# 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

E-mail for correspondence: `sabine.schnabel@wur.nl`

**Abstract:** Through recent developments in genotyping (e.g. of plant populations) more information on genetic markers become available. In order to perform powerful genome-wide association studies (GWAS) it is important to analyse linkage disequilibrium (LD) through pairwise comparisons of genetic markers. Large numbers of these markers pose new problems in terms of analysis and visualization. In a case study for Maize we explore and quantify LD decay using monotone quantile regression.

**Keywords:** Quantile regression; smoothing; monotonicity; linkage; LD decay

## 1 Introduction and Motivation

Genome-wide association studies have emerged as a great tool for the localization of QTLs (quantitative trait loci) in plant and animal breeding programs. However, a crucial requirement to powerful GWAS is the investigation of the genetic relatedness (kinship matrix) (Astle and Balding, 2009). For an appropriate kinship matrix, insight into LD between genetic markers is necessary. This matrix is best based on a set of independent markers (Listgarten et al., 2012). To find such a set of suitable markers (e.g. single nucleotide polymorphism – SNP) we need to explore 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). As an example, we are using data from chromosome 1 of a Maize population consisting of 123 European Dent inbred lines and 114 European Flint inbred lines (Fischer et al., 2008). Figure 1(A) shows an LD decay plot for this part of the genome.

## 2 Analysis and Application

Chromosome 1 of the above described Maize genome has almost 5000 markers, resulting in more than 12 million pairwise comparisons between two markers on the same chromosome. Visualization of these large data sets is challenging. We used a scatterplot smoother (Eilers and Goeman, 2004) for depicting global LD decay on chromosome 1 (length ≈ 300 Mbp (Mega base pairs)) in Figure 1(A). As mentioned above, the most common measure for LD decay is the squared Pearson correlation $R^2$. In order to improve the quality of the fit as well as the visualization, we advocate to use $\sqrt{|R|}$ instead. Further investigation concerning this transformation will be reported elsewhere. 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. We use non-parametric quantile regression with a monotonicity constraint, $\mu_\tau = s_\tau(d)$, where $\mu_\tau$ is the quantile function at percentile $\tau$, $d$ is the SNP distance between pairs of markers and $s_\tau(\cdot)$ is a smooth and unknown function (Muggeo et al., 2013; Bollaerts et al., 2006). We choose $P-$splines for a smooth functional form. By imposing $b_k < b_{k-1}$, with $b_k$ as the coefficient of the $k$-th spline, a monotone decreasing curve is ensured (that is in line with the underlying biological assumptions). This analysis can be done for any quantile of interest. In Figure 1(B) quantile curves for $\tau = 0.25, 0.5, 0.75$ are plotted as well as a threshold in terms of LD decay (on the initial scale of $R^2$=0.1, therefore here at $\sqrt[4]{0.1} = 0.56$). For the exploration of local LD decay we might be interested in the distance $\Delta_{0.5}$ from which onwards the LD decay is falling beneath the threshold. This can be interpreted as the average distance within this window from which onwards two marker loci are considered to be independent of each other. Such two markers could, for example, be selected for the computation of a kinship matrix. In the example, local LD decay in terms of median distance at threshold $\sqrt[4]{0.1}$ is 249666 bp. On Chromosome 1 of this population we have about 1000 sliding windows of 2.5 Mbp width with an average of 2000 points falling into one window. The distances $\Delta_\tau$ for all sliding windows are collected and plotted in Figure 1(C) at the respective center of the window. We collect these data for $\tau$ values that are of interest. While these data points are an interesting result as such to quantify local LD decay, it is hard to judge by eye the relationship that is plotted in Figure 1(C). Therefore we used $P-$splines to fit a smooth curve to these results as displayed by the curve in Figure 1(D). For our example, we observe a bi-modal form of the relationship. The black vertical line in the graph indicates the so-called centromere. We have reason to believe that the left mode is around the centromere. However, this bi-modal phenomenon could not be observed for all of the 10 chromosomes in this data set.

## 3 Conclusion and Discussion

This is a case study of how to explore and quantify local LD decay patterns in Maize. We are using quantile regression with monotonicity constraints for a first summary of the LD decay. On top of that we are applying $P-$splines to smooth the median local LD decay. These curves are easier to interpret and to inspect for the collaborating biologists.

While the presented steps are a good tool to quantify local LD decay, they have also been instrumental 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 1 (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.

**Acknowledgments:** This case study was performed while the second and third author 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 SynbreedSynergistic plant and animal breeding (FKZ:0315528d).

### References

Astle, W. and Balding, D.J. (2009). Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statistical Science*, **24**, 451 − 471

Bollaerts, K., Eilers, P.H.C., and Aerts, M. (2006). Quantile regression with monotonicity restrictions using P-splines and the L1-norm. *Statistical Modelling*, **6**, 189 − 207

Bush, W.S. and J.H. Moore (2012). Chapter 11: Genome-wide association studies. *PLOS Computational Biology*, **8**, 1 − 11

Eilers, P.H.C. and Goeman, J.J. (2004). Enhancing scatterplots with smoothed density. *Bioinformatics*, **20**, 623 − 628.

Fischer, S., Möhring, J., Schön, C. C., Piepho, H.-P., Klein, D., Schipprack, W., Utz, H.F., Melchinger, A.E., and Reif, J.C. (2008). Trends in genetic variance components during 30years of hybrid maize breeding at the University of Hohenheim. *Plant Breeding*, **127**, 446 − 451

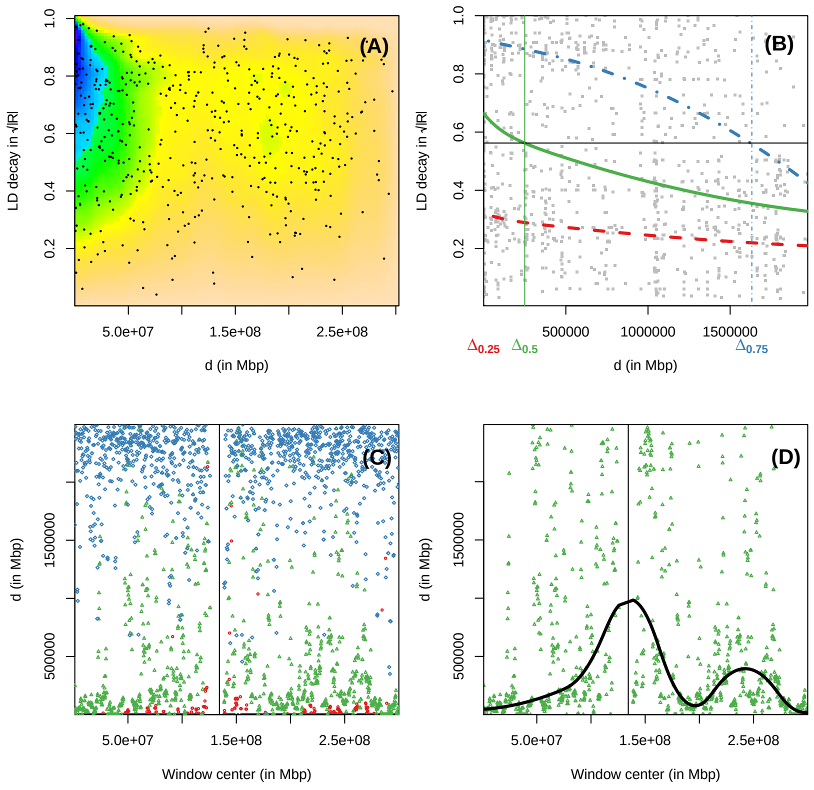

FIGURE 1. (A) Plot of LD decay for Chromosome 1 using `scattersmooth`. (B) Sample of data with threshold $R^2$=0.1($\sqrt{|R|} = 0.56$) and $\Delta_{0.25,0.5,0.75}$ shown. (C) Collection of $\Delta_{0.25,0.5,0.75}$ in Red/Green/Blue for Chromosome 1 at the center of the sliding window with indication of the centromere. (D) Smooth fit to $\Delta_{0.5}$.

Hill, W.G. and Robertson, A. (1968). Linkage Disequilibrium in Finite Populations. *Theoretical and Applied Genetics*, **38**, 226 − 231

Listgarten, J., Lippert, C., Kadie, C.M., Davidson, R.I., Eskin, E., and Heckerman, D. (2012). Improved linear mixed models for genome-wide association studies. *Nature Methods*, **9**, 525 − 526

Muggeo V. , Sciandra M., Tomasello A., and Calvo, S. (2013) Estimating growth charts via nonparametric quantile regression: a practical framework with application in Ecology. *Environmental and Ecological Statistics*, 20, 519 − 531.