Xb}\right)$$

$$\mathbf{g}_{\text{pred}} = \left(\mathbf{G}_{\text{f}_{\text{vt}}} \cdot \sigma^2_{\text{f}} + \mathbf{G}_{\text{r}_{\text{vt}}} \cdot \sigma^2_{\text{r}}\right)\left[\mathbf{G}_{\text{f}_{\text{tt}}} \cdot \sigma^2_{\text{f}} + \mathbf{G}_{\text{r}_{\text{tt}}} \cdot \sigma^2_{\text{r}} + \mathbf{I} \cdot \sigma^2_{\text{e}}\right]^{-1}\left(\mathbf{y}_{\text{b}} - \mathbf{Xb}\right)$$

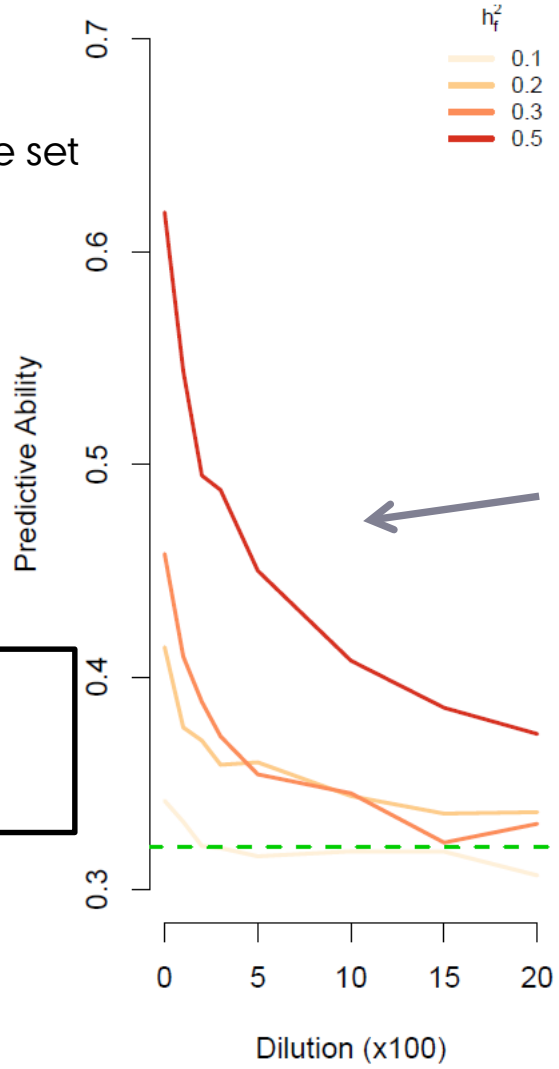
GFBLUP different weights on genomic relationships
=> differential shrinkage

Genomic Feature BLUP Models

Effect of dilution
- adding non-causal SNPs to feature set

Cross validation
- 10% test and 90% train
- predictions across lines
- repeated 50 times

$$PA = \frac{1}{50}\sum_{i=1}^{50} Corr(y_{pred}, y_{obs})$$



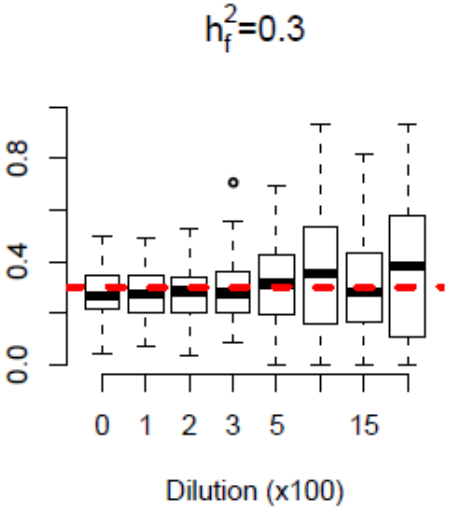
$$h_f^2 = \frac{\sigma_f^2}{\sigma_f^2 + \sigma_r^2}$$

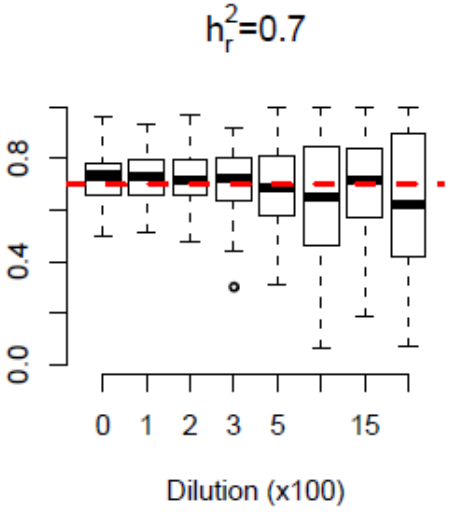
GFBLUP can increase accuracy

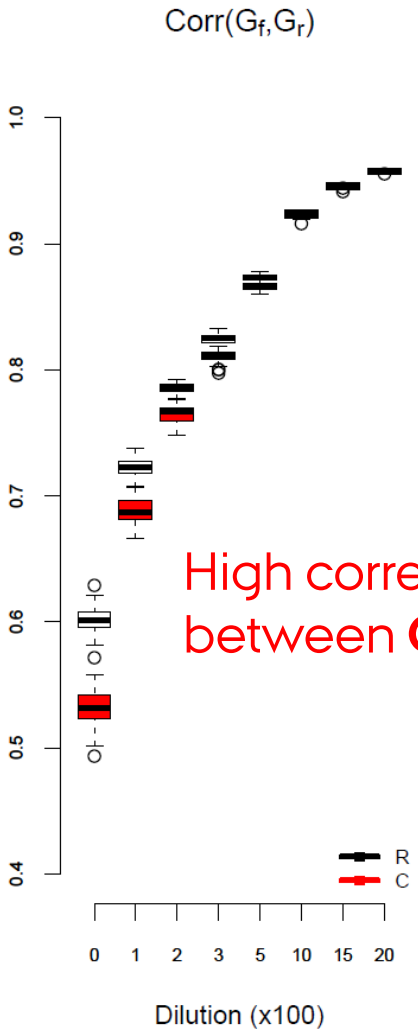
GBLUP

Genomic Feature BLUP Models

If “diluted” difficult to reliably estimate genomic parameters => lower accuracy







High correlation between $\mathbf{G}_f$ and $\mathbf{G}_r$

$$h_f^2 = \frac{\sigma_f^2}{\sigma_f^2 + \sigma_r^2}$$

$$h_r^2 = \frac{\sigma_r^2}{\sigma_f^2 + \sigma_r^2}$$

Genomic Feature BLUP Models

GFBLUP models can increase the accuracy of genomic predictions in populations of "unrelated" individuals.

- **DGRP:** increase prediction accuracy for 3 quantitative traits (startle response, starvation resistance and chill coma recovery)

- **Simulations:** increase prediction accuracy provided the genomic features are enriched for causal variants

GFBLUP (and Bayesian GF) models are computationally intensive

Need computationally fast and powerful modelling approaches
- allow us to rapidly analyze many different layers of genomic features

Genomic BLUP Set Test

- Searching for patterns in GBLUP-derived single-marker statistics, by including them in genetic marker set tests, that could reveal associations between a set of genetic markers (genomic feature) and a complex trait.

- Can GBLUP-derived single-marker statistics be used to develop better GFBLUP models?

- Simulation setup similar to one used to evaluate GFBLUP

Genomic BLUP Set Test

Step 1: Fit a single one component (or multi component) linear mixed model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{g} + \mathbf{e}$$

Step 2: Backsolve to get single markers effects and test statistics:

$$\hat{\mathbf{s}} = \mathbf{W}'(\mathbf{W}\mathbf{W}')^{-1}\hat{\mathbf{g}}$$

$$\mathbf{t}_{\hat{\mathbf{s}}_\mathbf{j}} = \frac{\hat{\mathbf{s}}_\mathbf{j}^2}{\mathbf{Var}(\hat{\mathbf{s}}_\mathbf{j})}$$

Step 3: For each genomic feature compute a summary statistic from the single marker statistics within feature

Genomic BLUP Set Test

Summary statistic from the single marker statistics such as:

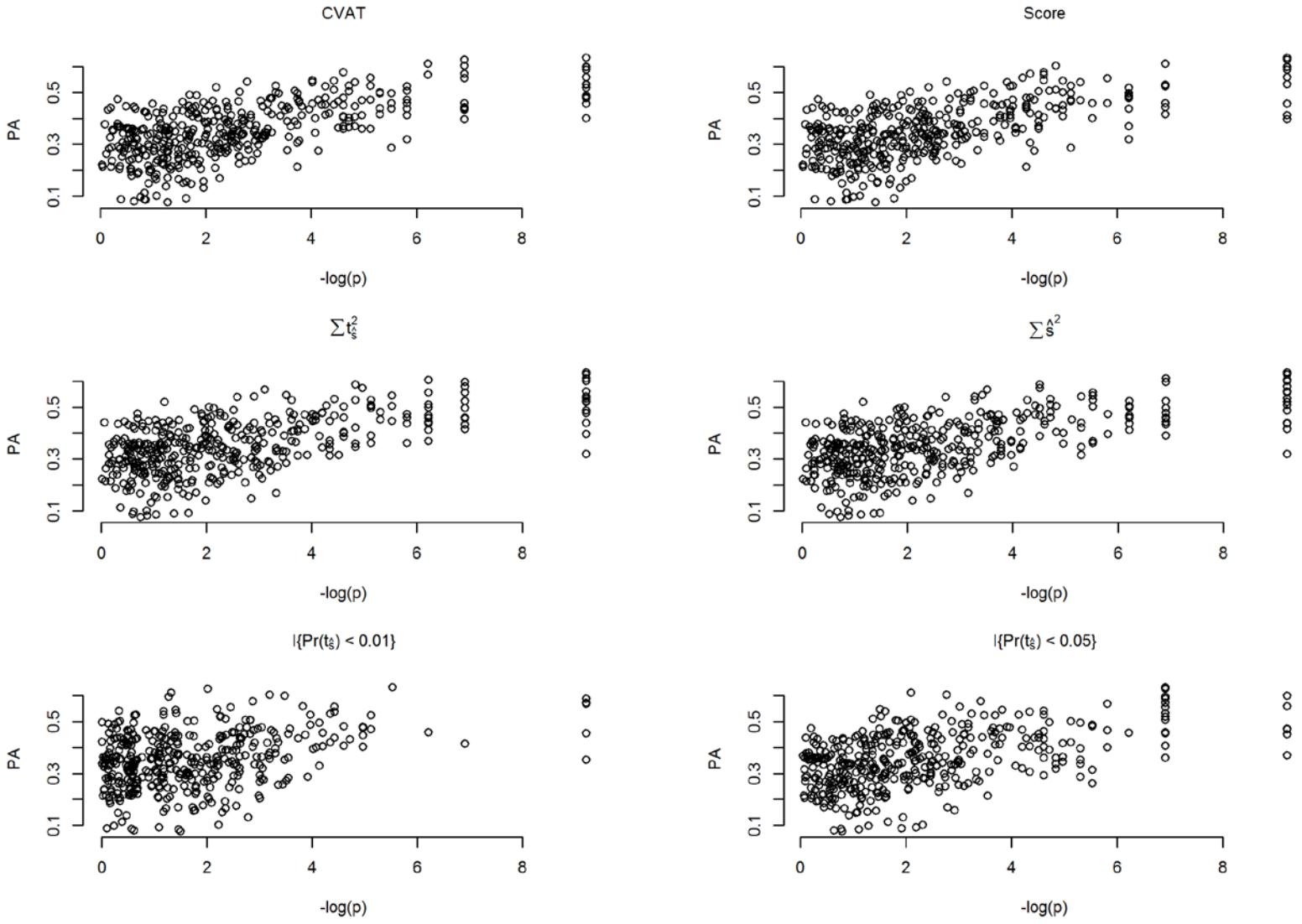
Count number of single marker test statistics above a certain threshold

$$T_{count} = \sum_{i=1}^{n_F} I(t_i > t_0)$$

Sum of all single marker test statistics

$$T_{sum} = \sum_{i=1}^{n_F} t_i$$

Genomic BLUP Set Test



Genomic BLUP Set Test

- There are patterns in GBLUP-derived single-marker statistics that can reveal associations between a set of genetic markers (genomic feature) and a complex trait.

- Improved inference and prediction accuracy of GFBLUP may be achieved by identifying genomic regions enriched for causal genetic variants.

Summary

- **GFBLUP models** can increase the accuracy of genomic predictions in populations of "unrelated" individuals.
- **GBLUP derived marker set test** can reveal genomic features associated with complex traits

- **Flexible modeling framework**
  - GFBLUP model analyses in dairy cattle (Fang et al. 2017)
  - GFBLUP model analyses in pigs (Sarup et al. 2016)
  - GFBLUP model analyses in humans (Rohde et al.; Sørensen et al.)
  - GFBLUP model using non-additive gene actions (Morgante et al.)

- **Computational intensive**
  - optimal computing strategy depends on structure of data
  - efficient algorithms (e.g. AI-REML)
  - multi-core, multi-node procedures

R package qgg: psoerensen.github.io/qgg/

Acknowledgements

Palle Duun Rohde

Stefan McKinnon Høj-Edwards

Pernille Sarup

Lingzhao Fang

Izel Fourie Sørensen

Trudy Mackay NC State

Fabio Morgante, NC State

Danish Strategic Research Council

GenSAP: Centre for Genomic Selection in Animals and Plants, contract no. 12-132452