Statistical perspectives on dependencies between genomic markers

zur Erlangung des akademischen Grades
doctor rerum naturalium habilitata (Dr. rer. nat. habil.)

an der Agrar- und Umweltwissenschaftlichen Fakultät
der Universität Rostock

vorgelegt von Dr. rer. nat. Dörte Wittenburg
aus Rostock

Verteidigung am 12. November 2021
1. Modelling of non-additive genetic effects by means of dominance is more important than epistasis for the accuracy of predicted breeding values or genetic values (Wittenburg et al., 2011).

2. A GBLUP approach can be adequately extended to jointly consider breeding values and dominance deviations (Wittenburg et al., 2015; Wittenburg et al., 2013).

3. Genetic effects captured by markers can be estimated with high precision and significance of billions of additive and non-additive marker effects can be tested in a reasonable amount of computing time (Wittenburg and Liebscher, 2018).

4. Only if the number of causal variants is low, and so the linkage disequilibrium between them, the bias of estimating genetic variance components with a classical model parameterisation is negligible (Wittenburg et al., 2011).

5. Dependencies between marker genotypes need to be taken into account in a statistical model for precise genomic evaluations (Wittenburg et al., 2020; Wittenburg et al., 2016).

6. Non-random mating influences the extent of dependence between marker genotypes (Wittenburg et al., 2020; Wittenburg et al., 2016).

7. The design of an experiment for fine-mapping of causal variants needs to incorporate dependencies between markers for calculating the minimum required sample size (Wittenburg et al., 2020).

8. Negative correlation between marker genotypes leads to an essentially inflated sample size in the design of fine-mapping experiments of two or more causal variants (Wittenburg et al., 2020).

9. Markers can be grouped with respect to the extent of dependence between them (Wittenburg et al., 2021).

10. Though the bovine genome assembly has high quality, putatively misplaced markers exist (Qanbari and Wittenburg, 2020).

11. Progeny genotypes in half-sib families are a valuable source for estimating recombination rate between pairs of markers, for assessing putatively misplaced markers in the genome assembly and for deriving a genetic map of marker positions (Hampel et al., 2018; Qanbari and Wittenburg, 2020).

12. Providing software solutions (R packages, Fortran programs) and scripts for data analysis ensures reproducibility of research.
Publications


Summary

My thesis is dedicated to statistical methods which have proved useful for high dimensional data analysis in genomic evaluations. The intention of such evaluations is to explain the association between genetic and phenotypic variation or to explore associations among molecular markers.

Molecular markers are used as predictor variables in a statistical model to investigate the genetic impact on a quantitative trait. Genetic effects are assumed to be attributable to different kinds of gene action which can be additive or non-additive. For a proper statistical treatment, the genotype codes for different kinds of genetic effects need to be statistically parameterised in such a way that the different sources of genetic variation constitute an orthogonal decomposition of the total genetic variation. In my studies, the accuracy of breeding value estimation was hardly affected by inclusion or exclusion of non-additive genetic effects though the accuracy of total genetic values could improve with non-additive effects considered. The identification of causal variants or genome regions associated with trait expression was difficult when interaction effects were included. I developed a testing approach to identify at least those markers with strongest additive, dominant or any kind of pairwise interaction effect in a computationally rapid manner. Though there is no direct link between statistical effect and gene action, the resulting significant markers give rise to be further studied for their biological relevance using specific breeding designs.

The orthogonal decomposition of genetic variation is actually only possible in the absence of linkage disequilibrium between markers which is not a realistic assumption. I studied the dependence between markers in populations with family stratification, starting with half-sib families which is a typical family structure in livestock and generalising the methodology to full-sib families. The covariance between SNPs was analytically derived with respect to the genotype codes for the additive effect of SNP loci. The resulting covariance matrix deserved special attention due to its multi-functionality in several fields of genomic evaluations. First, it was used to incorporate prior knowledge about the distribution of marker effects in regression models for a genome-wide association analysis.

Second, this matrix was key to design future experiments for fine-mapping loci that are associated with trait expression based on dense marker data. Knowing the genetic information of selected mates, the number of progeny could be determined to guarantee a certain power of association analysis later on.

Third, markers were grouped depending on the extent of covariance/correlation. Grouping is a universal strategy to cope with multicollinearity in genomic evaluations due to the (partly) tight linkage of markers. Groups of highly associated markers may be employed in a penalised regression approach that allows to differentiate genomic regions with impact on trait expression from neutral regions. Such an approach can be extended to also include non-additive genetic effects.
Additionally knowing the covariance between genotype codes for the dominance effect of SNP loci enabled the improvement of a likelihood-based approach for estimating paternal recombination rates and maternal linkage disequilibrium. This was a very thankful approach as it allowed the estimation of recombination rate between any SNP pair. Not only were recombination rates required to set up the covariance matrix as outlined above, they were also important for the derivation of a genetic map. The genetic distance between markers was estimated from recombination rates that were not restricted to neighbouring SNPs. Furthermore, SNPs that were putatively misplaced in the underlying genome assembly could identified from unusually large recombination rates to other SNPs in close proximity. The likelihood-based approach was applied to large-scale cattle data and thus the outcome contributed to further improving the bovine genome assembly.

This thesis clearly showed that the dependence among SNPs can be controlled from a statistical point of view and exploited for beneficial use in genomic evaluations.