2010
2011

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

Statistical analysis of selected yeast segregants

Promotor :
Prof. dr. Tomasz BURZYKOWSKI
De heer Jurgen CLAESEN

Leandro Garcia Barrado

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

2010
2011

FACULTY OF SCIENCES

Master of Statistics: Biostatistics

Masterproef

Statistical analysis of selected yeast segregants

Promotor :
Prof. dr. Tomasz BURZYKOWSKI
De heer Jurgen CLAESEN

Leandro Garcia Barrado

Master Thesis nominated to obtain the degree of Master of Statistics , specialization Biostatistics

Maastricht University

universiteit hasselt
UNIVERSITEIT VAN DE TOEKOMST

Acknowledgments

In an attempt to pay gratitude to everyone who has directly and in-directly helped or supported me in successfully finishing this master thesis, I would like to say thanks.

First and foremost I wish to say thank you to my two supervisors Prof. dr. Tomasz Burzykowski and Drs. Jürgen Claesen for their excellent guidance, ideas, critical comments and support. I really am grateful to them for giving me the opportunity of writing this master thesis in this exciting field of statistics.

In a second wave I would like to thank my girlfriend, parents, brother, family and friends, for putting up with me during this interesting journey. For showing interest in everything I do – this master thesis included – and the opportunity they gave me to commence this study.

Abstract

Introduction: *S. cerevisiae* is an economically important yeast strain because of its high-ethanol tolerance property. In order to develop strains expressing this characteristic even more, interest lies in identifying the genetic constitution of this trait. Based on bulk segregant high-resolution Quantitative Trait Loci-mapping, data is provided from which this constitution could be derived. This data set contains SNP frequencies of which higher frequencies possibly indicate regions of interest.

Methods: In order to get insight from the provided frequency data, wavelet shrinkage is used to smooth the data in order to find narrow chromosome regions of high-frequency. The wavelet shrinkage methodology shrinks the coefficients of the Discrete Wavelet Transform by a certain threshold in order to come up with a smoothed estimate of the curvature in the present data. Providing a wide variety of options; wavelet choice, threshold type and primary resolution level, the method can be fine-tuned to serve a specific purpose. Moreover, remedial measures to ensure assumptions are investigated, i.e. normalizing and variance stabilizing transformation and adaptive lifting.

Results: The results confirm that changing the different characteristics of the wavelet shrinkage approach provides different results concerning smoothness and shape of the estimated curves. Different wavelets lead to differently shaped and smoothed estimated curves, various threshold procedures lead to different thresholds which provide diverse levels of denoising and exists an inverse relationship between the primary resolution level and the extent of effect of threshold. Considering the *S. cerevisiae* data, no narrow chromosome position regions could be identified which could lead to an easy genetic constitution.

Conclusion: The present report acknowledges the versatility of the wavelet shrinkage approach in being applicable in a wide variety of cases through fine-tuning of its different characteristics. On the other hand, the possibly small signal-to-noise ratio contained in the data set did not enable the method to be successful in identifying some region of interest.

Keywords: Wavelet shrinkage, adaptive lifting, Bulk segregant QTL-mapping

Table of contents

1 Introduction

By virtue of its high-level ethanol tolerance trait, the yeast strain *Saccharomyces cerevisiae* is very important in several industrial fermentation processes. Its property is exploited, for example, in bio-ethanol production (Matsushika et al., 2009) and alcoholic beverages by fermentation such as beer and wine (Nevoigt et al., 2002). Thus, identification of the genetic constitution of this important trait enabling the development of increased high-ethanol tolerance strains would be economically very valuable.

Ethanol tolerance can be regarded as a genuine quantitative trait, controlled by many genetic elements, each contributing to it in varying ways. These genetic elements are called Quantitative Trait Loci (QTL) and comprise of specific parts of the genome – clustered or single genes – explaining some proportion of the variance observed in the quantitative trait. The interacting contributions in bringing forth ethanol tolerance in *S. cerevisiae* make it practically impossible to pinpoint these separate elements. In order to dissect the genetic basis of such quantitative traits, the Quantitative Trait Locus mapping approach can be adhered. This approach enables to genomically localize simultaneously the different loci contributing to a quantitative trait.

1.1 Quantitative Trait Locus Mapping

Quantitative Trait Locus mapping is based on the principles of meiotic recombination, i.e. the mechanism responsible for mixing genomic sequences during reproduction. Especially the inverse relation between the absolute distance separating two loci on the same chromosome and their recombination frequency, makes QTL-mapping possible. In short, the closer loci are positioned next to each other, the higher the probability of them being co-segregated (Lynch and Walsh, 1998). Based on this principle, QTL-mapping is the analysis of the extent of co-segregation of QTLs of which the positions are unknown (the loci of interest) and loci with positions which are known (genetic markers) in order to locate the QTLs of interest in the genome. Genetic markers represent genetic differences between individuals of an organism that are inherited in a Mendelian way and can therefore easily be followed over generations. Among the most widely used genetic markers are the molecular markers Single Nucleotide Polymorphisms (SNPs) (Strachan and Read, 1999).

Applied to *S. cerevisiae*, QTL-mapping is performed as illustrated in Figure 1. In a first step, two strains, the target strain, expressing the quantitative trait of interest, and the reference strain, lacking the quantitative trait of interest, are crossed. After mating and sporulation, haploid segregants of different genetical composition – a result of the aforementioned meiotic recombination – are observed. Next, all these segregants are phenotyped and only those expressing the trait are retained. After the remaining segregants are genotyped for the genetic markers, the proportions of co-segregation of the genetic markers with the quantitative trait of interest are used to infer the position of the QTL of interest. Because of the inverse relation between absolute genetic distance and co-segregation, the higher the proportion of selected segregants carrying the same genetic marker, the closer these markers are to the QTL of interest.

Applicability of QTL-mapping was, until recently, still limited due to low availability of molecular markers and the laboriousness of their genotyping, causing QTL to have low mapping resolution (Swinnen, 2011). The use of artificial markers as alternative for natural markers (Swinnen, 2011) together with recent advances in genotyping technologies allows for the simultaneous genotyping of

thousands of molecular markers. This leads to rapid, cost-effective methods to analyze genetic variation between strains. Winzeler et al. (1998) for example, proposed to use High-density oligonucleotide arrays, enabling high-resolution QTL-mapping.

Of course, to have higher genome coverage and mapping resolution, i.e. having many genetic markers densely spread over the whole genome, it is advisable to simultaneously identify and genotype thousands and thousands of markers. This could become laborious and expensive to do for each individual retained segregant. In order to make such an approach more feasible, bulk segregant analysis is a possible solution (Brauer et al., 2006). By mixing the genomic DNA from all of the selected segregants before applying high-throughput high-resolution QTL-mapping to the pool, bulk segregant analysis avoids the time and costs disadvantages of QTL-mapping each individual segregant.

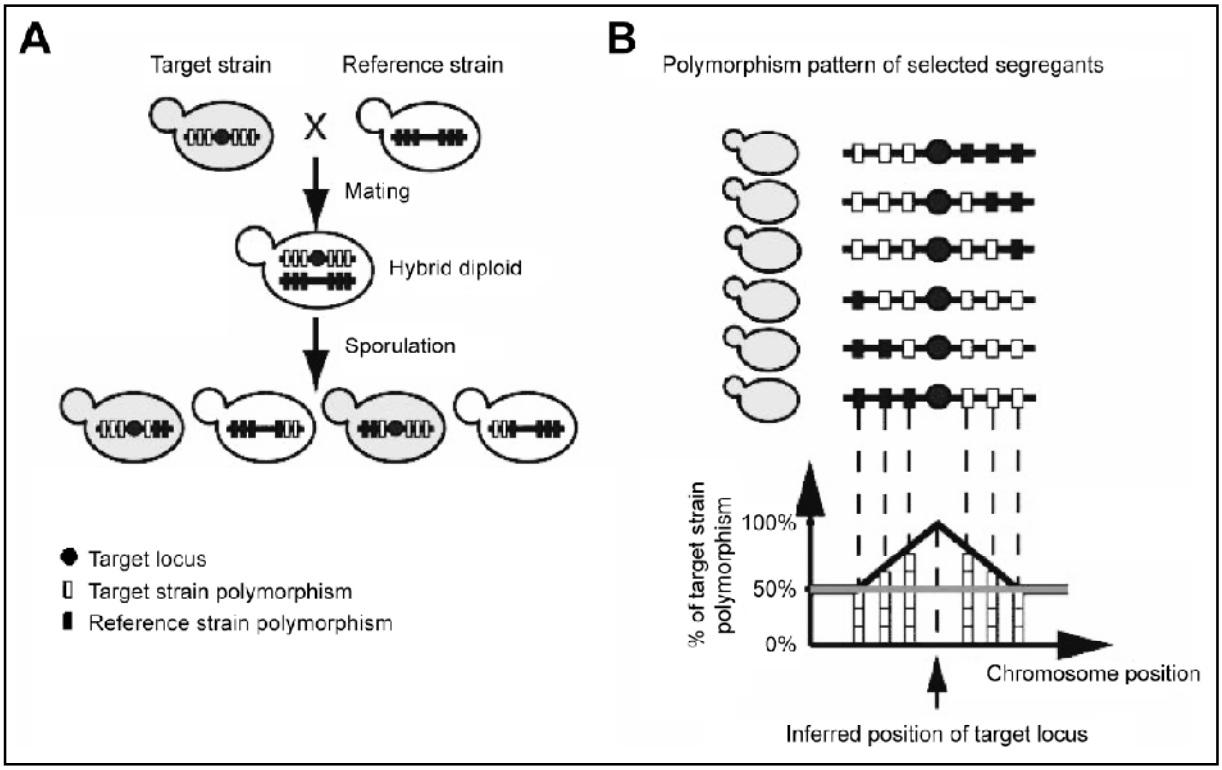


Figure 1: Schematic illustration of QTL-mapping in *S. cerevisiae*. **A.** Process of recombination and segregation of reproduction of target and reference strain. The black circle denotes QTL of interest, white and black squares denote markers for target and reference strain's genotype, respectively. Segregants depicted in grey express the phenotype of interest, so are retained for the next step. (for simplicity only 1 chromosome presented) **B.** By analysis of the proportion of selected segregants expressing the target strain's genotype, the position of the QTL of interest can be inferred. (Figure from Swinnen, 2011)

In creating the pool of selected segregants' DNA and replication of randomly cut DNA parts to be able to extract information concerning the markers, variability is introduced. Moreover, not all markers are genotyped on the same amount of DNA-strings leading to observations of different efficiency for different markers. This is especially the case in the extremes of the chromosome.

After application of the aforementioned techniques, the result is a frequency data set in which regions of chromosome positions, possibly containing the locus of interest, are indicated by higher than average frequencies – the frequencies expected with a recombination frequency of 50 percent. In order to identify these regions and possibly narrow them down even further – enabling fine-

mapping of these regions on gene level – the current report considers the use of wavelet shrinkage as a smoothing tool.

1.2 Wavelet Shrinkage

The use of wavelets allows decomposing functions, or sequences, retaining both frequency and space characteristics. Moreover, because of the wide range of wavelet functions to choose from, from very discontinuous to very smooth, the method is particularly well equipped in decomposing functions which express erratic behaviour; i.e. swift jumps, spikes and other irregular patterns. After decomposition, shrinkage algorithms can be applied to the obtained coefficients and, after the inverse decomposition, a smoothed – denoised – version of the input data can be obtained. Given the noisy QTL-mapping data and the fact that the goal is to identify a detailed region of a possibly erratic function, wavelet shrinkage could be a very efficient tool to give an informative, useful description of the underlying signal incorporated in the data.

The current report investigates the application of wavelet shrinkage methodology to high-resolution quantitative trait loci mapping of whole-genome sequence data. The choice of wavelet, the algorithm used to shrink the wavelet coefficients, the primary resolution level and the impact of the underlying assumptions will be scrutinized.

Concretely, the research question is twofold; firstly, how do the characteristics of the wavelet shrinkage methodology, wavelet function, shrinkage thresholds and underlying assumptions influence the results of smoothing high-density QTL-mapping frequencies? And secondly, can wavelet shrinkage methodology provide a narrow, detailed region in those frequencies for which it is useful to apply fine-mapping on the gene level?

2 Methodology

2.1 Data

The data stem from an experiment investigating the genetic characteristics of high-level ethanol tolerance in *Saccharomyces cerevisiae* (Swinnen, 2011). To this end, the commercial platform Illumina NGS was used to perform high-resolution quantitative trait loci mapping using whole-genome sequencing on *S. cerevisiae* segregants. The data used in the following report are the pre-filtered – only reliable SNPs are retained – SNP frequency data from three chromosomes, i.e. chromosome III, XIV and XVI. Due to the characteristics and intermediate steps of the platform to come up with the frequencies, variance and faults sneak in at several places causing some markers to be over- or under-genotyped, creating SNP frequencies based on different sample sizes.

As can be seen from Figure 2, chromosome III and XVI express a generally flat profile, and it is believed no significant curvature is herein contained. For this reason chromosome III and XVI are used as control data. Chromosome XIV, on the other hand, is the chromosome of interest, with three regions already known to influence high-level ethanol tolerance in *S. cerevisiae*. These three regions, MKT1, SWS2 and APJ1 are indicated between the red, blue and green dashed lines, respectively. The grey dots denote observations which are discarded whenever a data set is required to be dyadic, i.e. having a length of 2^J for some $J \in \mathbb{N}$. The black dots represent the selected data set. From Figure 2, it is already visible that the region surrounding the known important regions mounts out from the data, but in this region no particular narrow high-proportion regions can be distinguished.

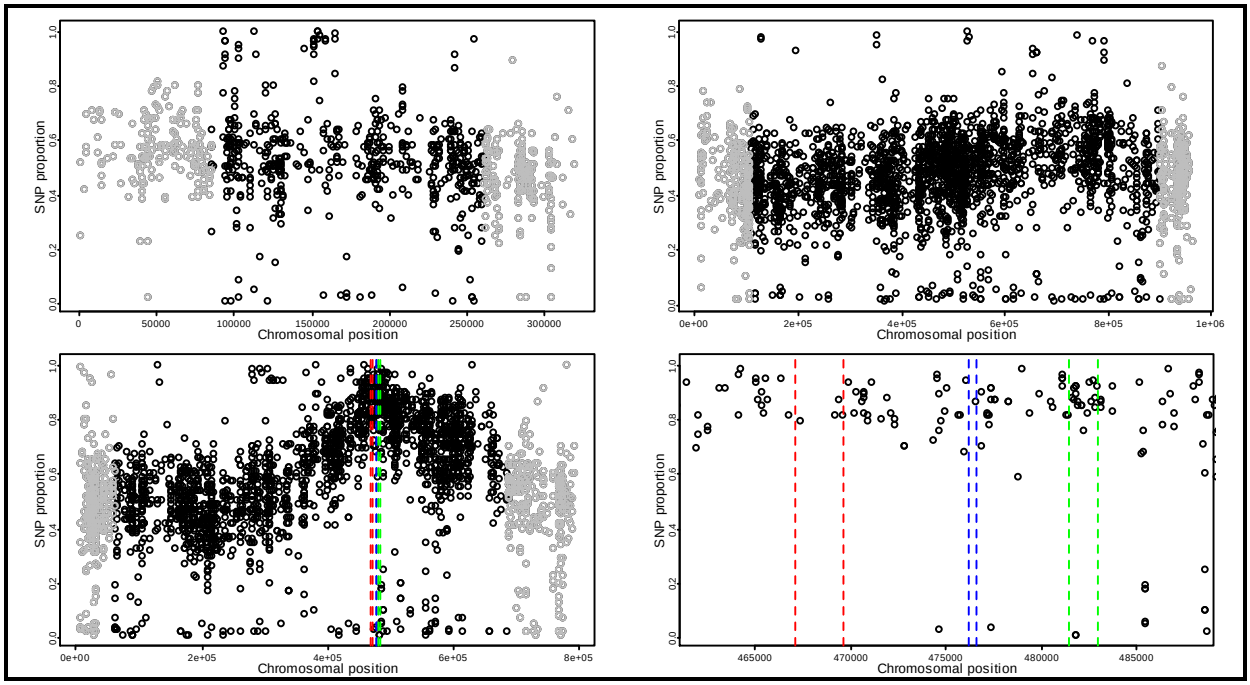


Figure 2: Pre-filtered data from chromosome III (top left), XVI (top right) and XIV full and zoomed view (bottom left and right, respectively). Dots denote observed SNP proportions for the corresponding chromosomal positions. Grey dots are observations left out when data of dyadic size; a power of 2, is required. Red, blue and green dashed lines in bottom plots denote three known genes linked to high-ethanol tolerance.

Considering chromosome XIV, 2581 frequency observations are available over a total chromosome length of 792354 base pairs. These observations are based on sample sizes of segregants' DNA parts ranging from 5 to 403. For the control chromosome III and XVI, 863 and 2478 observations are

available, respectively. Chromosome III has a total chromosome length of 319712 base pairs, chromosome XVI of 955644 base pairs. Segregants' DNA parts on which the SNPs were genotyped ranged between 2 to 107 and 9 to 104 for chromosome III and XVI, respectively.

We can thus assume the data to come from a binomial process with proportion of success vector **p** and sample size vector **n** so that each SNP frequency y_i can be said to be a realization of a binomial distributed $\boldsymbol{Y}$:

$$\boldsymbol{Y}_i \sim Bin(n_i, p_i) \; for \; i = 1, \dots N.$$

The main interest lies in the underlying function of proportions **p** bringing forth these observations. This function provides information on the location of possible genes affecting high-ethanol tolerance in *S. cerevisiae* and can be referred to as the underlying signal contained within the empirical observations.

2.2 Wavelet Shrinkage Methodology

Wavelet shrinkage is the application of wavelet methods as a form of non-parametric regression, also called curve estimation or wavelet regression. As described by Nason (2008), the idea behind wavelet shrinkage is the following. In case a function with additive noise is observed, take a wavelet transform, modify – shrink – its wavelet coefficients and take the inverse wavelet transform to estimate the underlying function.

When observations **y** = $(y_1, \dots, y_n)$ are considered to come from model:

$$y_i = g(x_i) + e_i \qquad for \; i = 1, \dots N, \tag{2.1}$$

the aim is to estimate the unknown function $g(x_i)$ using the observations y_i. Under the assumptions that:

1) The e_i are independent and identically normally distributed
2) The regression ordinates x_i are equally spaced

In the next sections, the methodology of wavelet decomposition will be exemplified where after the implemented shrinkage algorithms will be described. To check for the impact of the assumptions which apply to the data at hand, some normalizing and variance stabilizing techniques will be discussed as well as the method of adaptive lifting to resolve the unequally spaced and not ideally sized data.

2.2.1 Wavelets

In mathematics, a wave is defined as an oscillating function, working on time or space. Fourier analysis, for example, can thus be considered as wave analysis since it expands signals in terms of sinusoids. Waves, such as the sinusoid, are defined with equal amplitude over the entire real line, expressing infinite energy. This is desirable for periodic, time-invariant or stationary phenomena – hence the widely spread use of Fourier analysis (Campbell and Robson, 1968). Intuitively, *wavelet* is understood as 'little wave', a function which oscillates but on a compact support (decays to zero very rapidly). These characteristics cause wavelets to have energy concentrated in space around some point allowing for the analysis of, not only frequency, but also time (space) characteristics. Therefore,

wavelet analysis is of great use in the analysis of transient, non-stationary phenomena (Burrus, Gopinath and Guo, 1997).

Wavelet expansion or wavelet transform

In what follows, only measurable functions f on $\mathbb{R}$ belonging to the space of square integrable functions $L^2(\mathbb{R})$ satisfying:

$$\int_{-\infty}^{\infty} |f(x)|^2 \, dx < \infty$$

will be considered. As is the case for the Fourier's basis functions, the wavelets are constructed to form a complete orthonormal system.

The first step in analysing a signal or function is usually to express it as a linear decomposition, which allows for better description and processing. In the wavelet expansion case, such linear decomposition is defined as:

$$f(x) = \sum_k \sum_j d_{j,k} \psi_{j,k}(x),$$

a two-parameter system where j and k are integer indices. $\psi_{j,k}(x)$ are the wavelet expansion functions. $d_{j,k}$ are called the *Discrete Wavelet Transform* (DWT) or wavelet coefficients of $f(x)$. Compared to the Fourier transformation, the additional summation over space indicator k allows wavelets to also extract information localized in space. Since wavelets are constructed to form an orthonormal system, Burrus, Gopinath and Guo (1997) indicate that the wavelet coefficients $d_{j,k}$ can be computed by taking the inner product:

$$d_{j,k} = \langle \psi_{j,k}(x), f(x) \rangle = \int f(x) \psi_{j,k}(x) dx.$$

For some defined wavelet $\psi(x)$, the whole set of wavelet expansion functions can be generated for integers j and k by dilation and translation of this generating (mother) wavelet function:

$$\psi_{j,k}(x) = 2^{j/2} \psi\left(2^j x - k\right) \qquad for\ j, k \in \mathbb{Z}.$$

In this definition, the $2^{j/2}$ term ensures that a constant norm is maintained independent of scale j. In other words, it ensures that the 'output energy' of the transform is the same as the 'input energy'. Moreover, the space location and frequency or scale is parameterized by k and j, respectively (Burrus, Gopinath and Guo, 1997).

Multiscale transform

In this next section the wavelet function $\psi(x)$ will be derived from a scaling (father wavelet) function $\varphi(x)$ defined from the concept of resolution. The scaling functions defined in terms of integer dilation (scaling) and translation of the basic scaling function $\varphi(x)$ is:

$$\varphi_{j,k}(x) = 2^{j/2} \varphi\left(2^j x - k\right) \quad k \in \mathbb{Z} \quad \varphi \in L^2.$$

For all integers $k \in]-\infty, \infty[$ the subspace of $L^2(\mathbb{R})$ spanned by these scaling functions over k is defined as

$$v_j = \overline{Span_k\{\varphi_k(2^J x)\}} = \overline{Span_k\{\varphi_{J,k}(x)\}} \quad for\ all\ k \in \mathbb{Z}.$$

This can be understood as:

$$f(x) = \sum_k c_k \varphi_k(2^j x + k) \quad for\ any \quad f(t) \in v_j,$$

using the scaling functions $\varphi_{j,k}(x)$ as a basis, v_j is the total set of functions $f(x)$ which can be constructed using the stated linear combination and c_k are the scaling coefficients. The over-bar indicates that this span also includes the limits of the infinite sum in addition to the linear combination of the basis set. Moreover, the larger j, the finer detail can be presented by narrower scaling functions translating in smaller steps.

It is also possible to view wavelet decomposition in the light of multi-resolution analysis, where the decomposition of a signal is in terms of the resolution of detail (Burrus, Gopinath and Guo, 1997). The basic requirement of multi-resolution analysis is by providing the nesting of the spanned spaces:

$$v_j \subset v_{j+1} \quad \text{for all} \quad j \in \mathbb{Z}$$

with:

$$v_{-\infty} = \{0\}, \quad v_{\infty} = L^2.$$

The previous statements acknowledge that a space containing high resolution signals (larger j) will also contain lower resolution signals. Following this definition, the spaces have to satisfy a natural scaling condition ensuring that elements in a space are simply scaled versions of the elements in the next, higher resolution, space. This can be expressed as:

$$f(t) \in v_j \quad \Leftrightarrow \quad f(2t) \in v_{j+1}.$$

All together it can be stated that if $\varphi(x) \in v_0$, it also belongs to v_1, which is spanned by $\varphi(2x)$. This means that $\varphi(x)$ could be expressed in terms of a weighted sum of shifted $\varphi(2x)$. Leading to the following result

$$\varphi(x) = \sum_n h(n)\sqrt{2}\varphi(2x - n) \qquad n \in \mathbb{Z},$$

with *h(n)* a sequence called the scaling function coefficients (father wavelet coefficients or scaling filter) and $\sqrt{2}$ ensures maintaining the norm of the scaling function with the scale of 2.

It can be shown (Daubechies, 1992) that if such scaling (father wavelet) function exists, an associated orthonormal wavelet basis exists such that a (mother) wavelet function can be defined as:

$$\psi(x) = \sum_n (-1)^{n-1}\overline{h_{-n-1}}\varphi_{1,n}(x) \qquad n \in \mathbb{Z}.$$

All these results eventually lead to the fine-scale representation of a function $f(x)$ by:

$$f(x) = \sum_{k\in\mathbb{Z}} c_{j_0,k}\varphi_{j_0,k}(x) + \sum_{j=j_0}^{\infty}\sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k}(x). \quad (2.2)$$

In this equation, $c_{j_0,k}$ can be thought of as representing the average level of the function, where j_0 denotes the coarsest scale of the wavelet transform, while the $d_{j,k}$ express the amount of detail, increasing with scale j.

In practice, to transform a sequence $\boldsymbol{y}$ of length n, expected to come from some function f(x), the General Fast Discrete Wavelet Transform is performed. This transform is based on the fact that the coefficients denoted in equation 2.2, can be obtained by performing a dyadic decimation operation (Nason, 2008). Consider two quadrature mirror filters H – defined as a low pass filter, thought of as averaging – and G – a high pass filter, extracting detail by differencing. Then the scaling and wavelet coefficients $c_{j,k}$ and $d_{j,k}$ can be obtained in the following fashion:

Firstly, apply filter H to the sequence $\boldsymbol{y}$ then:

$$c^*_{J-1,k} = \sum_n h_{n-k}\, c_{J,n} \qquad with\ c_{J,n} = y_n.$$

Next, retain every other element of $c^*_{J-1,k}$; known as dyadic decimation. Then all the scaling coefficients $c_{j-1,k}$ can be defined as $c^*_{J-1,2k}$. The same approach holds for the wavelet coefficients $d_{j,k}$ where filter H is replaced by G. This way, as represented in Figure 3, the discrete wavelet transform can be performed by recursively applying filter operations H and G followed by dyadic decimation. The grey bars in this figure represent the scaling coefficients $c_{j-1,k}$ while the white bars depict the wavelet coefficients $d_{j-1,k}$.

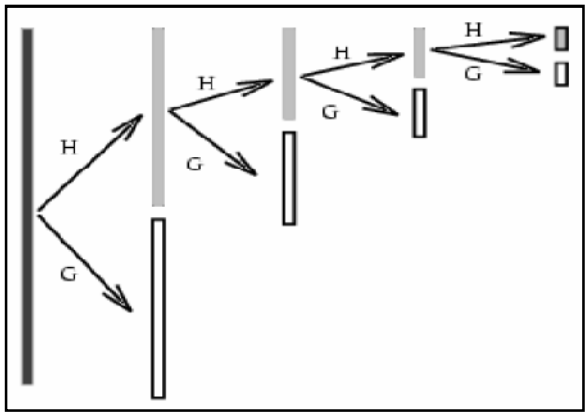


Figure 3: Graphical representation of the dyadic decimated discrete wavelet transform.

This highly efficient and fast way of discrete wavelet transform is applied throughout the software used in this report with as most important disadvantage that in order to have this dyadic decimation filter sequence to work, the data have to satisfy a length of 2^J for some $J \in \mathbb{N}$.

2.2.2 Wavelet functions

In the report, three wavelet functions and their accompanying scaling functions are used. These three wavelet functions all belong to the family of Daubechies' extremal phase compactly supported wavelets. As elaborated in Daubechies (1988), the wavelets in this family are constructed to be orthonormal wavelets each indexed by a number N, denoting twice the number of vanishing moments. Vanishing moments indicate the maximum degree of polynomials the wavelet will

decompose with all coefficients exactly equal to zero (Nason, 2008). The Daubechies' extremal phase compactly supported wavelets with 1, 3 and 9 vanishing moments will be applied in the report.

Haar wavelet

The Daubechies wavelet characterised by 1 vanishing moment, is better known as the Haar wavelet and is the simplest orthogonal wavelet system, already defined by Haar in 1910, found out to be a special case of the Daubechies' wavelet family. The mother and father Haar wavelet are shown in Figure 4. The Haar mother wavelet function is defined as:

$$\psi(x) = \begin{cases} 1 & x \in \left[0, \frac{1}{2}\right), \\ -1 & x \in \left[\frac{1}{2}, 1\right), \\ 0 & otherwise. \end{cases}$$

while the Haar father wavelet function is the simple unit-width, unit-height pulse function $\varphi(x)$.

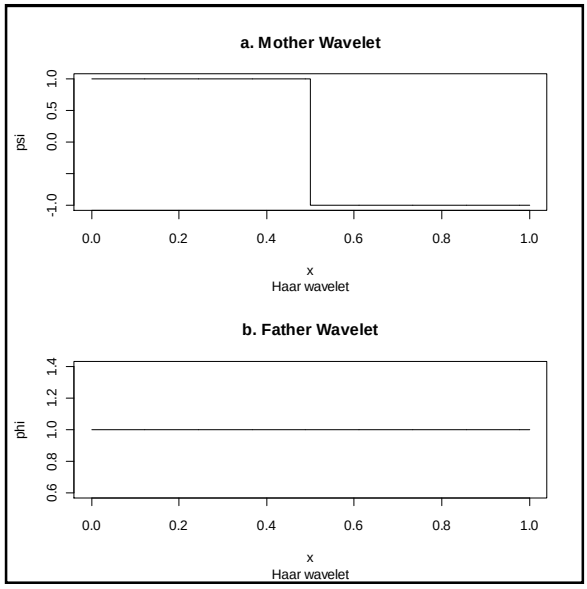


Figure 4: Mother and father Haar Wavelet function.

Figure 4 shows how discontinuous the Haar mother and father wavelet functions are. Such wavelets allow estimation of functions that are characterised by similar discontinuities, one of the hallmarks of the wavelet shrinkage methodology.

Daubechies (3 and 9 vanishing moments)

The remaining two wavelets are the Daubechies' wavelets with 3 and 9 vanishing moments. As apparent from their representation in Figure 5, increasing the number of vanishing moments, introduces smoother mother and father wavelet functions. Although the wavelets shown in Figure 5 have smoother transitions, their pulse-like shape enables them to capture erratic information patterns.

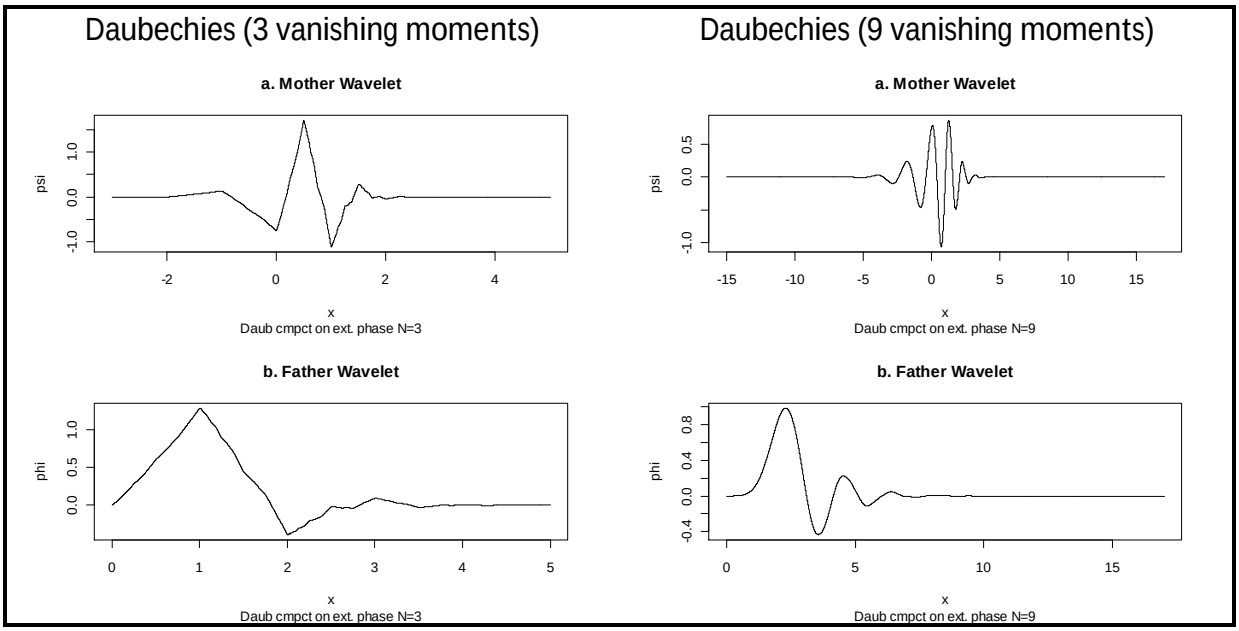


Figure 5: Mother and father Daubechies' 3 and 9 vanishing moments wavelet functions. Left hand side panels denote mother and father Daubechies 3 vanishing moments wavelet functions, right hand side panels Daubechies' wavelet with 9 vanishing moments.

2.3 Threshold Definitions

The next sections will describe the three different methods for selecting the wavelet shrinkage thresholds which were used to obtain smoothed estimates of the QTL data set. The general idea is that at this stage the data are wavelet transformed and the scaling coefficient of the coarsest scale – expressed as $c_{0,0}$ – along with the wavelet coefficients are available. Due to the characteristics of the wavelet transform mentioned in the previous section, the expectation is that large wavelet coefficients contain true signal and noise, while small ones are considered to only contain noise, hence can be discarded. The goal of the wavelet shrinkage algorithms is to come up with a threshold value below which a wavelet coefficient is considered invaluable and remove it from the wavelet coefficients.

We can rewrite model 2.1 by the wavelet-transformed model:

$$d^* = d + \varepsilon, \tag{2.3}$$

in which d^* represents the wavelet transform of the original data y, *d* denotes the wavelet transform of the underlying true data and ε represents the wavelet transform of the noise. By Parseval's identity of the preservation of squared error, estimating data in the wavelet coefficients space or in the data space resolves to the same thing. As introduced by Donoho and Johnstone (1994), the soft thresholding function will be applied in the current report. This function is denoted by:

$$\hat{d} = sgn(d^*)(|d^*| - \lambda)\mathbb{I}\{|d^*| > \lambda\},$$

where $\mathbb{I}$ is the indicator function, $\hat{d}$ the estimated coefficient of the wavelet transform of the true signal, d^* the empirical coefficient to be thresholded and λ denotes the threshold. Figure 6 represents an example of the soft thresholding function. In this function, the empirical coefficients smaller than the threshold are put to zero. But, in addition, the remaining coefficients are pulled towards zero by subtracting the threshold value. Since it is usually believed that the coarser scales

are less influenced by noise than finer scale levels, shrinkage is seldom applied to all coefficients. Usually wavelet shrinkage is applied starting from some coarse scale up to the finest scale level. The coarsest level at which thresholding is applied is referred to as the *Primary Resolution Level* (PRL) (Nason, 2008).

The threshold value λ can be defined in different ways, minimizing some error measure, measuring the expected distance between the estimate and the true value (Nason, 2008). In the next sections, the applied definitions for λ are denoted.

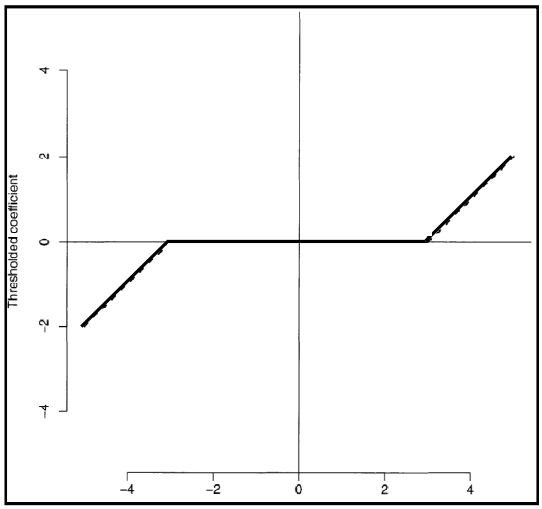


Figure 6: Example of a soft threshold function. (Figure adapted from Nason, 2008)

After the estimated wavelet coefficients $\hat{d}$ are computed based on the chosen shrinkage threshold, the final step is to apply the inverse discrete wavelet transform and transform the coefficients $\hat{d}$ back to obtain an estimate of the true function of interest denoted by $g(x)$ in (2.1).

2.3.1 SURE

Sure thresholding, also referred to as SureShrink, is a threshold computation method developed by Donoho and Johnstone (1995). The method is based on Stein's (1981) Unbiased Risk Estimation (SURE) technique for estimation of a quadratic loss involving the mean of a multivariate normal distribution. Applying Stein's result together with the assumption of normally distributed noise, Donoho and Johnstone show that:

$$SURE(\lambda; x) = n - 2 \cdot \#\{i: |x_i| \leq \lambda\} + \sum_{i=1}^{d} (|x_i| \wedge \lambda)^2$$

is an unbiased estimator of the expected value of the difference between the true and estimated wavelet coefficients. The optimal SURE threshold value is then defined as the value of λ minimizing the above stated equation.

2.3.2 FALSE DISCOVERY RATE (FDR)

A second possibility of deciding which of the noisy wavelet coefficients d^* are non-zero is proposed by Abramovich and Benjamini (1996). They formulate the problem as a multiple hypothesis testing problem where they test for each coefficient $d_{j,k}$ whether:

$$H_0: d_{j,k} = 0 \quad versus \quad H_A: d_{j,k} \neq 0 .$$

Intuitively, the method is constructed as follows. Consider R to be the number of coefficients which, by using some thresholding procedure, are not set to zero. Then, Q can be defined as the proportion of wrongly kept coefficients, $Q = \frac{V}{R}$, with V equal to the number of coefficients falsely kept. If then the expectation of Q is defined as the *False Discovery Rate of Coefficients* (FDRC), Abramovich and Benjamini (1996) suggest using a threshold value which maximizes the number of retained coefficients while controlling the FDRC by some level q.

In practice, the method is performed in the following four steps:

1. For each $d^*_{j,k}$ calculate two-sided p-value $p_{j,k}$ testing $H_{j,k}: d_{j,k} = 0$. $p_{j,k} = 2(1 - \Phi(\frac{|d^*_{j,k}|}{\sigma}))$
2. Order all $p_{j,k}$s to their size
3. Let i_0 be the largest *i* for which $p_{(i)} \leq (\frac{i}{m})q$.
4. Then $\lambda = \lambda_{i_0} = \sigma\, \Phi^{-1}(1 - \frac{p_{i_0}}{2})$

In which $\Phi()$ denotes the standard normal distribution function, since the noise is assumed to be normally distributed, while m denotes the number of coefficients to be thresholded. Moreover, in the application in the R-function WaveThresh, the σ will be estimated by the *Median Absolute Deviation* (MAD) of the wavelet coefficients on which shrinkage will be applied.

2.3.3 UNIVERSAL THRESHOLDING

Donoho and Johnstone (1994) propose to use the Universal threshold λ^u, defined as:

$$\lambda^u = \sigma\sqrt{2 \log n}.$$

In this definition, σ represents the standard deviation of the error ε and n the number of observations. According to Donoho and Johnstone (1994), it can be proved that using λ^u would eliminate the noise with high probability. As an estimate of σ they suggest to use the MAD of the finest-scale wavelet coefficients.

The default option incorporated in the R-function WaveThresh assumes the MAD of all the coefficients to which the shrinkage eventually will be applied. Both estimates are investigated in the current report.

2.4 Gaussianization and Variance Stabilizing

In the definitions of the shrinkage thresholds from the previous section, the noise contained in the observed data set is assumed to be normal. From the description of the data, however, it is clear that the noise is coming from a binomial distribution. To investigate the impact of violating this normality assumption, three gaussianization and variance stabilizing transformations were considered.

Anscombe

In order to normalize the noise and ensure a stabilized variance, the data could be transformed to improve approximation of their distribution by a normal distribution. Anscombe (1948) suggested for

$\{y_i\}$, realizations of $Y_i \sim Bin(n_i, p_i)$ for $i \in \{0, \ldots, N\}$, the following transformation to make the data 'more normally' distributed:

$$\mathcal{A}y_i = sin^{-1}\sqrt{\left(\frac{y_i + c}{n_i + 2c}\right)}.$$

Moreover, Anscombe (1948) suggests to use $c = \frac{3}{8}$ in order to ensure optimal performance.

Freeman and Tukey

In line with the Anscombe transformation, Freeman and Tukey (1950) propose a more general transformation based on an averaged inverse sine function defined as:

$$\beta y_i = sin^{-1}\sqrt{\left(\frac{y_i}{n_i + 1}\right)} + sin^{-1}\sqrt{\left(\frac{y_i + 1}{n_i + 1}\right)}.$$

Nunes and Nason Haar transformation (NN-Haar)

Nunes and Nason (2009) propose a normalizing and variance stabilizing transformation based on modification of the wavelet coefficients of the Haar discrete wavelet transform (using wavelet filters (1,1)/2 and (1,-1)/2) of the observations. Nunes and Nason (2009) suggest the following transformation scheme leading to a transformed data vector **u** which contains elements which are more normally distributed:

1) Obtain the vector of Haar discrete wavelet transform coefficients $(\boldsymbol{c_0}, \boldsymbol{d_0}, \boldsymbol{d_0}, \ldots, \boldsymbol{d_{J-1}})$ and modify these coefficients as follows:

$$f_{j,k} = \frac{d_{j,k}}{\sqrt{\frac{(c_{j,k}(n_{j+1,k-1} + n_{j+1,k} - 2^{J-j}c_{j,k}))}{n_{j+1,k-1} + n_{j+1,k}}}}.$$

2) Perform the inverse Haar discrete wavelet transform using the modified vector $(\boldsymbol{c_0}, \boldsymbol{f_0}, \boldsymbol{f_0}, \ldots, \boldsymbol{f_{J-1}})$ to obtain the transformed data vector **u**.

As noted by Nunes and Nason (2009), by undoing these steps 1 and 2, the transform can be inverted.

2.5 Adaptive Lifting

Next to the assumption of normally distributed noise, two other assumptions are made to be able to apply the Fast Wavelet Transform algorithm and the shrinkage definitions described above. First of all, a very important and influencing requirement is that the data set should be dyadic. As clearly can be seen in Figure 2, a large part of the data set is not used in the analyses due this restriction needed to use the pyramidal algorithm to come up with the DWT. Secondly, in the wavelet shrinkage methodology the assumption is made that the regression ordinates x_i are equally spaced. Also this assumption is not satisfied since the observed SNP frequencies are not observed at chromosome positions which are a fixed distance apart.

One way to resolve these two issues at the same time is the adaptive lifting method suggested by Nunes, Knight and Nason (2006). By making the 'lifting one coefficient at a time' method, elaborated by Jansen, Nason and Silverman (2001), adaptive, they are able to resolve the non-equispaced and size problem of a data set. Adaptive lifting can be explained by six steps in which Nunes, Knight and Nason (2006) built some adaptive characteristics. The general idea is to lift the sampled function values one by one into a set of detail and scaling coefficients, not unlike the discrete wavelet transform. Consider the observed values as function values on an irregular grid and use them as the initial scaling coefficients.

1) From the initial scaling coefficients, select the point c_{j_n} with the finest detail as the first point to be lifted. This point is found by means of the minimum scaling function integral.
2) Identify the set of neighbours of c_{j_n} denoted by I_n.
3) Make use of these neighbours to predict the value of the function at position j_n by means of simple regression techniques.
4) Compute the detail coefficient d_{j_n} by computing the difference between c_{j_n} and that prediction.
5) Remove the selected point c_{j_n} and update the identified neighbours in order to keep the total 'energy' constant (recall wavelet transform, Jansen, Nason and Silverman, 2001).
6) Repeat this process

Nunes, Knight and Nason (2006) introduce adaptive characteristics to steps 2 and 3. The definition of the neighbourhood is important because it will influence the prediction of the candidate to be lifted and thus the smoothness of the eventual fit. Considering the simple regression technique used in step 3, Nunes, Knight and Nason (2006) suggest to use polynomials of different orders to come up with the prediction, suited to the pattern in the data at hand. After repeating this procedure r times, the function f can be represented as a linear combination of n-r+1 wavelet coefficients, generated during each 'step 4', and the remaining, during each 'step 5' updated, scaling coefficients. This could be regarded as a kind of wavelet transform but without orthogonality of the wavelet and scaling functions (Jansen, Nason and Silverman, 2001). Because of this, the assumption of independent normally distributed noise will not hold and the aforementioned shrinkage techniques will not be applicable to the lifted coefficients. For this reason, Nunes, Knight and Nason (2006) make use of a modified version of the empirical Bayesian wavelet shrinkage approach.

Modified empirical Bayesian wavelet shrinkage

Since for a wide class of functions it is known that their DWT coefficients lead to a sparse representation in which few coefficients are larger than zero (Nunes, Knight and Nason, 2006). It is suggested by Johnstone and Silverman (2005) to use the following prior distribution for a DWT coefficient:

$$d_{j,k}^{*} \sim (1-\pi)\delta_0 + \pi\gamma,$$

where π denotes the prior probability of the true wavelet coefficient $d_{j,k}^{*}$ to be non-zero. γ represents the distribution of this true wavelet coefficients conditional on it being non-zero. Based on nice theoretical results and practical advantages Nunes, Knight and Nason (2006) propose a quasi-Cauchy prior as a choice for γ.

Since the approach of lifting one coefficient at a time does not lead naturally to discrete dyadic scales, Nunes, Knight and Nason (2006) suggest to estimate the π_j using level-wise marginal-maximum likelihood but then on artificially created scales, obtained from quantizing the detail coefficients. To overcome the problem of noise which is not independent and identically normally distributed, the empirical Bayes method is applied to the normalized detail coefficients. With this modification the likelihood of $d_{j.}|d_{j.}^*$ is again given by $d_j \sim N(d_{j.}^*, \sigma^2)$. This way the detail coefficients can be estimated by the median of the posterior distribution. Here after the estimated coefficients are denormalized and the adaptive lifting transform can be inverted to obtain the estimated true function (Nunes, Knight and Nason, 2006).

2.6 Software

All computations, as well as most of the figures in the report at hand were performed and created using R 2.13.1 (R Development Core Team, Vienna, Austria). As an example, part of the codes is contained in Appendix6.

3 Results

In order to investigate the impact of the choice of wavelet, the wavelet shrinkage threshold, the primary resolution level and violation of the assumptions of normality and equally spaced observations in a dyadic data set of length 2^J for some $J \in \mathbb{N}$, wavelet shrinkage methodology was applied to the high-ethanol level tolerance QTL-mapping data set.

In the application of the aforementioned methodology, several aspects and characteristics were varied in order to investigate their effect. First of all, the type of wavelets used to obtain the discrete wavelet transform was investigated by comparing results for the Haar wavelet and Daubechies wavelet with 3 and 9 vanishing moments, respectively. Secondly, it was examined how the choice of threshold in the shrinkage procedure affects the estimated curves. In this light, four different approaches were followed: SURE, FDR and Universal thresholding were applied. In the case of the Universal threshold, also the difference between using the MAD of only the finest scale coefficients or the MAD of all coefficients to be thresholded, was studied. Finally, the primary resolution level was also varied between scale 3, 6 or 9.

To explore the consequences of violating the assumption of normality of the noise distribution, the Anscombe, Freeman and Tukey and NN-Haar transformations where applied to the raw data before applying the wavelet shrinkage methodology with all the varying characteristics just described.

Finally, adaptive lifting with empirical Bayesian wavelet shrinkage was applied to the normalized and variance stabilized data set. Comparing these results to the previous ones provides an indication of how the violation of equally space data affects results.

In all instances, methods were applied to the chromosome of interest, XIV, in which a clear estimated curvature would be expected, as well as to chromosomes III and XVI, which should show a flat signal. This enabled testing the method for applicability to the problem at hand.

3.1 Wavelet choice

Considering the effect of the choice of wavelet used in the DWT, the Haar wavelet was compared to the Daubechies wavelet with 3 and 9 vanishing moments. To this end, the middle $2^{11} = 2048$ data points were discrete wavelet transformed using the different wavelets of interest. Next, wavelet shrinkage was applied using a Universal threshold approach with primary resolution level equal to 3 and the MAD of all wavelet coefficients to be thresholded as an estimate for σ. Finally, the shrunken coefficients were back transformed to obtain a smoothed estimate for the signal in the data.

Some results are plotted in Figure 7. From this figure it becomes apparent that the Daubechies wavelets produce smoother estimates over chromosomal position. The discontinuous jumps seen in the function estimated with the Haar wavelet are less pronounced in the estimated function obtained with the Daubechies wavelet with 3 vanishing moments. Moreover, the estimates from the two Daubechies wavelets differ in terms of smoothness of the curves. The more vanishing moments a wavelet possesses, the smoother the estimated curve.

Considering the information coming from the zoomed plots, no apparent small regions of higher proportions can be noticed, not even in the coloured intervals already known to affect high-ethanol tolerance. Overall, the estimates seem smoothed to a good extent, but still unsmoothed spikes are

observed. Moreover, also the smoother wavelets with a higher number of vanishing moments, show these unsmoothed parts, albeit to a lesser extent. Comparing the results of chromosome XIV with those for control chromosomes III and XVI (results presented in Appendix 1), reveals that indeed some curvature is observed in the estimated function for chromosome XIV and picked up by the method. The estimated functions for the control chromosomes are characterised by much flatter curves.

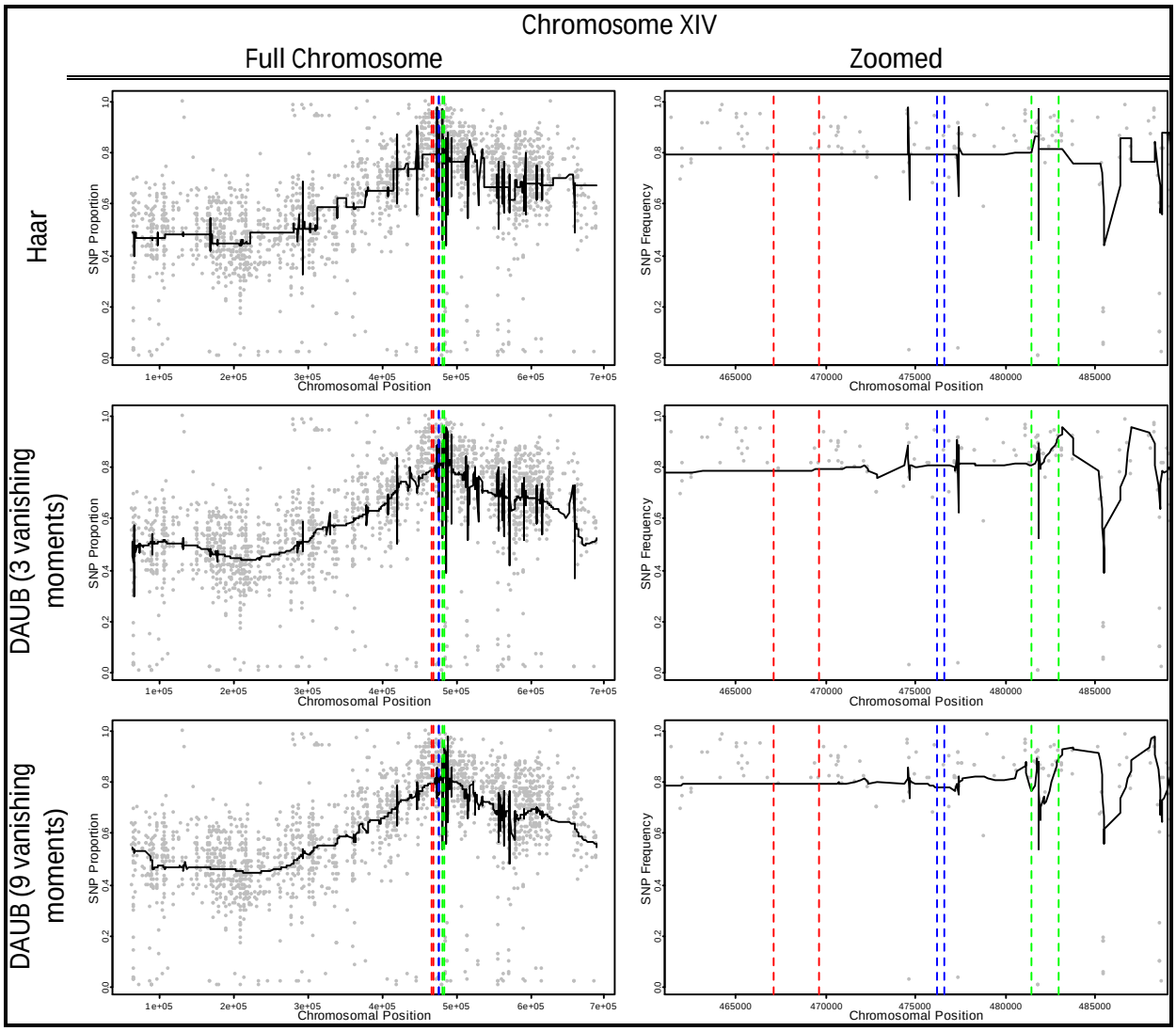


Figure 7: Comparison of wavelet shrinkage method for three different wavelets. Solid black lines denote estimate of underlying function using Universal threshold with MAD estimator of all coefficients from primary resolution level = 3 onwards. Different wavelets used are Haar and Daubechies waveletes with 3 or 9 vanishing moments. Left panels show estimation on full chromosome, right panels zoom in on region already known to contain important genes.

3.2 Wavelet shrinkage thresholds

Next, we focussed on the wavelet shrinkage threshold. Four different threshold types were compared; the SURE, FDR, Universal with MAD computed on the finest detail scale and Universal with MAD computed on all wavelet coefficients to be thresholded. Results are presented for the case when the selected data set was first expanded using the Haar wavelet transform. The four different wavelet shrinkage methods were applied using primary resolution level 3 where after the shrunken wavelet coefficients were back transformed to obtain the estimates.

From the results depicted in Figure 8, it can be seen that the different thresholds produce different results. As the thresholds progress from SURE to FDR to the Universal thresholds, the estimated functions get smoother. Moreover, the "unsmoothed" spikes also seem to decrease in size, providing a generally smoother curve. This result seems to carry over to the region of interest. Also in this region the peaks are compressed when treading from SURE to the Universal threshold in which the standard deviation estimate is based on all wavelet coefficients to be thresholded.

Comparison of the results of the same setting to the control chromosomes reveals again that the approach is picking up some curvature in the chromosome of interest. Showing overall flat curves with occasional peaking also in those estimates (Appendix 2). In addition, it seems to be the case that the SURE threshold is well performing on the control chromosomes lacking a clear curvature, resulting in curves as smooth as provided by the Universal threshold approach.

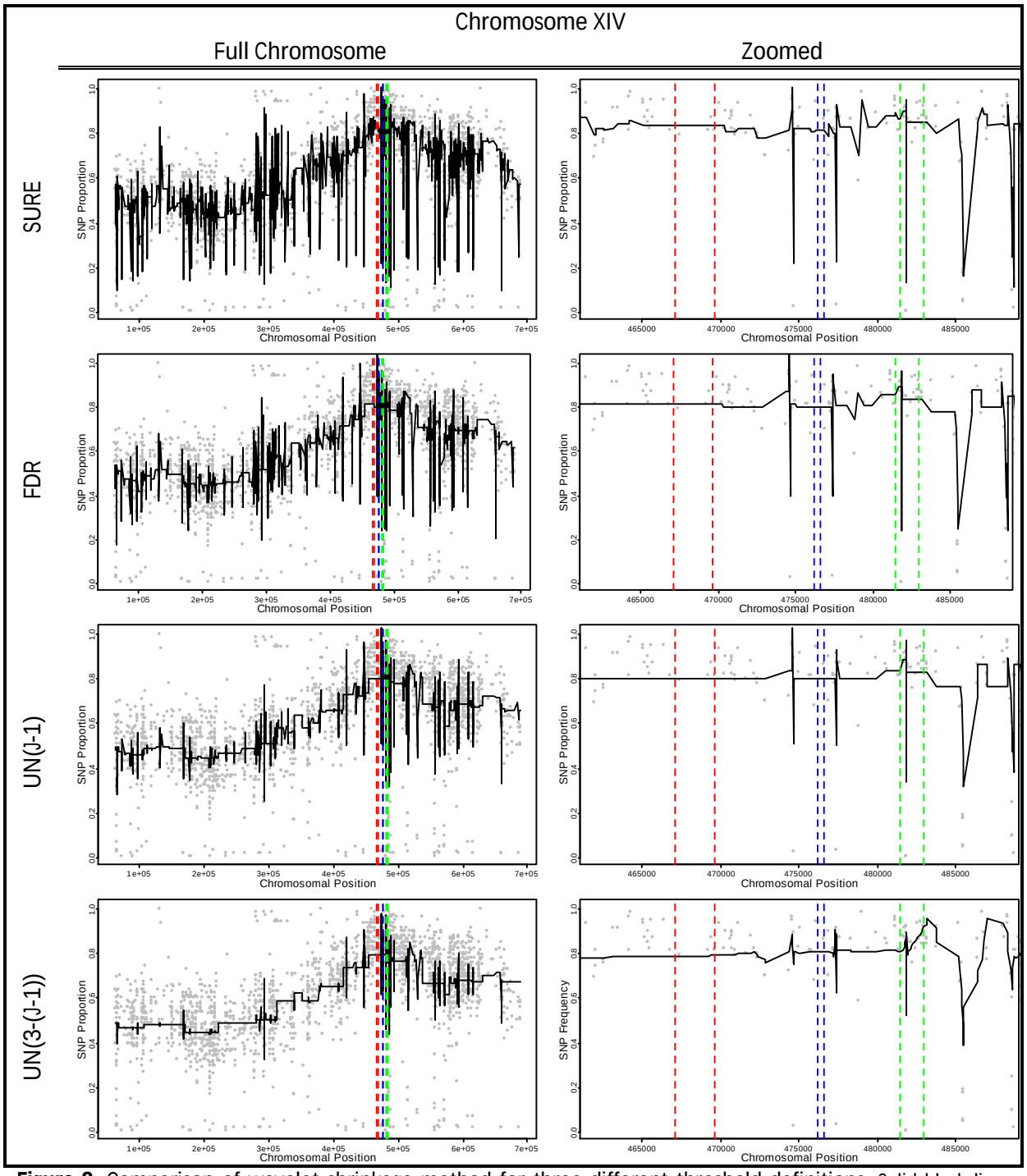


Figure 8: Comparison of wavelet shrinkage method for three different threshold definitions. Solid black lines denote estimate of underlying function using Haar DWT and PRL equal to 3. Comparison between the use of SURE, FDR or Universal, with MAD computed for different levels, thresholds. Left panels show estimation on full chromosome, right panels zoom in on region already known to contain important genes.

3.3 Primary Resolution Level (PRL)

Another question of interest was the influence of the primary resolution level on estimation. As before, this characteristic was varied and results compared. The coarsest level from which the wavelet coefficients were shrunken was set to scale level 3, 6 and 9, respectively. Figure 9 shows the results from the situation in which the selected data were expanded by discrete wavelet transform using the wavelet from the Daubechies family with 9 vanishing moments. The obtained wavelet coefficients were then shrunk by means of the Universal threshold where the estimate for σ was the MAD computed on all wavelet coefficients to be thresholded. After thresholding, the inverse transform was applied to obtain the estimated curves.

From the plots included in Figure 9 it can be seen that with increasing PRL the estimates are getting less compressed. Since more wavelet coefficients are retained and not put or pulled towards zero – charcteristic of the soft thresholding function, shrinkage will have less effect on the estimates and curves will be less smoothed. This observation can be made for the estimated function of the whole chromosome, as well as the results of the region of interest. In this region, no specific narrow-regions of increased proportion seem to stand out.

Comparison of the presented results with the results of the control chromosomes reveals the same characteristics with respect to increasing PRL but no curvature seems to be estimated. Again, this shows the approach to capture some curvature in the chromosome of interest which is not present in chromosomes III and XVI (Appendix 3).

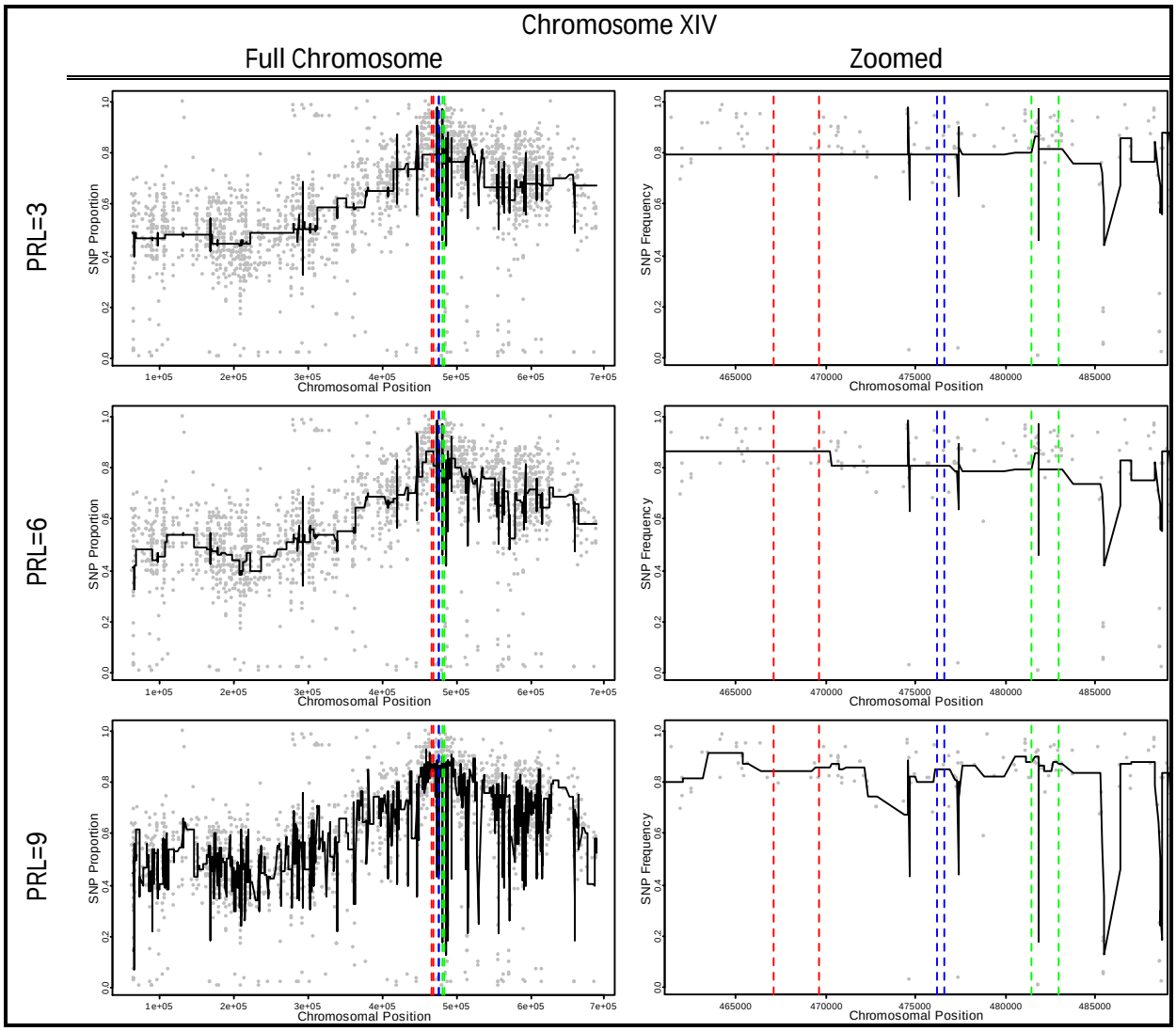


Figure 9: Comparison of wavelet shrinkage method for three different primary resolution levels. Solid black lines denote estimate of underlying function using DWT with Daubechies wavelet with 9 vanishing moments and Universal threshold with MAD computed for scaling levels from PRL to finest detail level. Different PRLs are 3, 6 and 9. Left panels show estimation on full chromosome, right panels zoom in on region already known to contain important genes.

3.4 Normality assumption

Concerning the violation of the normally distributed noise in the current data set, the three aforementioned transforms were applied and the results compared. The wavelet transform coefficients, adopting Daubechies wavelet with 9 vanishing moments, of the normalized selected data were shrunken using the Universal threshold with MAD estimate of all coefficients to be thresholded. Results are depicted in Figure 10. In addition, the untransformed curve estimate is also included in the figure.

The first observation that can be made from Figure 10 is that the NN-Haar transform proposed by Nunes and Nason (2009) does not seem to be adapted to incorporate a binomial sequence based on binomial processes with different sample sizes. The problem seems to be connected to the back transformation in which proportionality is not guaranteed and proportions are estimated to exceed 1. The transformations proposed by Anscombe (1948) and Freeman and Tukey (1950) are performing well, in providing smooth curves occasionally interrupted by spikes. Moreover, there seems to be little difference in the performance of both transformations, although the Freeman and Tukey (1950) transformation appears to provide peaks which are a fraction smaller. When comparing to the results of the untransformed case, it seems that ignoring the non-normality of the noise results in underestimating the error, causing smaller peaks and a possibly too smooth curve estimate. The same observations can be made for the region of interest.

These results can again be carried over to the control chromosomes in which any indication of a clear curvature is absent (Appendix 4).

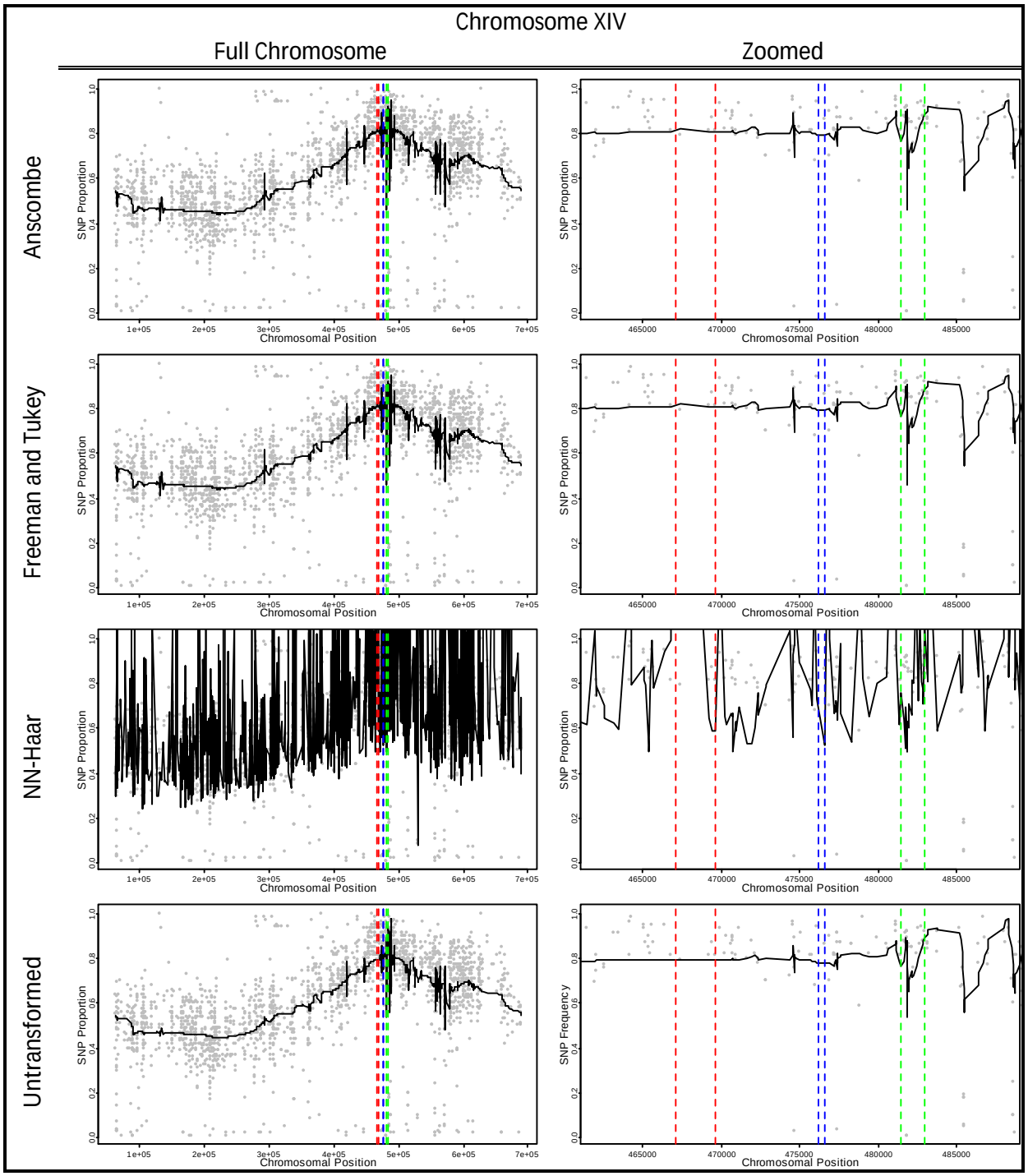


Figure 10: Comparison of wavelet shrinkage method after normalizing and variance stabilizing transformations. Solid black lines denote estimate using DWT with Daubechies' wavelet (9 vanishing moments) and Universal threshold with MAD of wavelet coefficients from PRL(3) to finest detail level. Transformations are Anscombe, Freeman and Tukey and NN-Haar normalization and variance stabilizing transformations accompanied by untransformed fit. Left panels show estimation on full chromosome, right panels zoom in on region already known to contain important genes.

3.5 Non-equally spaced observations

Adaptive lifting was performed to take the non-equally spaced observations in the data into account. At the same time, since the method does not incorporate the discrete wavelet transform, no restrictions on length of the data set are imposed. After transforming the data using the Freeman and Tukey (1950) normalization and variance stabilizing transformation, adaptive lifting with two orders of simple regression and three neighbourhood definitions was applied. Linear versus quadratic regression predictions with symmetric neighbourhood definitions of 2, 10 and 20 observations were compared. Thus, 2, 10 or 20 observations surrounding the to-be-lifted point on each side – 4, 20 or 40 observations in total – were used to come up with a linear or quadratic prediction.

The R-function *adlift* proposed by Nunes, Knigth and Nason (2006) was adjusted to be able to fit data sets larger than 500 observations. Still, the application was very computer intensive so adaptive lifting was solely applied to chromosome III and XIV. Moreover, in the quadratic case only neighbourhood lengths of 10 and 20 were computationally possible for the data of chromosome XIV.

The results provided in Figure 11 show the effects of both regression order and neighbourhood size. In the linear prediction case, the estimated function tends to be more linear. While the quadratic prediction, on the other hand, allows more curvature in the estimated curve. Increasing the size of the symmetric neighbourhood causes this observation to be intensified, levelling out the data. This is clearly shown in the case of 20 neighbours where the region of interest is smoothed strongly downwards, to create a practically linear curve. In all cases, unsmoothed parts with large spikes in both directions are observed. Moreover, no candidate regions for fine-mapping are provided in the results of the region of interest. Also, these regions are very comparable in all cases, with increased neighbourhood size leading to downward smoothed curves.

These observations are replicated in the results for chromosome III, albeit to a lesser extent since the observations already express a more flat profile (Appendix 5).

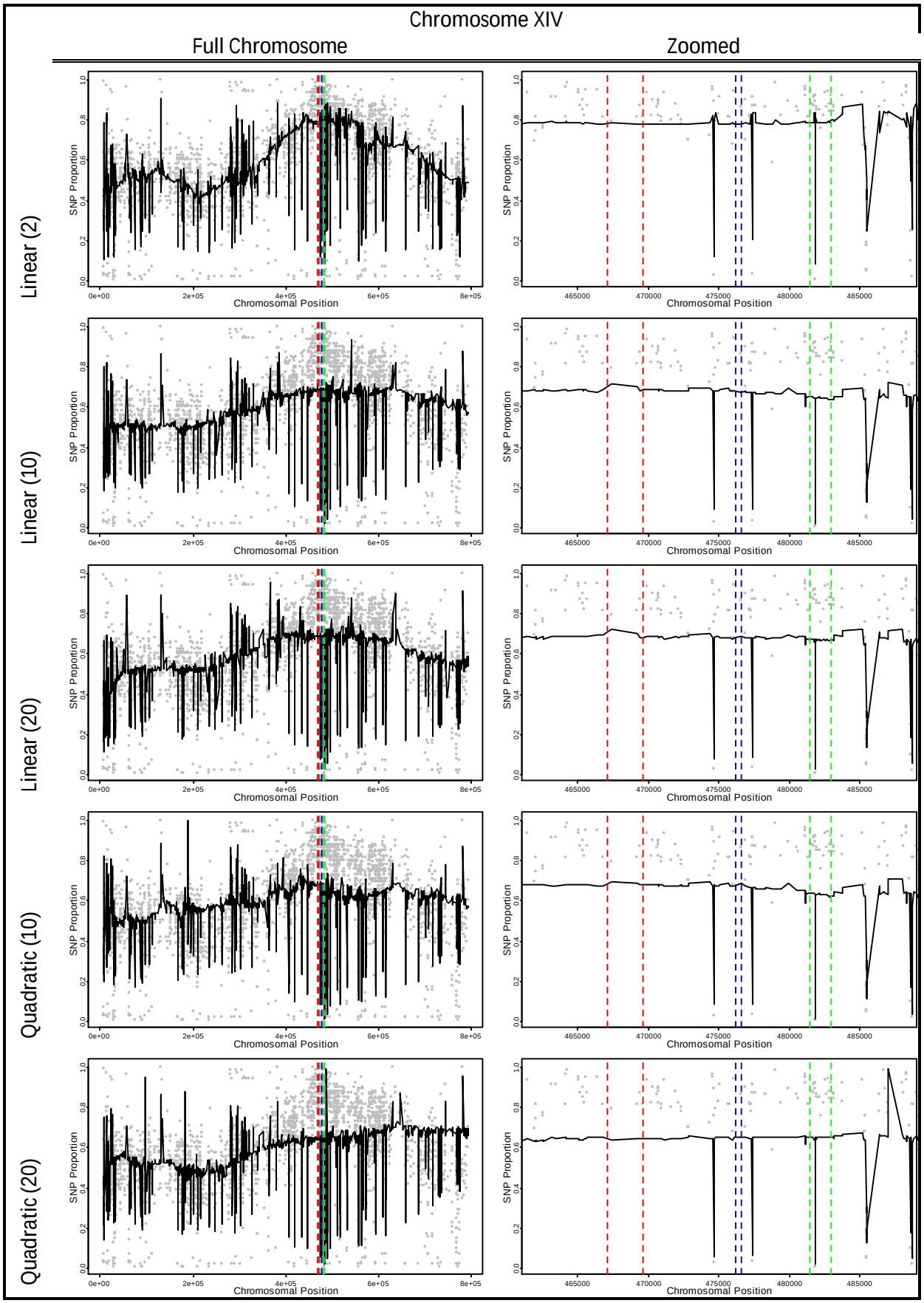


Figure 11: Adaptive lifting applied to chromosome XIV. Comparison of two different prediction orders (linear vs quadratic) and three neighbourhood sizes (2, 10 and 20 symmetrical neighbours) for full chromosome (left hand size plots) and region of interest (right hand side plots) smoothing.

4 Discussion and Conclusion

Wavelet function

Three different wavelet functions were employed to obtain smoothed estimates of the underlying probability curve generating the observed SNP frequencies, from the selected yeast segregants. Characterised by an increased number of vanishing moments, the Haar, Daubechies 3 and 9 wavelet functions are different in terms of smoothness; increased number of vanishing moments induces a smoother wavelet. Since the wavelet shrinkage methodology makes use of wavelet expansion, this behaviour is also found in the estimated curves. The estimated curves get softer, express less discontinuous jumps when wavelets with increasing vanishing moments are deployed. Though, unsmoothed parts with spikes are still observed, albeit to a lesser extent, in the case of smoother wavelets.

Applied to the selected yeast segregants problem, the idea was to find narrow plateau-like regions indicating increased co-segragation frequencies. This suggests that, based on its discontinuous characteristics, the Haar wavelet could provide good results. Nevertheless, due to the spiky behaviour of the fit, struggling to smooth out the curve, no such regions could be identified. The comparison to the control chromosomes, revealed that the wavelet shrinkage method is able to pick up some curvature out of the noise, since the estimated curves for the controls are overall flat.

Shrinkage threshold

The results of four different thresholds were compared to obtain an idea about their smoothing characteristics. As was suggested by Nason (2008) and is observed in the results, the SURE, FDR and Universal thresholds can be ordered providing the coarsest to smoothest estimates, respectively. Within the family of Universal thresholds, a different MAD estimate for σ leads to different results. In the case that the MAD of only the finest scale coefficients is used, the estimated curve is less smooth than in the case the MAD of all coefficients to be thresholded is computed. This is to be expected when the MAD is larger in the second case, implying extra noise is taken into account to compute the threshold.

In the problem of smoothing the selected yeast segregants data, the Universal method, which provides the smoothest estimated curve, is recommended. The results show that indeed the shrinkage method using the Universal threshold does a good job in retaining the available curvature in the data. Though, no useful narrow regions of increased proportions of co-segregation can be observed. In the case of the control chromosomes, the expected ordering of threshold methods is not observed. The SURE threshold does not differ very much from the largest Universal threshold of the two. This could indicate that applied to data in which no particular curvature is contained, the SURE threshold produces equally smooth estimate curve as the Universal threshold.

Primary resolution level

Considering the PRL, it can be concluded that the more wavelet coefficients are included in the shrinkage procedure, the more effect the shrinkage procedure has. Translated in terms of the primary resolution level, there exists an inverse relationship between primary resolution level and smoothness of the estimated curve.

For the problem at hand, the smallest of the selected primary resolution levels provides the smoothest results, but again no narrow region of increased proportions could be found.

Violation of normal noise assumption

To overcome the violation of the assumption that the noise incorporated in the data is independent identically distributed normal noise, three data transformations were investigated. The straightforward angular based transformations of Anscombe (1948) and Freeman and Tukey (1950) are performing well. Both transformations indicate that when not taking into account the binomial noise, wavelet shrinkage is over-smoothing the estimated curve. Between these two transformations, little difference in performance could be observed. Nevertheless, the estimates after the Freeman and Tukey (1950) transform were more compressed, leading to smoother curves.

Investigation of the recently proposed NN-Haar transform by Nunes and Nason (2009) provided degenerate results in which proportions were estimated to exceed 1. The problem seems to be occurring in the back transformation. The NN-Haar transformation is build to work on the frequency level instead of the proportion level. This means that some erratic parts in the frequency curve, caused by a small binomial sample size, will get smoothed upwards, even above their binomial sample size. The back transformation incorporates this smoothing and causes these upward smoothed parts eventually to exceed 1 on the proportion scale.

In terms of the selected yeast segregants problem, the results indicate that the Freeman and Tukey (1950) transformation is advisable. Compared to the untransformed case, there is not too much difference in results – probably due to sufficient binomial sample sizes – but this transformation ensures the assumption of normal distributed noise to be upheld.

Violation of non-equally spaced observations

To take the non-equally spaced observations in the data into account, adaptive lifting was applied. At the same time, since adaptive lifting does not impose a dyadic data set length, wavelet shrinkage could be performed on the whole data set. Results from two prediction orders, linear and quadratic regression, and three symmetric neighbourhood lengths, 2, 10 and 20 on each side of the to-be-lifted observation, were compared. The effect of prediction order reflects itself in an estimated curve featuring curvature close to the same order of the prediction order. Where the number of neighbours influences the extent to which this is the case. The more neighbours are taken to come up with the prediction, the more the prediction order will affect the estimated curve.

Considering the QTL-mapping data, no candidate regions of interest were provided. Overall, the estimated curves, of both chromosome XIV and III, are observed to have many unsmoothed spiky parts. This could be an indication that the empirical Bayesian shrinkage approach is under-smoothing the curve and perhaps different smoothing methods could be developed.

Conclusion

Considering the methodology, the current report has shown the versatility of the wavelet shrinkage approach in non-parametric curve estimations. Several characteristics can be tuned to fit specific needs of the researches in a very wide range. Wavelets can be chosen to fit an enormous range of different curves. Thresholds in combination with primary resolution levels can be selected based on

smoothing characteristics. Even when certain assumptions are violated, transformations in combination with different expansion algorithms – i.e. adaptive lifting – can be applied to ensure sound results.

Nevertheless, no clear narrow region of increased probability concerning the high-ethanol tolerance in *S. cerevisiae* could be extracted. Spiky behaviour of the estimated curves could not be excluded and in the region of interest an overall flat function was estimated. Probably a low signal-to-noise ratio imbedded in the data could be responsible for these observations.

In order to guide the selection of the wavelet shrinkage approach's characteristics and improve the application to the *S. cerevisiae* data, several suggestions for further research can be formulated. To give more insight in the efficiency of the curve estimations, confidence intervals could be constructed. This could also equip the researcher with a tool to fine-tune the wavelet functions and maximize the extracted information. Considering the application, careful thought in how to make the data collection less error and noise prone could potentially increase the signal-to-noise ratio in the data, providing better results from applying wavelet shrinkage.

Concerning, adaptive lifting, it seems like a promising method to apply wavelet shrinkage methodology to data sets containing non-equally spaced observations. Moreover, since no discrete wavelet transform is performed, the dyadic data length assumption no longer applies. Though, applicability of this method should be improved to ensure a more stable computer application. Preferably the method should allow the possibility to incorporate different threshold types, to have an alternative to the empirical Bayesian shrinkage, using a less computer intensive lifting algorithm.

5 References

Abramovich, F. and Benjamini, Y. (1996). Adaptive thresholding of wavelet coefficients. *Comp. Stat. Data anal.,* **22**, 351-361.

Anscombe, F. J. (1948). The transformation of poisson, binomial and negative-binomial data. *Biometrika,* **35**, 246-254.

Brauer, M. J., Christianson, C. M., Pai, D. A. and Dunham, M. J. (2006). Mapping novel traits by array-assisted bulk segregant analysis in Saccharomyces cerevisiae. *Genetics,* **173**, 1813-1816.

Burrus, C.S., Gopinath, R.A., and Guo, H. (1997). *Introduction to Wavelets and Wavelet Transforms: a Primer*, Prentice Hall, Upper Saddle River, NJ.

Campbell, F. W. and Robson, J. G. (1968). Application of Fourier analysis to the visibility of gratings. *Journal of physiology,* **197**, 551-566.

Daubechies, I. (1988). Orthonomal bases of compactly supported wavelets. *Comms. Pure Apll. Math.*, **41**, 909-996.

Daubechies, I. (1992). *Ten Lectures on Wavelets*, SIAM, Philadelphia, USA.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal denoising in an orthonormal basis chosen from a library of bases. *Compt. Rend. Acad. Sci. Paris A,* **319**, 1317-1322.

Freeman, M. F. and Tukey, J. W. (1950). Transformations related to the angular and the square root. *Ann. Math. Stat.,* **21**, 607-611.

Jansen, M., Nason, G. P. and Silverman, B. W. (2001). Scattered data smoothing by empirical Bayesian shrinkage of second generation wavelet coefficients. In Unser, M. and Aldroubi, A. (eds) *Wavelet applications in signal and image processing.* Proceedings of SPIE 4478, (87-97).

Johnstone, I. M. and Silverman, B. W. (2005). Empirical Bayes selection of wavelet thresholds. *Ann. Statistit.*, **33**, 1700-1752.

Lynch, M. and Walsh, J. B. (1998). *Genetics and analysis of quantitative traits,* Sinauer Associates, Massachusetts, USA.

Matsushika, A., Inoue, H., Watanebe, S., Kodaki, T., Makino, K. and Sawayama, S. (2009). Efficient bio-ethanol production by a recombinant flocculent *Saccharomyces cerevisiae* strain with a genome-integrated $NADP^+$-dependent Xylitol dehydrogenase gene. *Applied and environmental microbiology,* **75,** 3818-3822.

Nason, G.P. (2008). *Wavelet Methods in Statistics with R*, Springer, NY.

Nevoigt, E., Pilger, R., Mast-Gerlach, E., Schmidt, U., Freihammer, S., Eschenbrenner, M., Garbe, L. And Stahl, U. (2002). Genetic engineering of brewing yeast to reduce the content of ethanol in beer. *FEMS Yeast Research,* **2,** 225-232.

Nunes, M. and Nason, G. P. (2009). A multiscale variance stabilization for binomial sequence proportion estimation. *Statistica Sinica*, **19**, 1491-1510.

Nunes, M., Knight, M. I. and Nason, G. P. (2006). Adaptive lifting for nonparametric regression. *Stat. Comput,* **16**, 143-159.

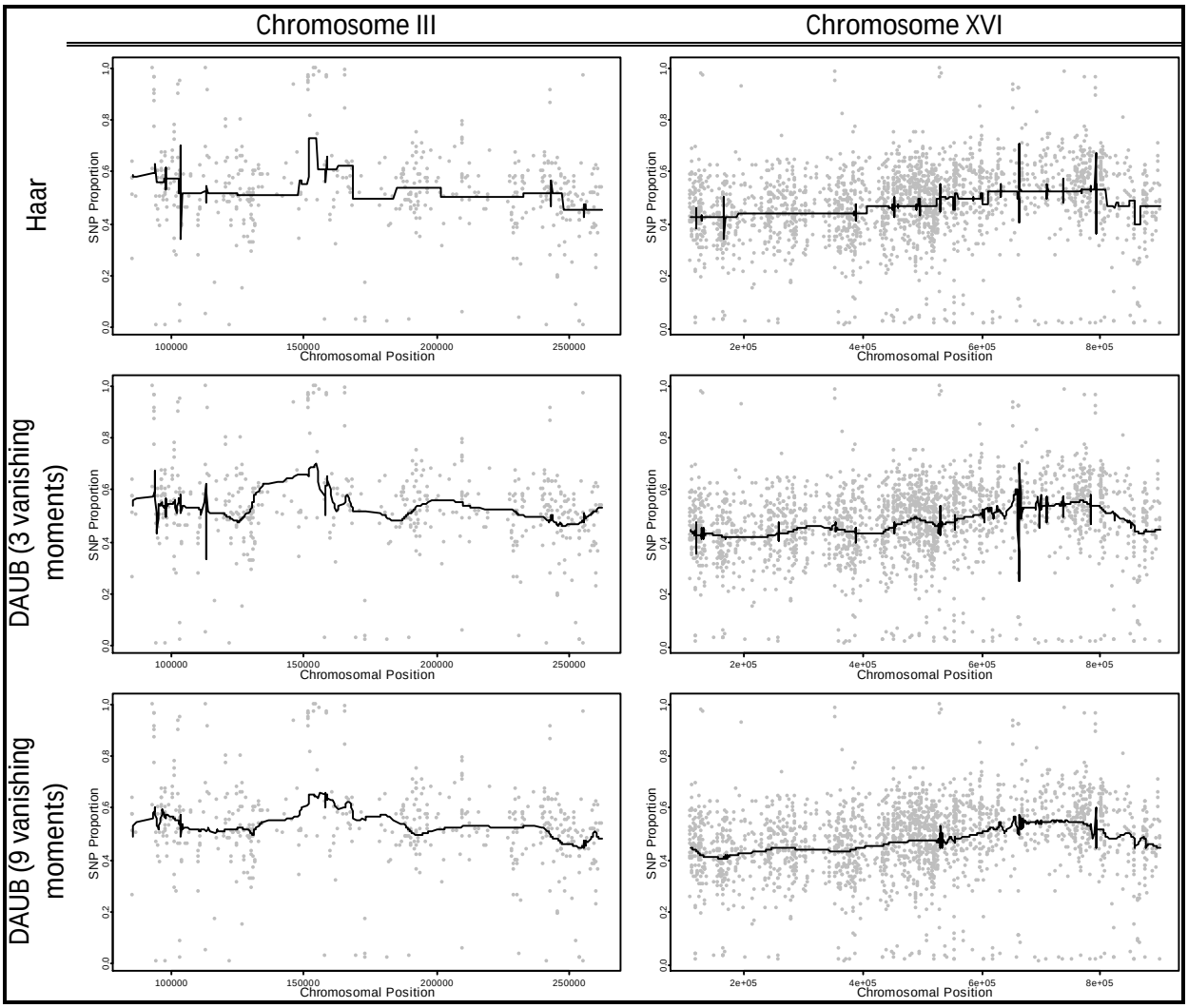
Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution*. Ann. Statist.,* **9**, 1135-1151.

Strachan, T. and Read, A. P. (1999). *Human molecular genetics*. BIOS Scientific Publishers, Ltd, Oxford, UK.

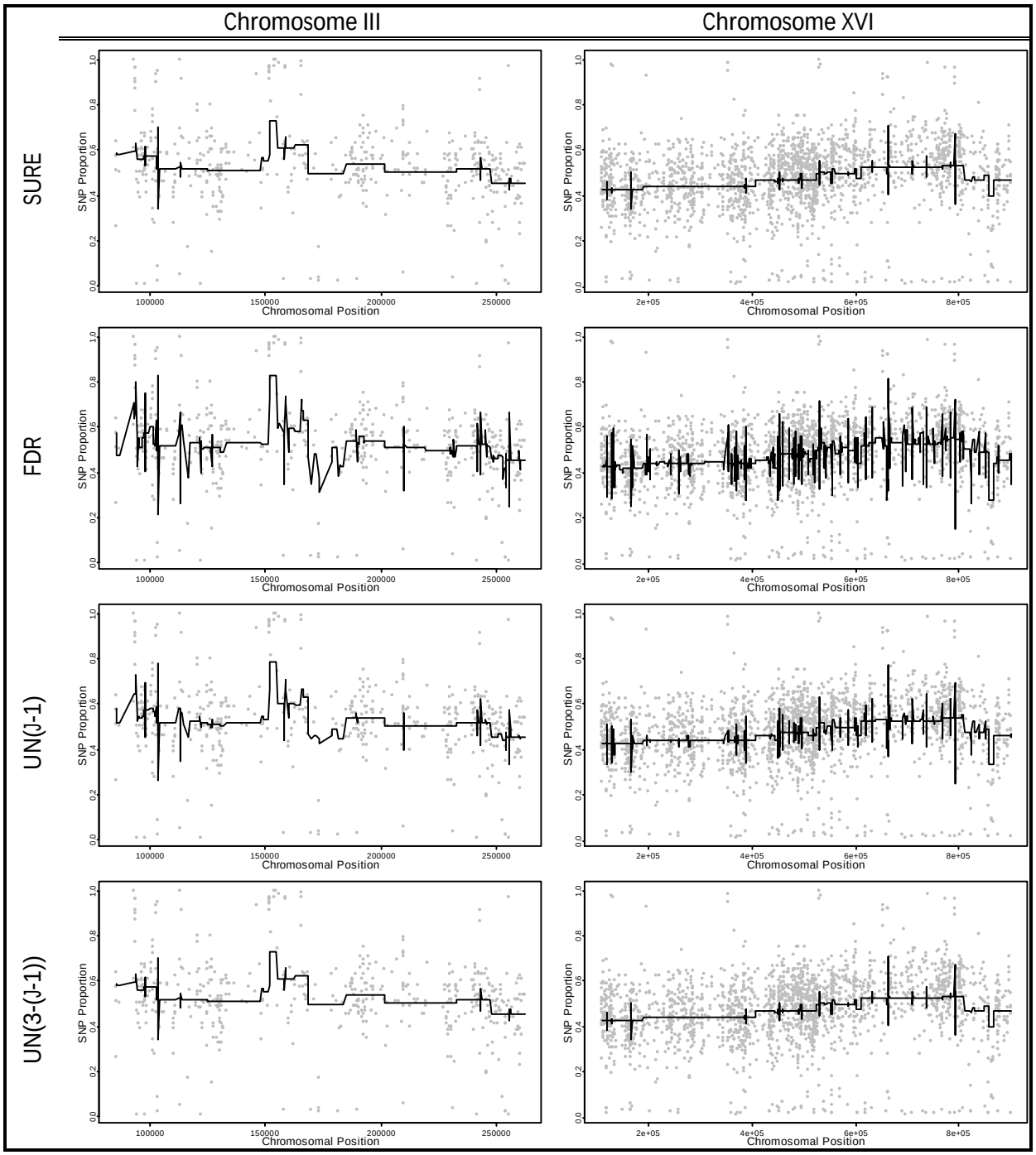
Swinnen, S. (2011). Polygenic analysis of high ethanol tolerance in *Saccharomyces cerevisiae*. *PhD thesis,* Katholieke Universiteit, Leuven [Unpublished].

Winzeler, E. A., Richards, D. R., Conway, A. R., Goldstein, A. L., Kalman, S., McCullough, M. J., McCusker, J. K., Stevens, D. A., Wodicka, L., Lockhart, D. J. and Davis, R. W. (1998). Direct allelic variation scanning of the yeast genome. *Science,* **281**, 1194-1197.

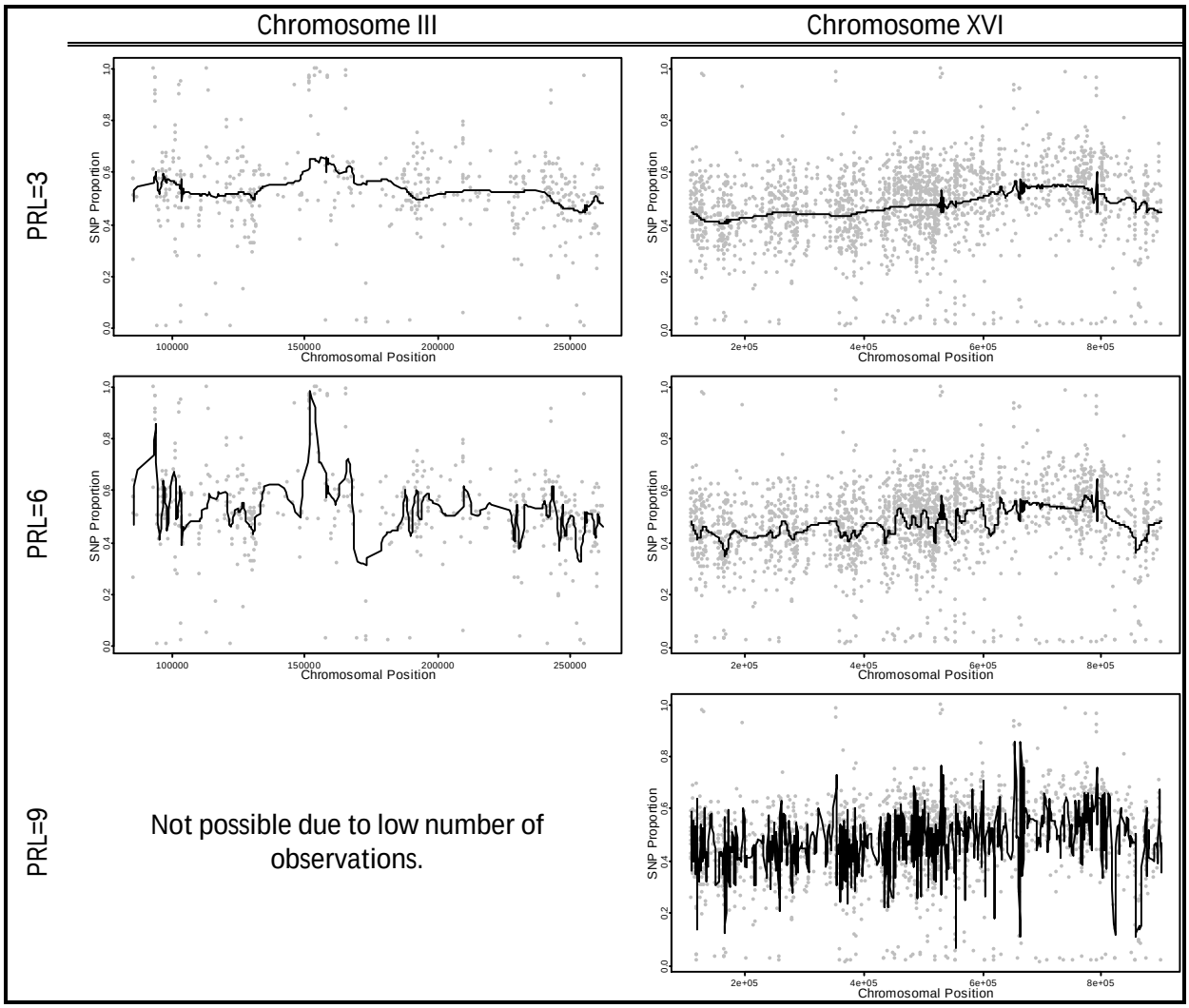
APPENDIX



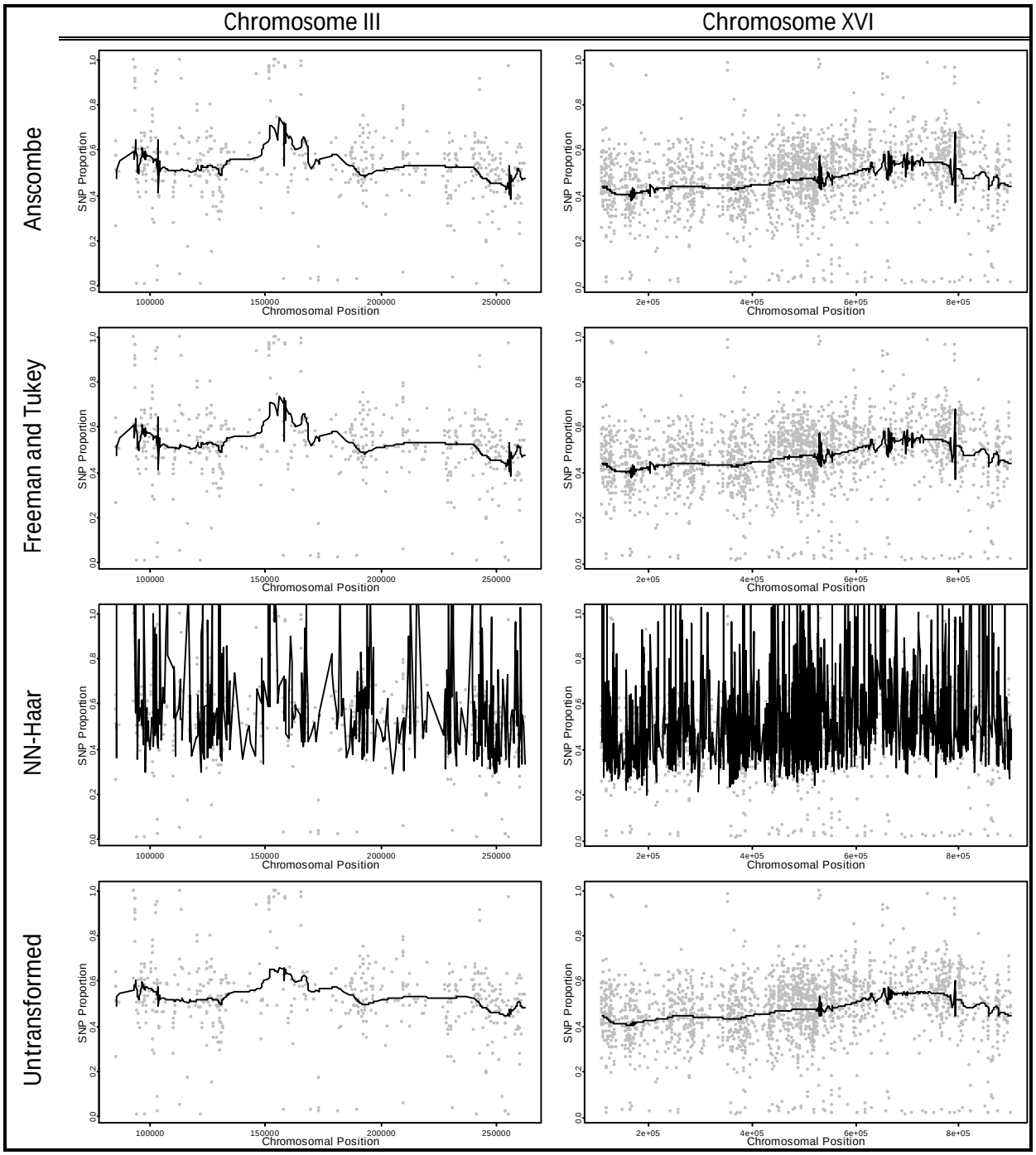
Appendix 1: Comparison of wavelet shrinkage method for three different wavelets. Solid black lines denote estimate of underlying function using Universal threshold using MAD of all coefficients from PRL(3) onwards. Comparison between the use of Haar and Daubechies wavelets with 3 or 9 vanishing moments. Left panels show estimation on full chromosome for chromosome III, right panels on chromosome XVI.



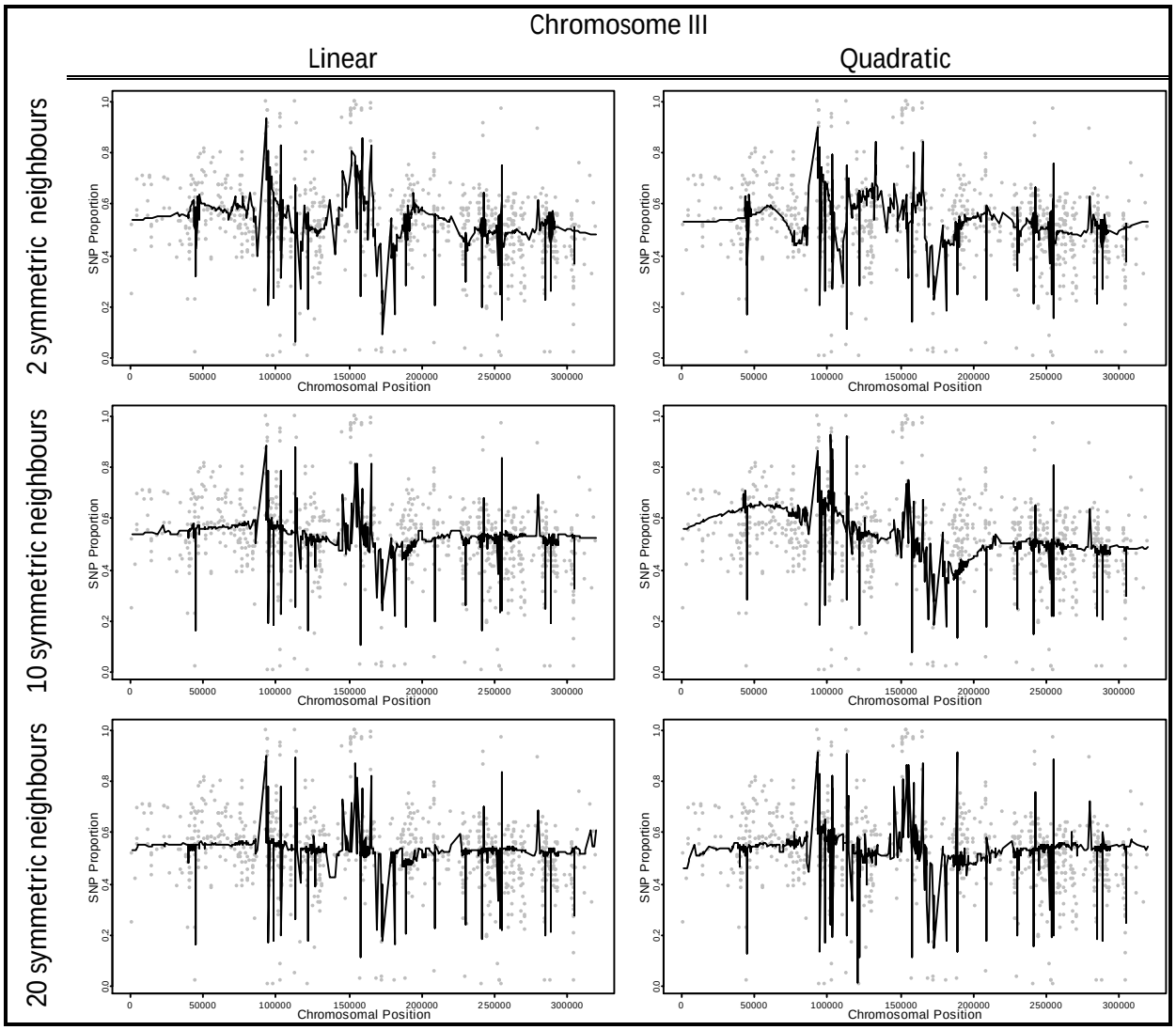
Appendix 2: Comparison of wavelet shrinkage method for three different threshold definitions. Solid black lines denote estimate of underlying function using Haar discrete wavelet transform and PRL 3. Comparison between the use of SURE, FDR or Universal, with MAD computed for different levels, thresholds. Left panels show estimation on full chromosome for chromosome III, right panels on chromosome XVI.



Appendix 3: Comparison of wavelet shrinkage method for three different primary resolution levels. Solid black lines denote estimate of underlying function using DWT with Daubechies wavelet with 9 vanishing moments and Universal threshold with MAD computed for wavelet coefficients from PRL to finest-detail level. Different PRLs are 3, 6 and 9. Left panels show estimation on full chromosome for chromosome III, right panels on chromosome XVI.



Appendix 4: Comparison of wavelet shrinkage method when assumption of normality is violated. Solid black lines denote estimate of underlying function using Daubechies' wavelet with 9 vanishing moments and Universal threshold with MAD computed for scaling levels from PRL to finest-detail level. Different transformations are Anscombe, Freeman and Tukey and NN-Haar accompanied by untransformed fit. Left panels show estimation on full chromosome for chromosome III, right panels on chromosome XVI.



Appendix 5: Adaptive lifting applied to chromosome III. Comparison of two different prediction orders (linear and quadratic) and three neighbourhood sizes (2, 10 and 20 symmetrical neighbours) for full chromosome (left hand size plots) and region of interest (right hand side plots) smoothing.

Appendix 6: R-Codes

Wavelet Denoising after Freeman and Tukey (1950) transformation.

```
# PACKAGES
library(wavethresh);
library(EbayesThresh);
library(binhf);

### DISCRETE WAVELET TRANSFORM
filtnr = 9;   # 1 = HAAR
fam = "DaubExPhase";
ThType = "soft";    # Threshold type: 'soft' or 'hard'
ThLev = 3;    # Primary Resolution Level (3 = default in WaveThresh)
type = 'wavelet';    # 'wavelet' = Decimated DWT, 'station' = time-ordered Non-Decimated DWT

# Removing extentive obs
J = floor(log2(N));
ex_ob = N-(2^J);
rm_obs = c(1:floor(ex_ob/2),(N-(ceiling(ex_ob/2)-1)):N);

prop = prop[-rm_obs];
position = position[-rm_obs];
n = n[-rm_obs];

counts <- as.matrix(Data[-rm_obs,14:18])
diMs <- dim(counts)
counts[which(counts=="-")] <- 0
counts <-
  matrix(as.numeric(counts),nrow=diMs[1],ncol=diMs[2])

#in case of an insertion, the column with deletions has to be set to 0. Otherwise the total number of successes equals the number of trials.
counts[which(Data[-rm_obs,6]=="-"),5] <- 0
SNP_freq <- rowSums(counts,)

new_N = 2^J;

# Apply Freeman and Tukey transformation
  SNP_freq_norm = free(SNP_freq,n);

# Fast Pyramidal Wavelet Transform
wave_trans = wd(SNP_freq_norm ,
filter.number=filtnr, family = fam, type = type);

## SURE thresholding of coefficients
Thsure_coeff = threshold(wave_trans, levels =
ThLev:(nlevels(wave_trans) - 1), type = ThType,
policy="sure", dev=madmad);

Thsure = threshold(wave_trans, levels =
ThLev:(nlevels(wave_trans) - 1), type = ThType,
policy="sure", dev=madmad,
return.threshold=T)[1];

## Take the inverse Discrete Wavelet Transform
freq_trans = wr(Thsure_coeff);

# Inverse Freeman and Tukey transform
freq_est = freeinv(freq_trans, n);
estimates = freq_est/n;

## False Discovery Rate thresholding of coefficients
Thfdr_coeff = threshold(wave_trans, levels =
ThLev:(nlevels(wave_trans) - 1), type = ThType,
policy="fdr", dev=madmad);

Thfdr = threshold(wave_trans, levels =
ThLev:(nlevels(wave_trans) - 1), type = ThType,
policy="fdr", dev=madmad, return.threshold=T)[1];

# Take the inverse Discrete Wavelet Transform
estimates = wr(Thfdr_coeff);

# Inverse Freeman and Tukey transform
estimates = freeinv(estimates,n);
estimates = estimates/n;

## Universal thresholding of coefficients (Donoho and Johnstone(1994b)) computed on finest scale
FineCoefs = accessD(wave_trans,
lev=nlevels(wave_trans)-1);
sigma = mad(FineCoefs);
UthDJ = sigma*sqrt(2*log(length(prop)));

Th_coeff = threshold(wave_trans, levels =
ThLev:(nlevels(wave_trans) - 1), type = ThType,
policy="manual", value=UthDJ,
return.threshold=F);
ThDJ = threshold(wave_trans, levels =
ThLev:(nlevels(wave_trans) - 1), type = ThType,
policy="manual", value=UthDJ,
return.threshold=T)[1];

# Take the inverse Discrete Wavelet Transform
estimates = wr(Th_coeff);

# Inverse Freeman and Tukey transform
estimates = freeinv(estimates,n);
estimates = estimates/n;

## Universal thresholding of coefficients
Thun_coeff = threshold(wave_trans, levels =
ThLev:(nlevels(wave_trans) - 1), type = ThType,
policy="universal", dev=madmad);
```

```
Thun = threshold(wave_trans, levels =
ThLev:(nlevels(wave_trans) - 1), type = ThType,
policy="universal", dev=madmad,
return.threshold=T)[1];

# Take the inverse Discrete Wavelet Transform
estimates = wr(Thun_coeff);

# Inverse Freeman and Tukey transform
estimates = freeinv(estimates,n);
estimates = estimates/n;
```

Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:
Statistical analysis of selected yeast segregants

Richting: **Master of Statistics-Biostatistics**
Jaar: **2011**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Garcia Barrado, Leandro

Datum: **12/09/2011**