Genomic Feature Prediction Models
- improved prediction accuracy by joint marker estimation

Palle Duun Rohde[1,*], Izel Fourie Sørensen[1] and Peter Sørensen[1]

[1]*Centre for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, Denmark.*
[*]*palle.d.rohde@mbg.au.dk*

GENOMIC FEATURE PREDICTION MODELS

Utilising prior biological information to create sets of SNP markers, i.e. genomic features (**Fig. 1**), can increase the accuracy of genomic predictions if the genomic feature is enriched for trait specific causal variants.

Genomic features can be genes, pathways, protein interactions, gene expression data, prior SNP associations etc. (**Fig. 1**).

Here, we show that when we jointly estimate marker effects, we obtain increased predictive performance of genomic feature models for human standing height (**Fig. 2**).

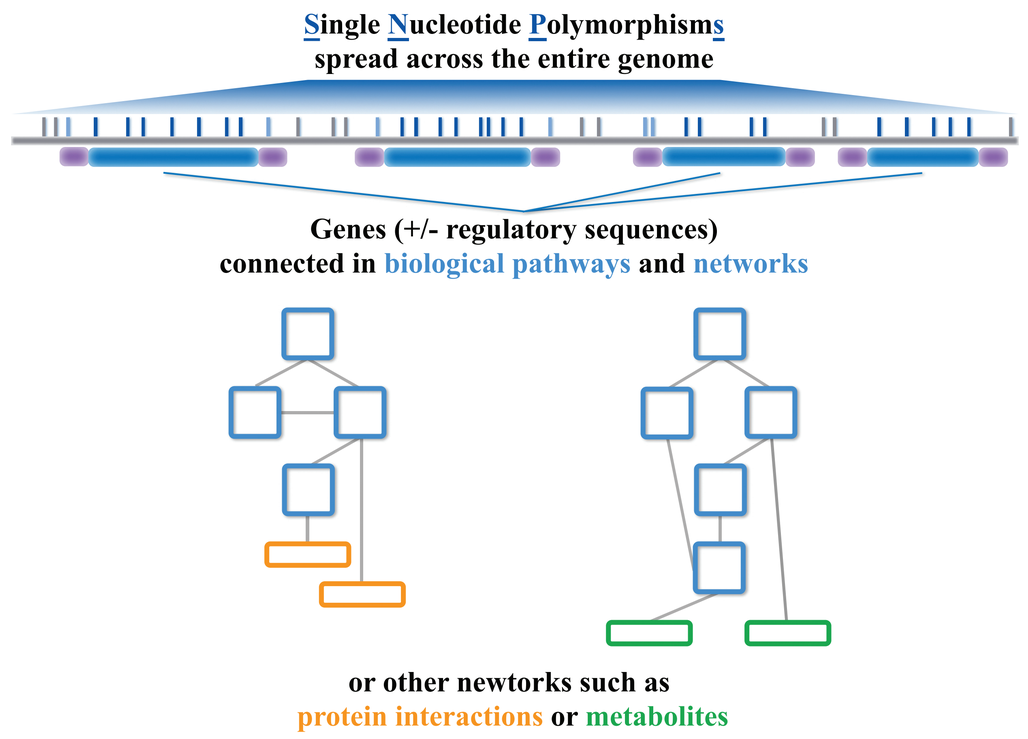

Fig. 1 Conceptual illustration of genomic features.

POLYGENIC RISK SCORE PREDICTION

Polygenic risk scores (PRS) are computed as,

$$\text{PRS} = \sum_{i=1}^{m} \mathbf{W}_i \hat{s}_i$$

where $\mathbf{W}_i$ is the centred and scaled genotype and $\hat{s}_i$ is the weight of the *i*-th SNP marker.

Commonly, the weight coefficients ($\hat{s}_i$) are obtained as **marginal effects** from single marker association models.

Alternatively, the weight coefficients can be obtained as **joint effects** from a multi-marker model, fitting all markers simultaneously.

The multi-marker model was based on a linear mixed model, implemented using a **Gauss Seidel residual update** (GSRU) algorithm.

We applied both methods to human standing height from the **UK Biobank** (335K individuals, white British unrelated). In both scenarios, marker estimation was done on a training population, and accuracy of PRS was assessed in the validation population.

For the **marginal effect analysis** SNPs were pruned for LD such that PRS was estimated using SNPs with r^2<0.7.

For the **joint effect analysis** top 30.000 SNPs from the marginal analysis were selected, and their effects re-estimated using GSRU.

INCREASED PREDICTIVE PERFORMANCE

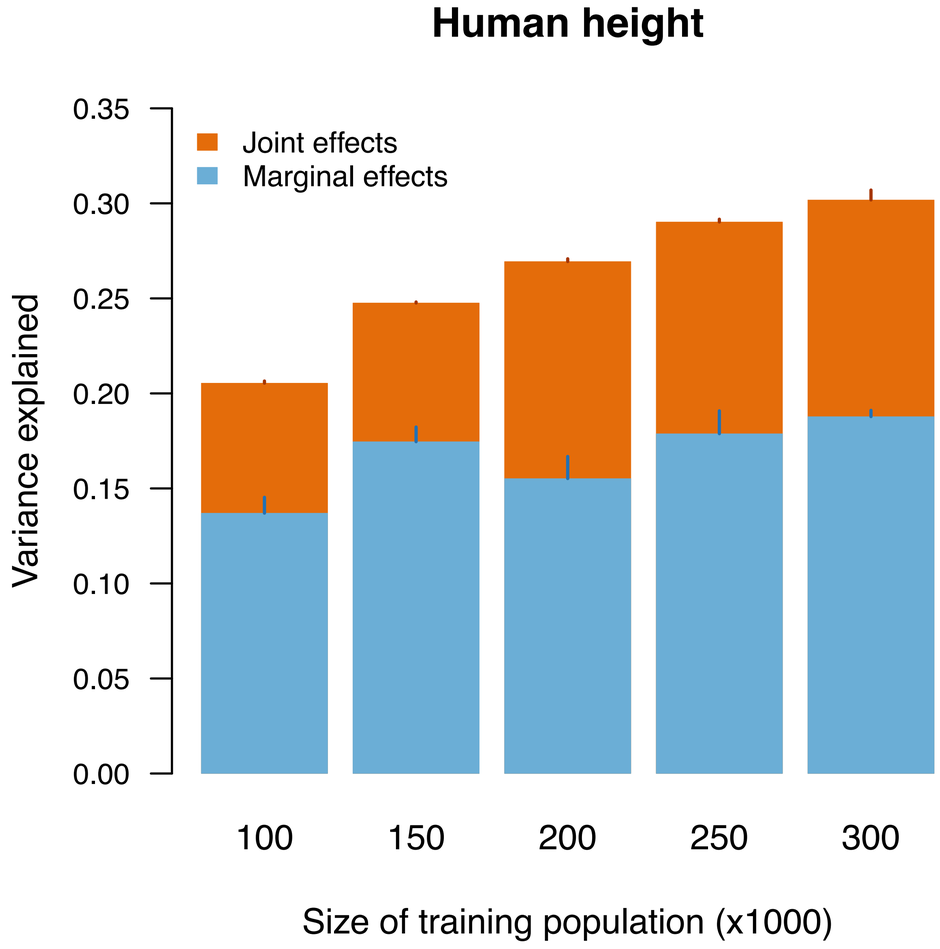

Fig. 2 Mean (+ standard error) variance explained (R^2) in validation population for PRS prediction utilising SNP weights from either marginal (blue) or joint estimation (orange).

CONCLUSION

When predictions were based on jointly estimated marker effects prediction accuracy (R^2 - variance explained) was up to 50% higher as compared to predictions based on marginal effects.

SOFTWARE

All analyses were done using the **R-package, *qgg***. This package provides a range of genomic feature modelling approaches designed to handle large-scale data.

DOWNLOAD

psoerensen.github.io/qgg/

Authors declared no competing interests

AARHUS UNIVERSITY
InnovationsFonden FORSKNING, TEKNOLOGI & VÆKST I DANMARK
GenSAP