Quantifying LD decay by quantile regression
A case study

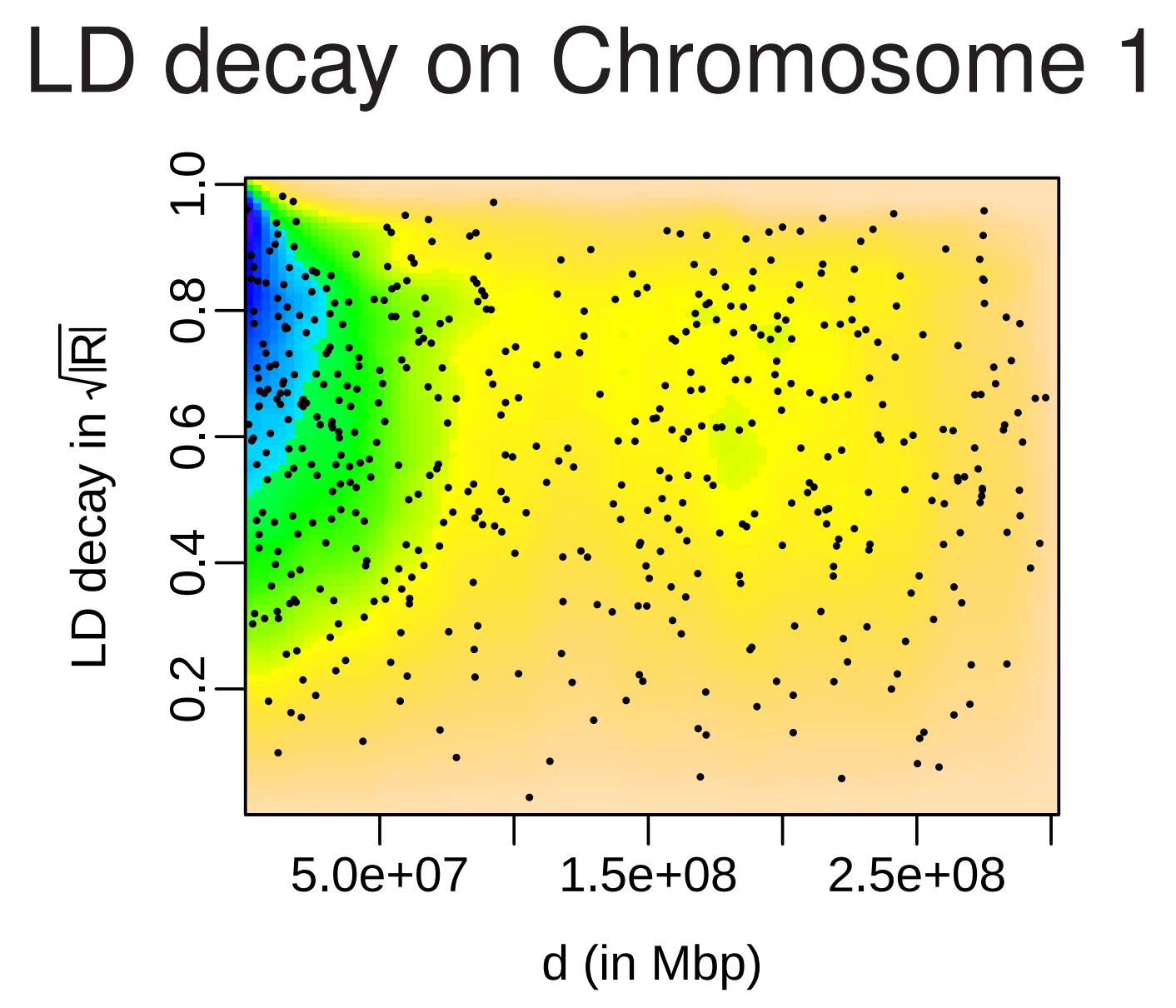
Sabine K.Schnabel[1], Federico Torretta[2] and Matthias Westhues[3]

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)


LD decay on Chromosome 1

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

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 $b_{k\tau}$ coefficient of the B_k-th spline of quantile τ, K dimension of the design matrix.

Global/local LD

- Usually analysis focusses on the global LD decay (per chromosome) → general picture about the linkage desequilibrium and linkage between the markers
- Here: emphasis on local LD decay to get more insight on a smaller scale
  - Overlapping sliding windows of 2.5 Mbp → around 1000 windows on chromosome 1 with on average 2000 points per window
  - Fit a set of quantile curves to each of the windows (here $\tau = 0.25, 0.5, 075$)
  - Choose threshold T in terms of R^2 and collect the associated distances $\Delta_.$ from whereon LD decay is lower than T

Results

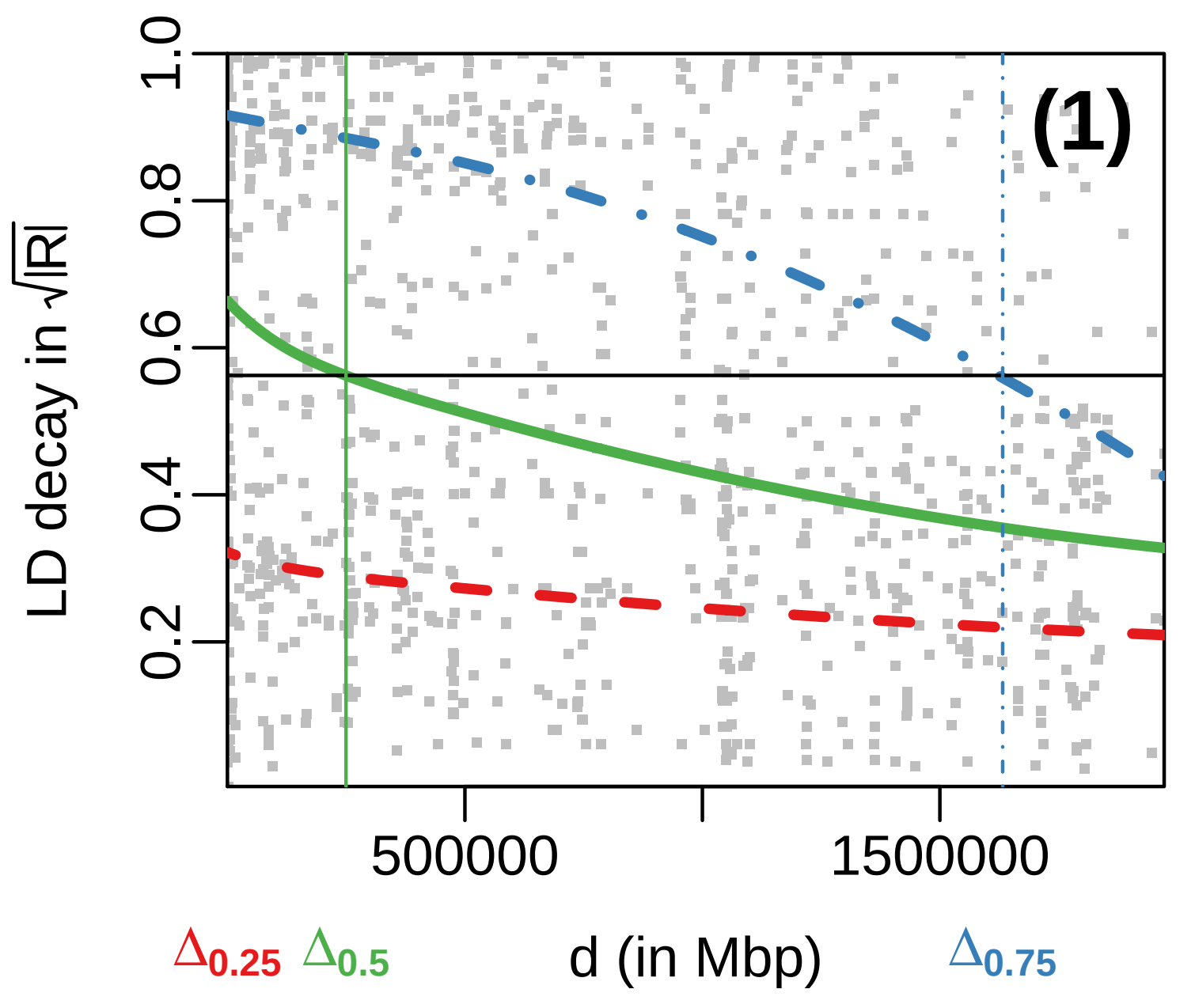
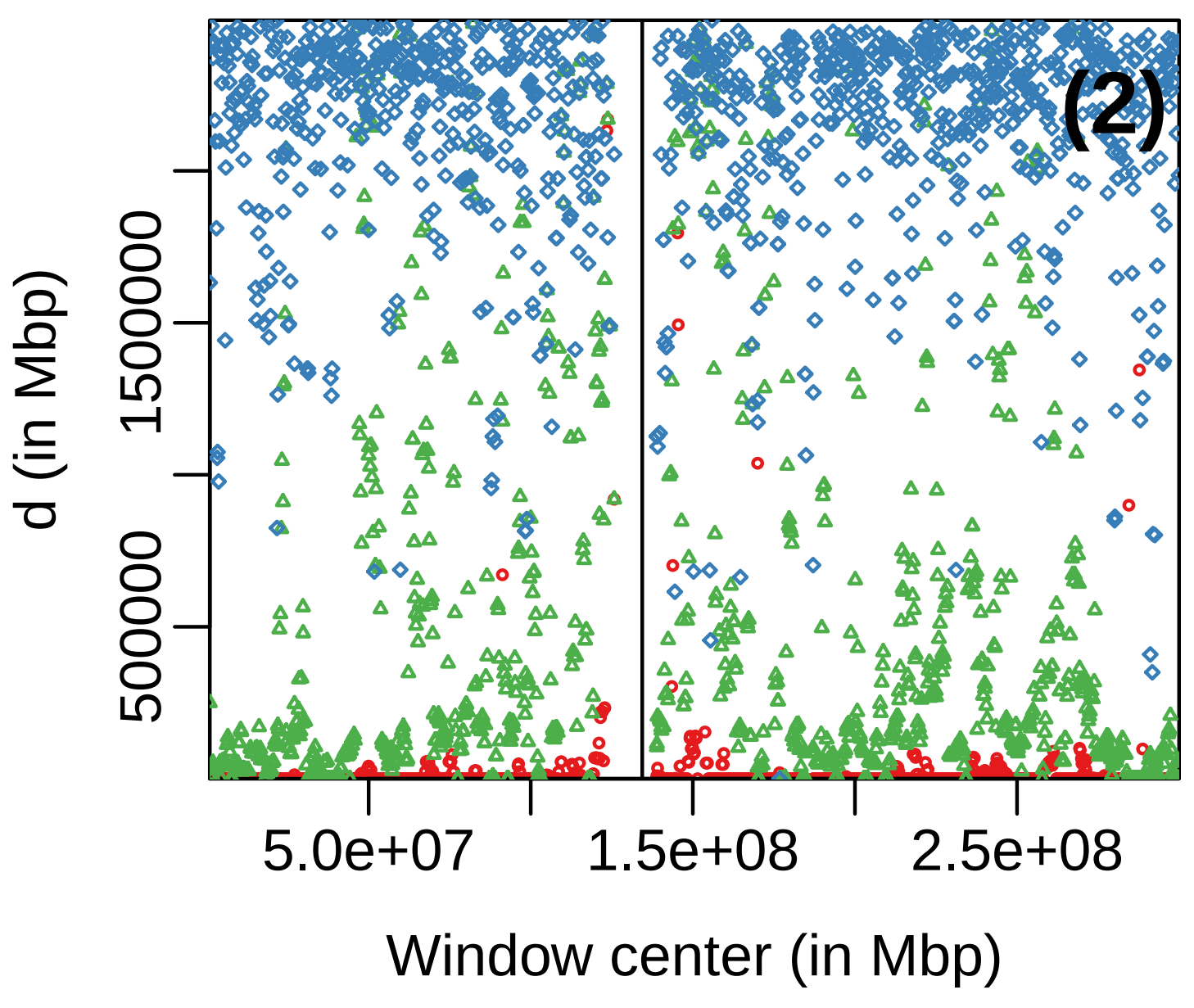
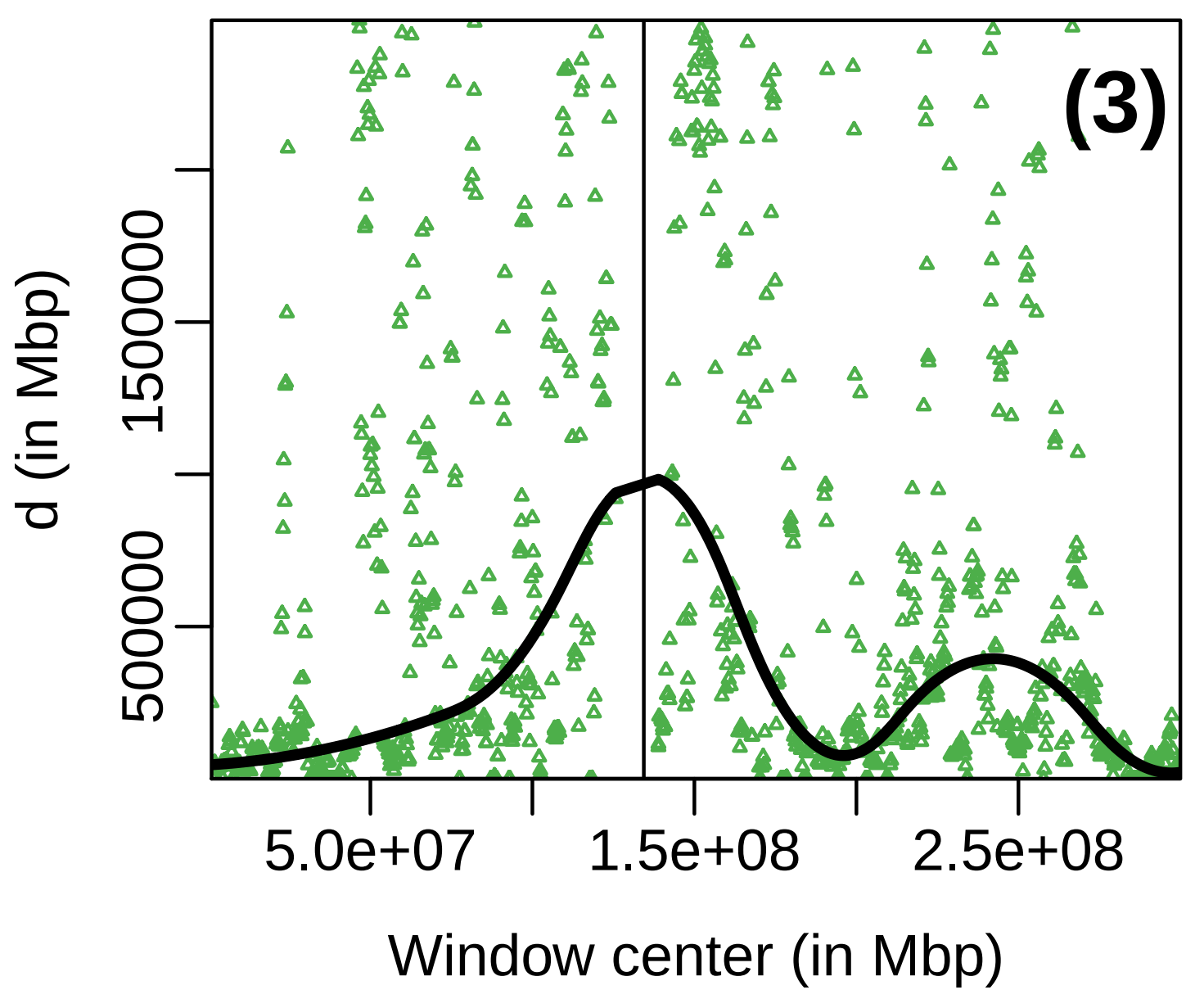


Figure (1) Quantile curves for $\tau = 0.25, 0.5, 0.75$ for a subsample. The distance associated to threshold T is indicated at $\Delta_{0.25,0.5,0.75}$.

Figure (2) Collection of $\Delta_{0.25,0.5,0.75}$ (with indication of centromere)

Figure (3) Smooth fit to $\Delta_{0.5}$





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 depth exploration of local LD decay (in comparison to global LD decay) leads to new insight.
- 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.
- Can serve as a diagnostic tool
  - Discovering of undercoverage through sliding windows with low sample sizes
  - Clustering of correlation values → unknown phenomenon in the data, adjustment in subsequent analysis

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.

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, Technische Universität München, for the SNP genotyping of the parental lines, funded by the German Federal Ministry of Education and Research (BMBF) within the AgroClustEr "Synbreed-Synergistic plant and animal breeding" (FKZ:0315528d).