QUANTIFYING LD DECAY BY QUANTILE REGRESSION A CASE STUDY

SABINE K.SCHNABEL[1], FEDERICO TORRETTA[2] AND MATTHIAS WESTHUES[3]

1: Biometris, Wageningen University and Research Centre, The Netherlands; 2: Università di Palermo, Italy; 3: Universität Hohenheim, Germany

INTRODUCTION

- Genome-wide association studies: great tool for the localization of QTLs (quantitative trait loci) in plant and animal breeding programs.
- Investigation of the genetic relatedness (kinship matrix) required for powerful GWAS

→ Insight into LD between genetic markers necessary (LISTGARTEN *et al.*, 2012)

Figure 1(A)

- Find suitable set of independent markers
- Exploration of LD decay over the whole genome
- LD is commonly measured in terms of the squared Pearson correlation coefficient R^2 between pairs of genetic markers (HILL and ROBERTSON, 1968).

We are proposing the following approach:

- Describe the steps

DATA AND VISUALIZATION

- Data from Maize population (FISCHER *et al.*, 2008), especially from Chromosome 1 with almost 5000 markers → more than 12 million pairwise comparisons
- For these large data: visualization is difficult in a scatterplot.
- Apply a scatterplot smoother (EILERS and GOEMAN, 2004)

→ computation of a two-dimensional histogram, smoothing of the counts and display with a color map

- In order to improve the quality of the fit and the visualization → use of $\sqrt{|R|}$ instead of Pearson's R^2.

QUANTILE REGRESSION

- Using non-parametric quantile regression with a monotonicity constraint (BOLLAERTS *et al.*, 2006; MUGGEO *et al.*, 2013)
- Monotone decreasing curve is in line with biological assumptions.
- $\mu_\tau = s_\tau(d)$, μ_τ quantile function at percentile τ, d SNP distance between pairs of markers and $s_\tau(\cdot)$ smooth and unknown function
- $P-$splines for a smooth functional form, therefore:
  $\min \sum_k^K b_{k\tau} B_k(d)$ subject to $b_k < b_{k-1}$ for $k = 2, \ldots, K$, with $b_{k\tau}$ the coefficient of the B_k-th spline, K the dimension of the design matrix.

GLOBAL/LOCAL LD

In our case, it is of interest to examine what happens to LD decay on a smaller scale. Therefore we investigate local LD decay in subsequent overlapping sliding windows of 2.5 Mbp width and fit a set of quantile curves to each of these sections of the whole plot.

- Description local LD

RESULTS

Figure 1B

Some text

Figure 1C,D

CONCLUSION AND DISCUSSION

- Case study of how to explore and quantify local LD decay patterns in Maize using quantile regression with monotonicity constraints for a first summary of the LD decay.
- Applying $P-$splines to smooth the median local LD decay → easy to interpret and inspect for the collaborating biologists
- In addition to a good tool to quantify local LD decay → also an instrument in identifying problems with the underlying genotypic data that have previously been overlooked.

In this sense they can serve as a diagnostic tool. On the one hand we discovered sliding windows with low sample sizes which suggests undercoverage in certain distances in LD decay. While fitting the smooth curves in Figure **??** (D), we observed a noticeable clustering in terms of correlation values in some of the subsets of the data. This phenomenon was unknown to date in this data set and has lead to adjustments in subsequent data analyses.
More results from this case study will be reported elsewhere.

ACKNOWLEDGEMENTS

This case study was performed while FT and MW were visiting at Biometris at Wageningen University and Research Centre in Winter 2014/2015. We are indebted to the group of Prof. Dr. Ruedi Fries, from Technische Universität München, for the SNP genotyping of the parental lines, which was funded by the German Federal Ministry of Education and Research (BMBF) within the AgroClustEr "Synbreed-Synergistic plant and animal breeding" (FKZ:0315528d).

WAGENINGEN UNIVERSITY
WAGENINGEN UR

UNIVERSITÀ DEGLI STUDI DI PALERMO

UNIVERSITY OF HOHENHEIM

REFERENCES

BOLLAERTS, K., P. H. C. EILERS, and M. AERTS, 2006 Quantile regression with monotonicity restrictions using p-splines and the l_1-norm. Statistical Modelling **6**: 189–207.

EILERS, P. H. C. and J. J. GOEMAN, 2004 Enhancing scatterplots with smoothed densities. Bioinformatics **20**: 623–628.

FISCHER, S., J. MÖHRING, C. C. SCHÖN, H.-P. PIEPHO, D. KLEIN, W. SCHIPPRACK, H. F. UTZ, A. E. MELCHINGER, and J. C. REIF, 2008 Trends in genetic variance components during 30 years of hybrid maize breeding at the university of hohenheim. Plant Breeding **127**: 446–451.

HILL, W. and A. ROBERTSON, 1968 Linkage disequilibrium in finite populations. Theoretical and Applied Genetics **38**: 226–231.

LISTGARTEN, J., C. LIPPERT, C. KADIE, R. DAVIDSON, E. ESKIN, and D. HECKERMAN, 2012 Improved linear mixed models for genome-wide association studies. Nature Methods **9**: 525–526.

MUGGEO, V., M. SCIANDRA, A. TOMASELLO, and S. CALVO, 2013 Estimating growth charts via nonparametric quantile regression: a practical framework with application in ecology. Environmental and Ecological Statistics **20**: 519–531.