Detecting binding sites in PAR-CLIP data using a Bayesian hierarchical model

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Eva-Maria Hüßler
aus Ahaus

Düsseldorf, Juli 2021

aus dem Mathematischen Institut
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Berichterstatter:

1. Prof. Dr. Holger Schwender
   Heinrich-Heine-Universität Düsseldorf

2. Prof. Dr. Jörg Rahnenführer
   Technische Universität Dortmund

Tag der mündlichen Prüfung:

22. September 2021

Abstract

MicroRNAs (miRNAs) play an important role in gene regulation by interacting with messenger RNA (mRNA) sites. To detect these binding sites on the mRNA, biochemical methods such as Photactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation (PAR-CLIP) can be conducted. The PAR-CLIP method introduces T-to-C substitutions at sequenced cDNA that help to distinguish between binding site positions in the PAR-CLIP data and noise. T-to-C substitutions could, however, also occur due to other reasons, such as SNPs or mismatches. Most of the few existing statistical procedures for detecting binding sites in PAR-CLIP data do not account for types of substitutions other than PAR-CLIP induced substitutions. None of the existing methods enable the inclusion of additional information that are relevant for the biology of miRNA binding sites, such as the type of mRNA region that can help to detect binding sites. Moreover, the focus of existing models lies in detecting binding sites from only one experiment.

To detect binding sites, BayMAP, a Bayesian hierarchical mixture model taking other sources of substitutions into account, will be presented in this thesis. This allows the incorporation of additional information as well as a structure for reflecting dependencies of substitution positions very close to each other. The incorporation of additional information and neighborhood dependencies allows a better detection of miRNA-binding sites. Additionally, it also offers a better understanding of the biology of binding sites. Moreover, a method to combine the results of several PAR-CLIP experiments is developed to state the prediction of binding sites more precisely. Finally, BayMAP is compared to existing models in applications to real PAR-CLIP data sets as well as simulated data sets.

Acknowledgment

First of all, I want to thank my supervisor Holger Schwender. He not only made this thesis possible for me, he was also always available with valuable advice while giving me the freedom I needed. The fruitful working atmosphere and his support in scientific related business trips were an important benefit. Thank you Holger, you made me evolve scientifically as well as personally. I also want to thank Pablo Landgraf, who supported me regarding the application field, e.g., genetics, the PAR-CLIP method and publicly available data sets. Martin Schäfer, a former member of Hogler Schwender's research group, contributed to this work with his experience on Bayesian statistics. The first three years of my thesis were funded by the Düsseldorf School of Oncology (DSO). I wish to thank the DSO, in particular Cornelia Höner, financing this project and organizing network events, such as the DSO retreat as well as the support in business trips and in the training of key qualifications. Through the DSO retreat, I met Andreas Klötgen, who shared his knowledge about aligning PAR-CLIP data and creating simulated data sets with me. I also want to thank Wolfgang, my first desk neighbor, who supported me in tackling diverse issues at the beginning of the project. I wish to thank our working group and the employees of the mathematical institute for any topic related or non-topic related discussion (with special thanks to Tobias and Felix) as well as my friends (in particular Kira, Virginie and Lisa). Last but not least, I thank my parents and my family who supported me in every direction I took (thanks again Claire and Stephan). There are countless other persons to whom I am very grateful and who contributed in one way or another to this work. Even if I do not mention you by name, this work would not have been possible without your help. Thank you!

Contents

Chapter 1

Introduction and Motivation

MicroRNAs (miRNAs) are non-coding RNAs in the length of about 22 nucleotides (nt) which play an important role in gene regulation [4]. With the aid of Argonaute proteins (Ago), miRNAs bind to target messenger RNA (mRNA) and thereby repress translation and destabilize mRNA [45]. Over the last years, their involvment in cancer has been studied (see, e.g.,[1, 16, 52]). Some miRNAs have been shown to be functional in leukemia of lymphoid origin in adults [34] as well as of myeloid origin in adults [21] and children [13]. Identification of Ago-binding sites on target mRNA is crucial for understanding miRNA functions. MiRNAs, mRNAs and their function will be discussed in Section 2.1.

Crosslinking and immunoprecipitation (CLIP) methods followed by high-throughput sequencing [24, 32, 38] is currently the standard method for the identification of target mRNAs. Gene regions that bind to a protein of interest, here Ago, can be identified or isolated by using a protein specific antibody. Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation (PAR-CLIP) is one of these CLIP procedures and was initially developed by Hafner et al. [24]. PAR-CLIP not only achieves a high efficacy but also induces a transition of thymine to cytosine (T-to-C) in the sequenced complementary DNA (cDNA). PAR-CLIP will be discussed in Section 2.2.

PAR-CLIP can identify genomic binding sites with high efficacy, but due to limited specificity, a secondary differentiation of real binding sites and noise is necessary. For the purpose of the distinction of real binding sites and noise via a statistical method, the PAR-CLIP induced T-to-C substitutions can be helpful. However, not all observed

T-to-C substitutions are due to PAR-CLIP. They can, for example, also be single nucleotide polymorphisms (SNPs), that are variations in the genome. In addition, not all reads that overlap a T position within a binding site show the specific T-to-C substitution. Very low and very high substitution rates and therefore positions, where either nearly all or nearly none of the observed reads contain the specific T-to-C substitution, are often due to other reasons than induction by PAR-CLIP [28, 49]. Very low substitution rates may occur due to sequencing errors or mismatches. Very high substitution rates can, e.g., result from homozygous SNPs or RNA highly similar, but not identical to the mapped reference genome [20]. The purpose of this thesis is to distinguish between PAR-CLIP induced T-to-C substitution positions and non-PAR-CLIP induced ones, so that binding sites are identified.

Only a few methods to analyze PAR-CLIP data currently exist. Hafner et al. [24] developed the PAR-CLIP method and discovered the T-to-C substitutions that are typical for this method. They were also the first to propose a method to detect binding sites by considering T-to-C substitutions. In their work, a genomic region is declared as a binding site, if at least five overlapping reads are observed, of which at least 20% consist of T-to-C substitutions. A window of 41 nucleotides is then centered around the position with the most frequent T-to-C substitutions. This method is very easy to apply but with choices more or less arbitrary so that there may be many findings of false positives or false negatives. Additionally, this simple approach would favor positions, like SNPs, with substitution rates close to 100%.

PARalyzer developed by Corcoran et al. [12] is currently the most cited and the first statistical tool for finding binding sites in PAR-CLIP data. This method uses a kernel density estimation approach, where the estimated density for T-to-C substitution rates is compared to the estimated density for non-substitutions, i.e. T-to-T. This tool will be described in more detail in Section 4.1.

Jaskiewicz et al. [28] developed two distinct tools for the identification of binding sites that were available on the website CLIPZ [43]. However, the website is no longer maintained. One tool is based on enrichment in relation to mRNA-seq and one on T-to-C substitutions. The first method relies on the idea that binding sites with a high affin-

ity to Ago have a higher probability to crosslink with Ago and thus a higher number of reads. Nevertheless, the number of reads also depends on the abundance of the mRNA. Therefore, they take the number of PAR-CLIP reads into account in relation to the number of expected reads, i.e. expected mRNA reads of sequencing without prior crosslinking and immunoprecipitation.

For the second proposed method Jaskiewicz et al. [28] suggest that the number of T-to-C substitutions can be modeled by a binomial distribution. To distinguish PAR-CLIP induced substitution positions from positions with mismatches or SNPs, they suppose that the probability of having a method-induced substitution position is high if a position's substitution rate is probably within a range of a prespecified upper and lower bound. This range, however, has to be defined in advance, e.g., by considering a trusted set of binding sites, like a known set of binding sites which have a high coverage in the data set. Their second method takes into account the specific T-to-C substitutions but depends on the choices for the trusted data set.

The wavClusteR method, proposed by Sievers et al. [49], enhanced by Comoglio et al. [11] and published at the same time as the methods by Jaskiewicz et al. [28], is based on the same idea as the previous method of ranking sites by T-to-C substitutions. They also presume that for one T-to-C substitution position, the number of substitutions follows a binomial distribution. Moreover, they also consider information from types of substitutions other than T-to-C to better distinguish between PAR-CLIP induced substitutions and other substitutions. For this purpose, they propose a two-component mixture model, that will be described in more detail in Section 4.2.

BMix, developed by Golumbeanu et al. [20] is another mixture model, that takes into account low- and high-frequency errors. BMix is based on the idea of wavClusteR but proposes a three-component mixture model. They additionally assume existence of dependencies between the different substitution probabilities of each mixture component. A mismatch or sequencing error can for example occur at positions that are supposed to be SNPs, so that the probability of observing a substitution at a SNP position depends on the probability of observing a mismatch. BMix is described in more detail in Section 4.3.

STAMMP developed by Torkler [51] is also a mixture model based on the same idea as wavClusteR [49]. The distribution of the number of substitutions due to mismatches or SNPs, is estimated by a two-component mixture model taking into account all substitutions except T-to-C. p-values for T-to-C substitution positions are then calculated for the null hypothesis that T-to-C substitutions are caused by errors. However, p-values are calculated in such a way, that they favor SNP positions to be declared as PAR-CLIP induced.

Additionally, there are a handfull of other methods. PARma developed by Erhard et al. [15] is a method, where clusters are built via overlapping reads in a similar manner to Hafner et al. [24], but allowing clusters to overlap. They also analyze T-to-C substitutions, but mainly focus on detection of miRNAs binding to detected clusters. PIPE-CLIP proposed by Chen et al. [8] detects enriched clusters via a zero truncated negative binomial model and then takes into account the substitution rate. MiClip by Wang et al. [53] and a method by Yun et al. [55] are based on hidden marcov models, where T-to-C substitutions are also taken into account.

Most of the described methods consider the specific T-to-C substitutions observed in PAR-CLIP data. Only few of them account not only for mismatches but also for SNPs for the non-method-induced substitutions (in particular CLIPZ, wavClusteR and BMix). None of the methods allow to incorporate additional information that can also be helpful for distinguishing PAR-CLIP induced T-to-C substitutions from non-PAR-CLIP induced ones.

It is, for example, well known that binding sites occur most often in the 3'UTR of the mRNA than in the CDS and than in the 5'UTR [5]. This information could be considered by using a Bayesian framework. wavClusteR is already set in a Bayesian context. However, a posterior density for the substitution rate is computed for each position separately before a mean of all positions' posteriors is calculated. This means that the posterior for a substitution rate highly depends on the total number of reads for the considered position whereas the estimation of the substitution rate's density could be much more precise by considering all positions simultaneously.

Here, a Bayesian method will be proposed, where all positions are considered simultaneously for the computation of the posterior density. This method distinguishes, in a similar way as BMix and wavCluster, between method-induced and non-method-induced T-to-C substitution positions and allows additionally for supplementary information, such as the 3'UTR, the CDS and the 5'UTR or other variables, that may be important for binding sites. Dependencies between the different substitution probabilities as proposed by BMix are also considered to provide a better estimation and a better distinction.

In this thesis a three-component mixture model fulfilling the above requirements will be presented and tested for the capability to distinguish between method-induced T-to-C substitution positions, mismatches and SNPs via a Bayesian approach. Therefore, Section 3 gives a theoretical background for Bayesian data analysis adapted to the requirements stated above, i.e. a mixture model that allows the incorporation of additional information. We already presented this method under the name of BayMAP (Bayesian hierarchical model for the analysis of PAR-CLIP data) in Huessler et al. [27]. The model and its results are represented as a main part of this thesis.

However, in Huessler et al. [27] it is assumed that every position is independent even if positions that are very close to each other are probably on the same binding site and therefore not independent. To capture this structure, a derivative of BayMAP will also be presented in this thesis. The model of BayMAP as published [27], here called BayMAP 1.0, will ne shown in Section 5.1. The advanced model of BayMAP, here called BayMAP 2.0, will be presented in Section 5.3.

BayMAP as well as several other models, including wavClusteR and BMix, is a method that is position based. This means that for each T-to-C substitution position, it decides if the substitutions are probably due to the PAR-CLIP method and if the substitution position therefore probably lies on a binding site. It is, however, also of interest not only if the position lies on a binding site but also which region spans the binding site. Hence, in Section 4, it will be described how binding site regions are identified in wavClusteR and BMix. In Section 5.2, a new method of identifying binding site regions will be presented.

None of the above presented methods for the analysis of PAR-CLIP data combines information from multiple PAR-CLIP data sets. The information of a PAR-CLIP data set is only based on one sample in one experiment. In order to decide if a position is a general ubiquitous binding site position, it would however, improve the prediction if data from several samples were used. In Section 5.4 a method combining the information of several PAR-CLIP data sets or experimental replicates using BayMAP will be presented.

In order to validate the model, an extensive simulation study is executed and presented in Section 6. Moreover, BayMAP 1.0 and BayMAP 2.0 are applied to several publicly available PAR-CLIP data sets. For PAR-CLIP data sets that were created under the same experimental conditions, the information of these data sets is combined. Results are then compared to other methods for the analysis of PAR-CLIP data (Section 7). Finally, in Section 8 results are discussed.

Chapter 2

Biological Background

The aim is to detect miRNA binding sites on the mRNA, as the interaction between miRNAs and mRNAs play an important role in gene regulation. In this section, first the interaction itself will be explained. Afterwards, a method that helps to detect these interactions on the mRNA, the PAR-CLIP method, will be presented. Finally, PAR-CLIP data will be shown in a descriptive way.

2.1 microRNA-mRNA interactions

To understand the miRNA-mRNA interaction, first a general genomic background will be given (Section 2.1.1) before the interaction will be described in more detail (Section 2.1.2).

2.1.1 Genomic background

The human genetic information is stored in every cell of the human body. Every cell consists usually of 23 pairs of chromosomes that were passed on by the parents. Females have a pair of the chromosomes 1 to 22 as well as two X chromosomes whereas males carry an X and a Y chromosome instead of the two X chromosomes [48]. Each chromosome forms a double helix by two strands of coiled deoxyribonucleic acid (DNA). The double helix consists of two sugar phosphate backbones and inward directed ladders of nucleotides, that is adenine (A), cytosine (C), guanine (G) or thymine (T). The two strands of DNA are bound together, where A is always bound to T and C always to G. It is therefore sufficient to know one of the two DNA strands to describe the other

one [48].

The DNA of two individual human subjects are more than 99% identical. Nevertheless, millions of bases with genetic variations remain [48]. The most common genetic variation is the single nucleotide polymorphism (SNP). A SNP is a variation of a DNA base that occurs in more than 1% of the population [48]. If for example C is observed for a specific DNA position in the majority of the population, but A is observed in a minority but not negligible part of the population, the variation is called SNP.

Parts of the DNA are copied into ribonucleic acid (RNA) by unwinding the double helix and then synthesizing complementary bases to the single DNA strand. In contrary to DNA, in RNA, uracil (U) is the complementary base pair to A instead of T [3]. When knowing the RNA, one can conclude from which nucleotide bases in the DNA the RNA was copied from. The process of copying the DNA into RNA is called transcription.

The RNA molecules can be divided into protein coding and non-coding RNA. Messenger RNAs (mRNAs) are protein coding. The segments of the DNA that encode for mRNAs are called genes. The information of the mRNA is translated into proteins. The process from the mRNA to a protein is therefore called translation. These proteins are essential for the cell structure and cellular activities [3].

Beside the protein coding RNA, i.e. mRNA, several types of non-coding RNA with different functions exist. This includes microRNAs (miRNAs). MiRNAs are responsible for gene expression regulation, i.e. the regulation of translation and the stability of mRNAs [6]. Therefore, over- or underexpression of miRNAs can be linked to diseases such as cancer [52].

2.1.2 mRNAs, microRNAs and their interaction

Before an mRNA can be translated into proteins, the primary RNA transcript is modified [3]. The two ends of an mRNA are called 5' end and 3' end. The primary mRNA contains two untranslated regions (UTR) at the two ends, the 5' UTR and the 3' UTR. In between the ends, there are exons and introns, where only exons are coding sequences (CDS), i.e. sequences that are translated into proteins. In the modification process of

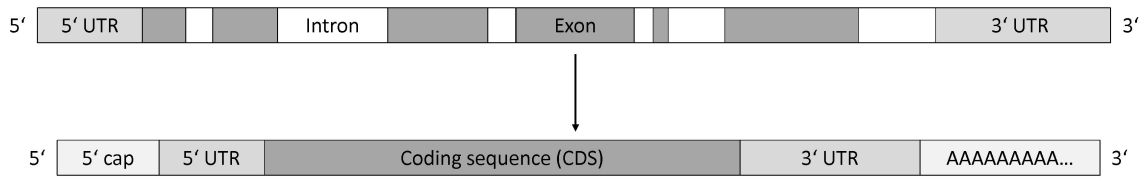

Figure 2.1: The process of primary mRNA to mature mRNA

the primary RNA transcript, introns are deleted by a process called splicing. At the 5’ end the 5’ cap is added and at the 3’ end the poly(A)-tail, a long RNA tail only containing adenine bases [3]. The result is the mature mRNA, where the exons are combined to the CDS in the middle of the mRNA (see Figure 2.1).

However, the translation into proteins, and therefore the levels of its protein product, also has to be regulated [3]. One important way of regulation is processed by the interaction of the mRNA with miRNAs.

MiRNAs are very small RNA sequences in the length of around 22 nucleotides (nt). They arise out of transcripts in the form of a hairpin, i.e. transcripts that fold in on themselves [4]. In a similar way to mRNAs, the primary miRNA is modified by splicing until only the 22 nt long mature miRNA remains. The mature miRNA is bound by Argonaute (Ago) proteins. The resulting complex is then termed an RNA-induced silencing complex (RISC). This complex can bind to target mRNA in a partial complementary fashion, which leads to destabilization, degradation or translational inhibition [54]. Only one Ago protein, Ago2, can directly cleave mRNA upon fully complementary binding to its target mRNA [3].

For the interaction between miRNAs and their target mRNA, one can distinguish between beneficial properties of the miRNA and the mRNA [5] as described in the following.

On the mRNA, the binding occurs most often on the 3’UTR [5]. However, it also occurs on the CDS and even less frequent on the 5’ UTR. The effectiveness of the binding, i.e. the power of inhibiting protein translation, is higher, the closer the binding site is to the poly(A)-tail. In regions on the mRNA with a lot of AU bases (AU-rich content) the interactions are also more effective [5]. It is also possible to have more than one binding

site on each mRNA even in close proximity. Those can exhibit concerted function.

In general, miRNAs, that are conserved across species (e.g., human, mouse, rat, dog, and chicken) have more target sites than unconserved ones [36]. Moreover, complementary base pairing (Watson-Crick pairing) is essential for binding of miRNA to mRNA. In binding sites, the Watson-Crick pairing most often comprises positions 2 – 7 of the 5' end of the miRNA, the so called seed [5]. A canonical target has complementary base pairing with the miRNA on positions $2-7$ plus an adenine base at position 1 or on positions $2-8$ or on positions $2-8$ plus an adenine base at position 1 [5]. There are, however, many different possibilities for the pairing, e.g., only on the seed or on positions $3-8$ or a pairing with one mismatch position on the seed. Additional Watson-Crick pairing on positions $13-16$ with at least three pairs can increase efficacy of the binding [5].

2.2 PAR-CLIP data

The biochemical method PAR-CLIP will be presented, as this method allows the identification of miRNA targets on the mRNA. Not only the PAR-CLIP method itself will be described in detail (Section 2.2.1), but also the process of preparing the data after the PAR-CLIP experiment (Section 2.2.2), so that further analyses can be conducted with the preprocessed data. Moreover, a descriptive analysis of the considered PAR-CLIP data sets will be realized in Section 2.2.3, as the descriptive analysis helps to understand the PAR-CLIP data and how the data could be used to detect binding sites.

2.2.1 Target recognition - PAR-CLIP

Several bioinformatics tools for the prediction of target mRNAs for a specific miRNA already exist, and work by looking for Watson-Crick pairing and other conventional properties. One of the most commonly used target prediction tools is TargetScan [2]. These tools, however, predict hundreds of mRNAs for each prespecified miRNA. Furthermore, they primarily predict canonical and conserved targets on the 3'UTR. However, other targets also exist [10, 29]. Instead of predicting targets, experimental approaches can be used in order to identify miRNA targets.

MiRNAs build a RISC complex with one of the four human Ago protein so that they can bind to the target mRNA [46]. To detect the binding sites on the mRNA, immunoprecipitation methods can be employed. These biochemical methods use protein specific antibodies, so that the protein complex can be isolated [7]. To identify binding sites on mRNAs, one can therefore use Ago-specific antibodies in order to isolate the RISC complex with the associated mRNA. Crosslinking and Immunoprecipitation (CLIP) methods expose the cultured cells to ultraviolet light (UV) prior to immunoprecipitation. UV crosslinking promotes covalent bonds between proteins and RNA, stabilizing the complex during the experimental isolation procedure [9]. The crosslinking can be enhanced by incorporation of photoreactive ribonucleoside (i.e. a nucleobase combined with a ribose sugar) analogs into the cells, namely 4-thiouridine (4-SU) or 6-thioguanosine (6-SG). Those are incorporated into mRNA instead of a uridine or guanosine , respectively during transcription. This method, developed by Hafner et al. [24] and called PAR-CLIP (Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation), achieves a high efficacy of crosslinked RNA and is therefore discussed in detail in this thesis.

Once crosslinking (in the cell) and immunoprecipitation (after lysis, i.e. after breaking down the cell) is applied, several steps have to be fulfilled. One is specific to the bound RNA, so that the ends of a bound RNA are digested and the background of non-crosslinked fragments reduced. Proteins are also degraded. The next step targets the remaining mRNAs, i.e. the mRNAs that are bound to Ago, by sequencing these fragments, i.e. determining the sequence of nucleotides of the RNA fragments. The RNA sequences cannot be sequenced directly, but have to be transformed and multiplied. First, adapters are added to the RNA fragments, then the RNA fragments are transcribed into complementary DNA (cDNA), as this cDNA can be amplified by Polymerase Chain Reaction (PCR), that is a method for generating DNA copies. Finally, these cDNA copies are sequenced. The final collection of these cDNA copies is called the cDNA library. The sequenced cDNA can then be aligned to the genome so that the RNA fragments, i.e. the binding sites of the mRNA, can be identified and associated with their genomic positions. The observed sequences are then called reads and one can calculate how many reads a genomic position has.

However, not all of the background or noise, i.e. fragments that were not bound to Ago, can be removed from the sample. It is therefore important to distinguish between noise and binding sites. The PAR-CLIP method not only has a higher accuracy in comparison to other CLIP methods by the incorporation of 4-SU (or 6-SG) but also induces a transition of thymine to cytosine (T-to-C) in the sequenced cDNA in 4-SU crosslinked samples (or guanine to adenine (G-to-A) in the case of 6-SG). This is introduced by the reverse transcriptase enzyme in the process of cDNA library generation. The presence of substitutions in crosslinked fragments, and therefore on binding sites, can aid to distinguish between noise and real binding sites. In the following, only 4-SU, and therefore T-to-C substitutions, are considered, as it is known for a higher efficacy than 6-SG and is, therefore, typically used [24].

2.2.2 Data preprocessing

Five publicly available PAR-CLIP data sets from three different studies are considered for further analyses [27]. Two data sets are from Kishore et al. [30] with SRA accession numbers SRR189784 and SRR189785 (here called Kishore A and Kishore B), one from Memczak et al. [41] with SRA accession number SRR650321 (here called Memczak) and two from Gottwein et al. [22] with SRA accession numbers SRR343336, SRR343337 (here called Gottwein A and Gottwein B). Only data sets with Ago2 as the protein of interest are chosen, as Ago2 is the most prominent Ago protein for gene silencing.

The two PAR-CLIP experiments by Kishore et al. [30] as well as the experiment by Memczak et al. [41] are conducted with cells from the human embryonic kidney (HEK) 293 cell line. HEK 293 cells are transformed primary cultures of HEK cells [23] that are widely used, e.g., for protein interaction studies [39] such as PAR-CLIP studies. Kishore et al. [30] used the HEK 293 cell line in order to compare CLIP with PAR-CLIP. Memczak et al. [41] analyzed the HEK 293 cell line to investigate if the circular RNA (that is RNA that forms a loop) CDR1-AS has miRNA binding sites. Gottwein et al. [22] wanted to analyze the role of miRNAs in primary effusion lymphoma. The two considered PAR-CLIP experiments from Gottwein et al. [22] are conducted with BC-3 cells, a commonly used cell line for the analysis of primary effusion lymphoma.

These data sets cannot be used directly but have to be preprocessed. During cDNA library preparation for RNA sequencing, adapters are added to the 3' end of the sequences. These adapters have to be removed from the data so that the observed reads contain only the sequences of interest. Adapters as well as read ends with low base quality are removed by using the bioinformatical tool cutadapt [40].

For the preprocessing of the data, one also has to decide which reads are discarded because of an inadequate read length. In wavClusteR and BMix, reads of a minimum length of 15 nt [49] and 14 nt [20] are considered for the analysis. Due to the crystallized structure of the RNA silencing complex [44], it can be assumed that at least 14 nt are well protected by the complex. Shorter reads are most likely degraded RNA fragments arbitrarily sticking to proteins and were thus considered noise in the data. Therefore, only reads with a minimum read length of 14 nt are considered.

Once the observed reads are trimmed and small reads discarded, the remaining reads have to be aligned against a reference genome so that one knows to which genes the observed reads belong. The PARA-suite aligner developed by Kloetgen et al. [31] additionally allows for exon-exon junctions by aligning not only against a genomic reference but also against a transcriptomic reference. The PARA-suite aligner is based on BWA [37], and also considers the fact that the substitution probability for T-to-C substitutions is higher than for the other substitutions in PAR-CLIP data. Reads were therefore aligned against the reference genome and transcriptome GRCh38.p7 using PARA-suite.

After alignment, it is, hence, known to which genomic positions each sequenced read belongs. In Figure 2.2 all reads that are observed in the Memczak data set on chromosome 1 from position 23506436 to 23506471 are plotted. One horizontal line represents one observed read in the figure. On positions 23506446 and 23506460 the reference genome shows a T, whereas 39% and 33%, respectively, display a T-to-C mutation. For the positions with a substitution, one can then count the total number of reads, that cover this position and the number of the observed specific substitutions (e.g., T-to-C substitutions).

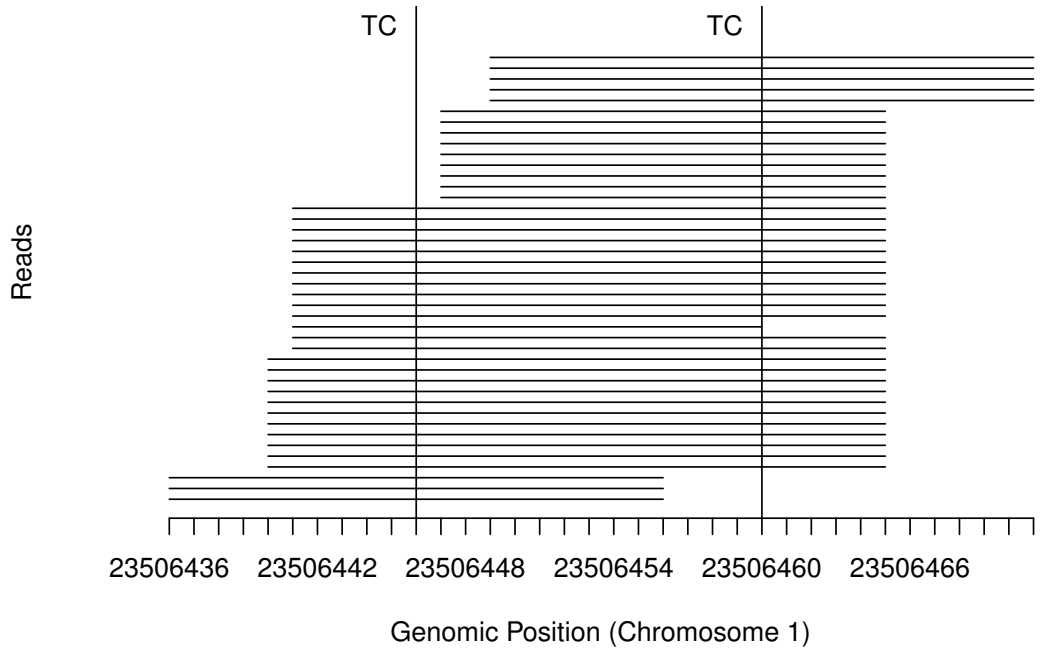

Figure 2.2: Observed reads for two T-to-C substitution positions on Chromosome 1 in the Memczak data set.

Using the wavClusteR package in R [47], all positions with a substitution can be identified. For every substitution position the total number of observed reads for this position, the number of substituted reads, the type of substitution, e.g., T-to-C, as well as the belonging chromosome, the position on the chromosome and the strand of the chromosome are calculated and stored. For the example in Figure 2.2, the corresponding data is represented in Table 2.1. As the two T-to-C substitution positions are the only positions with substitutions in this example, the data set only consists of two rows. Since the DNA has two strands, RNA can be copied from both strands. The two strands are here represented with the signs + and -. The variable substitution stands for the substitution type, e.g., T-to-C, but could be another substitution type, too, e.g., A-to-G. The variable count gives the number of reads that have the substitution at this position. Coverage is the total number of reads that cover this position.

However, positions with a very small number of observed reads, i.e. coverage, can also be discarded from the data set because of two reasons. First, these positions are not

likely to lie on binding sites otherwise they would be observed more often. Second, the data sets are already very large so that there is probably not an important loss of information by excluding these positions. This ensures that only positions one is really interested in are included and the computational analysis of the data will be faster by excluding low coverage positions. When sequencing depth (i.e. the total number of sequenced reads) is low it is reasonable to set a lower threshold than for data sets, where sequencing depth is high. Up to now, there is no universal optimal value for the minimum coverage for the analysis of PAR-CLIP data sets. Hafner et al. [24] and Corcoran et al. [12] require a minimum of five reads per read group. Corcoran et al. [12] argue that this minimum number could be higher for a higher read depth. For the analyses in wavClusteR [11, 49] a minimum coverage of 20 is taken.

In Table 2.2 the total number of mapped reads to chromosomes 1 to 22, X and Y by the preprocessing described in this section are displayed. For the data sets of Kishore et al. [30] and Memczak et al. [41], only positions with at least twenty reads per position are considered. For the data data sets of Gottwein et al. [22] only positions with at least five reads were considered, as sequencing depth was much smaller than in the other data sets. Finally, additional variables, such as the mRNA regions, i.e. 3'UTR, 5'UTR and CDS, are annotated and added to the data sets.

2.2.3 Descriptive analysis

Figure 2.3 shows the number of substitution positions per substitution type (e.g., T-to-C) for the first data set from Kishore et al. [30]. It is obvious that T-to-C substitutions are much more frequent than other types of substitutions. This is not surprising, as T-to-C substitutions are expected to be enriched in the PAR-CLIP experiment. However, the other types of substitutions are also observed even if they are not expected to be caused

Table 2.1: Data set for the observed reads in Figure 2.2.

Chromosome	Position	Strand	Substitution	Count	Coverage
1	23506446	-	TC	11	28
1	23506460	-	TC	13	39

Table 2.2: Total number of mapped reads for the five considered data sets.

Kishore A	Kishore B	Memczak	Gottwein A	Gottwein B
4,467,026	836,295	3,196,429	172,261	172,778

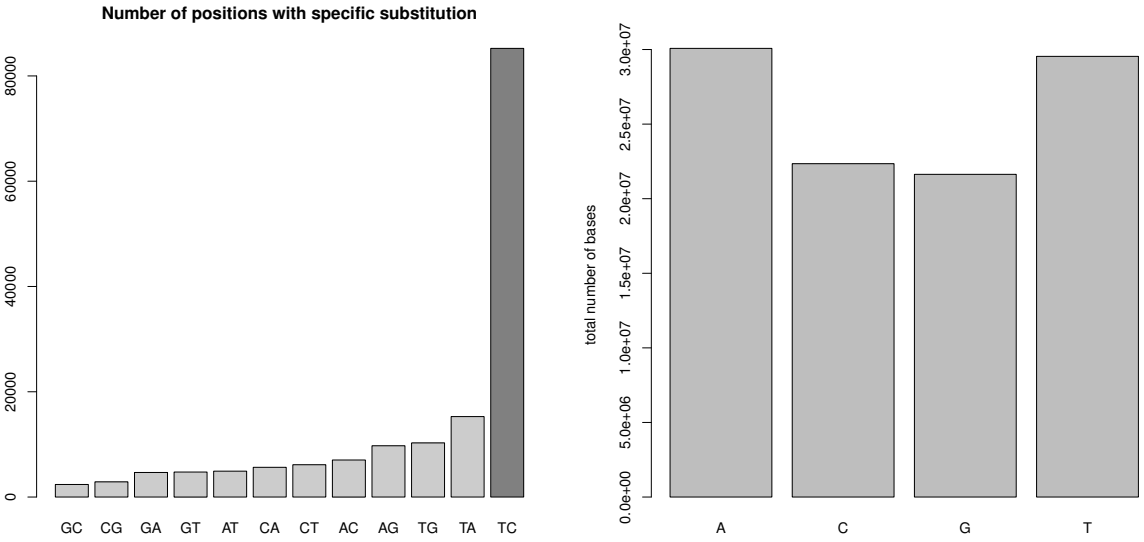

Figure 2.3: Left panel: Number of substitution positions per substitution type for the Kishore A data set. Right panel: Total number of nucleotides in all reads.

by the PAR-CLIP experiment. These substitutions could be due to several reasons, for example errors/mismatches or SNPs. One can therefore conclude, that a (small) part of the T-to-C substitutions is not induced by the PAR-CLIP experiment. The aim is therefore to distinguish between T-to-C substitution positions with substitutions induced by the PAR-CLIP method and T-to-C substitution positions with substitutions that are not method-induced.

It is here assumed, that observing positions with substitutions that are not method-induced, is equally likely for the different bases. However, in the data there seem to be small differences. It stands out that substitutions with T or A as reference are more frequent than with G or C (see Figure 2.3 right panel). This can be explained, when regarding the total number of observed bases in the data set, i.e. the number of all bases with or without substitution (see Figure 2.3 right panel), where it can be seen that A and T are more frequent than C and G.

When regarding the number of substitution positions per substitution type for the

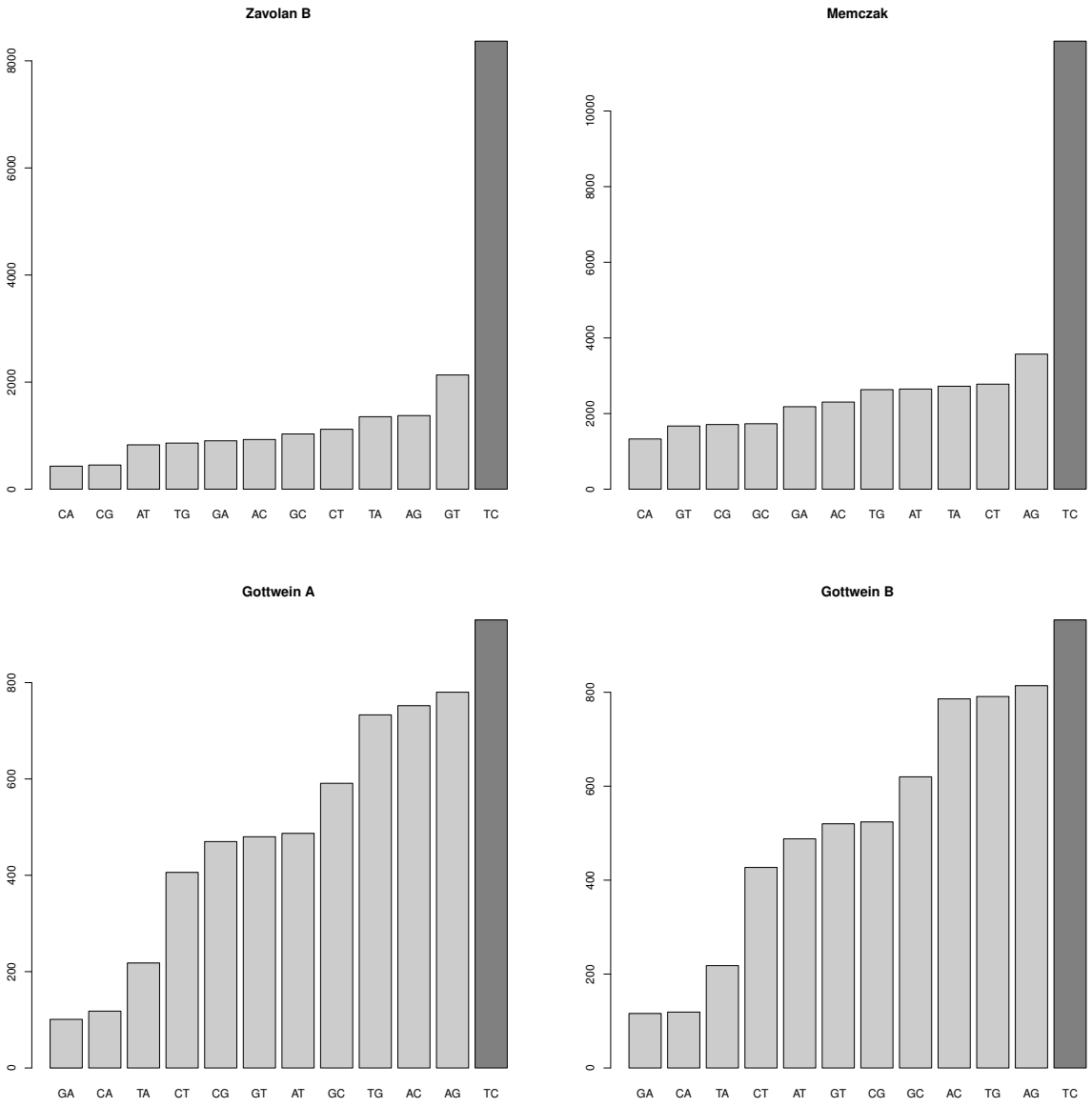

Figure 2.4: Number of substitution positions per substitution type.

other data sets in Figure 2.4, it stands out, that T-to-C substitutions are the most common substitution positions in all data sets. However, in the two data sets from Gottwein et al. [22], other substitution types, such as A-to-G, A-to-C or T-to-G are nearly as frequent as T-to-C substitution positions. As the read depth for those two data sets is not very high, a minimum coverage of five is applied instead of the twenty that are used for the other data sets (see Section 2.2.2). Even though this different minimum coverage is applied, the data sets from Gottwein et al. [22] only consist of around 6,000 substitution positions each, whereas the other data sets consist of around 20,000 positions (Kishore B), 40,000 positions (Memczak) and 160,000 positions (Kishore A) with a higher threshold. If the threshold for the Gottwein A data set were also a minimum cov-

Table 2.3: Number of T-to-C substitution positions divided by the total number of substitution positions.

Kishore A	Kishore B	Memczak	Gottwein A	Gottwein B
0.54	0.42	0.32	0.15	0.15

erage of twenty, only around one quarter of the original positions would remain. This means also, that the data set probably contains more noise due to the small minimum coverage threshold. This could also explain the higher fraction of non-T-to-C substitution positions. As an example, in the Gottwein A data set with five as a minimum coverage, the number of A-to-G substitution positions is equal to 84% of the number of T-to-C substitution positions. Whereas this percentage is only equal to 70% when applying a minimum coverage of twenty.

Table 2.3 shows the fraction of T-to-C substitution positions over all substitution positions. This proportion is relatively high for the data sets from Kishore et al. [30] with a value of even more than 50% in data set A, whereas the proportion is only equal to 15% in the data sets from Gottwein et al. [22]. The fraction is also an indicator for the noise level in the data. The higher the percentage of T-to-C substitution positions the more of the observed reads are probably reads from binding sites. Furthermore, the more noise the data have, the more it gets difficult to distinguish between PAR-CLIP induced T-to-C substitutions and non-PAR-CLIP induced ones. The highest noise level is therefore probably present for the data sets from Gottwein et al. [22].

One is not only interested in the number of positions that contain a specific substitution but also in the substitution rates for these positions, i.e. how many of the observed reads for one position represent a substitution (e.g., a T-to-C substitution). In Figure 2.5 two histograms of substitution rates are plotted for each data set, one for T-to-C substitution rates and one for all other substitution rates, except T-to-C.

When looking at the histograms of non-T-to-C substitution rates of the data sets from Kishore et al. [30] and Memczak et al. [41], it stands out, that mostly substitution rates very close to zero or to one are observed. As these are not T-to-C substitutions, these substitutions must be due to other reasons than the induction by PAR-CLIP [28, 49].

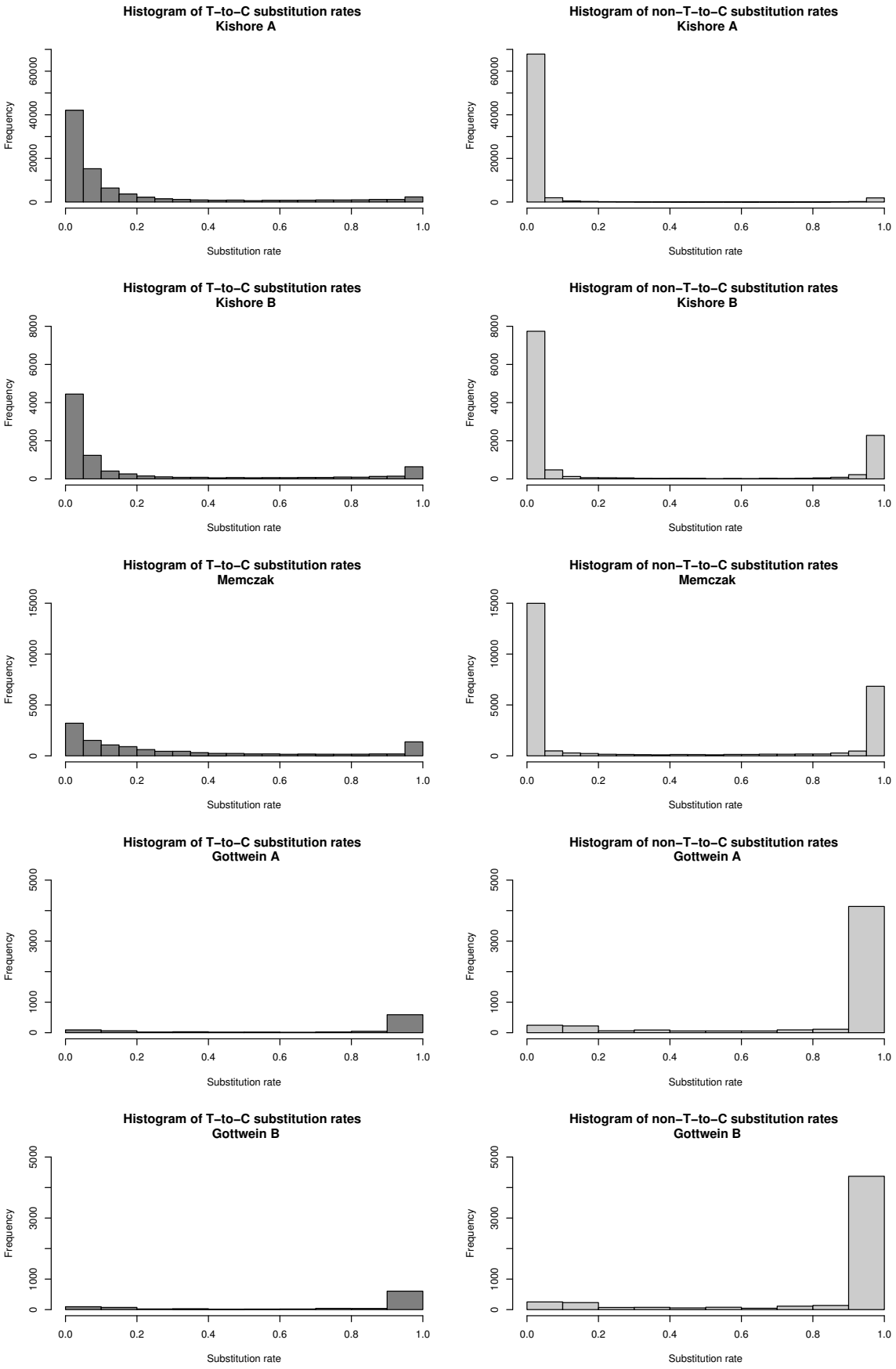

Figure 2.5: Left panel: Histograms of T-to-C substitution rates, Right panel: Histograms of non-T-to-C substitution rates.

Reasons for very low substitution rates can be the appearance of sequencing errors or mismatches. High substitution rates can be due to homozygous SNPs (i.e. a SNP at both chromosomes of the chromosome pair) or RNA highly similar, but not identical to the mapped reference genome [20, 27].

Comparing these substitution rates to the substitution rates of the T-to-C positions, it is obvious, that the T-to-C positions also have substitution rates very close to zero and one, but that the substitution rates are more divided over the whole range than the non-T-to-C substitutions.

The histograms for the data sets from Gottwein et al. [22] show especially a high number of positions with substitution rates very close to one. Substitution rates close to zero only seem to occur slightly more often than other substitution rates. Again, this artifact can be explained by the smaller threshold of five for the minimal coverage for one position. On the one hand, a position, that is a SNP, will also be observed for positions with a small coverage with less than twenty reads, since (almost) 100% of the positions are expected to be substituted. On the other hand, it is very unlikely to observe a mismatch at a position with only few reads, as the probability for an error or mismatch for one read should be very small. By including positions, here, with a coverage between 5 and 19, it is therefore expected, that mainly positions that have either method-induced substitutions or very high substitutions due to SNPs are included. This can also be seen in the data. 80% of the positions with a substituion rate of more than 0.9 in the Gottwein A data set are from positions with a coverage smaller than twenty.

T-to-C substitution positions that have either very low or very high substitution rates are therefore probably the result of mechanisms other than the PAR-CLIP method. Method-induced T-to-C substitution rates are expected to lie somewhere in between this range. This means, that not all reads with T positions that are within a binding site show a T-to-C substitution. This is probably due to incomplete incorporation of thio-uridine into the mRNA, so that some reads do not contain the specific T-to-C substitution [27]. The aim of this work is therefore to distinguish between method-induced T-to-C substitution positions and non-method-induced ones.

Chapter 3

Statistical background

The substitution positions can be divided into the three groups: mismatch positions, SNP positions and crosslinked positions, i.e. positions with PAR-CLIP induced T-to-C substitutions, as described in the previous section. In order to predict to which of the groups a substitution position belongs, BayMAP was developed as a three component mixture model set into a Bayesian context. In this section, basics of Bayesian data analysis are therefore presented along with Bayesian mixture models. Moreover, sampling methods with which the parameters' distributions can be estimated are shown.

A density will be denoted by $f(\cdot)$ and a conditional density by $f(\cdot \mid \cdot)$. If the considered variable is discrete, then $f(\cdot)$ is also used as notation for the probability of observing the value of the variable. For a concrete event, $P(\cdot)$ may also be used as notation for the probability of this event, e.g., $P(T_i = m)$. Background in Bayesian data analysis presented in this Section is mainly based on Gelman et al. [18].

3.1 Bayesian data analysis

Let $\boldsymbol{y} := (y_1 \ \ldots \ y_N)^\top \in \mathbb{R}^N$ be a vector of observed data and $\boldsymbol{\theta} := (\theta_1 \ \ldots \ \theta_M)^\top \in \mathbb{R}^M$ be an unknown parameter vector on which the distribution, from which $\boldsymbol{y}$ is drawn, depends. In frequentist statistics, $\boldsymbol{\theta}$ is supposed to be a vector with fixed but unknown values. By contrast, in Bayesian statistics, the uncertainty about $\boldsymbol{\theta}$ is represented by a distribution so that $\boldsymbol{\theta}$ is not supposed to be fixed but rather a random variable. Since it is wished to learn more about the unknown parameter vector $\boldsymbol{\theta}$ given the data, the density $f(\boldsymbol{y} \mid \boldsymbol{\theta})$

is of interest in Bayesian statistics. This density can be written with Bayes' theorem as

$$f(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{f(\boldsymbol{y} \mid \boldsymbol{\theta})\ f(\boldsymbol{\theta})}{f(\boldsymbol{y})} \propto f(\boldsymbol{y} \mid \boldsymbol{\theta})\ f(\boldsymbol{\theta})\ . \tag{3.1}$$

Since $\boldsymbol{y}$ is observed, therefore fixed, and $f(\boldsymbol{y}) > 0$ does not depend on θ, the constant $f(\boldsymbol{y})$ can be omitted and $f(\boldsymbol{\theta} \mid \boldsymbol{y})$ is then called proportional (represented by the sign $\propto$) to $f(\boldsymbol{y} \mid \boldsymbol{\theta})\ f(\boldsymbol{\theta})$. The density $f(\boldsymbol{\theta} \mid \boldsymbol{y})$ is called posterior density, $f(\boldsymbol{y} \mid \boldsymbol{\theta})$ is the likelihood of the data and $f(\boldsymbol{\theta})$ the prior density that is not dependent on the data.

The prior density represents the uncertainty about the unknown parameter $\boldsymbol{\theta}$ before collecting data. When there is no prior knowledge about $\boldsymbol{\theta}$, a vague prior distribution can be chosen, such as the uniform distribution or normal distribution with a large variance. The prior distribution can also depend on data from previous studies or on additional unknown parameters.

Often conjugate priors are chosen. A prior distribution is conjugate for the class of likelihood distributions if the posterior distribution arises from the same class of distributions as the prior. Formally, this means that if the chosen class of prior distributions is large enough, e.g., the class of all distributions, then every prior distribution would be conjugate [18]. The class of conjugate prior distributions is thus often restricted to distributions that have the same functional form as the likelihood, e.g., the exponential family. A prior distribution fulfilling this restriction is called natural conjugate prior. In the following, only natural conjugate priors are considered when writing about conjugacy. The advantage of conjugate priors lies in the interpretability and in easier computation of the posterior distribution.

One advantage of Bayesian statistics is therefore that additional information and additional structure can be implemented by the prior distribution. If the parameter vector $\boldsymbol{\theta}$ depends on another parameter vector $\boldsymbol{\phi} = (\phi_1\ \ldots\ \phi_S)^\top \in \mathbb{R}^S$, the unknown parameters can be modeled by a hierarchical model

$$f(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \boldsymbol{y}) \propto f(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\phi})\ f(\boldsymbol{\theta}, \boldsymbol{\phi}) = f(\boldsymbol{y} \mid \boldsymbol{\theta})\ f(\boldsymbol{\theta} \mid \boldsymbol{\phi})\ f(\boldsymbol{\phi})\ , \tag{3.2}$$

where $f(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \boldsymbol{y})$ is then called the joint posterior density with the joint prior density

$f(\boldsymbol{\theta}, \boldsymbol{\phi})$ and hyper prior density $f(\boldsymbol{\phi})$.

3.2 Bayesian mixture models

When the data $\boldsymbol{y}$ arise from different data-generating processes, this can be modeled by a mixture model. A three component mixture model is here required for BayMAP to distinguish between method-induced substitution positions, SNP and mismatch positions. Hence, only finite mixture models with a known number of components are considered here. Let $M \in \mathbb{N}^+$ be the number of components, θ_m with $m = 1, \ldots, M$ the unknown parameter on which the distribution of the m-th component depends and π_m the probability that an arbitrary observation arises from the m-th distribution. The density for the mixture model can then be written as

$$f(y_i \mid \boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{m=1}^{M} \pi_m f(y_i \mid \theta_m) \tag{3.3}$$

for $i = 1, \ldots, N$ with $\boldsymbol{\theta} = (\theta_1 \ \ldots \ \theta_M)^\top$, $\boldsymbol{\pi} = (\pi_1 \ldots \pi_M)^\top$ and $\sum_{m=1}^{M} \pi_m = 1$, $\pi_m \leq 0$. The likelihood of the data can then, under the independence assumption, be written as

$$f(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{N} \left(\sum_{m=1}^{M} \pi_m f(y_i \mid \theta_m) \right). \tag{3.4}$$

In a mixture model, y_i arises from one of the mixture components. For this purpose, an allocation variable T could be defined, to indicate to which population an observation belongs, with $T_i = m$ if observation i is drawn from the m-th mixture component. The probability π_m is then equal to $P(T_i = m \mid \boldsymbol{\pi})$ and $f(y_i \mid \theta_m) = f(y_i \mid T_i = m, \boldsymbol{\theta})$, so that (3.4) gets

$$f(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{N} \left(\sum_{m=1}^{M} P(T_i = m \mid \boldsymbol{\pi}) f(y_i \mid T_i = m, \boldsymbol{\theta}) \right). \tag{3.5}$$

It is, thus, reasonable to assume that T_i follows a categorical distribution $\text{Cat}(\pi_1, \ldots, \pi_M)$, so that the density $f(\boldsymbol{t} \mid \boldsymbol{\pi})$ is given by

$$f(\boldsymbol{t} \mid \boldsymbol{\pi}) = \prod_{i=1}^{N} \prod_{m=1}^{M} \pi_m^{\mathbb{1}_m(t_i)} = \prod_{m=1}^{M} \pi_m^{\sum_{i=1}^{n} \mathbb{1}_m(t_i)}, \tag{3.6}$$

where $\boldsymbol{t} = (t_1 \ \ldots \ t_N)^\top$ with $t_i \in \{1, \ldots, M\}$, $i = 1, \ldots N$ and $\mathbb{1}_m(t_i)$ is the indicator function

with

$$\mathbb{1}_m(t_i) := \begin{cases} 1, & \text{if } t_i = m \\ 0, & \text{if } t_i \neq m \end{cases}.$$

3.3 Sampling

When fitting the model, one is interested in the posterior distribution. If the model is simple enough, the posterior distribution could be a well known distribution, such as the normal or beta distribution, and no further estimation of the posterior is needed. This is for example the case for binomial data y, i.e. $Y \sim \text{Bin}(n,\theta)$, with a beta distribution as prior since

$$\begin{aligned} f\left(\theta \mid y,n\right) &\propto f\left(y \mid \theta,n\right)\ f\left(\theta\right) \\ &= \text{Bin}\left(y \mid n,\theta\right)\ \text{Beta}\left(\theta \mid a,b\right) \\ &\propto \theta^{y}\ (1-\theta)^{n-y}\ \theta^{a-1}\ (1-\theta)^{b-1} \\ &= \theta^{y+a-1}\ (1-\theta)^{n-y+b-1} \\ &\propto \text{Beta}\left(\theta \mid y+a, n-y+b\right)\ , \end{aligned}$$

where $\text{Bin}\left(y \mid n,\theta\right)$ is the discrete density function of the binomial distribution for the observed number of successes y with n the number of trials and θ the success probability, and $\text{Beta}\left(\theta \mid a,b\right)$ the density of the beta distribution with parameters a and b. This means, that the distribution of θ is known since $\left(\theta \mid y,n\right) \sim \text{Beta}\left(y+a, n-y+b\right)$.

If the posterior distribution is not a known distribution, one could sample different values for $\boldsymbol{\theta}$ from the posterior distribution, so that the shape of the distribution can be estimated by drawing random values. For instance, in simple non-hierarchical models, the shape of the posterior density can be estimated by sampling directly from the posterior distribution, e.g., by approaches as the rejection sampling approach. In hierarchical models, such as the one in (3.2), it can get more complex to sample from the joint posterior distribution, as it is a multiparameter model with a joint distribution for several parameters. Other sampling strategies are therefore required.

The standard procedure to fit such models are Marcov Chain Monte Carlo (MCMC)

Algorithm 1: Metropolis-Hastings algorithm

1. Choose or draw a starting point $\boldsymbol{\theta}^{(0)}$ with $f\left(\boldsymbol{\theta}^{(0)} \mid \boldsymbol{y}\right) > 0$.
2. For $h = 1, 2, \ldots$:
   (a) Draw a candidate $\boldsymbol{\theta}^*$ from a jumping distribution with density $\mathcal{J}\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(h-1)}\right)$.
   (b) Calculate
   $$r = \frac{f\left(\boldsymbol{\theta}^* \mid \boldsymbol{y}\right) \Big/ \mathcal{J}\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(h-1)}\right)}{f\left(\boldsymbol{\theta}^{(h-1)} \mid \boldsymbol{y}\right) \Big/ \mathcal{J}\left(\boldsymbol{\theta}^{(h-1)} \mid \boldsymbol{\theta}^*\right)}. \tag{3.7}$$
   (c) Set
   $$\boldsymbol{\theta}^{(h)} = \begin{cases} \boldsymbol{\theta}^* & \text{with probability } \min(r, 1) \\ \boldsymbol{\theta}^{(h-1)} & \text{otherwise} \end{cases}$$
   by drawing u from $U(0,1)$ and setting $\boldsymbol{\theta}^{(h)}$ to $\boldsymbol{\theta}^*$ if $u < r$ and to $\boldsymbol{\theta}^{(h-1)}$ otherwise.

algorithms. If one is interested in drawing $\boldsymbol{\theta}$ from its distribution with density $f\left(\boldsymbol{\theta} \mid \boldsymbol{y}\right)$, the idea is to start with an initial draw $\boldsymbol{\theta}^{(0)}$ and to draw then new samples for $\boldsymbol{\theta}$, where the h-th draw $\boldsymbol{\theta}^{(h)}$ is dependent on $\boldsymbol{\theta}^{(h-1)}$. The sequence $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \ldots)$ is then a Markov chain since

$$f\left(\boldsymbol{\theta}^{(h)} \mid \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(h-1)}\right) = f\left(\boldsymbol{\theta}^{(h)} \mid \boldsymbol{\theta}^{(h-1)}\right).$$

The density $f\left(\boldsymbol{\theta}^{(h)} \mid \boldsymbol{\theta}^{(h-1)}\right)$ has to be constructed in such a way that it converges to the density of the target distribution $f\left(\boldsymbol{\theta} \mid \boldsymbol{y}\right)$.

An MCMC algorithm widely used for this task is the Metropolis-Hastings algorithm [25] which is a basic algorithm that has many special cases. Let the parameter vector $\boldsymbol{\theta}$ be the parameter of interest. The basic algorithm can then be written as shown in Algorithm 1.

When it is not possible to draw $\boldsymbol{\theta}$ directly from its posterior distribution, the idea is to draw $\boldsymbol{\theta}$ iteratively. First, a candidate $\boldsymbol{\theta}^*$ is drawn from a jumping distribution with density $\mathcal{J}\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(h-1)}\right)$. Then, the target density of the candidate $\boldsymbol{\theta}^*$ given the data is compared to the target density of $\boldsymbol{\theta}^{(h-1)}$ from iteration $(h-1)$ given the data. When the candidate is more likely under the target distribution, it should remain in the Markov chain. Otherwise, the candidate should only remain in the Markov chain with a certain probability, that is the ratio of the two densities. However, when the jumping density

is not symmetric and it is not equally likely to jump from the old $\boldsymbol{\theta}^{(h-1)}$ to the new candidate $\boldsymbol{\theta}^*$ than the other way around, the asymmetry should be taken into account. In (3.7) the target densities are therefore divided by the jumping density. The Markov chain of Algorithm 1 is a Markov chain with a unique stationary distribution, where the stationary distribution is equal to the target density $f\left(\boldsymbol{\theta} \mid \boldsymbol{y}\right)$. The Markov chain converges, therefore, to the posterior distribution (see, e.g., Gelman et al. [18]).

Once the Markov chain has converged to its stationary distribution with enough draws for $\boldsymbol{\theta}$, early iterations should be discarded, as they highly depend on the starting point rather than the target distribution. The elimination of the first simulated values of the Markov chain is called burn-in. The Markov chain can be highly autocorrelated. In order to reduce autocorrelation of the chain and to reduce computer storage, only every o-th iteration could be kept in the chain. This procedure is called thinning. Thinning is, however, not necessary if enough draws of the target distribution are present.

In this thesis the normal distribution, with the last value of the Markov chain as a mean parameter, is used as a jumping distribution. Since the density of the normal distribution is symmetric, notably $\mathscr{J}\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(h-1)}\right) = \mathscr{J}\left(\boldsymbol{\theta}^{(h-1)} \mid \boldsymbol{\theta}^*\right)$, the ratio in (3.7) simplifies to

$$r = \frac{f\left(\boldsymbol{\theta}^* \mid \boldsymbol{y}\right)}{f\left(\boldsymbol{\theta}^{(h-1)} \mid \boldsymbol{y}\right)}\,. \tag{3.8}$$

When a symmetric density for the jumping distribution is used, the Metropolis-Hastings algorithm is called Metropolis algorithm. When using for example the normal distribution as a jumping distribution, a crucial step is to define the variance that should be employed. If the chosen variance is too small, the random walk moves too slowly and therefore needs too many iterations for convergence. If the variance is too large, the drawn candidates could be rejected too often, so that the random walk does not move most of the time (see, e.g., Gelman et al. [18]). The acceptance rate of the candidates should, thus, neither be too small nor too high. A reasonable acceptance rate could be in the range between 0.2 and 0.5.

In Algorithm 1, the parameter vector $\boldsymbol{\theta}$ is viewed in its entirety. It can, however, be useful to split $\boldsymbol{\theta}$ in subvectors and to update $\boldsymbol{\theta}$ componentwise. Let $\boldsymbol{\theta}$ be the parameter

Algorithm 2: Componentwise Metropolis-Hastings algorithm

1. Choose or draw a starting point $\boldsymbol{\theta}^{(0)}$ with $f\left(\boldsymbol{\theta}^{(0)} \,\middle|\, \boldsymbol{y}\right) > 0$.

2. For $h = 1, 2, \ldots$:
   For $j = 1, \ldots, J$:

   (a) Draw a candidate $\boldsymbol{\theta}_j^*$ from a jumping distribution with density

   $$\mathcal{J}\left(\boldsymbol{\theta}_j^* \,\middle|\, \boldsymbol{\theta}_j^{(h-1)}, \boldsymbol{\theta}_{-j}^{(h-1)}\right).$$

   (b) Calculate

   $$r = \frac{f\left(\boldsymbol{\theta}_j^* \,\middle|\, \boldsymbol{\theta}_{-j}^{(h-1)}, \boldsymbol{y}\right) \Big/ \mathcal{J}\left(\boldsymbol{\theta}_j^* \,\middle|\, \boldsymbol{\theta}_j^{(h-1)}, \boldsymbol{\theta}_{-j}^{(h-1)}\right)}{f\left(\boldsymbol{\theta}_j^{(h-1)} \,\middle|\, \boldsymbol{\theta}_{-j}^{(h-1)}, \boldsymbol{y}\right) \Big/ \mathcal{J}\left(\boldsymbol{\theta}_j^{(h-1)} \,\middle|\, \boldsymbol{\theta}_j^*, \boldsymbol{\theta}_{-j}^{(h-1)}\right)}. \tag{3.9}$$

   (c) Set

   $$\boldsymbol{\theta}^{(h)} = \begin{cases} \boldsymbol{\theta}^* & \text{with probability } \min(r, 1) \\ \boldsymbol{\theta}^{(h-1)} & \text{otherwise} \end{cases}$$

   by drawing u from $U(0,1)$ and setting $\boldsymbol{\theta}^{(h)}$ to $\boldsymbol{\theta}^*$ if $u < r$ and to $\boldsymbol{\theta}^{(h-1)}$ otherwise.

vector in which one is interested, that can be divided into $J \in \mathbb{N}^+$, components of subvectors so that $\boldsymbol{\theta} = \left(\boldsymbol{\theta}_1^\top \ldots \boldsymbol{\theta}_J^\top\right)^\top$. Let $\boldsymbol{\theta}_{-j}^{(h-1)}$ represent all components of $\boldsymbol{\theta}$ except for $\boldsymbol{\theta}_j$ with their last drawn values, so that

$$\boldsymbol{\theta}_{-j}^{(h-1)} = \left(\boldsymbol{\theta}_1^{(h)^\top} \ldots \boldsymbol{\theta}_{j-1}^{(h)^\top} \boldsymbol{\theta}_{j+1}^{(h-1)^\top} \ldots \boldsymbol{\theta}_J^{(h-1)^\top}\right)^\top .$$

The componentwise Metropolis-Hastings algorithm is presented in Algorithm 2.

If it is possible to draw $\boldsymbol{\theta}_j$ directly from its full conditional distribution with density $f\left(\boldsymbol{\theta}_j \mid \boldsymbol{\theta}_{-j}^{(h-1)}, \boldsymbol{y}\right)$, the jumping distribution can be set to the full conditional distribution, so that

$$\mathcal{J}\left(\boldsymbol{\theta}_j^* \mid \boldsymbol{\theta}_j^{(h-1)}, \boldsymbol{\theta}_{-j}^{(h-1)}\right) = f\left(\boldsymbol{\theta}_j^* \mid \boldsymbol{\theta}_{-j}^{(h-1)}, \boldsymbol{y}\right) .$$

This implies, that the ratio r is equal to

$$r = \frac{f\left(\boldsymbol{\theta}_j^* \mid \boldsymbol{\theta}_{-j}^{(h-1)}, \boldsymbol{y}\right) \Big/ f\left(\boldsymbol{\theta}_j^* \mid \boldsymbol{\theta}_{-j}^{(h-1)}, \boldsymbol{y}\right)}{f\left(\boldsymbol{\theta}_j^{(h-1)} \mid \boldsymbol{\theta}_{-j}^{(h-1)}, \boldsymbol{y}\right) \Big/ f\left(\boldsymbol{\theta}_j^{(h-1)} \mid \boldsymbol{\theta}_{-j}^{(h-1)}, \boldsymbol{y}\right)} = 1 ,$$

Algorithm 3: Gibbs sampler

1. Choose or draw a starting point $\boldsymbol{\theta}^{(0)}$ with $f\left(\boldsymbol{\theta}^{(0)} \,\middle|\, \boldsymbol{y}\right) > 0$.
2. For $h = 1, 2, \ldots$:
   For $j = 1, \ldots, J$:
   Draw $\boldsymbol{\theta}_j^{(h)}$ from the full conditional distribution with density
   $$f\left(\boldsymbol{\theta}_j \,\middle|\, \boldsymbol{\theta}_{-j}^{(h-1)}, \boldsymbol{y}\right).$$

so that the candidate will be accepted in every iteration. This special case of the componentwise Metropolis-Hastings algorithm is called Gibbs sampler [19] (see Algorithm 3), and is probably the most frequently used MCMC method in Bayesian modeling. It is the simplest of the MCMC algorithms due to the direct sampling of the full conditional and the advantage of less computation time because of the acceptance rate equal to one.

When using conjugate priors, the full conditional distributions arise from the same family so that they are known and the Gibbs sampler can be applied. Whenever possible, it is therefore reasonable to use conjugate priors. However, it is not always possible to use conjugate priors or other reasons could speak against it. The Gibbs sampler and the Metropolis-Hastings algorithm can nevertheless be combined in the componentwise Metropolis-Hastings algorithm by using the Gibbs sampler for those components, where direct sampling of the full conditional distribution is possible.

When possible and reasonable, conjugate priors are used for parameters in BayMAP. Since there are also parameter components for which no conjugate priors are embedded, Algorithm 2, that is combined for Gibbs sampler and Metropolis algorithm, is implemented in this thesis. For those components without conjugate prior, the normal distribution is used as jumping distribution, so that (3.9) simplifies to

$$r = \frac{f\left(\boldsymbol{\theta}_j^{*} \mid \boldsymbol{\theta}_{-j}^{(h-1)}, \boldsymbol{y}\right)}{f\left(\boldsymbol{\theta}_j^{(h-1)} \mid \boldsymbol{\theta}_{-j}^{(h-1)}, \boldsymbol{y}\right)}.$$

Chapter 4

A subset of existing statistical methods for the analysis of PAR-CLIP data

The idea of BayMAP is based on the methods wavClusteR [11, 49] and BMix [20]. Both methods are positon based methods, which means that it is first predicted for a substitution position if this position is a crosslinked position and therefore belongs to a binding site. However, a binding site is not only a position but a genomic region. Hence, in a second step the full binding site region is then estimated around identified crosslinked positions. In this section, the methods for detecting crosslinked positions as well as binding site regions around the identified positions are presented for wavClusteR and for BMix. Moreover, PARalyzer [12] will be described here in more detail, since it is the first statistical tool for detecting binding sites and commonly used. However, PARalyzer is not a position based method such as wavClusteR and BMix.

Notations can differ in this section from the other notations in this thesis, in particular in Section 3 and Section 5. This is to prevent the notations in this section from differing substantially from the original published methods.

4.1 PARalyzer

In PARalyzer [12], a kernel density estimation approach is used to compare the estimated density for T-to-C substitution rates to the estimated density for non-substitutions, i.e. non substituted T positions that are here called T-to-T. First, all reads that

are overlapping by at least a single nucleotide are grouped together. For further analyses, only read groups with at least five reads and two T-to-C substitution positions are considered. The idea is to then compare T-to-C density with non-substituted T-to-T density via gaussian kernel-density estimate. For any position $j = 1,\dots,L$ of the read group, where $L \in \mathbb{N}^+$ is the length of the considered read group, the density $f_{T\to C}(j)$ for T-to-C substitutions at position j is estimated by

$$f_{T\to C}(j) = \sum_{i=1}^{L} \frac{k_{T\to C}^{(i)}}{n_{T\to C}} \cdot \frac{1}{\lambda\sqrt{2\pi}} e^{-\frac{(i-j)^2}{2\lambda^2}}$$

and the density $f_{T\to T}(j)$ for non-substitutions T-to-T at the same position j by

$$f_{T\to T}(j) = \sum_{i=1}^{L} \frac{k_{T\to T}^{(i)}}{n_{T\to T}} \cdot \frac{1}{\lambda\sqrt{2\pi}} e^{-\frac{(i-j)^2}{2\lambda^2}} ,$$

where $k_{T\to C}^{(i)}$ and $k_{T\to T}^{(i)}$ are the numbers of observed substitutions and non-substitutions at position $i = 1,\dots,L$, and $n_{T\to C}$ and $n_{T\to T}$ are the total numbers of T-to-C substitutions and non-substitutions in the read group. The parameter λ is the bandwidth of the gaussian kernel density estimate, that is fixed for this method to $\lambda = 3$, as Corcoran et al. [12] argue that this value leads to robust results.

The kernel density estimate for position j takes therefore all T-to-C (or T-to-T) substitutions that belong to this read group into account. The closer the substitution position is to position j, the more are the substitutions (or non-substitutions) counted for the density estimate for position j. Corcoran et al. [12] then normalize the estimated densities by

$$g_{T\to C}(j) = \frac{f_{T\to C}(j)}{\sum_{i=1}^{L} f_{T\to C}(i)} ,$$

$$g_{T\to T}(j) = \frac{f_{T\to T}(j)}{\sum_{i=1}^{L} f_{T\to T}(i)} ,$$

so that $\sum_{j=1}^{L} g_{T\to C}(j) = 1$ and $\sum_{j=1}^{L} g_{T\to T}(j) = 1$. If $g_{T\to C}(j) > g_{T\to T}(j)$, position j is considered to be a binding site.

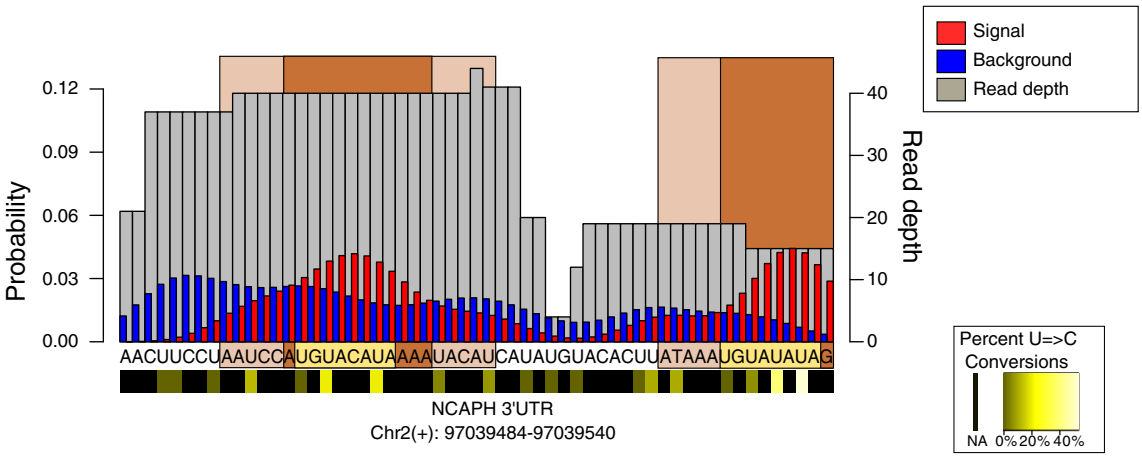

Figure 4.1: Example of PARalyzer binding site identification published in Corcoran et al. [12] for the reference genome GRCh37

.

In Figure 4.1 an example can be seen how a binding site is identified. The grey bars represent the observed read depth, the red and blue bars represent the kernel density estimates for the T-to-C substitutions and the T-to-T non-substitutions. A binding site is identified at those positions, where the red bars are larger than the blue bars (highlighted by the dark orange region). The identified boundaries of the binding site are then extended up to five nt in each direction if the coverage at these positions is still greater or equal to five (highlighted by the light orange region). A binding site is then declared for the whole orange region.

The advantage of this method is that neighborhood information is taken into account. This means that not only a single T-to-C substitution position is regarded but also T-to-C substitution positions in the same region, as it is likely that several T-to-C substitution positions exist on one binding site. However, it is not discussed in Corcoran et al. [12] why it is reasonable to compare the estimated densities $g_{T\to C}(j)$ with $g_{T\to T}(j)$, in which settings these comparisons provide meaningful results and in which not. For instance, estimates for $g_{T\to C}(j)$ for a read group with only T-to-C substitution positions, that have similar substitution rates, are not dependent on the substitution rates level (i.e. if the rates are high or not) due to the normalization. Depending on the context, it could, thus, be even difficult to distinguish method-induced substitutions from mismatches. Furthermore, PARalyzer favors positions with a very high rate of T-to-C substitutions in comparisons to the other T-to-C positions that can also be due to

SNPs.

4.2 wavClusteR

wavClusteR that will be presented in this section, is a position-based method for the detection of PAR-CLIP induced substitution positions. First, crosslinked positions are identified in wavClusteR (see Section 4.2.1) before binding site regions are identified around these detected positions (see Section 4.2.2).

4.2.1 Detection of crosslinked positions

wavClusteR [11, 49] is the first method that takes types of substitutions other than T-to-C into account. First, they assume that the number of substitutions K for one substitution position follows a binomial distribution with $n \in \mathbb{N}^+$ the number of observed reads for this position and $\mu \in [0,1]$ the substitution probability

$$\left(K \mid n, \mu\right) \sim \text{Bin}\left(n, \mu\right).$$

The idea is then, that T-to-C substitutions can either be due to the PAR-CLIP method or due to other reasons such as SNPs, mismatches, and sequencing errors. Substitution types other than T-to-C are probably due to the second group of reasons, the non-method-induced substitutions.

The distribution of these substitutions can therefore be helpful to distinguish between method-induced and non-method-induced T-to-C substitutions. It is assumed that the parameter μ follows a mixture distribution of two components, one for the method-induced substitutions and one for the non-method-induced ones in a Bayesian framework

$$\mu \sim p\left(\mu\right) = \lambda \; p_1\left(\mu\right) + (1-\lambda) \; p_2\left(\mu\right) \; , \tag{4.1}$$

where $p_1\left(\mu\right)$ is the densitiy for μ for the non-method-induced substitutions and $p_2\left(\mu\right)$ for the method-induced ones, and $\lambda \in [0,1]$ is the weight for the non-method-induced substitutions distribution. In the case of a substitution other than T-to-C, the weight λ would be equal to one, since the substitutions are not supposed to be method-induced,

but only non-method-induced. For T-to-C substitution positions, λ represents the probability of a non-crosslinked position.

First, for every substitution position $i, i = 1,\ldots,N$, with $N \in \mathbb{N}^+$ the total number of substitution positions, a posterior is calculated with a uniform prior, i.e. Beta$(1,1)$, so that the posterior for μ_i is given by

$$\left(\mu_i \mid n_i, k_i\right) \sim \text{Beta}\left(k_i + 1, n_i - k_i + 1\right) ,$$

where $k_i \in \mathbb{N}^+$ is the observed number of substitutions at position i. Then, $p_1\left(\mu\right)$ is estimated by taking the average of all posterior densities $f\left(\mu_i \mid k_i, n_i\right)$ except for T-to-C positions with

$$\hat{p}_1\left(\mu\right) = \frac{1}{N_{nTC}} \sum_{i=1}^{N_{nTC}} f\left(\mu_i \mid k_i, n_i\right),$$

where $N_{nTC} \in \mathbb{N}^+$ is the number of all substitution positions except for T-to-C. Afterwards, $p_{TC}\left(\mu\right)$ is estimated, where $p_{TC}\left(\mu\right)$ is the density of μ for all T-to-C substitutions. Note, that this density is not equal to $p_2\left(\mu\right)$ but to $p\left(\mu\right)$ in (4.1) for T-to-C substitutions, since $p_2\left(\mu\right)$ only represents T-to-C substitution positions that are method-induced and excludes the non-method-induced ones. $p_{TC}\left(\mu\right)$ is estimated in the same manner as $p_1\left(\mu\right)$. The weight parameter λ is estimated by taking the type of substitution, that has the most of substitution positions after T-to-C, and dividing this number by the number of T-to-C positions. By rearranging Equation (4.1), the density $p_2\left(\mu\right)$ is then estimated by

$$\hat{p}_2 = \frac{\hat{p}_{TC}\left(\mu\right) - \hat{\lambda}\ \hat{p}_1\left(\mu\right)}{1 - \hat{\lambda}}.$$

For evaluation, either the posterior class probability is taken with

$$\frac{(1-\lambda)\ p_2\left(\mu\right)}{\lambda\ p_1\left(\mu\right) + (1-\lambda)\ p_2\left(\mu\right)}$$

or the log-odds ratio with

$$\log\left(\frac{(1-\lambda)\ p_2\left(\mu\right)}{\lambda\ p_1\left(\mu\right)}\right).$$

After analyzing these functions, the positions with a posterior class probability larger than a specific value are not declared as binding site positions, but rather all T-to-C substitution positions with a substitution rate within a specific interval are considered

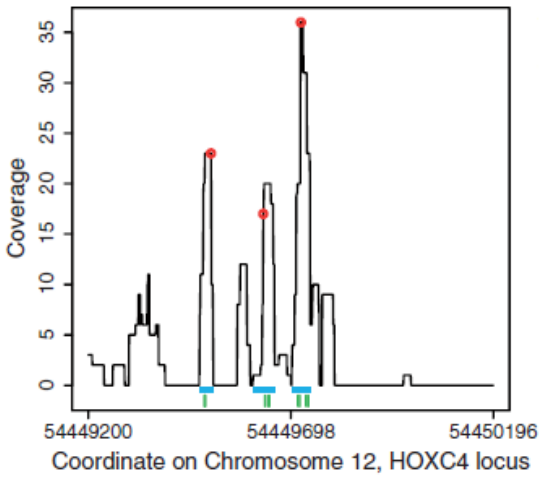

Figure 4.2: Example of coverage function with highly confident T-to-C substitutions (green bars), wavelet peaks (red circles) and the belonging binding sites (blue lines) published in Sievers et al. [49].

as crosslinked positions. In Sievers et al. [49] the interval between 0.2 and 0.7 is chosen, as T-to-C positions with substitution rates within this interval are likely crosslinked. This choice, however, is not justified any further by Sievers et al. [49], so that the selected interval boundaries seem to be arbitrary.

4.2.2 Detection of binding site boundaries

In a second step, after the identification of T-to-C substitution positions that are probably crosslinked and therefore positions on binding sites, binding sites' boundaries for these positions have to be estimated. In wavClusteR, two distinct methods for this estimation are proposed, wherein boundaries are either estimated via wavelet-based peak calling [49] or mini-rank norm [11].

First, Sievers et al. [49] developed a method for the detection of the boundaries of binding sites based on continuous wavelet transforms (CWT). CWT is a method that can be used to detect peaks in functions. Here, it is supposed that on binding sites local peaks should be observed, as binding site reads should be amplified during the PAR-CLIP method. Once peaks are detected by CWT, boundaries are identified by regarding differences in the coverage function. Sievers et al. [49] suppose that binding sites have only one peak. They define, therefore, the left boundary as the closest position to the peak, that still has a positive coverage difference, whereas the previous position has to

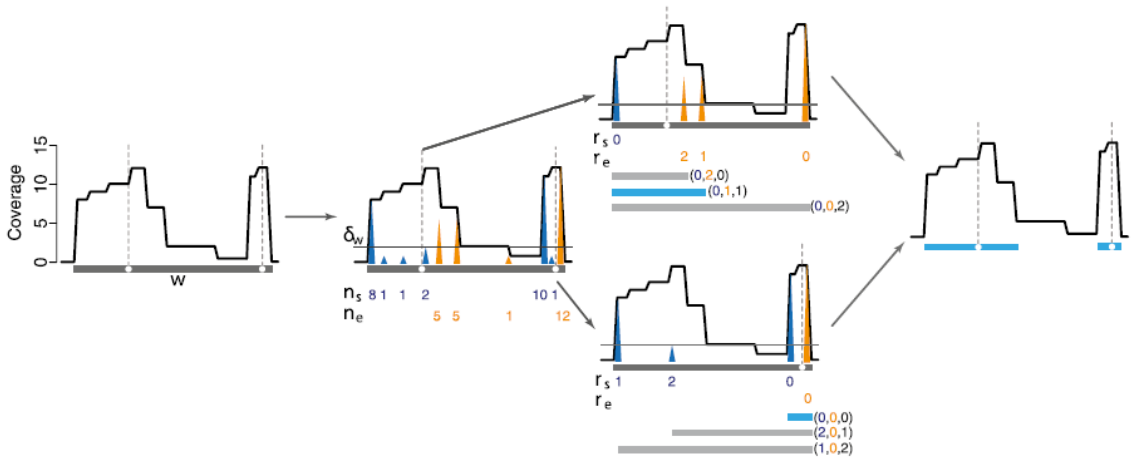

Figure 4.3: Example of the MRN algorithm published in Comoglio et al. [11].

have a negative one. For the right boundary, the same procedure is applied, so that it is the closest position to the peak with a negative coverage difference, whereas the following position has a positive one (see example in Figure 4.2).

However, Comoglio et al. [11] argue that their previously developed method based on CWT, risks having a high number of T-to-C substitution positions identified as method-induced positions (highly confident T-to-C substitution positions) that cannot be assigned to a binding site, as peaks cannot be detected. This could for example happen, when the coverage geometry of the considered genomic region is complex. They propose, therefore, to use another method for the identification of binding sites around highly confident T-to-C substitution positions, the mini-rank norm (MRN).

Around each highly confident T-to-C substitution position, a window w is spanned that contains all non-zero coverage positions (see first coverage function in Figure 4.3). Following this, all possible starting and ending positions are regarded. A possible starting position is a position with a positive difference in the coverage function, i.e. a position, where the coverage is higher than for the position left to it. A possible ending position is a position with a negative difference in the coverage function. For these starting and ending positions the absolute differences in the coverage function (blue and orange triangles in Figure 4.3) are in the two vectors $\boldsymbol{n_s}$ and $\boldsymbol{n_e}$. For this window w a local coverage threshold δ_w is calculated (see second coverage function in Figure 4.3). All values in the two vectors $\boldsymbol{n_s}$ and $\boldsymbol{n_e}$, that are smaller than δ_w are removed.

For each highly confident T-to-C substitution position all possible combinations of starting and ending positions for this T-to-C substitution position are regarded separately. In the example in Figure 4.3 there are two highly confident T-to-C substitution positions, so that first all possible combinations for the first (upper coverage function in Figure 4.3) and then for the second highly confident position (lower coverage function) are regarded.

For one highly confident position, all remaining coverage differences for the starting positions are ranked, then all remaining coverage differences for the ending positions are ranked and then the width of the putative cluster is ranked, so that there are three rank values for each putative cluster candidate. In the upper coverage function of Figure 4.3, one can see that there is only one possible starting position for the first highly confident T-to-C substitution position, but three possible ending positions. For the three combinations of starting and ending positions, the rank of the coverage difference for the starting position is equal to zero, as there is only one starting position. The rank of the coverage difference for the ending positions is equal to zero for the longest cluster here, as it has the highest difference in the coverage function. The two remaining ending positions have the same difference in coverage function, but to the one closer to the highly confident position a greater rank is associated. Then the width of the clusters is ranked, where the shortest width is associated to the smallest rank. The cluster with the rank vector closest to the vector $(0,0,0)$ in terms of the euclidean norm is then chosen as the binding site region (blue lines in Figure 4.3).

4.3 BMix

BMix that will be presented in this section, is a position-based method for the detection of PAR-CLIP induced substitution positions. First, crosslinked positions are identified in BMix (see Section 4.3.1) before binding site regions are identified around these detected positions (see Section 4.3.2).

4.3.1 Detection of crosslinked positions

BMix [20] is a mixture model with three binomial components for substitutions due to mismatches, SNPs and the PAR-CLIP method. The probability that the number of substitutions K_i is equal to k_i at position i for $i = 1, \ldots, N$, can be written as a mixture model with discrete density function $f\left(k_i \mid n_i\right)$

$$\begin{aligned} f\left(k_i \mid n_i\right) = \left(1-p\right)\left(1-q\right) \operatorname{Bin}\left(k_i \mid n_i, \mu_{\text{mm}}\right) \; + \; & q \operatorname{Bin}\left(k_i \mid n_i, \mu_{\text{SNP}}\right) \\ + \; p\left(1-q\right) \operatorname{Bin}\left(k_i \mid n_i, \mu_{\text{exp}}\right), & \end{aligned} \tag{4.2}$$

where $k_i \in \mathbb{N}^+$ is the observed number of T-to-C substitutions at position i, $n_i \in \mathbb{N}^+$ the number of reads at position i, $q \in [0,1]$ the probability that the position is a SNP, $p \in [0,1]$ the probability that the position is crosslinked given that it is not a SNP. The discrete density function of the binomial distribution for k_i substitutions with parameters n_i and $\mu_j \in [0,1]$, $j \in \{\text{mm, SNP, exp}\}$ is denoted by $\operatorname{Bin}\left(k_i \mid n_i, \mu_j\right)$, where μ_{mm}, μ_{SNP} and μ_{exp} are the substitution probabilities for mismatch positions, SNP positions and method induced substitution positions.

Golumbeanu et al. [20] assume that the probability parameters of the binomial distributions μ_{mm}, μ_{SNP} and μ_{exp} are not independent of each other. For a T-to-C SNP position, a T-to-C substitution is expected. However, sequencing errors could also occur for SNP positions, where an error could be an observed T-to-A substitution, T-to-G substitution or a T, each with probability μ_{mm}. The probability μ_{SNP} is therefore assumed to be

$$\mu_{\text{SNP}} = 1 - 3\mu_{\text{mm}}.$$

This error, however, could also occur on crosslinked positions. A part of the reads of the crosslinked positions are expected to contain the specific T-to-C substitution. On these reads, again, the three errors T-to-A, T-to-G or T could happen. On the not-substituted reads of the crosslinked position, only the error T-to-C will be noticed, as only the number of T-to-C substitutions is of interest. With γ being the probability that a read of a crosslinked position contains the T-to-C substitution, the probability μ_{exp} is therefore

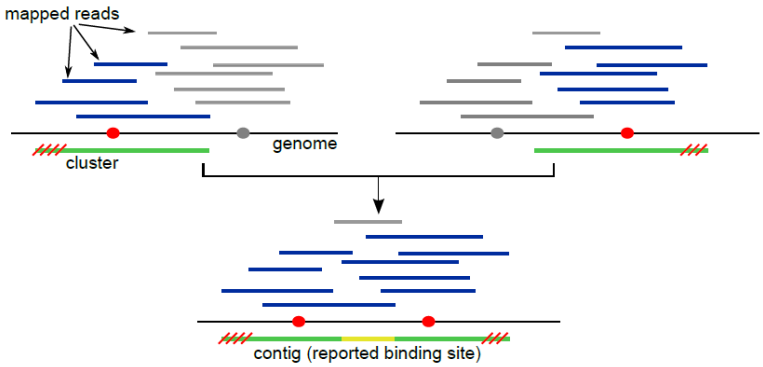

Figure 4.4: Example of a binding site construction from reads published in Golumbeanu et al. [20]. The blue lines represent reads that cover the considered T-to-C substitution position (red point), whereas grey lines are reads that do not cover the red T-to-C substitution position and grey points represent other (not yet considered) T-to-C substitution positions. The green lines are clusters that are built based on the observed reads.

assumed to be

$$\mu_{\text{exp}} = (1-\gamma)\,\mu_{\text{mm}} + (1-3\mu_{\text{mm}})\,\gamma.$$

Furthermore, it is supposed in BMix that μ_{exp} is bounded between μ_{mm} and $1-3\mu_{\text{mm}}$, which leads to the conclustion $\mu_{\text{mm}} \leq 0.25$.

Golumbeanu et al. [20] do not only want to use T-to-C substitution positions for the estimation of the unknown parameters of the mixture model, but also take into account A-to-C and G-to-C substitution positions. As it is supposed that only T-to-C substitutions can be experimentally induced by PAR-CLIP, the parameter p in (4.2) is set to zero for A-to-C and G-to-C substitution positions. The unknown parameters are estimated by maximizing the likelihood function. Classification is done by calculating the posterior probabilities of the group to which the substitution belongs (mismatch, SNP or crosslink).

4.3.2 Detection of binding site boundaries

Positions that are identified as PAR-CLIP induced with a posterior probability of 95% are then used for identifying binding sites around these positions. All aligned sequencing reads that include this identified PAR-CLIP induced position are merged whereas boundaries with a coverage of only one are left out. If such a cluster overlaps with another cluster, both are grouped together into one binding site (see Figure 4.4).

Chapter 5

Bayesian model for the analysis of PAR-CLIP data

As described in the previous sections, the aim of this project is to develop a statistical method that allows us to distinguish between method and non-method induced T-to-C substitution positions, so that miRNA binding sites on the mRNA can be identified. wavClusteR [11, 49] and BMix [20] are two models that enable the classification of T-to-C substitution positions to one of these groups. However, they do not allow the incorporation of supplementary information, such as the 3'UTR, the CDS and the 5'UTR. In order to allow for these additional variables, which are relevant for the biology of binding sites, a fully Bayesian hierarchical model is proposed here, and will hereinafter be referred to as BayMAP (Bayesian hierarchical Model for the Analysis of PAR-CLIP data). BayMAP is a three component mixture model, that distinguishes between method-induced substitution positions and for the non-method induced positions between SNP and mismatch positions.

In Section 5.1, the model, here called BayMAP 1.0, is presented as it has been published in the context of this work in Huessler et al. [27]. After the detection of method-induced substitution positions, one is interested in not only knowing the crosslinked positions but also the binding site regions belonging to these positions. A method for identifying the sequence of positions belonging to a binding site by combining reads to a cluster, is presented in Section 5.2. This method is not published in Huessler et al. [27] and thus firstly described in this thesis.

In Section 5.1, it is assumed that substitution positions are independent. However, T-to-C substitution positions very close to each other are probably either both crosslinked if on the same binding site or both not crosslinked if not on a binding site. This regional dependence can be included to the model in the adjusted version of BayMAP presented in Section 5.3. This adjusted version is here called BayMAP 2.0 and is also not published in Huessler et al. [27] and hence firstly described in this thesis.

For the application of BayMAP 2.0 it is necessary to know in advance, i.e. prior to application, if two substitution positions are close neighbors or not. Two positions are here defined as being close neighbors, if they are lying on the same potential binding site. As this is not known in advance, it has to be estimated. The method that will be presented in Section 5.2 can not only be employed to determine binding site regions around highly confident T-to-C substitution positions but also to determine potential binding site regions for every T-to-C substitution position. It can thus also be applied prior to the application of the model, so that this information can be used as input for BayMAP 2.0. BayMAP 2.0 depends therefore on the identification of potential binding site regions as input.

In order to not only use data from one PAR-CLIP experiment but several and in order to increase precision, a method for combining results of several PAR-CLIP data sets is described in Section 5.4. This method is also firstly presented in this thesis. Up to now and to the best of the author's knowledge, it is the first method specialized for PAR-CLIP, that allows the combination of the data from several PAR-CLIP experiments.

5.1 BayMAP 1.0: Detection of PAR-CLIP induced T-to-C substitution positions

Section 5.1 is divided into the presentation of the model itself (Section 5.1.1), the derivation of the full conditional distributions, that are necessary for sampling (Section 5.1.2), the sampling scheme (Section 5.1.3) and the determination of method-induced substitution positions by estimating the probability of the position being crosslinked given the data (Section 5.1.4). The content of the Sections 5.1.1 and 5.1.4 has already been

published in Huessler et al. [27], whereas Sections 5.1.2 and 5.1.3 are newly presented in this thesis.

5.1.1 The model

One data set consists of $N \in \mathbb{N}^+$ substitution positions, where for every substitution position i, $i = 1,\ldots,N$, the observed number of substitutions $k_i \in \mathbb{N}^+$, the total number of observed reads $n_i \in \mathbb{N}^+$ as well as the substitution type (e.g., T-to-C) are known (see also Section 2.2.2). It is here assumed that the random variable K_i for the number of substitutions, follows a distribution with parameters n_i and $\mu \in [0,1]$, where μ is the probability of a substitution. A natural choice for this distribution is the binomial distribution, as for example in wavClusteR. However, only positions with at least one substitution are considered. Therefore, it is here supposed that K_i follows a binomial distribution truncated for zeros, the zero truncated binomial distribution

$$K_i \sim \mathrm{ZTB}\left(n_i, \mu\right) \text{ with}$$

$$P\left(K_i = k_i \mid K_i > 0, n_i, \mu\right) = \frac{P\left(K_i = k_i, K_i > 0 \mid n_i, \mu\right)}{P\left(K_i > 0 \mid n_i, \mu\right)} = \frac{\mathrm{Bin}\left(k_i \mid n_i, \mu\right)}{1 - \mathrm{Bin}\left(0 \mid n_i, \mu\right)}, \tag{5.1}$$

where $\mathrm{Bin}\left(k_i \mid n_i, \mu\right)$ is the discrete density function of the binomial distribution for k_i substitutions with n_i reads and substitution probability μ.

In a Bayesian framework, it is supposed that the unknown parameters, here μ, also follow a distribution. The posterior density for the unknown parameter μ can be written as

$$f\left(\mu \mid \boldsymbol{k}, \boldsymbol{n}\right) = \frac{f\left(\boldsymbol{k} \mid \mu, \boldsymbol{n}\right)\, f\left(\mu\right)}{\int_0^1 f\left(\boldsymbol{k} \mid \mu, \boldsymbol{n}\right)\, f\left(\mu\right)\, d\mu} \propto f\left(\boldsymbol{k} \mid \mu, \boldsymbol{n}\right)\, f\left(\mu\right), \tag{5.2}$$

where $f\left(\boldsymbol{k} \mid \mu, \boldsymbol{n}\right)$ is the likelihood of the data and $\boldsymbol{k} := (k_1,\ldots,k_N)$, $\boldsymbol{n} := (n_1,\ldots,n_N)$. Nevertheless, the substitution probability μ for a read of a substitution position is different for each of the here considered three position types, i.e. mismatch positions, SNPs and crosslinked positions (i.e. positions with method-induced T-to-C substitutions). Similar to wavClusteR and BMix, μ thus follows a mixture model with three densities for the three position types. For this purpose, one can define the non-observed

allocation variable Z_i with

$$Z_i := \begin{cases} \text{“mm”}, & \text{if position } i \text{ is a mismatch position} \\ \text{“SNP”}, & \text{if position } i \text{ is a SNP position} \\ \text{“exp”}, & \text{if position } i \text{ is a crosslinked position} \end{cases}, \tag{5.3}$$

$$Z_i \sim \text{Cat}\big((1-p_i)\,q,\ (1-p_i)\,(1-q),\ p_i\big),$$

where $\text{Cat}(\cdot)$ is the categorical distribution. The probability that a position is crosslinked with experimentally induced substitutions is p_i. If a position is not crosslinked, it can either be a mismatch position or a SNP position. In case the position i is not crosslinked, the probability that this position is a mismatch position is equal to q. The probability $P(Z_i = \text{“SNP”})$ is therefore equal to $(1-p_i)(1-q)$. When position i is not a T-to-C substitution position but a position with another type of substitution, p_i is then set to zero, as only T-to-C substitutions are induced by the PAR-CLIP method.

A Dirichlet prior is usually used in mixture models for the parameters of a categorical distribution, as it is a conjugate prior. Here, however, additional information should be incorporated in the model, such as the mRNA region. The idea is to model this additional information via the parameter p_i. The probability p_i that a position is a crosslinked position should for example, be different for a position on the 3'UTR than for another position. The parameters p_i and q are thus regarded separately without the usual Dirichlet prior.

Let $\boldsymbol{\mu} := \big(\mu_{\text{mm}},\ \mu_{\text{SNP}},\ \mu_{\text{exp}}\big)^\top$, where μ_{mm} is the probability of a mismatch substitution, μ_{SNP} is the probability of a SNP substitution and μ_{exp} is the probability of an experimentally induced, i.e. crosslinked substitution. The mixture density for position i can then be written as

$$\begin{aligned} f\left(k_i \mid \boldsymbol{\mu}, q, p_i, n_i\right) =& (1-p_i)\ \left(q \cdot f\left(k_i \mid \mu_{\text{mm}}, n_i\right) + (1-q) f\left(k_i \mid \mu_{\text{SNP}}, n_i\right)\right) \\ &+ p_i f\left(k_i \mid \mu_{\text{exp}}, n_i\right), \end{aligned} \tag{5.4}$$

In BMix, it is assumed that the probability parameters μ_{mm}, μ_{SNP} and μ_{exp} are not in-

dependent of each other, since, e.g., a SNP position can also have reads with errors or mismatches. This will be explained in more detail in the following.

The parameter μ_{mm} is the probability to observe a C at a position of a read, where a T is expected. It is, therefore, the probability of observing by error a T-to-C substitution. However, other errors, such as A-to-G for a position, where an A is expected, could occur, too. Here, it is assumed that the error probabilities are the same for all substitution types, i.e. that μ_{mm} is the probability of observing by error any of the substitution types, e.g., A-to-G. This error can occur at every position including SNP positions and crosslinked positions. At a homozygous T-to-C SNP position, one would expect to observe the T-to-C substitution for every read. Nevertheless, it is possible to observe by error either a T-to-A, a T-to-G or a non-substitution, i.e. a T-to-T. In the same way as in BMix, the observed T-to-C substitution probability for a SNP is, hence, in this work assumed to be

$$\mu_{\text{SNP}} = 1 - 3\mu_{\text{mm}}. \tag{5.5}$$

This type of error can also appear at a T-to-C substitution position with experimentally induced substitutions. However, not all reads are crosslinked at such a position and one has, thus, to distinguish between crosslinked reads, where a T-to-C substitution is expected, and non-crosslinked reads, where the T-to-C substitution is not expected. If a read is crosslinked, a T-to-A, a T-to-G or a T could be observed by error instead of the expected T-to-C substitution. If the read is not crosslinked, only the error T-to-C would count for the number of T-to-C substitutions. In BMix, it is, therefore, assumed, that

$$\mu_{\text{exp}} = \left(1-\gamma\right)\mu_{\text{mm}} + \left(1-3\mu_{\text{mm}}\right)\gamma\,, \tag{5.6}$$

where $\gamma \in [0,1]$ is the probability that a read is crosslinked. In BayMAP, however, this error is not modeled for μ_{exp}. This is justified by the fact that μ_{mm} is expected to be very close to zero and when applying (5.5), μ_{SNP} is forced to be very close to one. Thus, when applying Equation (5.6), γ and μ_{exp} should be almost equal, since μ_{mm} should be very close to zero. The model would therefore be more complicated without a noteworthy benefit. Equation (5.5) is thus implemented in BayMAP whereas (5.6) is not.

Another assumption in BMix is that $\mu_{\text{mm}} < \mu_{\text{SNP}}$, so that with (5.5) $\mu_{\text{mm}} < (1 - 3\mu_{\text{mm}})$. Solving this equation leads to $\mu_{\text{mm}} < 0.25$. In order to avoid label switching problems, e.g., the interchanging of the two groups SNPs and mismatches in the analysis, the assumption that $\mu_{\text{mm}} < 0.25$ is also implemented in BayMAP.

It could also be assumed, that $\mu_{\text{mm}} < \mu_{\text{exp}} < \mu_{\text{SNP}}$ so that μ_{exp} is forced to be somewhere in the range between a very small substitution rate close to zero and a very large substitution rate close to one, as it was already expected in Section 2.2.3. On the one hand, this restriction would complicate the model, since μ_{exp} would then be dependent on other model parameters. On the other hand, the restrictions of $\mu_{\text{mm}} < 0.25$ together with $\mu_{\text{SNP}} = 1 - 3\mu_{\text{mm}}$ should already force μ_{mm} to be close to zero and μ_{SNP} to be close to one, so that restriction for μ_{exp} should not be necessary.

Besides the number of substitutions, more information is available that can offer valuable clues about the likeliness of a binding site. For example, it is already known, that a binding site occurs most often in the 3'UTR, but can also appear in the CDS and, less often, in the 5'UTR [5]. In order to use this information, a hierarchical model can be constructed, which also models the probability p_i that the substitution position i is a binding site and has therefore experimentally induced T-to-C substitutions. Note, that p_i is set to zero if the substitution type is not a T-to-C substitution. For $i = 1, \ldots, N$ only those i are considered here for which the substitution type is T-to-C, that is noted here with $i = 1, \ldots, N_{\text{TC}}$, where $N_{\text{TC}} \in \mathbb{N}^+$ is the total number of T-to-C substitution positions.

The probability p_i is equal to $P\left(Z_i = \text{“exp”}\right)$. Instead of Z_i, the binary response variable $\mathbb{1}_{\text{“exp”}}(Z_i)$ can be considered here. Note that $E\left(\mathbb{1}_{\text{“exp”}}(Z_i)\right) = P\left(Z_i = \text{“exp”}\right) = p_i$. When having a binary response variable, often generalized linear models are applied, where a function of the expected value is modeled via a generalized linear model

$$g\left(p_i\right) = \beta_0 + \beta_1 \cdot x_{i1} + \cdots + \beta_L \cdot x_{iL} = \boldsymbol{x_i}^\top \boldsymbol{\beta} \tag{5.7}$$

for $i = 1, \ldots, N_{\text{TC}}$ where $\boldsymbol{x_i} := (1\ x_{i1}\ \ldots\ x_{iL})^\top \in \mathbb{R}^{(L+1)}$ is a vector with values for covariates of length $L+1$ for position i and $\boldsymbol{\beta} = \left(\beta_0\ \beta_1\ \ldots\ \beta_L\right)^\top \in \mathbb{R}^{(L+1)}$ is the corresponding

vector of length $L+1$ of unknown parameters. The function $g(\cdot)$ is called the link function that has to be chosen. The logit and the probit link are two very common link functions. The logit link is defined as

$$\text{logit}(p_i) := \log\left(\frac{p_i}{1-p_i}\right)$$

whereas the probit link is the inverse cumulative distribution function of the standard normal distribution

$$\Phi^{-1}\left(p_i\right) = \boldsymbol{x_i}^\top \boldsymbol{\beta}\,.$$

Both link functions provide very similar results except for the tails. In classical frequentist inference, the logit link is very often chosen, as the logit is the canonical link for binomial data. The canonical link has some properties that simplify computation in comparison to any other link function. This regression model is called logistic regression [14].

In Bayesian statistics, however, the probit regression model has computational advantages over the logistic regression model, as conjugate priors exist for the probit model, so that Gibbs sampling can be applied [18]. The probability p_i that a position has experimentally induced substitutions (without knowledge of the number of substitutions) is therefore here modeled via the probit model with

$$p_i = g^{-1}\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right) = \begin{cases} \Phi\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right), & \text{if } i \text{ is a T-to-C substitution position} \\ 0 & \text{otherwise} \end{cases}. \tag{5.8}$$

The model for BayMAP 1.0 consists of the unknown parameters $\boldsymbol{\beta}$, μ_{mm}, μ_{exp} and q, for which prior distributions have to be specified. The distribution of μ_{SNP} can be calculated directly once the distribution of μ_{mm} is known. As already mentioned, the probit model allows conjugate priors for the parameter vector $\boldsymbol{\beta}$. Here independent vague conjugate priors for each $\beta_\ell, \ell = 0,\ldots,L$ are applied, i.e. a normal distribution with a large variance. Since q is the parameter of a Bernoulli distribution for non-experimentally induced positions (either mismatch or SNP), a conjugate prior for the Bernoulli (and the binomial) distribution is applied, that is the beta distribution. Here,

the beta distribution with parameters both equal to one is used, i.e. a standard uniform prior. No conjugate prior exists for the zero truncated binomial distribution, so that the conjugate prior for the binomial distribution chosen, is a uniform prior. More precisely, the following priors for $\boldsymbol{\beta}$, μ_{mm}, μ_{exp} and q are used

$$\begin{aligned} \beta_\ell &\sim N\left(0, 10^6\right), \quad \ell = 0, \ldots, L, \\ \mu_{\text{mm}} &\sim U\left(0, 0.25\right), \\ \mu_{\text{exp}} &\sim U\left(0, 1\right), \\ q &\sim U\left(0, 1\right). \end{aligned} \tag{5.9}$$

The joint posterior distribution of the full model as well as the full conditionals are derived in the next section.

5.1.2 Full conditional distributions

In Bayesian statistics, one is interested in the analysis of the posterior distribution, that is the distribution of all model parameters given the data. Let $\boldsymbol{z} := (z_1 \ \ldots \ z_N)^\top$ be a vector with $z_i \in \{$"mm", "SNP", "exp"$\}$, $i = 1, \ldots, N$ with values for the random latent variable $\boldsymbol{Z} := (Z_1 \ \ldots \ Z_N)^\top$ with Z_i, $i = 1, \ldots, N$, defined in (5.3). Let $\boldsymbol{X} := \left(\boldsymbol{x_1} \ \ldots \ \boldsymbol{x}_{\boldsymbol{N}_{\text{TC}}}\right)^\top \in \mathbb{R}^{N_{\text{TC}} \times (L+1)}$ be the design matrix of the generalized linear model with $N_{TC} \in \mathbb{N}^+$ the number of T-to-C substitution positions. For BayMAP 1.0 the joint posterior distribution of all unknown parameters including $\boldsymbol{z}$ can be written as

$$\begin{aligned} f\left(\boldsymbol{\mu}, \boldsymbol{z}, q, \boldsymbol{\beta} \mid \boldsymbol{k}, \boldsymbol{n}, \boldsymbol{X}\right) &\propto f\left(\boldsymbol{k} \mid \boldsymbol{\mu}, \boldsymbol{n}, \boldsymbol{z}\right) \cdot f\left(\boldsymbol{\mu}, \boldsymbol{z}, q, \boldsymbol{\beta} \mid \boldsymbol{X}\right) \\ &= f\left(\boldsymbol{k} \mid \boldsymbol{\mu}, \boldsymbol{n}, \boldsymbol{z}\right) \cdot f\left(\boldsymbol{\mu}\right) \cdot f\left(\boldsymbol{z}, q, \boldsymbol{\beta} \mid \boldsymbol{X}\right) \\ &= f\left(\boldsymbol{k} \mid \boldsymbol{\mu}, \boldsymbol{n}, \boldsymbol{z}\right) \cdot f\left(\boldsymbol{\mu}\right) \cdot f\left(\boldsymbol{z} \mid \boldsymbol{p} = g^{-1}\left(\boldsymbol{X\beta}\right), q\right) \cdot f\left(q, \boldsymbol{\beta} \mid \boldsymbol{X}\right) \\ &= f\left(\boldsymbol{k} \mid \boldsymbol{\mu}, \boldsymbol{n}, \boldsymbol{z}\right) \cdot f\left(\boldsymbol{\mu}\right) \cdot f\left(\boldsymbol{z} \mid \boldsymbol{p} = g^{-1}\left(\boldsymbol{X\beta}\right), q\right) \cdot f\left(q\right) \cdot f\left(\boldsymbol{\beta}\right), \end{aligned} \tag{5.10}$$

The first line in (5.10) is an application of the Bayes theorem, where the denominator $f\left(\boldsymbol{k} \mid \boldsymbol{n}, \boldsymbol{X}\right)$ is omitted since it does not depend on the parameters of interest and the latent variable $\boldsymbol{z}$, so that the proportionality holds. Let the vector $\boldsymbol{K} := (K_1 \ \ldots \ K_N)^\top$ be the random vector of the numbers of substitutions, that belongs to the observed data

$\boldsymbol{k}$. The first factor of the first line is true, as $\boldsymbol{K}$ does not depend on q and $\boldsymbol{X\beta}$, if $\boldsymbol{z}$ is given, since the group probabilities are not necessary if the group to which each substitution belongs is known. The second line is true because of the definition of the conditional probability where $f(a,b) = f\left(a \mid b\right) f(b)$ for a joint density for a and b. Furthermore, the probability parameter $\boldsymbol{\mu}$ does not depend on $\boldsymbol{Z}$ directly. The last line holds, since q and $\boldsymbol{\beta}$ are supposed to be independent and both not dependent on $\boldsymbol{X}$.

Note, that the data for BayMAP 1.0 is not only determined by $\boldsymbol{k}$, but also includes $\boldsymbol{n}$ and $\boldsymbol{X}$. Let $f\left(\boldsymbol{X} \mid \boldsymbol{\psi}\right)$ and $f\left(\boldsymbol{n} \mid \boldsymbol{\xi}\right)$ be the densities for $\boldsymbol{X}$ and $\boldsymbol{n}$ dependent on parameter vectors $\boldsymbol{\psi}$ and $\boldsymbol{\xi}$. A full Bayesian model would thus include these parameters $\boldsymbol{\psi}$ and $\boldsymbol{\xi}$. However, when it is assumed that $\boldsymbol{\psi}$ and $\boldsymbol{\xi}$ are independent of the other model parameters, then $\boldsymbol{\psi}$ and $\boldsymbol{\xi}$ do not contribute to the distribution of all parameters other than $\boldsymbol{\psi}$ and $\boldsymbol{\xi}$ [18] since

$$f\left(\boldsymbol{\psi}, \boldsymbol{\xi}, \boldsymbol{\mu}, \boldsymbol{z}, q, \boldsymbol{\beta} \mid \boldsymbol{k}, \boldsymbol{n}, \boldsymbol{X}\right) = f\left(\boldsymbol{\psi}, \boldsymbol{\xi} \mid \boldsymbol{n}, \boldsymbol{X}\right) \cdot f\left(\boldsymbol{\mu}, \boldsymbol{z}, q, \boldsymbol{\beta} \mid \boldsymbol{k}, \boldsymbol{n}, \boldsymbol{X}\right). \tag{5.11}$$

This means, that it is sufficient to only consider the second factor, that is equal to (5.10) when only interested in $\boldsymbol{\mu}$, $\boldsymbol{z}$, q and $\boldsymbol{\beta}$.

Since the multiparameter model is too complex to sample directly from the posterior in (5.10), the componentwise Metropolis-Hastings algorithm is thus implemented in BayMAP 1.0 (see Algorithm 2). First, starting values for all unknown parameters have to be chosen or drawn, i.e. $\boldsymbol{\mu}^{(0)}$, $\boldsymbol{z}^{(0)}$, $q^{(0)}$, $\boldsymbol{\beta}^{(0)}$. For the next step, the full conditional distributions of all parameters are needed, i.e. the distribution of a parameter given all other parameters, so that the distribution of the parameters can be estimated by randomly sampled values based on the full conditional distributions. If the full conditional distribution of a parameter is a known distribution from which it is possible to sample directly, Gibbs sampling can be used for this component (see Algorithm 3). Otherwise, a normal jumping distribution is employed for sampling as described in Algorithm 2. In this section, the derivations of all full conditional densities, that are necessary for the implementation of the algorithm, are shown.

The first parameter of interest is $\boldsymbol{\mu}$. The density of $\boldsymbol{\mu}$ given all other parameters can be

written as

$$f\left(\boldsymbol{\mu} \mid z, q, \boldsymbol{\beta}, \boldsymbol{k}, \boldsymbol{n}, \boldsymbol{X}\right) = f\left(\boldsymbol{\mu} \mid z, \boldsymbol{k}, \boldsymbol{n}\right) \propto f\left(\boldsymbol{k} \mid \boldsymbol{\mu}, \boldsymbol{n}, z\right) \cdot f\left(\boldsymbol{\mu}\right). \tag{5.12}$$

The first equality $f\left(\boldsymbol{\mu} \mid z, q, \boldsymbol{\beta}, \boldsymbol{k}, \boldsymbol{n}, \boldsymbol{X}\right) = f\left(\boldsymbol{\mu} \mid z, \boldsymbol{k}, \boldsymbol{n}\right)$ holds, since it is not necessary to know q, $\boldsymbol{\beta}$ and $\boldsymbol{X}$ if z is known. This is the case because q and $\boldsymbol{p} = g^{-1}\left(\boldsymbol{X}\boldsymbol{\beta}\right)$ are parameter vectors to determine the probability for the group to which the substitutions belong. If these groups are given, there is no need in knowing these probabilities.

The second parameter of interest is the latent variable $\boldsymbol{Z}$, with full conditional density for z

$$\begin{aligned} f\left(z \mid \boldsymbol{\mu}, q, \boldsymbol{\beta}, \boldsymbol{k}, \boldsymbol{n}, \boldsymbol{X}\right) &\propto f\left(\boldsymbol{\mu} \mid z, \boldsymbol{k}, \boldsymbol{n}\right) \cdot f\left(z \mid \boldsymbol{p} = g^{-1}\left(\boldsymbol{X}\boldsymbol{\beta}\right), q, \boldsymbol{k}, \boldsymbol{n}\right) \\ &= \frac{f\left(\boldsymbol{k} \mid \boldsymbol{\mu}, \boldsymbol{n}, z\right) \cdot f\left(\boldsymbol{\mu}\right)}{f\left(\boldsymbol{k} \mid \boldsymbol{n}, z\right)} \cdot \frac{f\left(\boldsymbol{k} \mid \boldsymbol{n}, z\right) \cdot f\left(z \mid \boldsymbol{p}, q\right)}{f\left(\boldsymbol{k} \mid \boldsymbol{n}\right)} \\ &\propto f\left(\boldsymbol{k} \mid \boldsymbol{\mu}, \boldsymbol{n}, z\right) \cdot f\left(z \mid \boldsymbol{p}, q\right), \end{aligned} \tag{5.13}$$

where the fraction can be reduced by $f\left(\boldsymbol{k} \mid \boldsymbol{n}, z\right)$. The terms $f\left(\boldsymbol{\mu}\right)$ as well as $f\left(\boldsymbol{k} \mid \boldsymbol{n}\right)$ can be canceled out, as they do not depend on z.

The full conditional density for the parameter q, that is the probability of having a mismatch position given that the position is not crosslinked, can be written as

$$f\left(q \mid \boldsymbol{\mu}, z, \boldsymbol{\beta}, \boldsymbol{k}, \boldsymbol{n}, \boldsymbol{X}\right) = f\left(q \mid z, \boldsymbol{p} = g^{-1}\left(\boldsymbol{X}\boldsymbol{\beta}\right)\right) \propto f\left(z \mid q, \boldsymbol{p}\right) \cdot f\left(q\right), \tag{5.14}$$

since q is independent of $\boldsymbol{\mu}$ and $\boldsymbol{K}$ if z is given, as explained above. Moreover, it is assumed, that q is independent of $\boldsymbol{\beta}$, so that $f\left(q \mid \boldsymbol{\beta}, \boldsymbol{X}\right) = f\left(q\right)$.

The full conditional distribution of the last parameter of interest $\boldsymbol{\beta}$, that is the regression parameter for the generalized linear model, can be written as

$$f\left(\boldsymbol{\beta} \mid \boldsymbol{\mu}, z, q, \boldsymbol{k}, \boldsymbol{n}, \boldsymbol{X}\right) = f\left(\boldsymbol{\beta} \mid \mathbb{1}_{\text{“exp”}}\left(z\right), \boldsymbol{X}\right) \propto f\left(\mathbb{1}_{\text{“exp”}}\left(z\right) \mid \boldsymbol{X}, \boldsymbol{\beta}\right) \cdot f\left(\boldsymbol{\beta}\right), \tag{5.15}$$

since the parameter vector $\boldsymbol{\beta}$ for the generalized linear model is only dependent on the

data $\boldsymbol{X}$ for the generalized linear model and on the binary variable vector $\mathbb{1}_{\text{“exp”}}(\boldsymbol{z}) := \left(\mathbb{1}_{\text{“exp”}}(z_1) \ \ldots \ \mathbb{1}_{\text{“exp”}}(z_N)\right)^\top$, that determines if substitutions for positions i, $i = 1, \ldots, N$, are experimentally induced substitutions.

In BayMAP 1.0 the chains are sampled from full conditional distributions with densities (5.12) - (5.15), so that a sample from the joint posterior distribution with density (5.10) exists. In the following, concrete densities are inserted in full conditional distributions, e.g., the density of the zero truncated binomial distribution for the density of $\boldsymbol{k}$.

For simplicity reasons, $\boldsymbol{\mu} = \left(\mu_{\text{mm}}, \mu_{\text{SNP}}, \mu_{\text{exp}}\right)^\top$ is defined in this section as $\left(\mu_1, \mu_2, \mu_3\right)^\top$, so that $m = 1,2,3$ represents here the three groups of mismatches, SNPs and experimentally induced substitutions. This is also applied for z_i, $i = 1, \ldots, N$, so that $z_i \in \{1, 2, 3\}$. Additionally, z_{im} for $i = 1, \ldots, N$ and $m = 1,2,3$ is newly defined to be equal to one if substitution position i belongs to group m and equal to zero otherwise.

Full conditional distribution of probability parameter μ

In Section 5.1.1, K_i follows a zero truncated binomial distribution and the restriction $\mu_{\text{SNP}} = 1 - 3\mu_{\text{mm}}$ is considered. However, instead of a zero truncated binomial distribution, easier options with a binomial distribution and with no restriction could have also been considered. In this section only the conditional distributions for the model as presented in 5.1.1 are derived. The full conditionals for the other here stated combinations are presented in Appendix A.1.

The advantage of the easiest model with a binomial distribution and no restriction on μ_{SNP} is, that the Gibbs sampler can be implemented (see Appendix A.1). However, without the restriction of $\mu_{\text{SNP}} = 1 - 3\mu_{\text{mm}}$, μ_{SNP} is no longer forced to be very close to one. This can lead to label switching problems, e.g., that the group labeled “SNP” actually represents the experimentally induced substitutions and the other way around. Moreover, if the binomial distribution were embedded, it would not be taken into account that positions with zero substitutions are not considered.

The density for $\boldsymbol{\mu}$ given all parameters with K_i following a zero truncated binomial

distribution and $\mu_2 = 1 - 3\mu_1$, is given by (5.12) with

$$f\left(\boldsymbol{\mu} \mid \boldsymbol{z}, \boldsymbol{k}, \boldsymbol{n}\right) \propto f\left(\boldsymbol{k} \mid \boldsymbol{\mu}, \boldsymbol{n}, \boldsymbol{z}\right) \cdot f\left(\boldsymbol{\mu}\right).$$

With the priors for μ_1 and μ_3 given in (5.9), the second factor of the conditional density of $\boldsymbol{\mu}$ is

$$\begin{aligned} f\left(\boldsymbol{\mu}\right) &= f\left(\mu_1 \mid \mu_2\right) \cdot f\left(\mu_2\right) \cdot f\left(\mu_3\right) \\ &\propto \mathbb{1}_{\mu_2}(1 - 3\mu_1) \cdot \mathbb{1}_{[0.25,1]}(\mu_2) \cdot \mathbb{1}_{[0,1]}(\mu_3)\,, \end{aligned} \tag{5.16}$$

where $\mathbb{1}_{[0.25,1]}(\mu_2)$ is the indicator function defined to be equal to one if $\mu_2 \in [0.25, 1]$ and zero otherwise.

The first factor of the conditional density of $\boldsymbol{\mu}$, that is the likelihood, can be written as

$$\begin{aligned} f\left(\boldsymbol{k} \mid \boldsymbol{\mu}, \boldsymbol{n}, \boldsymbol{z}\right) &= \prod_{i=1}^{N} f\left(k_i \mid \boldsymbol{\mu}, n_i, z_i\right) \\ &= \prod_{i=1}^{N} \prod_{m=1}^{3} \mathrm{ZTB}\left(k_i \mid \mu_m, n_i\right)^{z_{im}} \\ &= \prod_{i=1}^{N} \prod_{m=1}^{3} \left(\frac{\binom{n_i}{k_i} \mu_m^{k_i} \left(1 - \mu_m\right)^{n_i - k_i}}{1 - \left(1 - \mu_m\right)^{n_i}} \right)^{z_{im}}, \end{aligned} \tag{5.17}$$

with the assumption of independent entries of $\boldsymbol{K}$. The parameter μ_2 can then be replaced by $1 - 3\mu_1$. Moreover, the term can be simplified by using

$$\prod_{i=1}^{N} \mu_1^{k_i z_{i1}} = \mu_1^{\sum_{i=1}^{N} k_i z_{i1}} = \mu_1^{N_1 \bar{k}_1}\,,$$

where $\bar{k}_m$ is the arithmetic mean of all k_i belonging to group m, $m = 1, 2, 3$ and $N_m \in \mathbb{N}^+$ is the total number of substitution positions in group m. Let n_m^{total} be the sum over

all n_i belonging to group m, the likelihood then becomes

$$
\begin{aligned}
&f\left(\boldsymbol{k} \mid \boldsymbol{\mu}, \boldsymbol{n}, \boldsymbol{z}\right) \\
&= \prod_{i=1}^{N} \binom{n_i}{k_i} \left(\frac{\mu_1^{k_i} \left(1-\mu_1\right)^{n_i-k_i}}{1-\left(1-\mu_1\right)^{n_i}} \right)^{z_{i1}} \left(\frac{\left(1-3\mu_1\right)^{k_i} \left(3\mu_1\right)^{n_i-k_i}}{1-\left(3\mu_1\right)^{n_i}} \right)^{z_{i2}} \left(\frac{\mu_3^{k_i} \left(1-\mu_3\right)^{n_i-k_i}}{1-\left(1-\mu_3\right)^{n_i}} \right)^{z_{i3}} \\
&= \frac{\mu_1^{N_1\bar{k}_1} \left(1-\mu_1\right)^{n_1^{\text{total}}-N_1\bar{k}_1} \left(1-3\mu_1\right)^{N_2\bar{k}_2} \left(3\mu_1\right)^{n_2^{\text{total}}-N_2\bar{k}_2} \mu_3^{N_3\bar{k}_3} \left(1-\mu_3\right)^{n_3^{\text{total}}-N_3\bar{k}_3}}{\prod_{i=1}^{N} \frac{1}{\binom{n_i}{k_i}} \left(1-\left(1-\mu_1\right)^{n_i}\right)^{z_{i1}} \left(1-\left(3\mu_1\right)^{n_i}\right)^{z_{i2}} \left(1-\left(1-\mu_3\right)^{n_i}\right)^{z_{i3}}} . \qquad (5.18)
\end{aligned}
$$

In the denominator of (5.18), however, the product sign cannot be replaced by a sum in the exponent, as the bases with $\left(1-\left(1-\mu_1\right)^{n_i}\right)$ are not identical for all i, $i = 1, \ldots, N$.

The full conditional density of $\boldsymbol{\mu}$ is thus the product of the likelihood in (5.18) and the prior of $\boldsymbol{\mu}$ in (5.16), so that

$$
\begin{aligned}
f\left(\boldsymbol{\mu} \mid \boldsymbol{z}, \boldsymbol{k}, \boldsymbol{n}\right) &\propto f\left(\boldsymbol{k} \mid \boldsymbol{\mu}, \boldsymbol{n}, \boldsymbol{z}\right) \cdot f\left(\boldsymbol{\mu}\right) \\
&\propto \frac{\mu_1^{N_1\bar{k}_1} \left(1-\mu_1\right)^{n_1^{\text{total}}-N_1\bar{k}_1} \left(1-3\mu_1\right)^{N_2\bar{k}_2} \left(3\mu_1\right)^{n_2^{\text{total}}-N_2\bar{k}_2} \mu_3^{N_3\bar{k}_3} \left(1-\mu_3\right)^{n_3^{\text{total}}-N_3\bar{k}_3}}{\prod_{i=1}^{N} \frac{1}{\binom{n_i}{k_i}} \left(1-\left(1-\mu_1\right)^{n_i}\right)^{z_{i1}} \left(1-\left(3\mu_1\right)^{n_i}\right)^{z_{i2}} \left(1-\left(1-\mu_3\right)^{n_i}\right)^{z_{i3}}} \\
&\quad \cdot \mathbb{1}_{\mu_2}\left(1-3\mu_1\right) \cdot \mathbb{1}_{[0.25,1]}\left(\mu_2\right) \cdot \mathbb{1}_{[0,1]}\left(\mu_3\right) \qquad (5.19)
\end{aligned}
$$

Full conditional distribution of allocation variable $\boldsymbol{Z}$

The density of the full conditional distribution of $\boldsymbol{Z}$ is given by (5.13) with

$$
f\left(\boldsymbol{z} \mid \boldsymbol{p}, q, \boldsymbol{\mu}, \boldsymbol{k}, \boldsymbol{n}\right) \propto f\left(\boldsymbol{k} \mid \boldsymbol{\mu}, \boldsymbol{n}, \boldsymbol{z}\right) \cdot f\left(\boldsymbol{z} \mid \boldsymbol{p}, q\right) .
$$

As $Z_1, \ldots, Z_N$ are here assumed to be independent and Z_i follows a categorical distribution, the second factor is then

$$\begin{aligned} f\left(\boldsymbol{z} \mid \boldsymbol{p}, q\right) &= \prod_{i=1}^{N} \mathrm{Cat}\left(z_i \mid p_i, q\right) \\ &= \prod_{i=1}^{N} \left((1-p_i)q\right)^{z_{i1}} \left((1-p_i)(1-q)\right)^{z_{i2}} p_i^{z_{i3}} \,. \end{aligned} \tag{5.20}$$

The first factor of the full conditional distribution is the likelihood of the data $\boldsymbol{k}$ and depends on the choice of the distribution of K_i, $i = 1, \ldots, N$, i.e. the binomial distribution or the zero truncated binomial distribution. The density for the full conditional distribution of $\boldsymbol{Z}$, where K_i follows a binomial distribution is given in Appendix A.2.

From (5.17), it is known, that the likelihood of the data $\boldsymbol{k}$ for K_i following the zero truncated binomial distribution is given by

$$f\left(\boldsymbol{k} \mid \boldsymbol{\mu}, \boldsymbol{n}, \boldsymbol{z}\right) = \prod_{i=1}^{N} \prod_{m=1}^{3} \left(\frac{\binom{n_i}{k_i} \mu_m^{k_i} \left(1-\mu_m\right)^{n_i-k_i}}{1-\left(1-\mu_m\right)^{n_i}} \right)^{z_{im}} .$$

With the weight

$$w_{im}^{\text{zero}} := \left(\frac{\binom{n_i}{k_i} \mu_m^{k_i} (1-\mu_m)^{n_i-k_i}}{1-(1-\mu_m)^{n_i}} \right),$$

the density of the full conditional distribution of $\boldsymbol{Z}$ can be written as

$$\begin{aligned} f\left(\boldsymbol{z} \mid \boldsymbol{p}, q, \boldsymbol{\mu}, \boldsymbol{k}, \boldsymbol{n}\right) &\propto f\left(\boldsymbol{k} \mid \boldsymbol{\mu}, \boldsymbol{n}, \boldsymbol{z}\right) \cdot f\left(\boldsymbol{z} \mid \boldsymbol{p}, q\right) \\ &= \prod_{i=1}^{N} \prod_{m=1}^{3} \left(\frac{\binom{n_i}{k_i} \mu_m^{k_i} (1-\mu_m)^{n_i-k_i}}{1-(1-\mu_m)^{n_i}} \right)^{z_{im}} \left((1-p_i)q\right)^{z_{i1}} \left((1-p_i)(1-q)\right)^{z_{i2}} p_i^{z_{i3}} \\ &= \prod_{i=1}^{N} \left(w_{i1}^{\text{zero}}(1-p_i)q\right)^{z_{i1}} \left(w_{i2}^{\text{zero}}(1-p_i)(1-q)\right)^{z_{i2}} \left(w_{i3}^{\text{zero}} p_i\right)^{z_{i3}} \,. \end{aligned} \tag{5.21}$$

The last line of (5.21) already has the structure of the product over the density functions of the categorical distribution. However, for a categorical distribution, the weights for each group have to be probabilities that sum up to one. The weights in (5.21) can therefore be divided by their sum

$$\widetilde{w}^{\text{zero}} = w_{i1}^{\text{zero}}(1-p_i)q) + w_{i2}^{\text{zero}}(1-p_i)(1-q) + w_{i3}^{\text{zero}}p_i \ , \tag{5.22}$$

so that they sum up to one and thus represent probabilities. This can be done by multiplying (5.21) by the constant factor $\left(\frac{1}{\widetilde{w}^{\text{zero}}}\right)^N$, since then

$$\begin{aligned} f\left(\boldsymbol{z} \mid \boldsymbol{p}, q, \boldsymbol{\mu}, \boldsymbol{k}, \boldsymbol{n}\right) &\propto \left(\frac{1}{\widetilde{w}^{\text{zero}}}\right)^N \prod_{i=1}^{N} \left(w_{i1}^{\text{zero}}(1-p_i)q)\right)^{z_{i1}} \left(w_{i2}^{\text{zero}}(1-p_i)(1-q)\right)^{z_{i2}} \left(w_{i3}^{\text{zero}}p_i\right)^{z_{i3}} \\ &= \prod_{i=1}^{N} \left(\frac{w_{i1}^{\text{zero}}(1-p_i)q)}{\widetilde{w}^{\text{zero}}}\right)^{z_{i1}} \left(\frac{w_{i2}^{\text{zero}}(1-p_i)(1-q)}{\widetilde{w}^{\text{zero}}}\right)^{z_{i2}} \left(\frac{w_{i3}^{\text{zero}}p_i}{\widetilde{w}^{\text{zero}}}\right)^{z_{i3}} . \end{aligned}$$

The latent variable Z_i, conditional on the other parameters and the data, hence follows a categorical distribution with

$$f\left(\boldsymbol{z} \mid \boldsymbol{p}, q, \boldsymbol{\mu}, \boldsymbol{k}, \boldsymbol{n}\right) \propto \prod_{i=1}^{N} \text{Cat}\left(\frac{w_{i1}^{\text{zero}}(1-p_i)q)}{\widetilde{w}^{\text{zero}}}, \frac{w_{i2}^{\text{zero}}(1-p_i)(1-q)}{\widetilde{w}^{\text{zero}}}, \frac{w_{i3}^{\text{zero}}p_i}{\widetilde{w}^{\text{zero}}}\right). \tag{5.23}$$

The variable Z_i can therefore be sampled directly from its full conditional using the Gibbs sampler.

Full conditional distribution of the probability q

The density of the full conditional distribution of q, that is the probability for non-experimentally induced substitution positions to be mismatch positions, is given with (5.14) by

$$f\left(q \mid \boldsymbol{z}, \boldsymbol{p}\right) \propto f\left(\boldsymbol{z} \mid q, \boldsymbol{p}\right) \cdot f\left(q\right) \ .$$

The prior density $f\left(q\right)$ is the density of the uniform distribution on $[0,1]$ (see (5.9)) and the density $f\left(\boldsymbol{z} \mid q, \boldsymbol{p}\right)$ is the product over the densities of categorical distributions as illustrated in (5.20).

The full conditional density for q therefore becomes

$$f\left(q \mid \boldsymbol{z}, \boldsymbol{p}\right) \propto \prod_{i=1}^{N} \left((1-p_i)q\right)^{z_{i1}} \left((1-p_i)(1-q)\right)^{z_{i2}} p^{z_{i3}} \mathbb{1}_{[0,1]}\left(q\right)$$

$$= q^{N_1}(1-q)^{N_2}\, \mathbb{1}_{[0,1]}\left(q\right) \prod_{i=1}^{N} (1-p_i)^{z_{i1}+z_{i2}} p_i^{z_{i3}}\,. \tag{5.24}$$

As the terms behind the product sign do not depend on q, the density can be simplified to

$$f\left(q \mid \boldsymbol{z}, \boldsymbol{p}\right) \propto q^{N_1}(1-q)^{N_2}\mathbb{1}_{(0,1)}\left(q\right)$$

$$\propto \frac{\Gamma\left(N_1+N_2+2\right)}{\Gamma\left(N_1+1\right)\Gamma\left(N_2+1\right)} q^{N_1}(1-q)^{N_2}\mathbb{1}_{(0,1)}\left(q\right)$$

$$= \text{Beta}\left(q \mid N_1+1, N_2+1\right)\,, \tag{5.25}$$

where $\Gamma(\cdot)$ is the gamma function. The probability q, conditional on $\boldsymbol{z}$ and $\boldsymbol{p}$, follows therefore a beta distribution with parameters (N_1+1), i.e. the number of mismatch positions plus 1, and (N_2+1), i.e. the number of SNP positions plus 1. A Gibbs sampler can thus also be used in order to sample q from its full conditional distribution.

Full conditional distribution of the parameter vector $\boldsymbol{\beta}$

In (5.15) the distribution of $\boldsymbol{\beta}$ is only dependent on $\boldsymbol{X}$ and $\boldsymbol{Z}$ with density $f\left(\boldsymbol{\beta} \mid \mathbb{1}_{\text{“exp”}}(\boldsymbol{z}), \boldsymbol{X}\right)$ $\propto f\left(\mathbb{1}_{\text{“exp”}}(\boldsymbol{z}) \mid \boldsymbol{X}, \boldsymbol{\beta}\right) \cdot f\left(\boldsymbol{\beta}\right)$. Instead of directly using

$$f\left(\mathbb{1}_{\text{“exp”}}(\boldsymbol{z}) \mid \boldsymbol{X}, \boldsymbol{\beta}\right) = \prod_{i=1}^{N_{\text{TC}}} P\left(Z_i = \text{“exp}'' \mid \boldsymbol{X}, \boldsymbol{\beta}\right) = \prod_{i=1}^{N_{\text{TC}}} \Phi\left(\boldsymbol{x_i}^{\top}\boldsymbol{\beta}\right), \tag{5.26}$$

a random latent variable Y_i for position $i = 1, \ldots, N_{\text{TC}}$ can be added with

$$\left(Y_i \mid \boldsymbol{x_i}^{\top}\boldsymbol{\beta}\right) \sim N\left(\boldsymbol{x_i}^{\top}\boldsymbol{\beta}, 1\right) \text{ and} \tag{5.27}$$

$$Z_{i3} = \mathbb{1}_{(0,\infty)}(Y_i). \tag{5.28}$$

An advantage of using the model with the latent variable is, that conjugate priors for $\boldsymbol{\beta}$ can be implemented, so that the Gibbs sampler can be used [18]. This means, that given the values $\boldsymbol{y} := \left(y_1 \ \ldots \ y_{N_{\text{TC}}}\right)^{\top} \in \mathbb{R}^{N_{\text{TC}}}$ for the random latent variable $\boldsymbol{Y} := \left(Y_1 \ \ldots \ Y_{N_{\text{TC}}}\right)^{\top}$,

a linear regression is modeled. The variance of Y_i in (5.27) is set to one, so that

$$\begin{aligned} P\left(Z_i = \text{“exp”} \mid \boldsymbol{X}, \boldsymbol{\beta}\right) &= P\left(Y_i > 0 \mid \boldsymbol{X}, \boldsymbol{\beta}\right) \\ &= P\left(Y_i - \boldsymbol{x_i}^\top \boldsymbol{\beta} > -\boldsymbol{x_i}^\top \boldsymbol{\beta}\right) \\ &= P\left(Y_i - \boldsymbol{x_i}^\top \boldsymbol{\beta} < \boldsymbol{x_i}^\top \boldsymbol{\beta}\right) = \Phi\left(\boldsymbol{x_i}^\top \boldsymbol{\beta}\right), \end{aligned} \tag{5.29}$$

because of the symmetry of the normal distribution and $\left(Y_i - \boldsymbol{x_i}^\top \boldsymbol{\beta}\right) \sim N(0,1)$.

The full conditional density in (5.15) can then be represented conditionally on $\boldsymbol{y}$, since

$$\begin{aligned} f\left(\boldsymbol{\beta} \mid \mathbb{1}_{\text{“exp”}}(\boldsymbol{z}), \boldsymbol{X}\right) &= \int_{\mathbb{R}^{N_{\text{TC}}}} f\left(\boldsymbol{\beta} \mid \boldsymbol{y}, \mathbb{1}_{\text{“exp”}}(\boldsymbol{z}), \boldsymbol{X}\right) f\left(\boldsymbol{y} \mid \mathbb{1}_{\text{“exp”}}(\boldsymbol{z}), \boldsymbol{X}\right) d\boldsymbol{y} \\ &= \int_{\mathbb{R}^{N_{\text{TC}}}} f\left(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}\right) f\left(\boldsymbol{y} \mid \mathbb{1}_{\text{“exp”}}(\boldsymbol{z}), \boldsymbol{X}\right) d\boldsymbol{y}\,. \end{aligned} \tag{5.30}$$

In the second line $f\left(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}\right)$ is not dependent on $\mathbb{1}_{\text{“exp”}}(\boldsymbol{Z})$ anymore, as $\boldsymbol{y}$ already includes the information if the entries of $\boldsymbol{z}$ are equal to “exp”. The aim is here to sample from the distribution of $\left(\boldsymbol{\beta} \mid \mathbb{1}_{\text{“exp”}}(\boldsymbol{z})\right)$. This aim can be reached by alternating between sampling $\boldsymbol{\beta}$ conditionally on $\boldsymbol{y}$ from density $f\left(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}\right)$ and sampling $\boldsymbol{y}$ conditionally on $\boldsymbol{\beta}$ from density $f\left(\boldsymbol{y} \mid \mathbb{1}_{\text{“exp”}}(\boldsymbol{z}), \boldsymbol{X}, \boldsymbol{\beta}\right)$ [17].

Since Y_i has to be greater than zero if Z_i is equal to “exp” and smaller or equal to zero otherwise, y_i can be sampled from a truncated normal distribution with mean parameter $\boldsymbol{x}_i^\top \boldsymbol{\beta}$ and variance equal to one.

The regression parameter $\boldsymbol{\beta}$ can therefore be sampled conditioned on $\boldsymbol{y}$ from

$$f\left(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}\right) \propto f\left(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}\right) f\left(\boldsymbol{\beta}\right)\,. \tag{5.31}$$

Since the prior for β_ℓ, $\ell = 0, \ldots, L$, is a normal distribution with mean parameter zero and variance 10^6 (see (5.9)) and $\beta_0, \ldots, \beta_L$ are assumed to be independent, the distri-

bution of $\boldsymbol{\beta}$ can also be illustrated by a multivariate normal distribution with

$$\begin{aligned} f(\boldsymbol{\beta}) &= N(\boldsymbol{\beta} \mid 0, \sigma^2 \boldsymbol{I}_{L+1}) \\ &= (2\,\pi\,\sigma^2)^{-\frac{(L+1)}{2}} \exp\left(-\frac{1}{2}\frac{\boldsymbol{\beta}^\top \boldsymbol{\beta}}{\sigma^2}\right) \\ &\propto \exp\left(-\frac{1}{2}\frac{\boldsymbol{\beta}^\top \boldsymbol{\beta}}{\sigma^2}\right), \end{aligned} \tag{5.32}$$

where σ^2 is here fixed to 10^6 in order to have a vague prior, π is the circle constant and $\boldsymbol{I}_{L+1}$ is the identity matrix of size $(L+1)$. The last line holds, as only the term in the exponential function includes the parameter of interest $\boldsymbol{\beta}$.

The first factor in (5.31) $f(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta})$ can also be represented by a multivariate normal distribution, since Y_i follows a normal distribution (see (5.27)) with $Y_1, \ldots, Y_{N_{\text{TC}}}$ assumed to be independent, so that

$$\begin{aligned} f(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}) &= N(\boldsymbol{y} \mid \boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{I}_{N_{\text{TC}}}) \\ &\propto \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right). \end{aligned} \tag{5.33}$$

The density for $\boldsymbol{\beta}$ conditional on $\boldsymbol{y}$ can thus be calculated as

$$\begin{aligned} f(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}) &\propto f(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta})\, f(\boldsymbol{\beta}) \\ &\propto \exp\left(-\frac{1}{2}\frac{\boldsymbol{\beta}^\top \boldsymbol{\beta}}{\sigma^2}\right) \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right) \\ &= \exp\left(-\frac{1}{2}\left((\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \frac{\boldsymbol{\beta}^\top \boldsymbol{\beta}}{\sigma^2}\right)\right) \\ &= \exp\left(-\frac{1}{2}\left(\boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{\beta}^\top \left(\boldsymbol{X}^\top \boldsymbol{X} + \frac{1}{\sigma^2}\boldsymbol{I}_{L+1}\right)\boldsymbol{\beta}\right)\right). \end{aligned} \tag{5.34}$$

The two summands including $\boldsymbol{\beta}$ in the last line can be written as a quadratic form if the term $\boldsymbol{y}^\top \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X} + \frac{1}{\sigma^2}\boldsymbol{I}_{L+1}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$ is added to the sum. If adding, it naturally has to be subtracted of the sum, too, so that the last term of the previous equation is identical to

$$f(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}) \propto \exp\Bigg(-\frac{1}{2}\Bigg(\boldsymbol{\beta} - \Bigg(\boldsymbol{X}^\top \boldsymbol{X} + \frac{1}{\sigma^2}\boldsymbol{I}_{L+1}\Bigg)^{-1} \boldsymbol{X}^\top \boldsymbol{y}\Bigg)^\top \Bigg(\boldsymbol{X}^\top \boldsymbol{X} + \frac{1}{\sigma^2}\boldsymbol{I}_{L+1}\Bigg)$$

$$\cdot \Bigg(\boldsymbol{\beta} - \Bigg(\boldsymbol{X}^\top \boldsymbol{X} + \frac{1}{\sigma^2}\boldsymbol{I}_{L+1}\Bigg)^{-1} \boldsymbol{X}^\top \boldsymbol{y}\Bigg)$$

$$+ \boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{X}\Bigg(\boldsymbol{X}^\top \boldsymbol{X} + \frac{1}{\sigma^2}\boldsymbol{I}_{L+1}\Bigg)^{-1} \boldsymbol{X}^\top \boldsymbol{y}\Bigg). \tag{5.35}$$

The terms of the last line do not depend on $\boldsymbol{\beta}$, so that there is no loss of information, when omitting the factor $\exp\Big(\boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{X}\Big(\boldsymbol{X}^\top \boldsymbol{X} + \frac{1}{\sigma^2}\boldsymbol{I}_{L+1}\Big)^{-1} \boldsymbol{X}^\top \boldsymbol{y}\Big)$, and therefore the density simplifies to

$$f(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}) \propto \exp\Bigg(-\frac{1}{2}\Bigg(\boldsymbol{\beta} - \Bigg(\boldsymbol{X}^\top \boldsymbol{X} + \frac{1}{\sigma^2}\boldsymbol{I}_{L+1}\Bigg)^{-1} \boldsymbol{X}^\top \boldsymbol{y}\Bigg)^\top \Bigg(\boldsymbol{X}^\top \boldsymbol{X} + \frac{1}{\sigma^2}\boldsymbol{I}_{L+1}\Bigg)$$

$$\cdot \Bigg(\boldsymbol{\beta} - \Bigg(\boldsymbol{X}^\top \boldsymbol{X} + \frac{1}{\sigma^2}\boldsymbol{I}_{L+1}\Bigg)^{-1} \boldsymbol{X}^\top \boldsymbol{y}\Bigg)\Bigg), \tag{5.36}$$

which is proportional to the multivariate normal distribution, so that

$$f(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}) \propto N\Bigg(\boldsymbol{\beta} \,\Bigg|\, \Bigg(\boldsymbol{X}^\top \boldsymbol{X} + \frac{1}{\sigma^2}\boldsymbol{I}_{L+1}\Bigg)^{-1} \boldsymbol{X}^\top \boldsymbol{y}, \Bigg(\boldsymbol{X}^\top \boldsymbol{X} + \frac{1}{\sigma^2}\boldsymbol{I}_{L+1}\Bigg)^{-1}\Bigg). \tag{5.37}$$

The paramter $\boldsymbol{\beta}$ and the variable $\boldsymbol{y}$ can therefore be sampled directly from their full conditional distributions, i.e. the normal distribution and the truncated normal distribution, respectively, using the Gibbs sampler.

5.1.3 Sampling scheme

In order to sample from the joint posterior distribution to learn more about the unknown parameters, a combination of the componentwise Metropolis-Hastings algorithm (see Algorithm 2) and the Gibbs sampler (see Algorithm 3) is used. This means, that for those components, where it is possible to sample directly from its full conditional distribution, the Gibbs sampler is implemented and that the other components are sampled using the normal distribution as a jumping distribution for the componentwise Metropolis-Hastings algorithm.

First, start values for the algorithm have to be drawn or chosen. This initialization can for example be done by the following scheme:

1. Draw $\boldsymbol{\beta}^{(0)}$ from $N(\mathbf{0}_{L+1}, \boldsymbol{I}_{L+1})$,

   where $\mathbf{0}_{L+1}$ is the null vector with length $(L+1)$.

2. Calculate $\boldsymbol{p}^{(0)} = \Phi\left(\boldsymbol{X}\boldsymbol{\beta}^{(0)}\right)$ and

   set those entries of $\boldsymbol{p}^{(0)}$ to zero, where the substitution type is not T-to-C.

3. Draw $q^{(0)}$ from $U(0,1)$.

4. Draw $z_i^{(0)}$ for $i = 1, \ldots, N$ from $\text{Cat}\left(\left(1-p_i^{(0)}\right)q^{(0)}, \left(1-p_i^{(0)}\right)\left(1-q^{(0)}\right), p_i^{(0)}\right)$.

5. Set $\boldsymbol{\mu}^{(0)} = (0.05, 0.85, 0.5)$.

These starting values can, however, also be based on other starting distributions or they can also be chosen in advance, for example with the aim of reaching convergence in a shorter amount of time (see e.g., Gelman et al. [18]). The parameter $\boldsymbol{\mu}^{(0)}$ is here set to a fixed value, so that at the beginning $\mu_{\text{mm}}^{(0)} \ll \mu_{\text{exp}}^{(0)} \ll \mu_{\text{SNP}}^{(0)}$. This has the advantage that label switching problems between μ_{exp} and either μ_{mm} or μ_{SNP} are unlikely.

With these starting values the parameters of interest can then be sampled based on their full conditional distributions. For iteration h, with $h = 1, 2, \ldots$, in BayMAP 1.0 this is done by the following scheme:

1. Determine $\mu_{\text{mm}}^{(h)}$ and $\mu_{\text{SNP}}^{(h)}$ by

   (a) Drawing μ_{mm}^* from the jumping distribution $N\left(\mu_{\text{mm}}^{(h-1)}, \rho_{\text{mm}}^2\right)$,

   (b) Calculating $\mu_{\text{SNP}}^* = 1 - 3\mu_{\text{mm}}^*$,

   (c) Setting $\boldsymbol{\mu}^* = \left(\mu_{\text{mm}}^*, \mu_{\text{SNP}}^*, \mu_{\text{exp}}^{(h-1)}\right)$,

   (d) Calculating $r = \dfrac{f\left(\boldsymbol{\mu}^* \mid \boldsymbol{z}^{(h-1)}, \boldsymbol{k}, \boldsymbol{n}\right)}{f\left(\boldsymbol{\mu}^{(h-1)} \mid \boldsymbol{z}^{(h-1)}, \boldsymbol{k}, \boldsymbol{n}\right)}$ with density in (5.19),

   (e) Drawing u from $U(0,1)$,

   (f) Setting $\mu_{\text{mm}}^{(h)} = \begin{cases} \mu_{\text{mm}}^*, & \text{if } u \le r \\ \mu_{\text{mm}}^{(h-1)}, & \text{if } u > r \end{cases}$ and $\mu_{\text{SNP}}^{(h)} = \begin{cases} \mu_{\text{SNP}}^*, & \text{if } u \le r \\ \mu_{\text{SNP}}^{(h-1)}, & \text{if } u > r \end{cases}$.

2. Determine $\mu_{\text{exp}}^{(h)}$ by

   (a) Drawing μ_{exp}^{*} from the jumping distribution $N\left(\mu_{\text{exp}}^{(h-1)}, \rho_{\text{exp}}^{2}\right)$,

   (b) Setting $\boldsymbol{\mu}^{*} = \left(\mu_{\text{mm}}^{(h)}, \mu_{\text{SNP}}^{(h)}, \mu_{\text{exp}}^{*}\right)$,

   (c) Calculating $r = \dfrac{f\left(\boldsymbol{\mu}^{*} \mid \boldsymbol{z}^{(h-1)}, \boldsymbol{k}, \boldsymbol{n}\right)}{f\left(\left(\mu_{\text{mm}}^{(h)}, \mu_{\text{SNP}}^{(h)}, \mu_{\text{exp}}^{(h-1)}\right) \mid \boldsymbol{z}^{(h-1)}, \boldsymbol{k}, \boldsymbol{n}\right)}$ with density in (5.19),

   (d) Drawing u from $U(0,1)$,

   (e) Setting $\boldsymbol{\mu}^{(h)} = \left(\mu_{\text{mm}}^{(h)}, \mu_{\text{SNP}}^{(h)}, \mu_{\text{exp}}^{(h)}\right) = \begin{cases} \boldsymbol{\mu}^{*}, & \text{if } u \leq r \\ \left(\mu_{\text{mm}}^{(h)}, \mu_{\text{SNP}}^{(h)}, \mu_{\text{exp}}^{(h-1)}\right), & \text{if } u > r \end{cases}$.

Note that the parameter $\boldsymbol{\mu}$ is here divided into two subcomponents including on the one side μ_{mm} as well as μ_{SNP} and on the other side μ_{exp}. It is nevertheless possible to use the full conditional distribution of the combined $\boldsymbol{\mu}$, since for a fixed μ_{exp}

$$\begin{aligned} f\left(\mu_{\text{mm}}, \mu_{\text{SNP}} \mid \boldsymbol{z}, \boldsymbol{k}, \boldsymbol{n}\right) &\propto f\left(\mu_{\text{mm}}, \mu_{\text{SNP}} \mid \boldsymbol{z}, \boldsymbol{k}, \boldsymbol{n}\right) f\left(\mu_{\text{exp}} \mid \boldsymbol{z}, \boldsymbol{k}, \boldsymbol{n}\right) \\ &= f\left(\mu_{\text{mm}}, \mu_{\text{SNP}}, \mu_{\text{exp}} \mid \boldsymbol{z}, \boldsymbol{k}, \boldsymbol{n}\right) = f\left(\boldsymbol{\mu} \mid \boldsymbol{z}, \boldsymbol{k}, \boldsymbol{n}\right) . \end{aligned} \tag{5.38}$$

The same is true for a fixed μ_{mm}, so that $f\left(\mu_{\text{exp}} \mid \boldsymbol{z}, \boldsymbol{k},\right) \propto f\left(\boldsymbol{\mu} \mid \boldsymbol{z}, \boldsymbol{k}, \boldsymbol{n}\right)$.

First, a candidate for μ_{mm} and μ_{SNP} (or for μ_{exp}) is drawn. If the full conditional density of the candidate is larger than that of the last stored value, i.e. if it speaks more for the candidate than for the old value, then the candidate is stored as new value, since then $r > 1$. If it speaks more for the old value, then $r < 1$ and the candidate is stored as a new value with probability r and otherwise the old value. This means, that candidates are always accepted, if they are increasing the density and sometimes if they are not.

The variance parameters ρ_{mm}^{2} and ρ_{exp}^{2} of the jumping distribution have to be specified in such a way, that the acceptance rate is neither too small nor too large (see Section 3.3).

The sampling scheme for the other parameters for $h = 1, 2, \ldots$ continues with:

3. Draw $z_i^{(h)}$ for $i = 1, \ldots, N$ from the categorical distribution defined in (5.23) with

$$\text{Cat}\left(\frac{w_{i1}^{\text{zero}^{(h)}}\left(1-p_i^{(h-1)}\right)q^{(h-1)})}{\widetilde{w}^{\text{zero}^{(h)}}}, \frac{w_{i2}^{\text{zero}^{(h)}}\left(1-p_i^{(h-1)}\right)\left(1-q^{(h)}\right)}{\widetilde{w}^{\text{zero}^{(h)}}}, \frac{w_{i3}^{\text{zero}^{(h)}}p_i^{(h-1)}}{\widetilde{w}^{\text{zero}^{(h)}}}\right).$$

4. Draw $q^{(h)}$ from $\text{Beta}\left(N_{\text{mm}}^{(h)}+1, N_{\text{SNP}}^{(h)}+1\right)$.

5. Draw $y_i^{(h)}$ for $i = 1, \ldots, N_{\text{TC}}$ from the truncated normal distribution with density

$$\begin{cases} f\left(y_i^{(h)} \mid \boldsymbol{x_i}^\top \boldsymbol{\beta}^{(h-1)}, 1, Y_i^{(h)} > 0\right), & \text{if } z_i^{(h)} = \text{“exp”} \\ f\left(y_i^{(h)} \mid \boldsymbol{x_i}^\top \boldsymbol{\beta}^{(h-1)}, 1, Y_i^{(h)} \leq 0\right), & \text{if } z_i^{(h)} \neq \text{“exp”} \end{cases},$$

   where $\boldsymbol{x_i}^\top \boldsymbol{\beta}^{(h-1)}$ is the mean parameter of the normal distribution and one the variance parameter.

6. Draw $\boldsymbol{\beta}^{(h)}$ from $N\left(\left(\boldsymbol{X}^\top \boldsymbol{X} + \frac{1}{\sigma^2}\boldsymbol{I}_{L+1}\right)^{-1}\boldsymbol{X}^\top \boldsymbol{y}^{(h)}, \left(\boldsymbol{X}^\top \boldsymbol{X} + \frac{1}{\sigma^2}\boldsymbol{I}_{L+1}\right)^{-1}\right)$ (see (5.37)).

7. Calculate $\boldsymbol{p}^{(h)} = \Phi\left(\boldsymbol{X}\boldsymbol{\beta}^{(h)}\right)$ and

   set those entries of $\boldsymbol{p}^{(h)}$ to zero, where the substitution type is not T-to-C.

For notations and derivations of the distributions see Section 5.1.2.

The sampling process should be repeated until convergence is reached. Only samples for iterations should be kept for which the Markov chain has converged. This means, that convergence has to be checked, e.g., by a graphical representation of the chains, and early iterations, i.e. the burn-in, should be discarded. In order to reduce autocorrelation, only every o-th iteration could be kept. The final number of stored iterations is here called $N_{\text{iter}} \in \mathbb{N}^+$.

5.1.4 Identification of method-induced substitution positions

In order to evaluate if a T-to-C substitution position is crosslinked and therefore a binding site position, the probability of the position being crosslinked given the data, can be estimated with the results of the MCMC algorithm. If position i, $i = 1, \ldots, N_{\text{TC}}$, has experimentally induced substitutions and is hence crosslinked, the latent variable Z_i is equal to “exp”. As Z_i is a latent variable, it is not observed, but realizations are sampled for every iteration step given all other parameters (for more details see Section 5.1.2).

Once the MCMC chain has converged, the remaining iterations for the converged chain

can be used to estimate the probability of position $i, i = 1,\ldots,N_{\text{TC}}$, presenting experimentally induced T-to-C substitutions. As shown in Section 5.1.2, Z_i given the data and all parameters except for Z_i, follows a Categorical distribution. In order to estimate the probability, one can, therefore, just count the number of times Z_i is chosen to be experimentally induced and divide it by the number of stored iterations of the converged chain, so that

$$\hat{P}\left(Z_i = \text{“exp"} \mid k_i, n_i\right) = \frac{\sum_{h=1}^{N_{\text{iter}}} \mathbb{1}_{\text{“exp"}}\left(z_i^{(h)}\right)}{N_{\text{iter}}}, \tag{5.39}$$

where $N_{\text{iter}} \in \mathbb{N}^+$ is the number of stored iterations in the algorithm and $z_i^{(h)}$ is the h-th entry of the MCMC sample for Z_i, $h = 1,...,N_{\text{iter}}$ and $i = 1,\ldots,N_{\text{TC}}$. Nevertheless, this estimation can only be done for positions that were used in the algorithm for estimating the distributions. Moreover, all the $z_i^{(h)}$ have to be stored, that means $N_{\text{TC}} \cdot N_{\text{iter}}$ values, where N_{TC} is the number of all positions with T-to-C substitutions. If it is not possible to store all the values for $z_i^{(h)}$ or if it is wished to estimate the probability for positions that were not used in the algorithm, another estimation method has to be used.

When comparing two models or hypotheses in Bayesian statistics, the Bayes factor is often used. The Bayes factor is the ratio of the marginal likelihood of two models, that is the density of the data assuming model 1 divided by the density of the data assuming model 2. Here, model 1 indicates that Z_i = “exp” and model 2 indicates that Z_i = “mm” or Z_i = “SNP”, so that the Bayes Factor BF_i can be calculated as

$$BF_i = \frac{f\left(k_i \mid Z_i = \text{“exp"}, n_i\right)}{f\left(k_i \mid Z_i \neq \text{“exp"}, n_i\right)} = \frac{\int_0^1 f\left(k_i \mid \mu, Z_i = \text{“exp"}, n_i\right) f\left(\mu \mid Z_i = \text{“exp"}\right) d\mu}{\int_0^1 f\left(k_i \mid \mu, Z_i \neq \text{“exp"}, n_i\right) f\left(\mu \mid Z_i \neq \text{“exp"}\right) d\mu}. \tag{5.40}$$

One way to estimate these integrals is Monte Carlo integration, where the mean of the likelihood over the sampled parameters of the MCMC chain is calculated [42], so that

$$\hat{f}\left(k_i \mid Z_i = \text{“exp"}, n_i\right) = \frac{1}{N_{\text{iter}}} \sum_{h=1}^{N_{\text{iter}}} f\left(k_i \mid \mu_{\text{exp}}^{(h)}, n_i\right).$$

The Bayes factor for position i can then be estimated as

$$\widehat{BF}_i = \frac{\sum_{h=1}^{N_{\text{iter}}} f\left(k_i \mid \mu_{\text{exp}}^{(h)}, n_i\right)}{\sum_{h=1}^{N_{\text{iter}}} \left(q^{(h)} \cdot f\left(k_i \mid \mu_{\text{mm}}^{(h)}, n_i\right) + (1-q^{(h)}) \cdot f\left(k_i \mid \mu_{\text{SNP}}^{(h)}, n_i\right)\right)}, \tag{5.41}$$

where $f\left(k_i \mid \cdot\right)$ is the likelihood of the zero truncated binomial distribution for $i = 1, \ldots, N_{\text{TC}}$.

If using the Bayes factor, the advantage is that only N_{iter} entries for μ_{exp}, μ_{mm} and q have to be stored instead of $N_{\text{TC}} \cdot N_{\text{iter}}$ entries for Z_i. However, the prior information of the additional variables, such as the mRNA region, is not taken into account. To this end, one can regard the posterior odds PoO_i that compare the probability of model 1 given the data to the probability of model 2 given the data

$$\begin{aligned} PoO_i &= \frac{P\left(Z_i = \text{“exp"} \mid k_i, n_i\right)}{P\left(Z_i \neq \text{“exp"} \mid k_i, n_i\right)} \\ &= \frac{f\left(k_i \mid Z_i = \text{“exp"}, n_i\right) \cdot P\left(Z_i = \text{“exp"}\right)}{f\left(k_i \mid Z_i \neq \text{“exp"}, n_i\right) \cdot P\left(Z_i \neq \text{“exp"}\right)} \\ &= BF_i \cdot \frac{P\left(Z_i = \text{“exp"}\right)}{P\left(Z_i \neq \text{“exp"}\right)}. \end{aligned} \tag{5.42}$$

The last factor in Equation (5.42) is called the prior odds. The posterior odds are the product of the Bayes factor and the prior odds. The prior odds PrO_i can be rewritten as

$$\begin{aligned} PrO_i &= \frac{P\left(Z_i = \text{“exp"}\right)}{P\left(Z_i \neq \text{“exp"}\right)} \\ &= \frac{\int_{\mathbb{R}^{L+1}} P\left(Z_i = \text{“exp"} \mid p_i = g^{-1}(\boldsymbol{x_i}^\top \boldsymbol{\beta})\right) \cdot f\left(\boldsymbol{\beta}\right) d\boldsymbol{\beta}}{\int_{\mathbb{R}^{L+1}} P\left(Z_i \neq \text{“exp"} \mid p_i = g^{-1}(\boldsymbol{x_i}^\top \boldsymbol{\beta})\right) \cdot f\left(\boldsymbol{\beta}\right) d\boldsymbol{\beta}}. \end{aligned} \tag{5.43}$$

The posterior odds include the information about the supplementary data such as the

mRNA region. They can be estimated as

$$\widehat{PoO}_i = \widehat{BF}_i \cdot \frac{\frac{1}{N_{\text{iter}}} \sum_{h=1}^{N_{\text{iter}}} p_i^{(h)}}{1 - \frac{1}{N_{\text{iter}}} \sum_{h=1}^{N_{iter}} p_i^{(h)}}, \tag{5.44}$$

where $p_i^{(h)}$ is similar to (5.8) equal to $\Phi\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}^{(h)}\right)$ for a T-to-C substitution position, so that the additional information for each position is also taken into account.

The estimated posterior odds can be easily converted to

$$\hat{P}(Z_i = \text{“exp"} \mid k_i, n_i) = \frac{\widehat{PoO}_i}{1 + \widehat{PoO}_i}$$

since

$$PoO_i = \frac{P\left(Z_i = \text{“exp"} \mid k_i, n_i\right)}{P\left(Z_i \neq \text{“exp"} \mid k_i, n_i\right)} = \frac{P\left(Z_i = \text{“exp"} \mid k_i, n_i\right)}{1 - P\left(Z_i = \text{“exp"} \mid k_i, n_i\right)}. \tag{5.45}$$

If the posterior odds for position i are larger than one, it implies therefore that

$$\hat{P}\left(Z_i = \text{“exp"} \mid k_i, n_i\right) > 0.5\,. \tag{5.46}$$

Hence, if $PoO_i > 1$ it is more likely that the T-to-C substitutions at position i are experimentally induced. The threshold of one is thus used in the following for reporting crosslinked T-to-C substitution positions.

5.2 Identification of binding site regions

BayMAP focuses on detecting positions that probably lie on a binding site. However, a binding site is composed of a sequence of genomic positions. For determining the binding sites around the found positions, different approaches have already been established. E.g., in BMix, a cluster is specified around a highly confident position by the sequence covered by all reads that include this position. Overlapping clusters are combined to a binding site (see Section 4.3.2). wavClusteR does not take into account the observed reads directly but rather addresses their coverage. They present two dif-

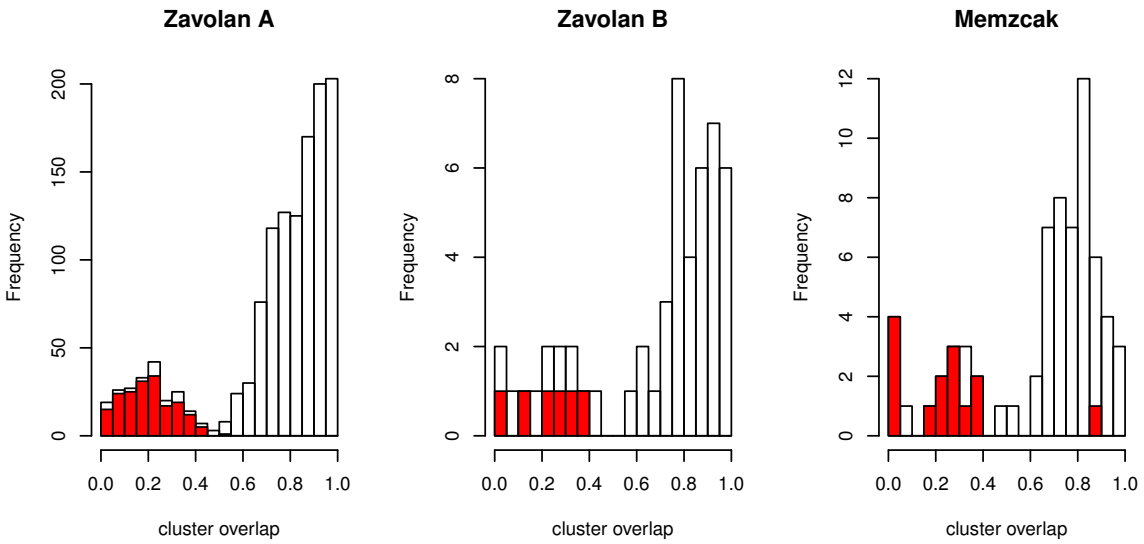

Figure 5.1: Histogram of cluster overlaps that are greater than zero and smaller than one for three different data sets. The observation is marked as red when just the cluster ends are overlapping but not the T-to-C substititution positions.

ferent peak finding methods, where a peak in the coverage function is assumed to be the result of a binding site (see Section 4.2.2).

All highly confident T-to-C substitution positions are assumed to be crosslinked positions and therefore positions on binding sites. It is therefore quite natural to look at all the reads that cover a highly confident position and to combine them to a binding site as done in BMix. When there are two overlapping clusters, it is either possible that these observations come from just one binding site or from more than one. In BMix overlapping potential binding sites are just declared as one binding site. In this thesis, a method is proposed for looking in more detail whether one or several binding sites should be declared.

In a first step cluster start and end points are defined by the furthest positions of reads that still cover the highly confident position. In a second step, the overlap of two clusters or potential binding sites g and j is calculated as

$$o_{gj} = \frac{\#\text{ overlapping positions between } g \text{ and } j}{\min\big(\text{length}(g), \text{length}(j)\big)}\,.$$

If $o_{gj} = 0$, the two clusters are not overlapping and therefore supposed to be two independent potential binding sites. If $o_{gj} = 1$, one of the two clusters is completely em-

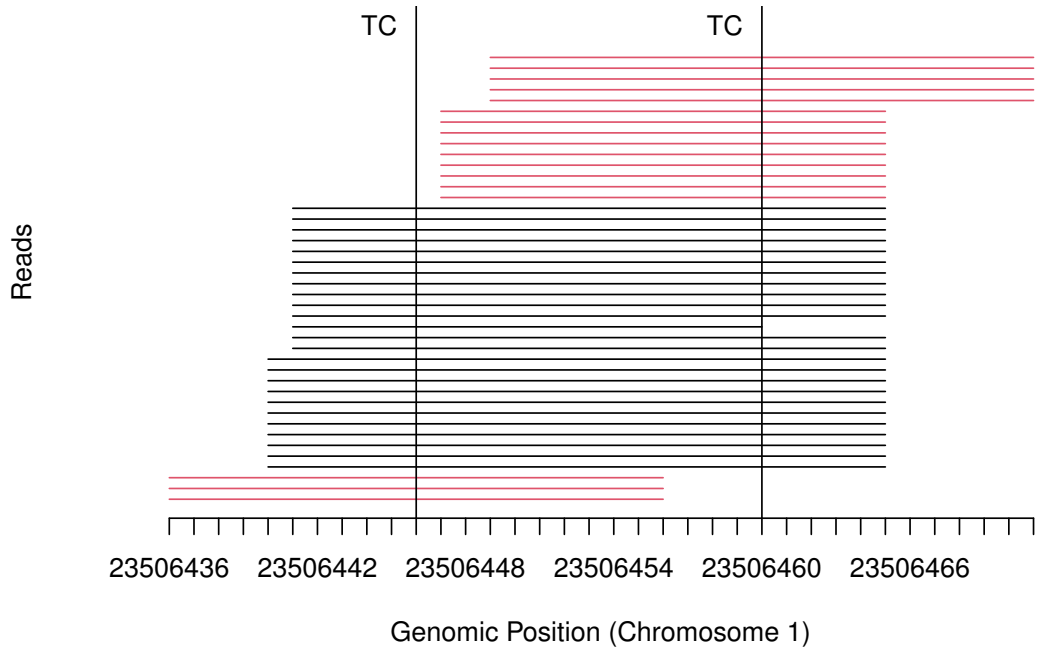

Figure 5.2: Observed reads for two T-to-C substitution positions on Chromosome 1 in the Memczak data set. Reads are marked in red if they only contain one of the two considered T-to-C substitution positions.

bedded in the other one. These potential binding sites are then combined to only one. The question remains how to proceed with two clusters g and j, where $0 < o_{gj} < 1$.

In Figure 5.1 the histograms of all overlaps, where $0 < o_{gj} < 1$, are represented. When having a closer look to these overlaps in the three different data sets from Kishore and Memczak, there seems to be a gap around an overlap of 0.5. The red histograms represent the subset of overlaps, where the T-to-C substitution position of the neighbor cluster is not part of the overlap. For these subsets, there seems to be a gap, too. When the overlap is greater than 0.5, the overlap of two neighbors is especially likely to include both T-to-C substitution positions. Since it is found, that only 1.5% of the clusters of all T-to-C substitution positions in the Kishore A data set have an overlap with a neighbor cluster with $0 < o_{gj} < 1$, it seems reasonable to apply a simple method in order to combine two clusters to one potential binding site or not. In the other data sets this percentage is even smaller (0.6% in Kishore B as well as in Memczak, 0.0% in Gottwein A and Gottwein B). In this thesis, it is therefore proposed to combine two

clusters to one potential binding site, when the cluster overlap is greater than 0.5.

In Figure 5.2 the reads for two T-to-C substitution positions are plotted. In a first step two clusters would be built around the two T-to-C substitution positions. In the first cluster only the reads that contain the first T-to-C substitution position are considered and in the second only the reads that contain the second T-to-C substitution position. In this case the two clusters overlap, as the reads of the two clusters are overlapping. However, the clusters are not the same, as there are reads that only contain one of the two T-to-C substitution positions. Here, the clusters even overlap over both T-to-C substitution positions, as there are reads that contain both T-to-C substitution positions. The first cluster starts at position 23506436 and ends at position 23506465. The second cluster starts at position 23506440 and ends at position 23506471. The number of overlapping positions is therefore 26 and the length of the smaller cluster is 30 so that the overlap in this case is $26/30 = 0.867 > 0.5$. The two clusters would, hence, be combined to one potential binding site.

Another example of overlapping clusters is given in Figure 5.3. There are three T-to-C substitution positions and it seems that the two positions towards the right belong together. The cluster of the middle T-to-C substitution position indeed has identical start and end positions as the cluster around the T-to-C substitution position at the right. The two T-to-C substitution positions represent hence only one potential binding site. The cluster on the left on the other hand only has few overlapping positions to the cluster at the right, that is 25%. Thus, these clusters are not combined to one potential binding site in contrast to the method proposed in BMix.

5.3 BayMAP 2.0: Detection of PAR-CLIP induced T-to-C substitutions using read cluster

In BMix and wavClusteR, first high-confident T-to-C substitution positions are identified. Then, in a second step, binding sites are built around these highly confident positions. BayMAP can also be used to identify binding site regions after the detection of binding site positions. However, it can also be of interest to first define potential

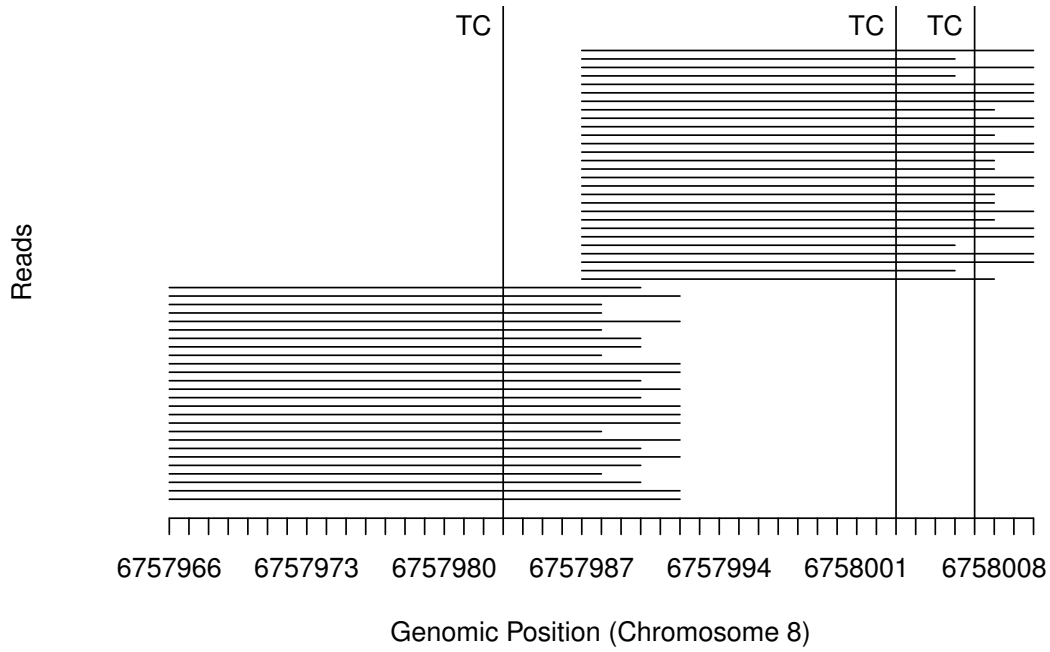

Figure 5.3: Observed reads for three T-to-C substitution positions on Chromosome 8 in the Memczak data set.

binding sites and to then use this information for the detection of binding site positions or binding site regions. For this purpose, BayMAP 2.0 is presented in this section, where potential binding sites are initially built around all T-to-C substitution positions, so that region information can be added to the model.

Section 5.3 is divided into the presentation of the model of BayMAP 2.0 (Section 5.3.1), the derivation of the full conditional distributions (Section 5.3.2), the sampling scheme (Section 5.3.3) and the determination of method-induced substitution positions by estimating the probability of the position being crosslinked given the data (Section 5.3.4).

5.3.1 The model

As discussed in the previous section, a binding site is composed of a sequence of genomic positions and can have several T-to-C substitution positions. Corcoran et al. [12] explain, that in PAR-CLIP experiments with Ago, the T-to-C substitution rate is the highest for T positions directly upstream, i.e. in the direction of the 5' end of the mRNA,

of the seed match region, but that the experimentally induced T-to-C substitutions also occur on the seed match region itself and on positions downstream of the seed match region. If T-to-C substitutions on a binding site are experimentally induced, it is thus likely that T-to-C substitutions at another position on the same binding site are also experimentally induced. T-to-C substitutions on the same potential binding site are ,therefore, not independent of each other. The data of neighbor T-to-C substitution positions could hence help for the detection of crosslinked positions. Here, it is proposed that these dependencies are modeled by adding a random effect to the generalized linear model in (5.7):

$$g\left(p_i\right) = \boldsymbol{x_i}^\top \boldsymbol{\beta} + \alpha_i \ , \ i = 1, \ \ldots, \ N_{\mathrm{TC}} \ . \tag{5.47}$$

This random effect could be different for every position i as is the case for conditional autoregressive (CAR) models. In CAR models, the random effects α_i are correlated to random effects of neighboring positions. In the simplest CAR model, i.e. the intrinsic CAR model, α_i follows a normal distribution with the mean of all neighboring random effects as mean parameter [35]:

$$\alpha_i \ \Big| \ \boldsymbol{\alpha}_{-i}, W, \tau^2 \sim N\left(\overline{\boldsymbol{\alpha}}_{-i}, \frac{\tau^2}{N_i}\right) , \tag{5.48}$$

where $\boldsymbol{\alpha}_{-i}$ is the vector of length N_i with all $\alpha_s, s \neq i$ that are in the neighborhood of α_i with its mean $\overline{\boldsymbol{\alpha}}_{-i}$. W is the neighborhood matrix with $w_{is} = 1$ if positions i and s are defined to be neighbors and $w_{is} = 0$ otherwise and τ^2 is a variance parameter. One has to decide which positions are neighbors and which are not. Are for example all positions in one potential binding site considered as neighbors or only those positions with no other T-to-C substitution points directly in between them. This model requires, however, a density estimation for every T-to-C substitution position i, $i = 1, \ldots, N_{\mathrm{TC}}$. These α_i are additionally highly correlated for those α_i that are on the same potential binding site. This correlation therefore causes autocorrelation for the MCMC chain of one α_i, so that a high amount of iterations is needed to achieve convergence.

An easier way to deal with the dependence between several substitution positions is to use normal random effects, where a random effect for every potential binding site

$j = 1,\ldots,N_{\text{cluster}}$ is added to (5.7), so that

$$g\left(p_{ij}\right) = \boldsymbol{x}_{ij}^{\top}\boldsymbol{\beta} + \alpha_j\,, \tag{5.49}$$

where $i = 1,\ldots,N_j$ with $N_j \in \mathbb{N}^+$ the number of T-to-C substitution positions on the potential binding site j. The model is in that way easier as there is only one additional parameter α_j for every potential binding site that has to be sampled, but not an α_i for every single T-to-C substitution position, so that, e.g., convergence can be achieved with a smaller amount of iterations.

As the covariates considered here, e.g., the 3' UTR, have the same values for the same potential binding site j, Equation (5.49) can even be more simplified by

$$g\left(p_j\right) = \boldsymbol{x}_j^{\top}\boldsymbol{\beta} + \alpha_j = \zeta_j. \tag{5.50}$$

p_j is, hence, the same for every position i that lies on the potential binding site j. In Bayesian statistics the random effect can be included by adding a new level to the generalized linear model [18] . Here, the new level is added by ζ_j, that is supposed to follow a normal distribution depending on the hyperparameters $\boldsymbol{\beta}$ and τ^2:

$$\zeta_j \mid \boldsymbol{x}_j^{\top}\boldsymbol{\beta}, \tau^2 \sim N\left(\boldsymbol{x_j}^{\top}\boldsymbol{\beta}, \tau^2\right). \tag{5.51}$$

In order to avoid autocorrelation, the latter method is the method of choice in BayMAP 2.0. However, application results for the intrinsic CAR model will also be presented to demonstrate that results are not appropriate.

There is, therefore, one additional parameter for which a prior distribution is needed, that is $\tau^2 > 0$. Here, a uniform prior for τ is chosen, as it leads to a conjugate form. A uniform prior for τ is equivalent to $f\left(\tau^2\right) \propto \tau^{-1}$ [18] as can be seen by transformation of the density. The joint posterior distribution as well as full conditional distributions for $\boldsymbol{\zeta}$, $\boldsymbol{\beta}$ and τ^2, that are used in BayMAP 2.0 for generating the MCMC chain, can be found in the next section.

5.3.2 Full conditional distributions

The model for BayMAP 2.0 has two parameters more than the model for BayMAP 1.0, that is $\boldsymbol{\zeta} = \left(\zeta_1 \ \ldots \ \zeta_{N_{\text{cluster}}}\right)^\top$ and τ^2, so that a new level for the random effect is added. The joint posterior distribution (see also (5.10)) then becomes

$$\begin{aligned} &f\left(\boldsymbol{\mu}, \boldsymbol{z}, q, \boldsymbol{\zeta}, \tau^2, \boldsymbol{\beta} \mid \boldsymbol{k}, \boldsymbol{n}, \boldsymbol{X}\right) \\ &\propto f\left(\boldsymbol{k} \mid \boldsymbol{\mu}, \boldsymbol{n}, \boldsymbol{z}\right) f\left(\boldsymbol{\mu}\right) f\left(\boldsymbol{z} \mid \boldsymbol{p} = g^{-1}\left(\boldsymbol{\zeta}\right), q\right) f\left(\boldsymbol{\zeta} \mid \boldsymbol{\beta}, \boldsymbol{X}, \tau^2\right) f\left(q\right) f\left(\boldsymbol{\beta}\right) f\left(\tau^2\right). \end{aligned} \tag{5.52}$$

Note that the covariates are here assumed to have the same values for the same potential binding site j, so that the matrix $\boldsymbol{X}$ of covariates reduces here to $\boldsymbol{X} := \left(\boldsymbol{x_1} \ \ldots \ \boldsymbol{x_{N_{\text{cluster}}}}\right)^\top \in \mathbb{R}^{N_{\text{cluster}} \times (L+1)}$. The full conditional distributions for $\boldsymbol{\mu}$, $\boldsymbol{z}$ and q that are derived in Section 5.1.2 do not change. Only $\boldsymbol{p} = g^{-1}\left(\boldsymbol{X\beta}\right)$ has to be replaced by $\boldsymbol{p} = g^{-1}\left(\boldsymbol{\zeta}\right)$. In this section, hence, only the densities for the full conditional distributions for $\boldsymbol{\zeta}$, $\boldsymbol{\beta}$ and τ^2 are derived. For the other full conditional distributions see Section 5.1.2.

Full conditional distribution of $\boldsymbol{\zeta}$

In a similar way to Section 5.1.2, the latent random variable $\boldsymbol{Y}$ can be introduced to sample the parameter vector $\boldsymbol{\zeta}$. This means that it can be alternated between sampling $\boldsymbol{\zeta}$ depending on $\boldsymbol{y}$, and sampling $\boldsymbol{y}$ conditionally on $\boldsymbol{\zeta}$ [17].

As stated in (5.51), ζ_j given both $\boldsymbol{x_j}^\top\boldsymbol{\beta}$ and τ^2 follows a normal distribution. The parameter ζ_j, $j = 1, \ldots, N_{\text{cluster}}$ determines the probability for position i on the potential binding site j with $p_{ij} = p_j = \Phi\left(\zeta_j\right)$, where $i = 1, \ldots, N_j$ with N_j the number of substitution positions on binding site j. As not every T position on a binding site has to have crosslinked substitutions, determination if the substitutions are crosslinked are drawn for each position separately. Hence, the latent variable $\boldsymbol{y}$ also has to be sampled for every T-to-C substitution position on a potential binding site j. In the same way as in (5.27), the random variable Y_{ij} belonging to the vector $\boldsymbol{Y}$ follows a normal distribution

with here

$$\left(Y_{ij} \mid \zeta_j\right) \sim N\left(\zeta_j, 1\right) \text{ and} \tag{5.53}$$

$$Z_{ij3} = \mathbb{1}_{(0,\infty)}(Y_{ij}). \tag{5.54}$$

Again, Y_{ij} given z_{ij} and ζ_j can thus be sampled from a truncated normal distribution. The information Y_{ij} for each cluster j can be combined to the cluster means

$$\overline{Y}_{.j} := \frac{1}{N_j} \sum_{i=1}^{N_j} Y_{ij} \text{ , where}$$

$$\overline{Y}_{.j} \sim N\left(\zeta_j, \frac{1}{N_j}\right), \tag{5.55}$$

so that $\overline{Y}_{.j}$ is independently, but not identical, distributed. The distribution of the full conditional distribution of $\boldsymbol{\zeta}$ can be written as

$$\begin{aligned} f\left(\boldsymbol{\zeta} \mid \boldsymbol{X\beta}, \tau^2, \boldsymbol{y}\right) &= \prod_{j=1}^{N_{\text{cluster}}} f\left(\zeta_j \mid \boldsymbol{x_j}^\top \boldsymbol{\beta}, \tau^2, \bar{y}_{.j}\right) \\ &\propto \prod_{j=1}^{N_{\text{cluster}}} f\left(\bar{y}_{.j} \mid \zeta_j\right) f\left(\zeta_j \mid \boldsymbol{x_j}^\top \boldsymbol{\beta}, \tau^2\right). \end{aligned} \tag{5.56}$$

Both terms in the last line of (5.56) are densities of the normal distribution (see (5.55) and (5.51)), where the parameter of interest, i.e. ζ_j, is only present inside the exponential function, so that the other factors can be omitted without loss of information. The product of the two terms in (5.56) is therefore

$$\begin{aligned} f\left(\bar{y}_{.j} \mid \zeta_j\right) f\left(\zeta_j \mid \boldsymbol{x_j}^\top \boldsymbol{\beta}, \tau^2\right) &\propto \exp\left(-\frac{1}{2}\left(\frac{\left(\bar{y}_{.j} - \zeta_j\right)^2}{\frac{1}{N_j}} + \frac{\left(\zeta_j - \boldsymbol{x_j}^\top \boldsymbol{\beta}\right)^2}{\tau^2}\right)\right) \\ &= \exp\left(-\frac{1}{2}\left(N_j \bar{y}_{.j}^2 - 2\zeta_j N_j \bar{y}_{.j} + N_j \zeta_j^2 + \frac{\zeta_j^2}{\tau^2} - 2\zeta_j \frac{\boldsymbol{x_j}^\top \boldsymbol{\beta}}{\tau^2} + \frac{(\boldsymbol{x_j}^\top \boldsymbol{\beta})^2}{\tau^2}\right)\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\zeta_j^2\left(N_j + \frac{1}{\tau^2}\right) - 2\zeta_j\left(N_j \bar{y}_{.j} + \frac{1}{\tau^2}\boldsymbol{x_j}^\top \boldsymbol{\beta}\right)\right)\right) \\ &= \exp\left(-\frac{1}{2}\left(\frac{\zeta_j^2 - 2\zeta_j\left(\frac{N_j \bar{y}_{.j} + \frac{1}{\tau^2}\boldsymbol{x_j}^\top \boldsymbol{\beta}}{N_j + \frac{1}{\tau^2}}\right)}{\frac{1}{N_j + \frac{1}{\tau^2}}}\right)\right). \end{aligned} \tag{5.57}$$

The proportionality of the third line in (5.57) holds, since the factor $\exp\left(-\frac{1}{2}N_j\bar{y}_{.j}^2\right)$ does not depend on ζ_j. The big nominator in the last line only misses an summand for application of the binomial theorem. This missing summand does also not depend on ζ_j. To the aim of resembling a normal distribution, it can thus be added via multiplication by

$$\begin{aligned} f\left(\bar{y}_{.j} \mid \zeta_j\right) f\left(\zeta_j \mid \boldsymbol{x_j}^\top\boldsymbol{\beta}, \tau^2\right) &\propto \exp\left(-\frac{1}{2}\left(\frac{\zeta_j^2 - 2\zeta_j\left(\frac{N_j\bar{y}_{.j}+\frac{1}{\tau^2}\boldsymbol{x_j}^\top\boldsymbol{\beta}}{N_j+\frac{1}{\tau^2}}\right)}{\frac{1}{N_j+\frac{1}{\tau^2}}}\right)\right)\exp\left(-\frac{1}{2}\frac{\left(\frac{N_j\bar{y}_{.j}+\frac{1}{\tau^2}\boldsymbol{x_j}^\top\boldsymbol{\beta}}{N_j+\frac{1}{\tau^2}}\right)^2}{\frac{1}{N_j+\frac{1}{\tau^2}}}\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\frac{\left(\zeta_j - \frac{N_j\bar{y}_{.j}+\frac{1}{\tau^2}\boldsymbol{x_j}^\top\boldsymbol{\beta}}{N_j+\frac{1}{\tau^2}}\right)^2}{\frac{1}{N_j+\frac{1}{\tau^2}}}\right)\right) \\ &\propto N\left(\zeta_j \,\middle|\, \frac{N_j\bar{y}_{.j}+\frac{1}{\tau^2}\boldsymbol{x_j}^\top\boldsymbol{\beta}}{N_j+\frac{1}{\tau^2}}, \frac{1}{N_j+\frac{1}{\tau^2}}\right). \end{aligned} \tag{5.58}$$

The parameter ζ_j thus follows, conditional on the data and the other parameters, a normal distribution. Hence, it can be sampled by using the Gibbs sampler.

Full conditional distribution of $\boldsymbol{\beta}$

The density of the full conditional distribution of $\boldsymbol{\beta}$ is as follows

$$\begin{aligned} f\left(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{\zeta}, \tau^2, \boldsymbol{X}\right) &\propto f\left(\boldsymbol{y} \mid \boldsymbol{\zeta}\right) f\left(\boldsymbol{\zeta} \mid \boldsymbol{X}, \boldsymbol{\beta}, \tau^2\right) f\left(\boldsymbol{\beta} \mid \tau^2\right) \\ &\propto f\left(\boldsymbol{\zeta} \mid \boldsymbol{X}, \boldsymbol{\beta}, \tau^2\right) f\left(\boldsymbol{\beta}\right), \end{aligned} \tag{5.59}$$

since $\boldsymbol{Y}$ does not depend directly on $\boldsymbol{\beta}$, but on $\boldsymbol{\zeta}$ and the regression parameter $\boldsymbol{\beta}$ is here supposed to be independent of τ^2. With the prior distribution of (5.9) and $\sigma^2 = 10^{-6}$, $f\left(\boldsymbol{\beta}\right)$ is then

$$f\left(\boldsymbol{\beta}\right) = N\left(\boldsymbol{\beta} \mid 0, \sigma^2\boldsymbol{I}_{L+1}\right) \propto \exp\left(-\frac{\boldsymbol{\beta}^\top\boldsymbol{\beta}}{2\sigma^2}\right). \tag{5.60}$$

As ζ_j given $\boldsymbol{x}_j^\top\boldsymbol{\beta}$ and τ^2 follows a normal distribution (see 5.51), the distribution of $\boldsymbol{\zeta}$

can be written as a multivariate normal distribution with density

$$f\left(\boldsymbol{\zeta} \mid \boldsymbol{X}, \boldsymbol{\beta}, \tau^2\right) = N\left(\boldsymbol{\zeta} \mid \boldsymbol{X}\boldsymbol{\beta}, \tau^2 \boldsymbol{I}_{N_{\text{cluster}}}\right) \propto \exp\left(-\frac{\left(\boldsymbol{\zeta} - \boldsymbol{X}\boldsymbol{\beta}\right)^\top \left(\boldsymbol{\zeta} - \boldsymbol{X}\boldsymbol{\beta}\right)}{2\tau^2}\right) \tag{5.61}$$

With (5.59), the densities (5.60) and (5.61) can be combined to

$$f\left(\boldsymbol{\beta} \mid \boldsymbol{\zeta}, \tau^2, \boldsymbol{X}\right) \propto \exp\left(-\frac{1}{2}\left(\frac{1}{\tau^2}\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} - \frac{2}{\tau^2}\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{\zeta} + \frac{1}{\tau^2}\boldsymbol{\zeta}^\top \boldsymbol{\zeta} + \frac{1}{\sigma^2}\boldsymbol{\beta}^\top \boldsymbol{\beta}\right)\right), \tag{5.62}$$

since $\boldsymbol{\beta}$ is only present inside the exponential function of the two normal distributions. In the same way as in (5.35), the summands including $\boldsymbol{\beta}$ can be written as a quadratic form, when adding a term not dependent on $\boldsymbol{\beta}$, so that

$$\begin{aligned}
f\left(\boldsymbol{\beta} \mid \boldsymbol{\zeta}, \tau^2, \boldsymbol{X}\right) = \exp\Bigg(-\frac{1}{2}\Bigg(&\left(\boldsymbol{\beta} - \left(\frac{1}{\tau^2}\boldsymbol{X}^\top \boldsymbol{X} + \frac{1}{\sigma^2}\boldsymbol{I}_{L+1}\right)^{-1} \frac{1}{\tau^2}\boldsymbol{X}^\top \boldsymbol{\zeta}\right)^\top \left(\frac{1}{\tau^2}\boldsymbol{X}^\top \boldsymbol{X} + \frac{1}{\sigma^2}\boldsymbol{I}_{L+1}\right) \\
&\cdot \left(\boldsymbol{\beta} - \left(\frac{1}{\tau^2}\boldsymbol{X}^\top \boldsymbol{X} + \frac{1}{\sigma^2}\boldsymbol{I}_{L+1}\right)^{-1} \frac{1}{\tau^2}\boldsymbol{X}^\top \boldsymbol{\zeta}\right) \\
&+ \frac{1}{\tau^2}\boldsymbol{\zeta}^\top \boldsymbol{\zeta} - \left(\frac{1}{\tau^2}\right)^2 \boldsymbol{\zeta}^\top \boldsymbol{X}^\top \left(\frac{1}{\tau^2}\boldsymbol{X}^\top \boldsymbol{X} + \frac{1}{\sigma^2}\boldsymbol{I}_{L+1}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{\zeta}\Bigg)\Bigg) \\
\propto \exp\Bigg(-\frac{1}{2}\Bigg(&\left(\boldsymbol{\beta} - \left(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\tau^2}{\sigma^2}\boldsymbol{I}_{L+1}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{\zeta}\right)^\top \left(\frac{1}{\tau^2}\boldsymbol{X}^\top \boldsymbol{X} + \frac{1}{\sigma^2}\boldsymbol{I}_{L+1}\right) \\
&\left(\boldsymbol{\beta} - \left(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\tau^2}{\sigma^2}\boldsymbol{I}_{L+1}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{\zeta}\right)\Bigg)\Bigg) \\
\propto N\Bigg(\boldsymbol{\beta} \Bigg|\ &\left(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\tau^2}{\sigma^2}\boldsymbol{I}_{L+1}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{\zeta}, \tau^2 \left(\boldsymbol{X}^\top \boldsymbol{X} + \frac{\tau^2}{\sigma^2}\boldsymbol{I}_{L+1}\right)^{-1}\Bigg).
\end{aligned} \tag{5.63}$$

The full conditional distribution is thus also a multivariate normal distribution, and the parameter $\boldsymbol{\beta}$ can therefore be sampled by a Gibbs sampler.

Full conditional distribution of τ^2

The last parameter is $\tau^2 > 0$, that is the variance of the random effect, and therefore of $\boldsymbol{\zeta}$ given $\boldsymbol{X\beta}$. The density of the full conditional distribution is proportional to

$$\begin{aligned} f\left(\tau^2 \mid \boldsymbol{y}, \boldsymbol{\zeta}, \boldsymbol{X}, \boldsymbol{\beta}\right) &\propto f\left(\boldsymbol{y} \mid \boldsymbol{\zeta}\right) f\left(\boldsymbol{\zeta} \mid \boldsymbol{X}, \boldsymbol{\beta}, \tau^2\right) f\left(\boldsymbol{\beta}\right) f\left(\tau^2\right) \\ &\propto f\left(\boldsymbol{\zeta} \mid \boldsymbol{X}, \boldsymbol{\beta}, \tau^2\right) f\left(\tau^2\right) . \end{aligned} \tag{5.64}$$

To the aim of getting the full conditional density of τ^2, the density of $\boldsymbol{\zeta}$ given $\boldsymbol{X\beta}$ and the prior density of τ^2 can be multiplied. As explained in the last passage of Section 5.3.1, the prior for τ^2 is here chosen to be

$$f\left(\tau^2\right) \propto \tau^{-1} = \left(\tau^2\right)^{-\frac{1}{2}} . \tag{5.65}$$

For the density of $\boldsymbol{\zeta}$ in (5.64), the factor of the normal distribution that is multiplied by the exponential function has now to be taken into account, as it includes τ^2. The two distributions can hence be combined to

$$\begin{aligned} f\left(\tau^2 \mid \boldsymbol{\zeta}, \boldsymbol{X}, \boldsymbol{\beta}\right) &\propto \left(\tau^2\right)^{-\frac{J}{2}} \exp\left(-\frac{1}{2\tau^2}\left(\boldsymbol{\zeta} - \boldsymbol{X\beta}\right)^\top \left(\boldsymbol{\zeta} - \boldsymbol{X\beta}\right)\right) \left(\tau^2\right)^{-\frac{1}{2}} \\ &= \left(\tau^2\right)^{-\frac{J+1}{2}} \exp\left(-\frac{1}{2\tau^2} \sum_{j=1}^{J} \left(\zeta_j - \boldsymbol{x}_j^\top \boldsymbol{\beta}\right)^2\right) \\ &= \left(\tau^2\right)^{-\left(\frac{J-1}{2}+1\right)} \exp\left(-\frac{J-1}{2\tau^2} \frac{1}{J-1} \sum_{j=1}^{J} \left(\zeta_j - \boldsymbol{x}_j^\top \boldsymbol{\beta}\right)^2\right) \\ &\propto \text{Inv-}\chi^2\left(\tau^2 \;\middle|\; J-1, \frac{1}{J-1} \sum_{j=1}^{J} \left(\zeta_j - \boldsymbol{x}_j^\top \boldsymbol{\beta}\right)^2\right) \end{aligned} \tag{5.66}$$

with $J := N_{\text{cluster}}$ for a better representation and $\text{Inv-}\chi^2\left(\tau^2 \mid \cdot, \cdot\right)$ the density of the scaled inverse chi-squared distribution. The conditional posterior distribution is therefore a scaled inverse chi-squared distribution with $J-1$ degrees of freedom and parameter

$$s^2 = \frac{1}{J-1} \sum_{j=1}^{J} \left(\zeta_j - \boldsymbol{x}_j^\top \boldsymbol{\beta}\right)^2 ,$$

so that τ^2 can be sampled using a Gibbs sampler.

5.3.3 Sampling scheme

In the same way as in BayMAP 1.0, first the parameters have to be initialized for the Markov chain (see Section 5.1.3). The initialization is here done by:

1. Draw $\boldsymbol{\beta}^{(0)}$ from $N(\mathbf{0}_{L+1}, \boldsymbol{I}_{L+1})$,

   where $\mathbf{0}_{L+1}$ is the null vector with length $(L+1)$.

2. Set $\boldsymbol{\zeta}^{(0)} = \boldsymbol{X}\boldsymbol{\beta}^{(\mathbf{0})}$.

3. Set $\tau^{2^{(0)}} = 1$.

4. Calculate $\boldsymbol{p}^{(0)} = \Phi\left(\boldsymbol{\zeta}^{(0)}\right)$ and

   set those entries of $\boldsymbol{p}^{(0)}$ to zero, where the substitution type is not T-to-C.

5. Draw $q^{(0)}$ from $U(0,1)$.

6. Draw $z_i^{(0)}$ for $i = 1, \ldots, N$ from $\text{Cat}\left(\left(1-p_i^{(0)}\right)q^{(0)}, \left(1-p_i^{(0)}\right)\left(1-q^{(0)}\right), p_i^{(0)}\right)$.

7. Set $\boldsymbol{\mu}^{(0)} = (0.05, 0.85, 0.5)$.

With this initialization, the sampling scheme for $h = 1, 2, \ldots$ is then:

1. Determine $\boldsymbol{\mu}^{(h)}$, $\boldsymbol{z}^{(h)}$ and $q^{(h)}$ as specified in the sampling scheme items (1) to (4) in Section 5.1.3.

2. Draw $y_{ij}^{(h)}$ for $j = 1, \ldots, N_{\text{cluster}}$ and $i = 1, \ldots, N_j$ from the truncated normal distribution with density

   $$\begin{cases} f\left(y_{ij}^{(h)} \mid \zeta_j^{(h-1)}, 1, Y_{ij}^{(h)} > 0\right), & \text{if } z_{ij}^{(h)} = \text{“exp”} \\ f\left(y_{ij}^{(h)} \mid \zeta_j^{(h-1)}, 1, Y_{ij}^{(h)} \leq 0\right), & \text{if } z_{ij}^{(h)} \neq \text{“exp”} \end{cases},$$

   where $\zeta_j^{(h-1)}$ is the mean parameter of the normal distribution and one the variance parameter.

3. Draw $\zeta_j^{(h)}$ for $j = 1, \ldots, N_{\text{cluster}}$ from

   $$N\left(\frac{N_j \bar{y}_{\cdot j}^{(h)} + \frac{1}{\tau^{2^{(h-1)}}} \boldsymbol{x_j}^\top \boldsymbol{\beta}^{(h-1)}}{N_j + \frac{1}{\tau^{2^{(h-1)}}}}, \frac{1}{N_j + \frac{1}{\tau^{2^{(h-1)}}}}\right) \text{ (see (5.58))}.$$

4. Calculate $\boldsymbol{p}^{(h)} = \Phi\left(\boldsymbol{\zeta}^{(h)}\right)$ and

   set those entries of $\boldsymbol{p}^{(h)}$ to zero, where the substitution type is not T-to-C.

5. Draw $\boldsymbol{\beta}^{(h)}$ from $N\left(\left(\boldsymbol{X}^\top\boldsymbol{X} + \frac{\tau^{2^{(h-1)}}}{\sigma^2}\boldsymbol{I}_{L+1}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{\zeta}^{(h)}, \tau^{2^{(h-1)}}\left(\boldsymbol{X}^\top\boldsymbol{X} + \frac{\tau^{2^{(h-1)}}}{\sigma^2}\boldsymbol{I}_{L+1}\right)^{-1}\right)$ (see (5.37)).

6. Draw $\tau^{2^{(h)}}$ from Inv-$\chi^2\left(J-1, \frac{1}{J-1}\sum_{j=1}^{J}\left(\zeta_j^{(h)} - \boldsymbol{x_j}^\top\boldsymbol{\beta}^{(h)}\right)^2\right)$ (see (5.66)).

5.3.4 Identification of method-induced substitution positions

For the identification of method-induced substitution positions, the calculation of the prior odds PrO_j for potential binding sites $j = 1, \ldots, N_{\text{cluster}}$ has to be slightly changed. In (5.43) one has to replace $\boldsymbol{x_i}^\top\boldsymbol{\beta}$ and $\boldsymbol{\beta}$ by ζ_j

$$PrO_j = \frac{P\left(Z_j = \text{“exp"}\right)}{P\left(Z_j \neq \text{“exp"}\right)} = \frac{\int_{\mathbb{R}} P\left(Z_j = \text{“exp"} \mid p_j = g^{-1}\left(\zeta_j\right)\right) \cdot f\left(\zeta_j\right) d\zeta_j}{\int_{\mathbb{R}} P\left(Z_j \neq \text{“exp"} \mid p_j = g^{-1}\left(\zeta_j\right)\right) \cdot f\left(\zeta_j\right) d\zeta_j}, \tag{5.67}$$

so that for Equation (5.44) $p_j^{(h)}$ is equal to $\Phi\left(\zeta_j^{(h)}\right)$ instead of $\Phi\left(\boldsymbol{x}_j^\top\boldsymbol{\beta}^{(h)}\right)$.

Up to now, the prior odds were used to calculate the posterior odds for position i. This is still possible by multiplying the Bayes factor for position i and the prior odds for the potential binding site j that is associated to position i.

However, here, it could also be of interest to account directly for the whole binding site instead of one position. This could be done by only considering the prior odds for binding site j without the Bayes factor for position i. This is possible, as Z_i is drawn depending on the data of k_i and n_i, and ζ_j depends indirectly on all realizations of Z_i for i belonging to binding site j. Consequently, the prior odds for position j depend on the data available for the potential binding site j.

5.4 Combining several PAR-CLIP data sets

PAR-CLIP experiments as well as the experimental validation of binding sites are labor and cost intensive. Therefore, it would be desirable to combine data sets of sev-

eral experiments. Furthermore, one would expect to receive more reliable results, if those were based on more experimental replicates, rather than on one single data set. For instance, some PAR-CLIP experiment are replicated under the same experimental conditions and with cells from the same cell line.

The data sets considered in this thesis from Kishore et al. [30] and Gottwein et al. [22] are each replicated once. In contrast, no replicate exists for the data set from Memczak et al. [41]. Note that one has to be careful when comparing different PAR-CLIP experiments that are not replicates, as the binding sites could then differ among experiments. Daschkey et al. [13] for example showed that the different Ago proteins can bind to different miRNAs and mRNAs.

The methods for analyzing PAR-CLIP data presented in Sections 1 and 4 as well as BayMAP 1.0 (Section 5.1) and BayMAP 2.0 (Section 5.3) are only focusing on the information of one data set. Up to now, if the results of several PAR-CLIP data sets should be compared, the overlap of discovered binding site regions has for example been regarded (see e.g., [22]). However, it can be the case, that some binding sites in one PAR-CLIP experiment are not detected as binding sites just because of a too small read depth, whereas they are detected in a second PAR-CLIP experiment because of the higher read depth. Those binding sites would not be covered by the overlap, since they are only detected in one out of two data sets.

First, it is of interest to combine the information of the experiments, since all data sets contain valuable information if the position is a binding site position or not. Second, it is hence also of interest to use a more elaborate method than the overlap of binding site regions. When using BayMAP, the information of as many data sets as possible can be easily combined by estimating the posterior odds.

Let D be the number of independent PAR-CLIP data sets, e.g., created under the same experimental conditions and k_{id} and n_{id} the number of substituted and the total number of reads for position i, $i = 1,\ldots,N_{\mathrm{TC}}$, in data set d, $d = 1,\ldots,D$. The posterior odds

for position i given the data of all D data sets can then be written as

$$\begin{aligned} PoO_i &= \frac{P\left(Z_i = \text{“exp"} \mid k_{i1}, n_{i1}, \ldots, k_{iD}, n_{iD}\right)}{P\left(Z_i \neq \text{“exp"} \mid k_{i1}, n_{i1}, \ldots, k_{iD}, n_{iD}\right)} \\ &= \frac{f\left(k_{i1}, \ldots, k_{iD} \mid Z_i = \text{“exp"}, n_{i1}, \ldots, n_{iD}\right) \cdot P\left(Z_i = \text{“exp"}\right)}{f\left(k_{i1}, \ldots, k_{iD} \mid Z_i \neq \text{“exp"}, n_{i1}, \ldots, n_{iD}\right) \cdot P\left(Z_i \neq \text{“exp"}\right)} \\ &= \frac{f\left(k_{i1} \mid Z_i = \text{“exp”}, n_{i1}\right) \cdot \ldots \cdot f\left(k_{iD} \mid Z_i = \text{“exp”}, n_{iD}\right)\ P\left(Z_i = \text{“exp”}\right)}{f\left(k_{i1} \mid Z_i \neq \text{“exp”}, n_{i1}\right) \cdot \ldots \cdot f\left(k_{iD} \mid Z_i \neq \text{“exp”}, n_{iD}\right) \cdot P\left(Z_i \neq \text{“exp”}\right)} \\ &= BF_{i1} \cdot \ldots \cdot BF_{iD} \cdot PrO_i \,. \end{aligned} \tag{5.68}$$

The third line for the equation can be written since independence is assumed. The prior odds PrO_i, when several data sets are present, can be written in the same way as in (5.43) with

$$PrO_i = \frac{P\left(Z_i = \text{“exp"}\right)}{P\left(Z_i \neq \text{“exp"}\right)} = \frac{P\left(Z_i = \text{“exp"}\right)}{1 - P\left(Z_i = \text{“exp"}\right)}\,, \tag{5.69}$$

where $P\left(Z_i = \text{“exp"}\right)$ is equal to

$$\begin{aligned} \int_{\mathbb{R}^{L+1}} \ldots \int_{\mathbb{R}^{L+1}} & P\left(Z_i = \text{“exp"} \mid p_{i1} = g^{-1}(\boldsymbol{x_i}^\top \boldsymbol{\beta}_1), \ldots, p_{iD} = g^{-1}(\boldsymbol{x_i}^\top \boldsymbol{\beta}_D)\right) \\ & \cdot f\left(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_D\right) d(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_D)\,. \end{aligned} \tag{5.70}$$

In (5.44) the prior odds are estimated by Monte Carlo integration. Here, however, it is not possible in the same way, since $P\left(Z_i = \text{“exp"} \mid p_{i1} = g^{-1}\left(\boldsymbol{x_i}^\top \boldsymbol{\beta}_1\right), \ldots, p_{iD} = g^{-1}\left(\boldsymbol{x_i}^\top \boldsymbol{\beta}_D\right)\right)$ is not known. This probability is estimated by its mean, so that

$$\widehat{PrO}_i = \frac{\frac{1}{N_{\text{iter}}} \sum_{h=1}^{N_{\text{iter}}} \frac{1}{D} \sum_{d=1}^{D} p_{id}^{(h)}}{1 - \frac{1}{N_{\text{iter}}} \sum_{h=1}^{N_{iter}} \frac{1}{D} \sum_{d=1}^{D} p_{id}^{(h)}}\,. \tag{5.71}$$

In order to estimate if a position has method-induced substitutions, one can therefore multiply the single Bayes factors times the combined prior odds. Since not all substitution positions are present in all replicates, for every position only the available information.

Chapter 6

Simulation study

In real PAR-CLIP data sets it is not known which positions have method-induced substitutions and which do not. A simulation study is, therefore, conducted to evaluate the performance in detecting method-induced substitution positions in comparison to other methods. Furthermore, this simulation study can be used to verify if the estimated densities reflect the true parameters, i.e. if the parameter estimation is unbiased. This is of special interest for the parameters of the additional variables, as BayMAP can also be used to analyze these effects for binding sites.

A simulation study is performed for BayMAP 1.0 (see Section 6.1), BayMAP 2.0 (see Section 6.2) as well as the method of BayMAP that allows to combine the results of several PAR-CLIP data sets (see Section 6.3).

For each of the methods, i.e. BayMAP 1.0, BayMAP 2.0, wavCluster and BMix, it has to be specified, when a T-to-C substitution position can be declared as crosslinked position. Here, the same probability cutoff of 0.5 is used for all methods [27]. This means that a position is declared as method-induced for a specific method, e.g., wavClusteR, if the probability of being crosslinked is estimated to be larger than 0.5 using this method.

6.1 BayMAP 1.0

For the simulation study of BayMAP 1.0, first the set up of the main simulation study is described (see Section 6.1.1). Then, the bias is analyzed for BayMAP 1.0 in Section

6.1.2 to verify the model. Afterwards, the performance in detecting crosslinked T-to-C substitution positions is evaluated and compared to simpler versions (see Section 6.1.3) of BayMAP 1.0 as well as to BMix and wavClusteR (see Section 6.1.4). Finally, for BayMAP 1.0 an additional way of simulating data is proposed to further validate the simulation results (see Section 6.1.5).

The simulation study and its results for BayMAP 1.0 presented in Section 6.1 are already published in Huessler et al. [27].

6.1.1 Set up of simulation study

The simulated data sets are based on the approach of Golumbeanu et al. [20]. For T-to-C substitution positions it is not known if the substitutions are method-induced or not. For all other substitution types it is supposed that observed substitutions are not linked to the PAR-CLIP method. The idea of Golumbeanu et al. [20] is to take a publicly available data set, to remove all T-to-C substitution positions and to artificially introduce method-induced substitutions for one substitution type. For this substitution type, it is therefore known which substitutions are method-induced. An advantage of this procedure is that a realistic PAR-CLIP data set is used with only very few changes at positions that are chosen to have PAR-CLIP induced substitutions. This simulation method is therefore implemented here in a similar way.

Here, the Kishore A data set is used for the simulation study. As discussed in Section 2.2.2, the data set consists of the number of substitutions, the number of reads, the type of substitutions, e.g., T-to-C, and additional information, i.e. the type of the mRNA region. First, all T-to-C substitution positions are deleted from the data set. A-to-G substitutions are chosen as new potential substitutions for binding sites. This choice for A-to-G is arbitrary, it could also have been all other types of substitution instead of A-to-G.

The A-to-G substitutions should represent the T-to-C substitutions in a normal PAR-CLIP data set. As discussed in Section 2.2.3, T-to-C substitution positions are much more frequent than any other type of substitution. Since A-to-G positions are now the new T-to-C substitutions, they should also occur more often than the others. For this

purpose, only as many of the other substitutions are kept, so that the ratio between their number and the number of A-to-G substitution positions is the same as it was before in comparison to the number of T-to-C substitution positions. In the original data set, the number of all non-T-to-C substitution positions together corresponds to 86.5% of the number of T-to-C substitution positions. This means concretely that only as many non-A-to-G substitution positions are chosen so that the number is equal to 86.5% of the number of A-to-G substitution positions. The deleted positions of the non-A-to-G substitutions are chosen randomly. Since the A-to-G substitutions represent T-to-C substitutions in the simulation, they are called T-to-C substitutions from now on.

Once all positions that should remain in the data set are selected, A-to-G substitution positions that should be declared as method-induced substitution positions have to be chosen. For each A-to-G substitution position it is drawn randomly whether this position should be a method-induced one or not with probability p_i. The parameter p_i is the probability that a position i has method-induced substitutions when the number of substitutions k_i is not given.

The percentage of method-induced T-to-C substitution positions of all T-to-C substitution positions, and therefore an indicator for p_i, can be estimated using the non-T-to-C substitutions. It is assumed that the number of non-method-induced T-to-C substitutions is approximately as high as the total number of substitution positions for another arbitrary substitution type. Consequently, the percentage of non-method-induced T-to-C substitution positions can here be estimated by dividing the total number of substitution positions for one specific substitution, that is not T-to-C, by the total number of T-to-C substitution positions. The percentage of method-induced T-to-C substitution positions can then be estimated by calculating one minus this fraction (see also Sievers et al. [49]). If using the substitution type (that is not T-to-C) with the largest number of substitution positions from the Kishore A data set, the estimated percentage of method-induced T-to-C substitution positions would be equal to 82.1%. For the substitution type with the smallest number, the estimation would be equal to 97.2%. These estimates could thus be an indication how to choose the values for p_i in the

Table 6.1: Different scenarios for $\boldsymbol{\beta}$, where the header represents the indices of β and p.

		0	3'UTR	CDS	5'UTR
Scenario large $\boldsymbol{\beta}$	β	0.5	1.85	1.15	0.75
	p	0.69	0.99	0.95	0.89
Scenario small $\boldsymbol{\beta}$	β	-0.5	1.5	1.0	0.5
	p	0.31	0.84	0.69	0.5
Scenario no effect	β	0.85	0.0	0.0	0.0
	p	0.8	0.8	0.8	0.8

simulation study.

In order to reflect the influence of the additional variables, the probability p_i is here chosen to be different for the different types of mRNA regions. In BayMAP this probability is modeled via a probit model. When defining fixed values for the parameters for the binary variables coding for the 3'UTR, the CDS or the 5'UTR, the probabilities p_i can be calculated by the probit model defined in (5.8) with the parameter vector $\boldsymbol{\beta}$. Three different settings for $\boldsymbol{\beta}$ are here employed, that are given in Table 6.1.

In the first setting, $\boldsymbol{\beta}$ is chosen in such a way, that $p_{3'\text{UTR}}$ is even higher than the large estimate of the previous passage of 97.2%, since a binding site is most likely in the 3'UTR. As the estimates of the previous passage are quite large, $\boldsymbol{\beta}$ is chosen in a way, that also p_{CDS} and $p_{5'\text{UTR}}$ are large, but smaller than $p_{3'\text{UTR}}$. In the second setting, $\boldsymbol{\beta}$ has smaller values, so that also a scenario with smaller values for p_i is reflected. In order to verify if BayMAP also works when there is no effect of the type of the mRNA region. In the last setting the parameters are set to zero with p_0 equal to 0.8 comparable to the smaller estimate of the percentage of method-induced T-to-C substitution positions in the previous passage.

Each position that is chosen as a position having method-induced substitutions has now to be modified. The idea is to change the number of substitutions for those positions. As discussed in Section 5.1.1, the number of substitutions is supposed to follow a zero truncated binomial distribution (see (5.1)). Given the probability of an experi-

Table 6.2: Different scenarios for μ_{exp}.

Scenario	Values
Whole range μ_{exp}	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9
Small μ_{exp}	0.05, 0.075, 0.1, 0.125, 0.15, 0.175, 0.2, 0.225 and 0.25
Beta distribution μ_{exp}	For each position randomly drawn from a Beta(2, 10) distribution

mentally induced substitution μ_{exp} and the number of reads n_i for position i, the number of substitutions K_i can be drawn from the zero truncated binomial distribution. Since n_i is given, only μ_{exp} has to be specified.

First, nine different settings with $\mu_{\text{exp}} = 0.1,\ 0.2,\ \ldots,\ 0.9$ are considered. However, as discussed in Section 2.2.3, it seems that the substitution rate for experimentally induced substitutions is close to the substitution rate for mismatches. It appears to be difficult to distinguish between mismatches and experimentally induced substitutions. As discussed in Section 5.1.1, μ_{mm} is assumed to be smaller than 0.25. A special interest lies therefore in the simulation of data sets with parameters for μ_{exp} smaller than 0.25. Therefore, nine additional settings are considered with $\mu_{\text{exp}} = 0.05,\ 0.075,\ \ldots,\ 0.25$. This means, that in one setting for one simulated data set the probability μ_{exp} is the same for each experimentally induced substitution position.

However, in a real PAR-CLIP data set, the substitution probability for positions on binding sites could differ for each position in one data set. Additional simulations are therefore also done for μ_{exp} different for each substitution position. For each position that is chosen as method-induced, a value between zero and one for μ_{exp} has hence to be drawn. Here a beta distribution is chosen to randomly draw values for μ_{exp} different for each substitution position, as the beta distribution is defined on the interval [0, 1] and because of its conjugacy to, e.g., the binomial distribution, an often used distribution to model probability parameters.

The density function of the beta distribution should be, on the one hand, close to the histogram of T-to-C substitution rates. On the other hand, the values for μ_{exp} should only reflect substitution rates for the crosslinked positions, so that the density function should not be high for substitution rates very small (close to zero) and very large

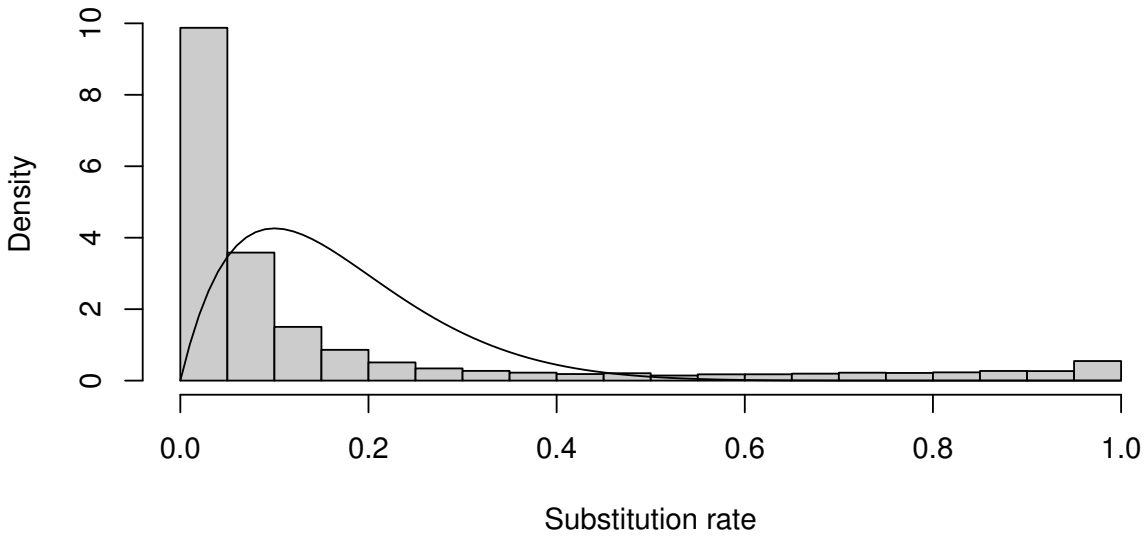

Figure 6.1: Histogram of T-to-C substitution rates in comparison to a Beta(2,10) density (black curve) for the Kishore A data set.

(close to one). Hence, a beta distribution with parameters two and ten is chosen as it is quite close to the histogram of T-to-C substitution rates for the first data set from Kishore et al. [30] but not too close, so that mismatch and SNP substitution rates are not reflected in the density function (see Figure 6.1).

There are thus three different scenarios for $\boldsymbol{\beta}$ (see Table 6.1) and three different scenarios for μ_{exp} (see Table 6.2). Not every scenario for $\boldsymbol{\beta}$ is mixed with every scenario for μ_{exp}, but only the first scenario for $\boldsymbol{\beta}$, i.e. the large $\boldsymbol{\beta}$, is mixed with every scenario for μ_{exp} and only the first scenario for the whole range μ_{exp} is mixed with every scenario for $\boldsymbol{\beta}$. Note that the scenarios for the whole range μ_{exp} and the small μ_{exp} each represent nine different settings, as for each different μ_{exp} data sets are simulated, whereas the beta distribution μ_{exp} only represents one setting, as μ_{exp} differs for every position in this scenario.

For all settings but those with the small values for μ_{exp}, i.e. $\mu_{\text{exp}} = 0.05, 0.075, \ldots, 0.25$, ten different data sets are simulated. For the settings in which μ_{exp} is small, twenty different data sets are simulated, since it is probably more difficult for the methods to distinguish between method-induced and non-method-induced substitution positions because of the smaller values for μ_{exp} that are closer to the expected value of μ_{mm}.

The results of the applications of BayMAP 1.0, wavClusteR and BMix are compared and

evaluated in terms of accuracy. The accuracy is the percentage of correctly identified positions. Sensitivity and specificity are also compared. Sensitivity, or true positive rate, is here the fraction of correctly identified method-induced T-to-C substitution positions out of all method-induced T-to-C substitution positions. Specificity, or true negative rate, is here the fraction of correctly identified non-method-induced T-to-C substitution positions out of all non-method-induced T-to-C substitution positions. Note that BayMAP is not compared to PARalyzer [12], since PARalyzer takes as input aligned reads, i.e. regions of positions, whereas the here simulated data sets only represent positions with substitutions, so that PARalyzer cannot be applied on these data sets.

Results in this section are obtained by applying BayMAP 1.0 employed in R [47] combined with WinBUGS [50]. The corresponding R package will be available online [26] (for R documentation see Appendix E.1).

The total number of iterations for the application of BayMAP 1.0 is here 15,000. The first 1,500 iterations of these 15,000 are discarded as burn-in. From the remaining iterations, every third iteration is used to the aim of autocorrelation reduction. This leads to 4,500 iterations that are kept for further analyses. Convergence is checked by trace plots (for example trace plots see Appendix B.1). The chains appear to have converged, as the mean and the variance of the chains seem to be stable with jumps large enough that can traverse the whole space.

The posterior odds (see Equation (5.44)) are employed in this Section as criterion if a position is method-induced or not. Equation 5.39 is not used as an alternative for the posterior odds, since the chains are run here in WinBUGS, that could not store every sampled value for the parameter Z_i.

6.1.2 Bias in estimation

Before analyzing the performance in terms of accuracy, sensitivity and specificity, it should be checked, if the estimations in BayMAP 1.0 are unbiased for two reasons. First, unbiased estimations show that the model is working. Second, it is here of special interest to analyze the estimated distributions for $\boldsymbol{\beta}$ to learn more about the biology of

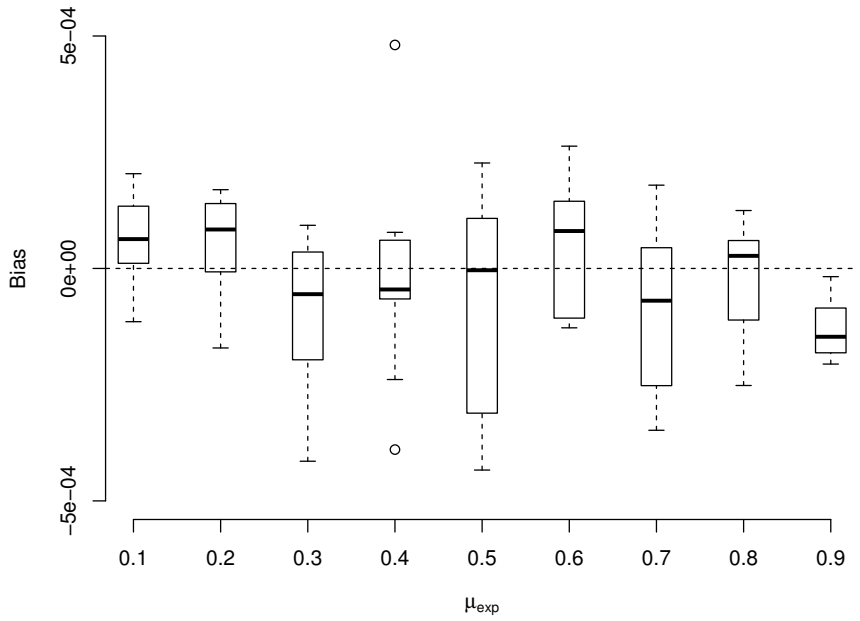

Figure 6.2: Bias of the mean estimate for μ_{exp} for nine different values of μ_{exp} for BayMAP for the simulation settings with $\beta_0 = 0.5$, $\beta_{3\text{'UTR}} = 1.85$, $\beta_{\text{CDS}} = 1.15$ and $\beta_{5\text{'UTR}} = 0.75$.

miRNA binding sites. Moreover, the estimates for μ_{exp} could give further clues about the substitution rates for method-induced substitutions. Such an analysis is however only reasonable if parameter estimations are more or less without bias.

In this section, the bias of μ_{exp} and $\boldsymbol{\beta}$ is therefore analyzed. The bias of μ_{mm}, μ_{SNP} and q is not analyzed, since their true values are not known. They are not given, since a real PAR-CLIP data set is taken for the simulations, where only some positions are changed that are chosen as method-induced. All other positions, that are not selected as method-induced, and that are therefore not crosslinked, are not edited, so that the parameters μ_{mm}, μ_{SNP} and q do not have to be specified for the simulation study, but reflect information from the underlying PAR-CLIP data set. These parameters will be analyzed in more detail in the applications to real PAR-CLIP data sets in Section 7.

In Figure 6.2 the bias of the mean estimate for μ_{exp} is shown for the simulation settings with a large $\boldsymbol{\beta}$ and the whole range μ_{exp}. For all of the 90 simulated data sets, the mean of the distribution of μ_{exp} vary less than 0.0005 from the true value of μ_{exp}. In the simulation settings with the small $\boldsymbol{\beta}$ (see Figure C.1), the mean estimates vary less than 0.0006 from the true value.

It seems, that there is a slight overestimation for small values of μ_{exp} and a slight under-

estimation for larger values of μ_{exp} in both settings, i.e. with the large $\boldsymbol{\beta}$ and the small $\boldsymbol{\beta}$. This finding can be confirmed, when looking at the bias for μ_{exp} considering the settings with the small μ_{exp} (see Figure C.2), where there seem to be very small overestimations for $\mu_{exp} \leq 0.175$. However, these differences to the true value of μ_{exp} are such small values, that one could speak of unbiased estimations for these simulation settings for μ_{exp}.

When μ_{exp} is drawn from the Beta$(2,10)$ distribution, it is not possible to determine the bias in the same way as above, as μ_{exp} is not fixed. This means, that the true μ_{exp} is different for each T-to-C substitution position, whereas μ_{exp} is estimated as one parameter in BayMAP 1.0. The mean estimates are all around 0.05 (from 0.0499 to 0.0502). In comparison, the 50% quantile of the Beta$(2,10)$ distribution from which μ_{exp} is drawn, is equal to 0.15 and the expected value to 0.1667. The peak (mode) of the distribution is at the value of 0.1. The estimates in the data set of 0.05 is therefore smaller than the expected value, the 50% quantile and the mode of the theoretical distribution.

A reason for this underestimation in comparison to the location parameters mentioned above, could be that it is in this setting particularly difficult to distinguish between mismatches and experimentally induced substitutions. For PAR-CLIP induced substitution positions with a medium substitution rate (for example 0.5), μ_{exp} could be underestimated, but BayMAP would nevertheless predict a method-induced substitution position. Whereas it is important for experimentally induced substitution positions with small substitution rates that μ_{exp} is estimated small enough so that it can be distinguished between mismatches and PAR-CLIP induced substitutions. In scenarios, where μ_{exp} is not fixed, i.e. the same for every position, it is therefore possible that μ_{exp} is underestimated. The underestimation, however, would lead to a higher accuracy thanks to a better sensitivity.

It is of special interest if $\boldsymbol{\beta}$ is estimated without bias, as $\boldsymbol{\beta}$ could not only be used for a better prediction of method-induced substitution positions but in particular for analyzing which factors could be important for a binding site. In contrast, even though it is also of interest to analyze the estimated values for μ_{exp}, it is for μ_{exp} especially important that it is estimated in a way that helps to distinguish crosslinked T-to-C sub-

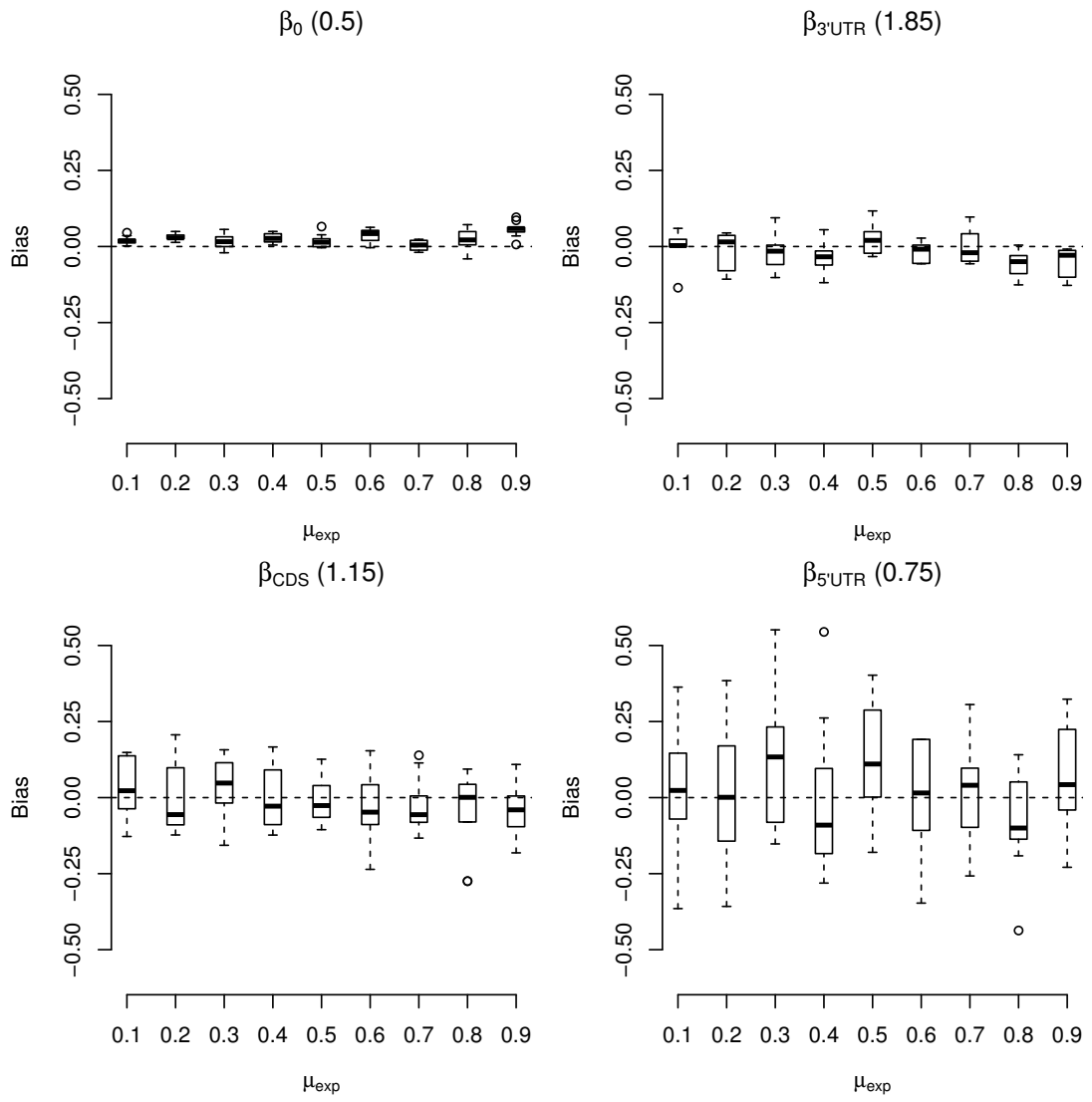

Figure 6.3: Bias of the mean estimates for the regression parameters in the probit model in BayMAP in relation to the mRNA position (3'UTR, CDS, 5'UTR) with $\beta_0 = 0.5$, $\beta_{3'\text{UTR}} = 1.85$, $\beta_{\text{CDS}} = 1.15$ and $\beta_{5'\text{UTR}} = 0.75$. The true values of the parameters for β are shown in parentheses. For $\beta_{5'\text{UTR}}$ two outliers with values 1.05 and 0.73 for $\mu_{\text{exp}} = 0.6$ are not displayed.

stitution positions from non-crosslinked ones.

Regarding the bias of the means for $\boldsymbol{\beta}$ of the simulation settings with a large $\boldsymbol{\beta}$ (Figure 6.3) and with a small $\boldsymbol{\beta}$ (Figure C.4), it is noticeable that the variances of the estimations is higher for $\beta_{5'\text{UTR}}$ and β_{CDS} than for β_0 and $\beta_{3'\text{UTR}}$. These results are not surprising, since there are less substitution positions annotated to the CDS and far less to the 5'UTR so that estimation could be imprecise.

In the simulation settings with the large $\boldsymbol{\beta}$ (Figure 6.3), the estimations seem to be more or less unbiased, as they vary around the zero line. Results for the bias of the large $\boldsymbol{\beta}$ in the settings in which the small μ_{exp} is considered are similar (see Figure C.3). Whereas

there seem to be a small overestimation of β_0 and small underestimation of $\beta_{3'\text{UTR}}$, β_{CDS} and $\beta_{5'\text{UTR}}$ when regarding the settings with the small $\boldsymbol{\beta}$ (see Figure C.4). This means however, that in this case only p_0 is slightly overestimated, since for $p_{3'\text{UTR}}$, p_{CDS} and $p_{5'\text{UTR}}$, the over- and underestimation are balanced out.

In the simulation setting, in which μ_{exp} is not fixed but drawn from a Beta(2,10) distribution instead, the variances of the mean estimates for $\boldsymbol{\beta}$ seem also to be higher for $\beta_{5'\text{UTR}}$ and β_{CDS} (Figure C.5). One could argue for a small underestimation of $\beta_{5'\text{UTR}}$. All in all, however, the estimations of $\boldsymbol{\beta}$ seem to be unbiased and therefore interpretable.

6.1.3 Comparison of BayMAP 1.0 to simpler versions

Before comparing the results of BayMAP 1.0 to other methods, it is here analyzed how BayMAP 1.0 performs and if also simpler versions of BayMAP 1.0 could be applied or if BayMAP 1.0 could be applied on reduced data sets. The ten simulated data sets with the whole range μ_{exp} and the large $\boldsymbol{\beta}$ are taken for these analyses.

After applying the ordinary BayMAP 1.0 on the whole simulated data sets of the here considered settings, BayMAP 1.0 is applied only to T-to-C substitution positions, i.e. without considering other substitution types such as G-to-A. BayMAP 1.0 is also applied to the whole data set, i.e. with all substitution positions, without considering the additional variables that had an influence on the probability that the T-to-C substitution position has substitutions induced by PAR-CLIP (termed no probit).

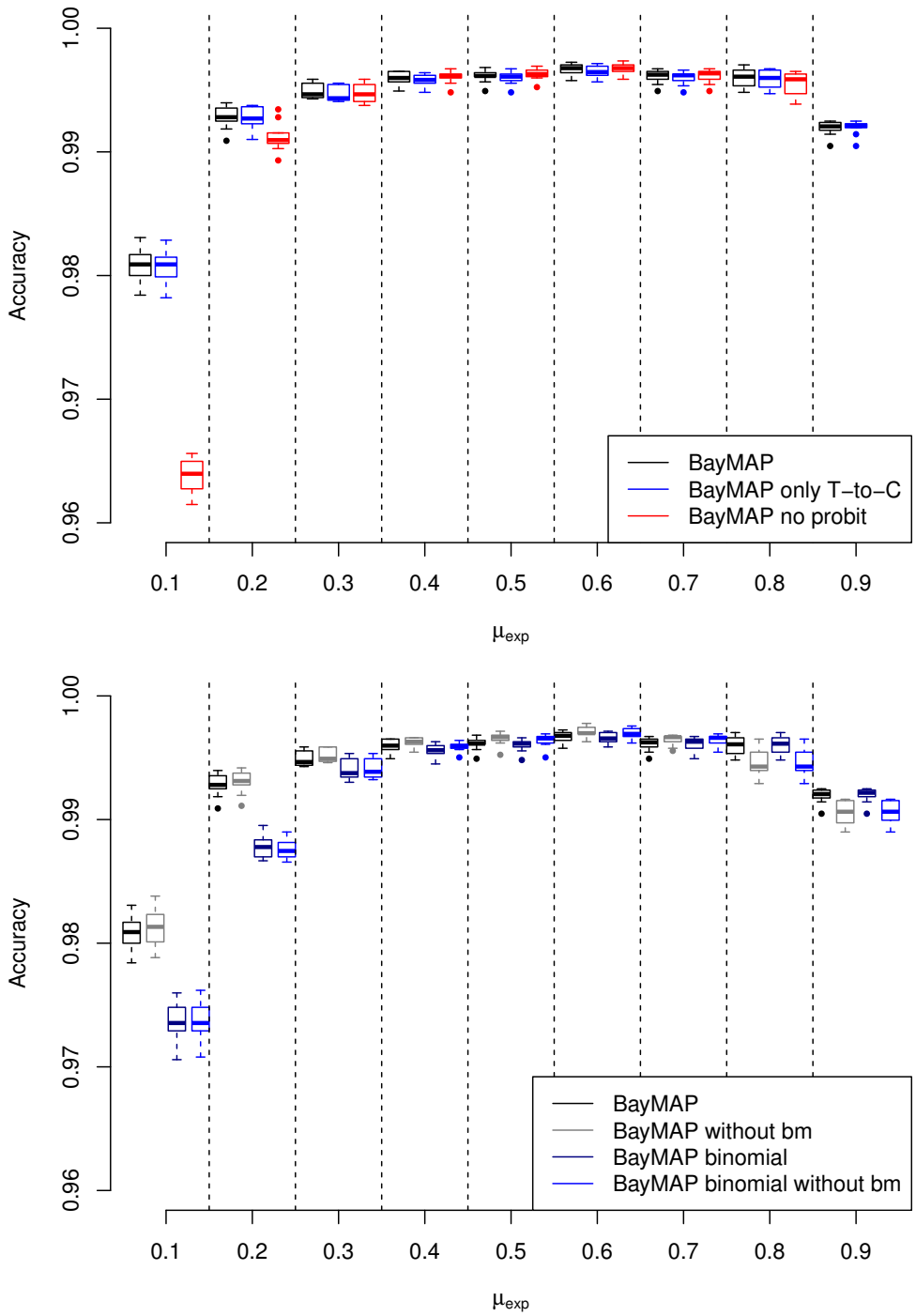

Figure 6.4: (BayMAP benchmark with simpler versions) Top panel: Distribution of the accuracy of BayMAP 1.0 (black), BayMAP 1.0 considering only T-to-C substitution positions (blue), and BayMAP 1.0 not considering additional variables (red), considering probabilities μ_{exp} of T-to-C substitutions at an experimentally induced position between 0.1 and 0.9. For a better graphical representation, a very low accuraciy of about 0.02 obtained in an application only with T-to-C substitutions with $\mu_{\text{exp}} = 0.1$ is not shown. The very low accuracies of about 0.11 obtained in all applications not considering additional variables with $\mu_{\text{exp}} = 0.9$ are also not shown. Bottom panel: Distribution of the accuracy of BayMAP 1.0 (black), BayMAP 1.0 not considering dependencies between μ_{mm} and μ_{SNP} (grey) , BayMAP 1.0 where a binomial distribution for K_i is modeled instead of a zero truncated binomial distribution (dark blue), and BayMAP 1.0 where a binomial distribution is modeled not considering dependencies between μ_{mm} and μ_{SNP} (light blue), considering probabilities μ_{exp} of T-to-C substitutions at an experimentally induced position between 0.1 and 0.9.

Then, BayMAP 1.0 is applied without considering the dependency between μ_{SNP} and μ_{mm} (termed without bm). Finally, BayMAP 1.0 is applied where the number of substitutions K_i follows a binomial distribution instead of a zero truncated binomial distribution (see (5.1)) with and without dependency of μ_{SNP} and μ_{mm}. The box plots for the accuracy of BayMAP 1.0 as well as simpler versions are shown in Figure 6.4.

Analyzing the accuracy of the original BayMAP 1.0 applied on the whole data set, it stands out, that BayMAP 1.0 has very high values of accuracy starting with values around 0.98, but mostly larger than 0.99.

First, only T-to-C substitution positions are taken as input data and compared (see top panel of Figure 6.4). Only taking T-to-C substitution positions would have the advantage, that the data set is smaller and analyzing would be faster. Results of BayMAP 1.0 applied to data sets only consisting T-to-C substitution positions are nearly equal to the results of BayMAP applied to the whole data sets. However, there exist one outlier for $\mu_{\text{exp}} = 0.1$ with an accuracy of only 0.02.

Second, ignoring the additional variables in the model also leads to high accuracy values but not as good as if the covariates were taken into account (see top panel of Figure 6.4). For $\mu_{\text{exp}} = 0.9$, results are not shown for the no probit application with a sensitivity close to zero, which means that experimentally induced substitution positions are mainly declared as SNPs and the other way around. Obviously, with high T-to-C substitution rates, additional variables are essential for prediction, although this situation rarely occurs in real experimental set ups.

When not considering the dependency between μ_{SNP} and μ_{mm}, results are nearly the same as for the ordinary BayMAP 1.0. Results seem to be slightly better for BayMAP 1.0 with dependency when μ_{exp} is equal to or larger than 0.8. With dependency, μ_{SNP} is estimated to be very close to one with medians from 0.994 to 0.996 whereas the medians vary from 0.840 to 0.968 when not considering the dependency. Although results in terms of accuracy seem to be comparable, it is nevertheless recommended to use dependencies, since estimations for μ_{SNP} seem to be more reasonable.

Since the data does not consist of positions with zero substitutions, the number of sub-

stitutions is supposed to follow a zero truncated binomial distribution. It is, nevertheless, possible to use the binomial distribution instead which would have the advantage that the easier Gibbs sampling could be employed. However, supposing the binomial distribution leads to slightly smaller values for the accuracy especially when μ_{exp} is equal to or smaller than 0.2.

In total, the comparison indicates, that it is possible to use simpler versions of BayMAP 1.0 or to apply BayMAP 1.0 on a reduced data set. However, the best results in terms of accuracy and the most stable results are obtained by applying the ordinary BayMAP 1.0. Only when not taking into account the dependency between μ_{exp} and μ_{mm}, the values for the accuracy are as good as for the ordinary BayMAP. The estimates for μ_{SNP} seem, however, to be less reasonable as explained above. Moreover, it increases the risk of label switching problems as they occurred for example for the data sets where only T-to-C substitutions are taken into account and where the additional variables are not used.

It is therefore recommended to use BayMAP 1.0 as originally proposed, i.e. on the whole data set considering additional variables assuming the zero truncated binomial distribution and dependencies between μ_{SNP} and μ_{mm}.

6.1.4 Comparison of BayMAP 1.0 to other methods

Although it is important that the parameter estimations are unbiased, it is of even more interest to analyze if BayMAP 1.0 is able to predict method-induced T-to-C substitution positions and better than existing methods, as this is BayMAP's main purpose. Thus, in this section, the performance of BayMAP 1.0 is analyzed and compared to BMix and wavClusteR.

The first nine considered simulation settings are those with the whole range μ_{exp} and large $\boldsymbol{\beta}$ (see Table 6.1 and Table 6.2). Results for this setting are shown in Figure 6.5.

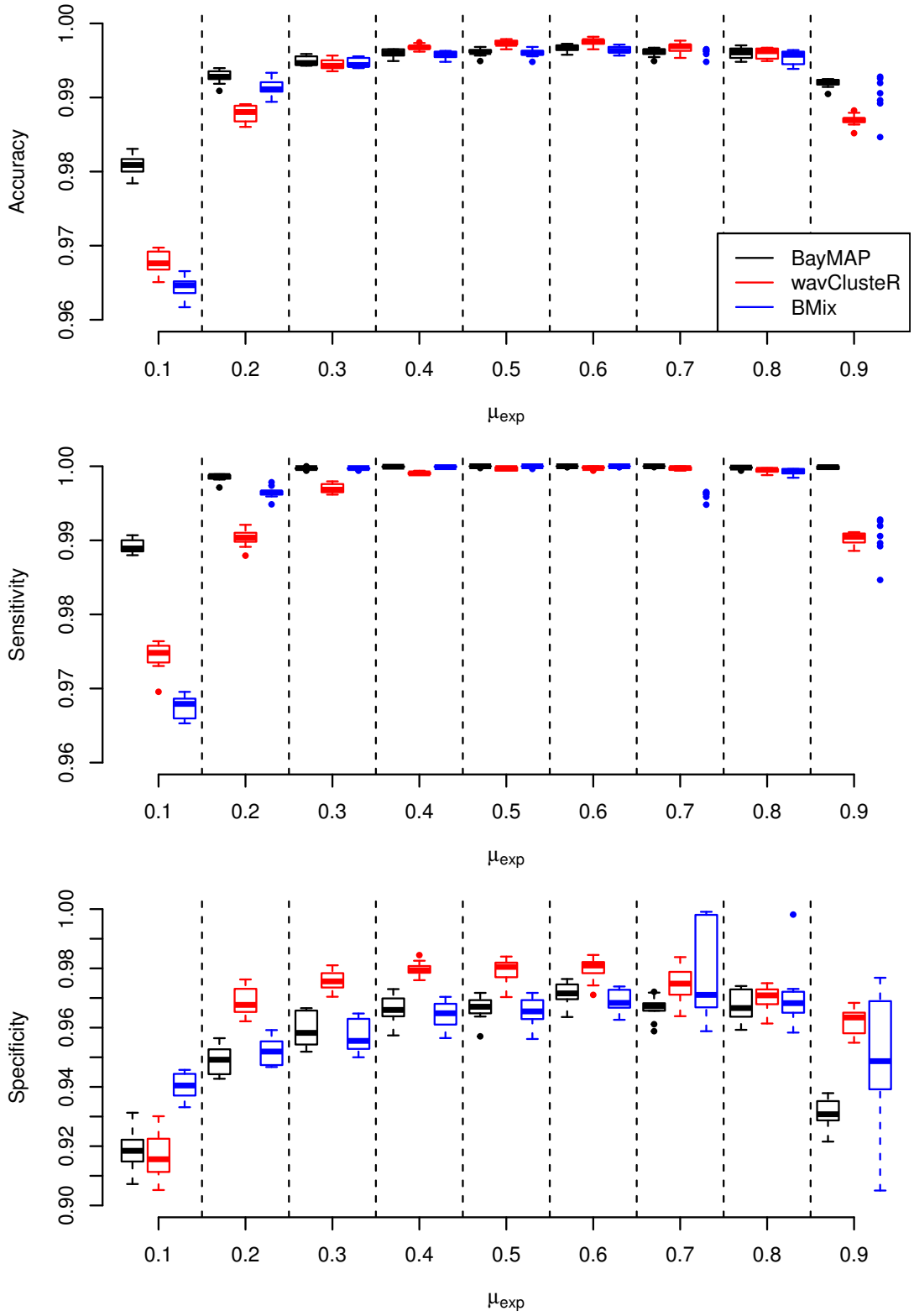

Figure 6.5: (Simulation whole rage μ_{exp} and large $\boldsymbol{\beta}$) Distribution of the accuracy (top panel), sensitivity (middle panel) and specificity (bottom panel) of BayMAP 1.0 (black box plots for each μ_{exp}), wavClusteR (red), and BMix (blue) for ten simulated data sets considering probabilities μ_{exp} of T-to-C substitutions at an experimentally induced position between 0.1 and 0.9 with $\beta_0 = 0.5$, $\beta_{3'UTR} = 1.85$, $\beta_{CDS} = 1.15$ and $\beta_{5'UTR} = 0.75$. For a better graphical representation, very low values of sensitivity close to zero and of accuracy of about 0.1 obtained in applications of BMix to data sets with $\mu_{exp} = 0.7$, 0.8, or 0.9 are not shown. As thus the 25% quantiles for $\mu_{exp} = 0.7$ and $\mu_{exp} = 0.9$ are about 0.11, the accuracies of BMix in these cases are only displayed as points, but not as box plots.

All three methods have a high performance with an accuracy higher than 0.99 for μ_{exp} between 0.3 and 0.6 (see Figure 6.5). Outliers for BMix are not shown for a better graphical representation. In total, BMix has eight outliers for $\mu_{\text{exp}} = 0.7, 0.8$ and 0.9 with a sensitivity close to zero and an accuracy of about 0.11. This shows that especially BMix has problems to distinguish between SNPs and method-induced substitutions when μ_{exp} is high.

When looking at the sensitivity and specificity, BayMAP 1.0 outperforms with a high sensitivity whereas wavClusteR and BMix have better values for the specificity. This means, that BayMAP 1.0 is good in detecting true method-induced substitution positions for these settings. It stands out, that BayMAP 1.0 performs better than wavClusteR and BMix especially when μ_{exp} is close to zero.

As discussed above, it seems to be crucial to distinguish between mismatch positions and method-induced substitution positions, since there is evidence in the considered data that PAR-CLIP induced substitution rates are not very high (see e.g., Figure 6.1). It is, thus, of special interest to have a closer look into small substitution rates smaller or equal than 0.25, as most positions seem to have rates in this range.

In Figure 6.6 results of the nine settings with the small μ_{exp} and large $\boldsymbol{\beta}$ are shown. The closer μ_{exp} is to zero the more difficult it gets to distinguish between mismatches and PAR-CLIP induced substitutions. Consequently, the accuracy decreases with decreasing μ_{exp} in the results. BayMAP 1.0 however, has a higher performance than the other methods in terms of accuracy, where the difference between BayMAP 1.0 and the other methods is higher the smaller μ_{exp} is.

This reveals BayMAP's high performance especially when it is hard to distinguish between method-induced substitutions and other substitutions.

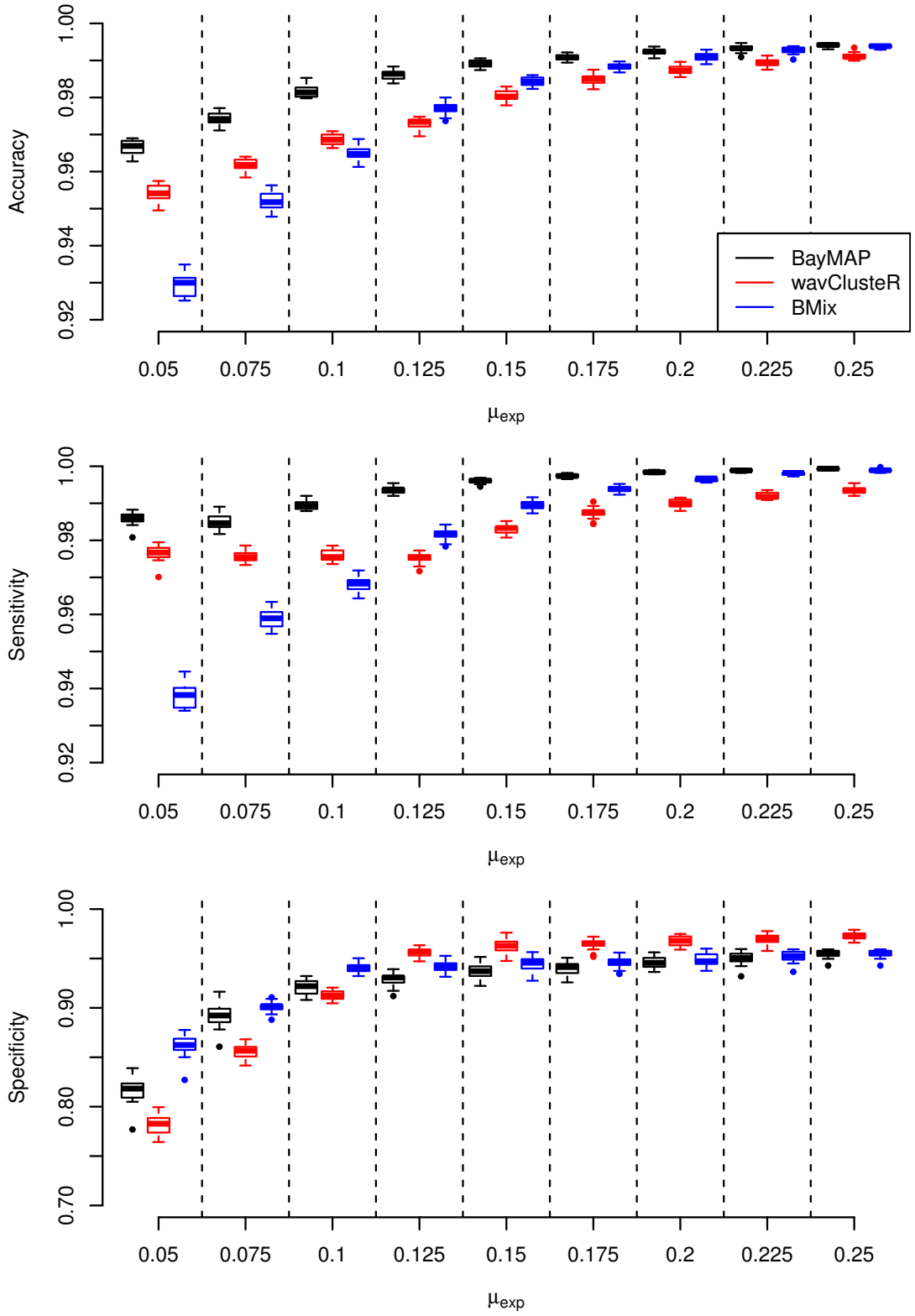

Figure 6.6: (Simulation with small μ_{exp} and large $\boldsymbol{\beta}$) Distribution of the accuracy (top panel), sensitivity (middle panel) and specificity (bottom panel) of BayMAP 1.0 (black box plots for each μ_{exp}), wavClusteR (red), and BMix (blue) for twenty simulated data sets considering probabilities of T-to-C substitutions at an experimentally induced position of $\mu_{\text{exp}} = 0.05, 0.075, \ldots, 0.225, 0.25$ with $\beta_0 = 0.5$, $\beta_{3\text{'UTR}} = 1.85$, $\beta_{\text{CDS}} = 1.15$ and $\beta_{5\text{'UTR}} = 0.75$.

It is not only of interest to show that BayMAP 1.0 works well for high values of $\boldsymbol{p}$ but also for smaller $\boldsymbol{p}$ which means that the data set consists of much more noise. Results for the settings with the smaller $\boldsymbol{\beta}$ and thus a smaller $\boldsymbol{p}$, are therefore presented in Figure 6.7.

BayMAP 1.0 and BMix both seem to perform very well with an overall mean of accuracy of 0.99 for BayMAP 1.0 and 0.95 for BMix. BMix has again outliers for $\mu_{\text{exp}} = 0.9$, where five values of sensitivity are close to zero so that the accuracy for these data sets is around 0.34. Even when not considering these five outliers, BayMAP 1.0 still performs slightly better than BMix in terms of accuracy.

When only looking at the sensitivity, wavClusteR seems to perform well with most values above 0.99 even for $\mu_{\text{exp}} = 0.1$, where wavClusteR outperforms the other methods. However, wavClusteR has much more problems in identifying the right positions with an overall mean of specificity of only 0.7. With values for the specificity smaller than 0.6 for $\mu_{\text{exp}} = 0.1$, the slightly better values in sensitivity is largely outweighed by the small values for specificity.

For now, BayMAP 1.0 had the advantage, that the influence of additional variables was included in the simulated data sets and that BayMAP is the only method that is able to model these covariates. Even when $\boldsymbol{\beta}$ with no effect is considered and when there is thus no influence of additional variables in the simulated data, BayMAP 1.0 performs well as can be seen in Figure C.6. Seven outliers of BMix for $\mu_{\text{exp}} = 0.7$ and 0.9 are not shown. Sensitivity of these outliers is close to zero so that accuracy is around 0.2. Even though there are no additional variables, BayMAP especially performs slightly better than wavClusteR and BMix for small or high values of the substitution rate ($\mu_{\text{exp}} = 0.1, 0.2$ and 0.9), i.e. when it is difficult to distinguish between method-induced and non-method-induced substitutions.

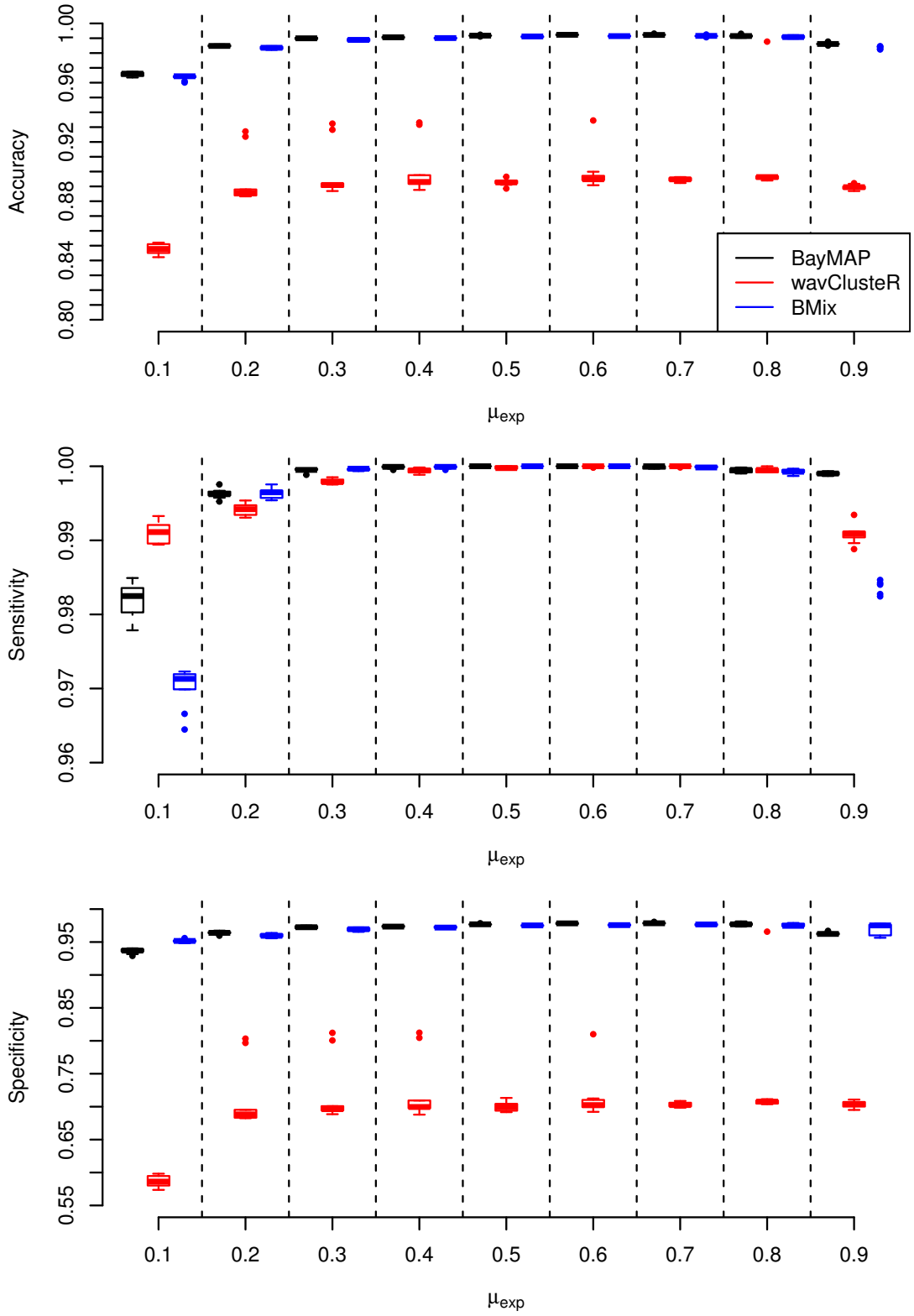

Figure 6.7: (Simulation whole range μ_{exp} and small $\boldsymbol{\beta}$) Distribution of the accuracy (top panel), sensitivity (middle panel) and specificity (bottom panel) of BayMAP 1.0 (black box plots for each μ_{exp}), wavClusteR (red), and BMix (blue) considering probabilities μ_{exp} of T-to-C substitutions at an experimentally induced position between 0.1 and 0.9 with $\beta_0 = -0.5$, $\beta_{3'\text{UTR}} = 1.5$, $\beta_{\text{CDS}} = 1.0$ and $\beta_{5'\text{UTR}} = 0.5$. For a better graphical representation, very low values of sensitivity very close to zero and of accuracy of about 0.34 obtained in applications of BMix to data sets with $\mu_{exp} = 0.9$ are not shown. As thus the 25% quantiles for $\mu_{exp} = 0.9$ are about 0.34, the accuracies of BMix in this case are only displayed as points, but not as a box plot.

As discussed above, μ_{exp} has not to be fixed to a value, but can also be drawn from a distribution, here the Beta(2,10) distribution. All three methods perform well (see Figure C.7). BayMAP 1.0 and wavClusteR have a slightly higher accuracy than BMix whereas BMix has here higher values specificity.

6.1.5 Simulation study with PARA-suite

In order to also have a second way of simulation, the simulation tool of PARA-suite is employed in an additional simulation study. PARA-suite takes a fasta file, that is a file format for, e.g., RNA sequences, as input and simulates short RNA reads out of it. Here, a fasta file of the 3'UTR of GRCh38.p7 is taken. Additional input parameters that are needed, such as substitution rates for the different substitution types (e.g., T-to-C, A-to-G), can be estimated by analyzing a real PAR-CLIP data set with PARA-suite. The data set from Memczak et al. [41] is used here for estimating these input parameters. A noise parameter also has to be given to the tool. Noise can be modeled by the fraction of reads that should be binding sites in the simulated data. If this fraction is set to 0.1, it means that around 90% of the reads are not reads of binding sites but noise. Data is here simulated for the nine fractions 0.1, 0.2, ..., 0.9. In PARA-suite reads are simulated before alignment, so that additional variables that are modeled in BayMAP by the probit model cannot be considered.

Simulated data sets from the PARA-suite tool for the nine different noise settings are analyzed (see Figure 6.8). In terms of accuracy, all three methods perform comparable. BayMAP 1.0 has a higher performance in terms of sensitivity when the fraction of binding sites is low, i.e. when noise is high. wavClusteR even has a sensitivity very close to zero when the fraction of binding sites is smaller or equal to 0.3. Since a small value of the fraction of binding sites means that only few reads are binding site reads, the accuracy is high nevertheless. For small values of the binding site fraction, it is in particular the specificity that determines the accuracy, as most positions are noise. By contrast, for large values of the binding site fraction, it is mainly the sensitivity that determines the accuracy. Due to this reason, the values for the accuracy are decreasing for larger values of the binding site fraction, since the large values of specificity influence decreasingly the accuracy and the smaller values of sensitivity get more weight.

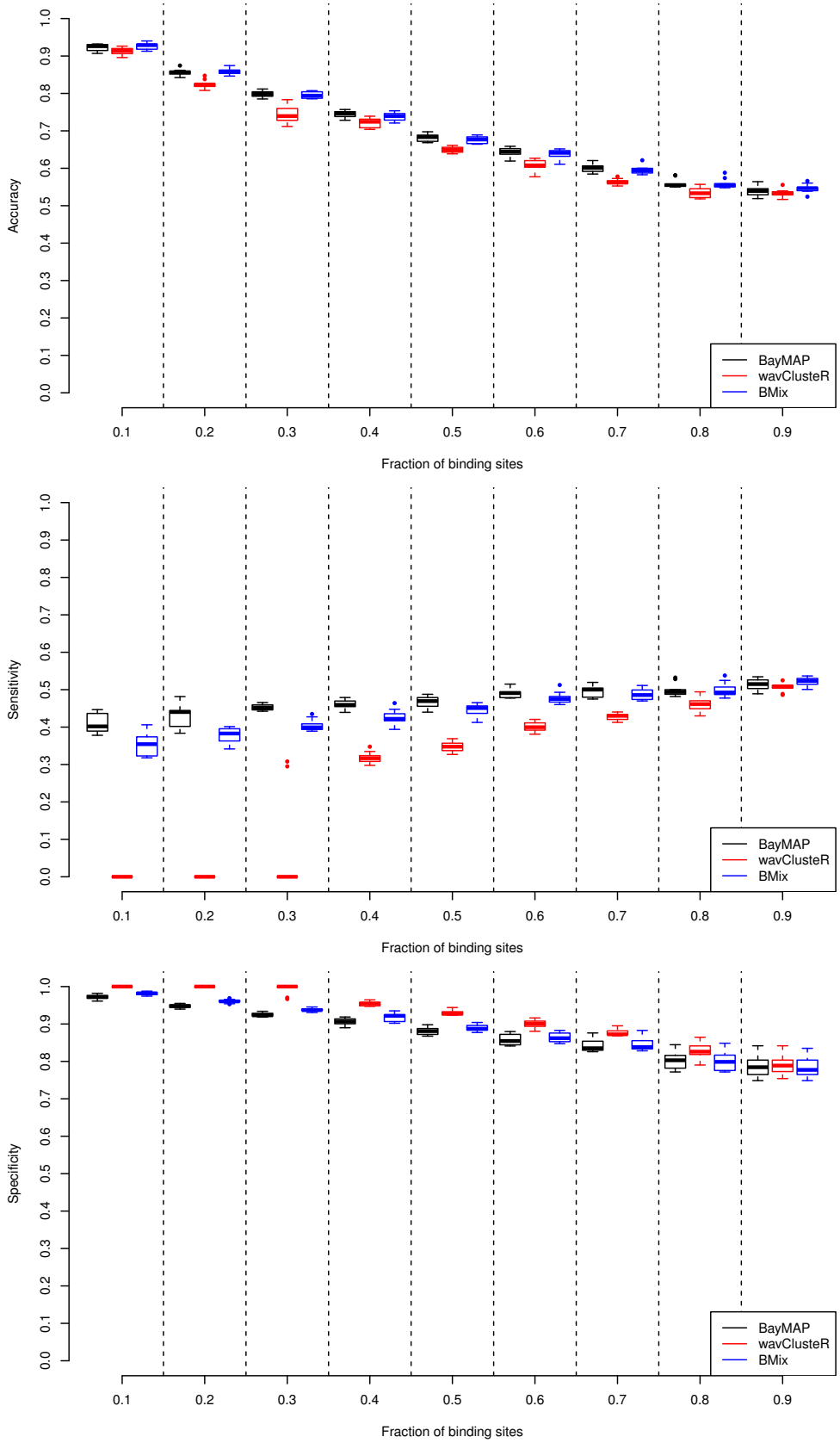

Figure 6.8: (Simulation PARA-suite) Distribution of the accuracy (top panel), sensitivity (middle panel) and specificity (bottom panel) of BayMAP 1.0 (black), wavClusteR (red), and BMix (blue) for ten simulated data sets considering fraction of binding sites between 0.1 and 0.9 simulated by PARA-suite. Note that the fraction of binding sites influences the number of reads that are noise in the data set, but not the number of T-to-C substitions at a binding site (as is the case for μ_{exp}).

6.2 BayMAP 2.0

For the simulation study of BayMAP 2.0, first the set up of the main simulation study is described (see Section 6.2.1). Then, the bias is analyzed for BayMAP 2.0 in Section 6.2.2 to verify the model. Finally, the performance in detecting crosslinked T-to-C substitution positions is evaluated and compared to simpler versions (see Section 6.2.3) of BayMAP 2.0 as well as to BMix and wavClusteR (see Section 6.2.4).

6.2.1 Set up of simulation study

In Section 6.1, it is assumed, that T-to-C substitution positions are independent of each other. However, T-to-C substitution positions very close to each other are usually either both on a binding site or both not on a binding site. This dependency is now added to the simulated data sets by an additional random effect for each potential binding site as explained in Section 5.3 with $p_j = \Phi\left(\boldsymbol{x_j}^\top \boldsymbol{\beta} + \alpha_j\right) = \Phi\left(\zeta_j\right)$.

The idea of simulations is here the same as in Section 6.1.1, where first T-to-C substitution positions are deleted of the first data set of Kishore et al. [30]. Then, a part of the A-to-G substitution positions is chosen randomly as method-induced substitution positions and their number of substitutions is changed artificially. As many of the non-A-to-G substitution positions are deleted as needed, so that the proportion of non-A-to-G substitution positions to A-to-G substitution positions is the same to the proportion of non-T-to-C substitution positions to T-to-C substitution positions in the original data set. In the following the A-to-G substitutions are named here again T-to-C substitutions, since they present the actual T-to-C substitutions.

When a cluster is not a binding site, then none of the T-to-C substitution positions should have experimentally induced substitutions. When a cluster is a binding site, however, not all of the positions are expected to be experimentally induced. In a first step, it has, hence, to be selected which of the clusters are binding sites, which is done here with the probability $p_j = \Phi\left(\boldsymbol{x_j}^\top \boldsymbol{\beta} + \alpha_j\right) = \Phi\left(\zeta_j\right)$.

In a second step, the positions with experimentally induced substitutions have to be chosen. Once a potential binding site j is selected to be a true binding site, it is drawn

randomly with probability p_j for every position on the j-th binding site if the number of substitutions is changed artificially.

Hence, p_j is not only taken here if cluster j is a binding site, but also if position i on the selected binding site j has method-induced substitutions. This is due to simplicity reasons, so that no further parameters have to be specified. If it is not very likely that a cluster is selected as a binding site, but it is picked nevertheless, positions on this binding site have thus not very likely experimentally induced substitutions in this simulation setting. Therefore, another advantage of this approach is, that most of the picked binding sites have crosslinked substitution on most of the T-to-C substitution positions, but that there exist also binding sites with only few or even none of the positions being method-induced, which could also happen in real PAR-CLIP data.

For the simulation, hence, not only $\boldsymbol{\beta}$ has to be specified but also a random effect α_j for every potential binding site as well as the potential binding itself.

The parameter $\boldsymbol{\beta}$ is set to the same values as in the simulation study in Section 6.1 (see Table 6.1). This means, that three scenarios are considered. In the first scenario with the large $\boldsymbol{\beta}$, the parameters are set to $\beta_0 = 0.5$, $\beta_{3'\text{UTR}} = 1.85$, $\beta_{\text{CDS}} = 1.15$ and $\beta_{5'\text{UTR}} = 0.75$. In the second scenario with the small $\boldsymbol{\beta}$, the parameters are set to $\beta_0 = -0.5$, $\beta_{3'\text{UTR}} = 1.5$, $\beta_{\text{CDS}} = 1.0$ and $\beta_{5'\text{UTR}} = 0.5$. Finally, the model is also tested on data, where there is no effect of the type of the mRNA region with parameters set to $\beta_0 = 0.85$ and $\beta_{3'\text{UTR}} = \beta_{\text{CDS}} = \beta_{5'\text{UTR}} = 0$.

Before choosing the random effect α_j and therefore ζ_j, it has first to be specified which of the positions belong to the same binding site. In order to determine potential binding sites, one could apply the method proposed in Section 5.2 to the data. However, A-to-G substitution positions are not distributed in the same way as T-to-C substitution positions, as T-to-C substitution positions are meant to appear much more frequent due to the PAR-CLIP experiment. This means, that it is less likely, that the read clusters of different A-to-G substitution positions overlap and that they can be combined, so that smaller regions for the potential binding sites are expected for A-to-G substitution positions.Thus, the distribution of the number of A-to-G substitution positions per cluster may be different than for T-to-C substitution positions.

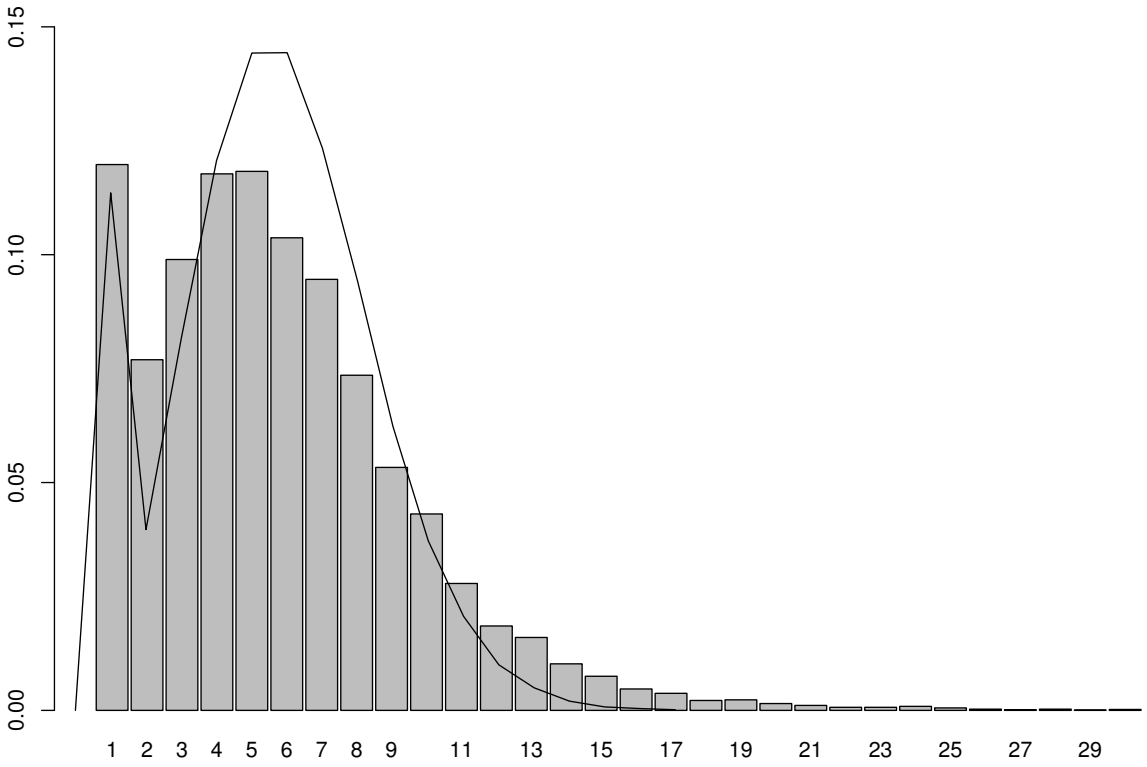

Figure 6.9: Barplot with relative frequencies of the number of T-to-C substitution positions per cluster for the Kishore A data set. The black line represents the density of the one inflated zero truncated Poisson distribution with parameters $\pi = 0.1$ and $\lambda = 6$.

Alternatively, the numbers of A-to-G substitution positions per cluster could be drawn randomly following the distribution of the number of T-to-C substitution positions per cluster in the original data set. The barplot in Figure 6.9 shows the distribution of the number of T-to-C substitution positions per cluster with relative frequencies for the first considered data set of Kishore et al. [30]. The clusters are determined using the here presented method in Section 5.2.

Since one is dealing here with the number of times an event occurs, the Poisson distribution that models such numbers, would be an obvious choice. Here, however, a cluster is only built when at least one T-to-C substitution position is present, so that it is not possible to observe zero T-to-C substitutions for one potential binding site. Instead of the Poisson distribution, the zero truncated Poisson distribution could therefore be taken, as it models the number of events, where it is known that at least one event is observed.

When having a look at the barplot in Figure 6.9, the high number of clusters with only

one T-to-C substitution position is outstanding. The observation of one seems therefore to be inflated. The zero inflated Poisson distribution is already described by Lambert [33]. This distribution can be transformed in such a way that it is a one inflated zero truncated Poisson distribution, that is firstly described and developed in this thesis to the best of the author's knowledge.

Let C be a random variable following the one inflated zero truncated Poisson distribution, its density can then be written as

$$P(C = c) = \pi \, \mathbb{1}_{c=1}(c) + (1-\pi) \, \mathrm{ZTP}\left(c \mid \lambda\right) \, \mathbb{1}_{c \in \mathbb{N}^+}(c) \,, \tag{6.1}$$

where π is the probability of the additional ones, $\mathrm{ZTP}\left(c \mid \lambda\right)$ is the density of the zero truncated Poisson distribution with parameter λ. The one inflated zero truncated Poisson distribution is thus a mixture model, where C is either equal to one with probability π or following a zero truncated Poisson distribution with probability $(1-\pi)$. Note, that

$$P(C = 1) = \pi + (1-\pi)\,\mathrm{ZTP}\left(1 \mid \lambda\right) \,,$$

since the observed value of a zero truncated Poisson distributed variable could also be equal to one.

When choosing $\pi = 0.1$ and $\lambda = 6$, the one inflated zero truncated Poisson distribution represents approximately the true distribution of the number of T-to-C substitution positions per cluster (see black line in Figure 6.9).

In order to determine to which cluster each A-to-G substitution position belongs, as many random numbers from the one inflated zero truncated Poisson distribution are drawn, so that the sum of all numbers is greater or equal to the number of A-to-G substitution positions. If this sum is greater than the number of A-to-G substitution positions, it would mean that the last cluster contains less positions than drawn.

Once the clusters are specified, an effect for every potential binding site has to be drawn. As described in (5.51), $\zeta_j = \boldsymbol{x_j}^\top \boldsymbol{\beta} + \alpha_j$ is supposed to follow a normal distri-

bution with

$$\zeta_j \mid \boldsymbol{x_j}^\top \boldsymbol{\beta}, \tau^2 \sim N(\boldsymbol{x_j}^\top \boldsymbol{\beta}, \tau^2)$$

given the variance parameter τ^2. As the true variance is unknown, τ^2 also has to be specified. Here, three different scenarios with $\tau^2 = 0.25$, $\tau^2 = 1$ and $\tau^2 = 4$ are considered, so that the standard deviations are $\tau = 0.5$, $\tau = 1$ and $\tau = 2$. As a comparison, the standard deviations in the real data set were estimated in the range from 0.63 to 1.84 as will be presented in Section 7.1.2. When these random values are drawn from the normal distribution, $p_j = \Phi(\zeta_j)$ can be calculated, so that it can be drawn with probability p_j for cluster j, whether it is a true binding site or not.

Finally, μ_{exp} has to be specified that is used for the random drawing of the number of substitutions for a position that is chosen as experimentally induced. As discussed in Section 6.1.1, μ_{exp} does not have to be the same for every position. The success of substitution by crosslinking could be linked to the location on the binding site, so that the substitution rate is assumed to be different for every position on a binding site.

The main simulated data sets in this section are therefore obtained with μ_{exp} drawn from a Beta(2,10) distribution (see Section 6.1.1). Additional, a fixed $\mu_{\text{exp}} = 0.2$ is also considered for the scenario with the large $\boldsymbol{\beta}$. For $\mu_{\text{exp}} = 0.2$, only the combination with the large $\boldsymbol{\beta}$ is examined to the aim of not oversizing the simulation study. The value $\mu_{\text{exp}} = 0.2$ is chosen, since one is in particular interested in smaller values of μ_{exp} as discussed in Section 6.1.1 and 0.2 is also relatively close to most of the estimated values for μ_{exp} as will be seen in Section 7.1.2 in Table 7.4.

For each of the different settings, first ζ_j is drawn ten times. Then, for each of the ten different entries of $\boldsymbol{\zeta} = (\zeta_1 \ldots \zeta_J)^\top$, 10 different data sets are simulated, so that in total 100 data sets exist for one setting. On all data sets, BayMAP 2.0, BayMAP 1.0, wavClusteR and BMix are applied.

Additionally, BayMAP 2.0 is also applied to the simulated data sets from Section 6.1, where μ_{exp} is drawn from Beta(2,10). In this way, the functionality of BayMAP 2.0 is also tested for settings, where there is no underlying effect of the different potential

binding sites. To estimate if a position has method-induced substitutions, (5.39) is employed to BayMAP 2.0 as well as BayMAP 1.0.

As described in Section 5.3.4, it is also reasonable to consider prior odds directly for the detection of binding sites in case of applying BayMAP 2.0, i.e. considering clusters. In the same manner, as in Section 6.1, the accuracy as well as the specificity and sensitivity are compared.

For calculating the accuracy, specificity and sensitivity, it has to be known for every position, if it is a binding site position or not. Here, not every position on a binding site has method-induced substitutions, since first it is chosen if a cluster is a binding site and then it is chosen for every position on a binding site separately if the position has PAR-CLIP specific substitutions or not. For the analysis if BayMAP 2.0 correctly identifies method-induced substitution positions, only positions with changed substitutions should therefore be taken into account. However, the overall aim is to detect binding sites and not only PAR-CLIP induced substitution positions. Moreover, BayMAP 2.0 is constructed in a way so that the data of neighbor positions on the same potential binding site influence the decision for a position. BayMAP 2.0 could therefore also predict a position as a binding site position just because of the neighbor positions. It is here hence more appropriate to check whether BayMAP 2.0 detects binding site positions. True positives for the accuracy, specificity and sensitivity are thus here defined as positions on chosen binding sites no matter if the position has method-induced substitutions or not.

Results in this section are obtained by applying BayMAP 2.0 employed in R [47]. In contrast to the previous section, here, WinBUGS [50] is not invoked for the sampling, but MCMC are obtained in R directly. The corresponding R package will be available online [26] (see Appendix E.2 for R documentation).

The total number of iterations for the application of BayMAP 1.0 is here 45,000. The first 15,000 iterations of these 45,000 are discarded as burn-in. From the remaining iterations, every sixth iteration is used to the aim of autocorrelation reduction. This leads to 5,000 iterations that are kept for further analyses. Convergence is checked

by trace plots (for example trace plots see Appendix B.2). The chains appear to have converged, as the mean and the variance of the chains seem to be stable with jumps large enough that can traverse the whole space.

6.2.2 Bias in estimation

For studying how BayMAP 2.0 is estimating the parameters, the bias is analyzed here for τ and $\boldsymbol{\beta}$ for the two settings with the large and the small $\boldsymbol{\beta}$. Since the true values of $\boldsymbol{\mu}$ and q are not known as explained in Section 6.1.2, they are not analyzed in this section.

In Figure 6.10 the bias of the mean estimate for τ is shown for the simulation settings with the large $\boldsymbol{\beta}$. It seems that τ is overestimated, when $\tau = 0.5$ or $\tau = 1$. However, this overestimation is here expected as discussed in the following.

In the simulation setting, it is first chosen randomly, if a cluster is a binding site. When a cluster is chosen as binding site, part of the T-to-C substitution positions of this binding sites are changed artificially. This means in particular that there is a group structure even if τ were chosen to be equal to zero. Consequently, τ is here expected to be overestimated, especially when τ is very small. The bias can therefore not be interpreted here as an ordinary bias, since true underlying τ is unknown.

However, the results show, that BayMAP 2.0 is able to detect group structures when group structures are present. In the simulation settings with the small $\boldsymbol{\beta}$ (see Figure C.8) results are similar but the overestimation of τ is larger. The reason for this is probably that the noise is larger in these data sets due to the smaller values for $\boldsymbol{p}$.

In Figure 6.11 the bias of the estimated means for $\boldsymbol{\beta}$ is represented for the large $\boldsymbol{\beta}$ and $\tau = 1$. It is noticeable, that β_0 is underestimated. This underestimation of β_0 can also be found in the other settings (see Figures C.9 - C.13).

An underestimation of β_0 would lead to an underestimation of the prior probability that a position is a binding site position that is neither lying on a 3'UTR nor lying on a CDS nor lying on a 5'UTR. It would, however, also lead to an underestimation of this prior probability for the single types of mRNA regions under the condition that

the other effects are well estimated, since this effect is composed by β_0 and the corresponding parameter for the region. This underestimation is probably due to the fact, that this smaller effect for the mRNA region can be equalized by the added random effect. For clusters that are not chosen as binding sites, there are no positions where the number of substitutions is changed artificially. This also means that there is somehow a group structure that can be illustrated by the random effect. Hence, this underestimation is not unexpected.

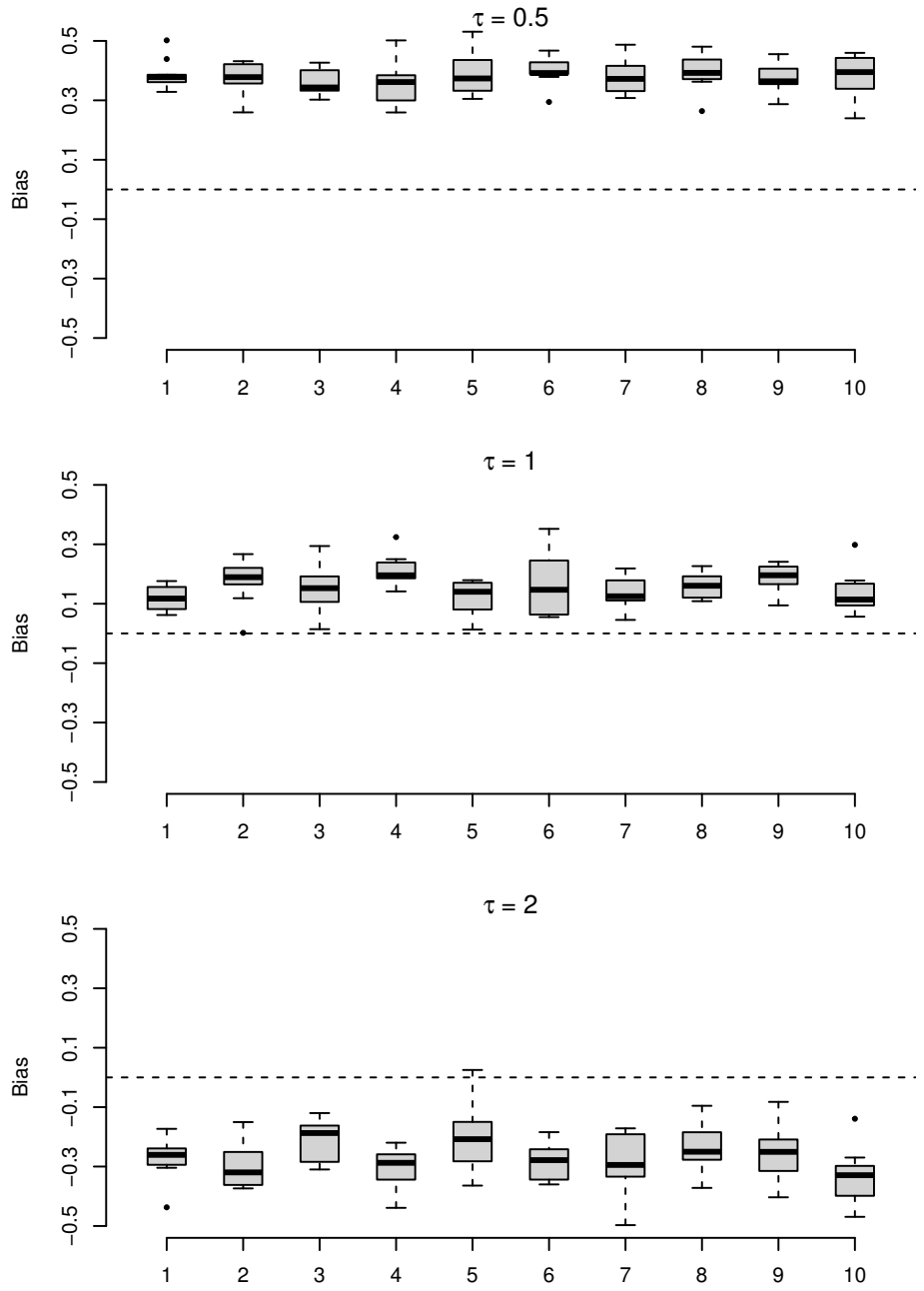

Figure 6.10: Bias of the mean estimates for τ considering ten draws of $\boldsymbol{\zeta}$ for each value of τ with $\beta_0 = 0.5$, $\beta_{3'\text{UTR}} = 1.85$, $\beta_{\text{CDS}} = 1.15$ and $\beta_{5'\text{UTR}} = 0.75$.

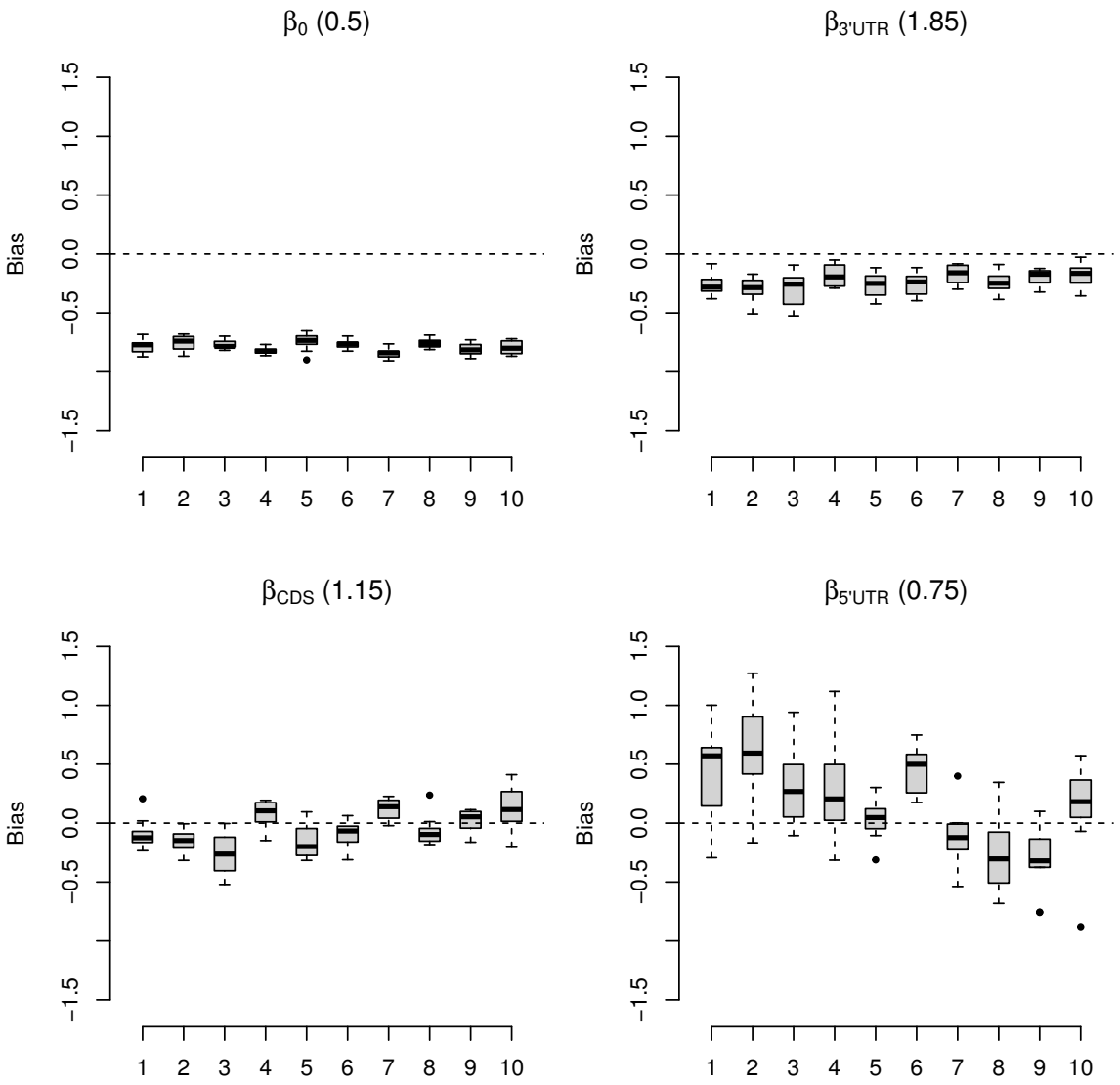

Figure 6.11: Bias of the mean estimates for the regression parameters in the probit model in BayMAP in relation to the mRNA position (3'UTR, CDS, 5'UTR) when $\tau = 1$. The true values of the parameters for β are shown in parentheses.

For $\beta_{3'\text{UTR}}$, β_{CDS} and $\beta_{5'\text{UTR}}$ small over- or underestimations can be detected in some settings (see in particular Figure C.11). In total, however, these parameters of $\boldsymbol{\beta}$ seem to be more or less unbiased. Even if the total effect of for example the 3'UTR is then underestimated, since it is composed by β_0 and $\beta_{3'\text{UTR}}$ as discussed above, the interpretation of these effects does not change. This is due to the fact that usually not the composed effect is analyzed, but only the effect itself, since a positive value for $\beta_{3'\text{UTR}}$, e.g., would mean in general that a method-induced substitution position is more likely on a 3'UTR than on a position that is neither on a 3'UTR, nor CDS nor 5'UTR.

6.2.3 Comparison of BayMAP 2.0 to BayMAP 1.0

Before comparing the simulation results of BayMAP 2.0 with other methods, it is of interest, whether BayMAP 2.0 performs well and better in comparison to BayMAP 1.0. Moreover, the simulation can be used to decide how to identify method-induced substitutions in BayMAP 2.0. As explained in Section 5.3.4 it could not only be reasonable in BayMAP 2.0 to estimate the posterior probability that position i has method-induced substitutions but also to use the prior odds as criterion. These two criteria are thus compared in this section. For estimating the posterior probability that position i is crosslinked, here the fraction of how often Z_i is sampled as experimentally induced in the MCMC chain.

In Figure 6.12 the distribution of the accuracy, the sensitivity and specificity is shown for the settings with the large $\boldsymbol{\beta}$, the beta distributed μ_{exp} and $\tau = 1$. The accuracy, sensitivity and specificity is always slightly higher for BayMAP 2.0 compared to BayMAP 1.0 when using the estimated posterior probabilities as criterion for BayMAP 2.0. This is also true for the other simulation settings (see Figures C.14 - C.24).

Looking at the results obtained with the prior odds in Figure 6.12, it stands out that, on the one hand, the sensitivity for the prior odds is better than for the ordinary criterion of BayMAP 2.0 with values mostly larger than 0.85, whereas the values for the ordinary BayMAP 2.0 are mostly only over 0.8. On the other hand, the results of the specificity for the prior odds vary largely with values between 0.77 and 0.96 whereas the ordinary BayMAP 2.0 has always very high values between 0.97 and 0.98. The values for the accuracy for the prior odds are larger here, since most of the positions are method-induced because of the large $\boldsymbol{\beta}$, so that the accuracy is mostly determined by the sensitivity.

The prior odds seem to have problems in terms of specificity in all settings with a large $\boldsymbol{\beta}$, in particular when τ is small (see Figures C.14, C.22 - C.24). On the contrary, in the settings with the no effect $\boldsymbol{\beta}$, the prior odds outperform the ordinary BayMAP 2.0 (see Figures C.19 - C.21). Results are comparable in the settings in which $\boldsymbol{\beta}$ is small (see Figures C.16 - C.18).

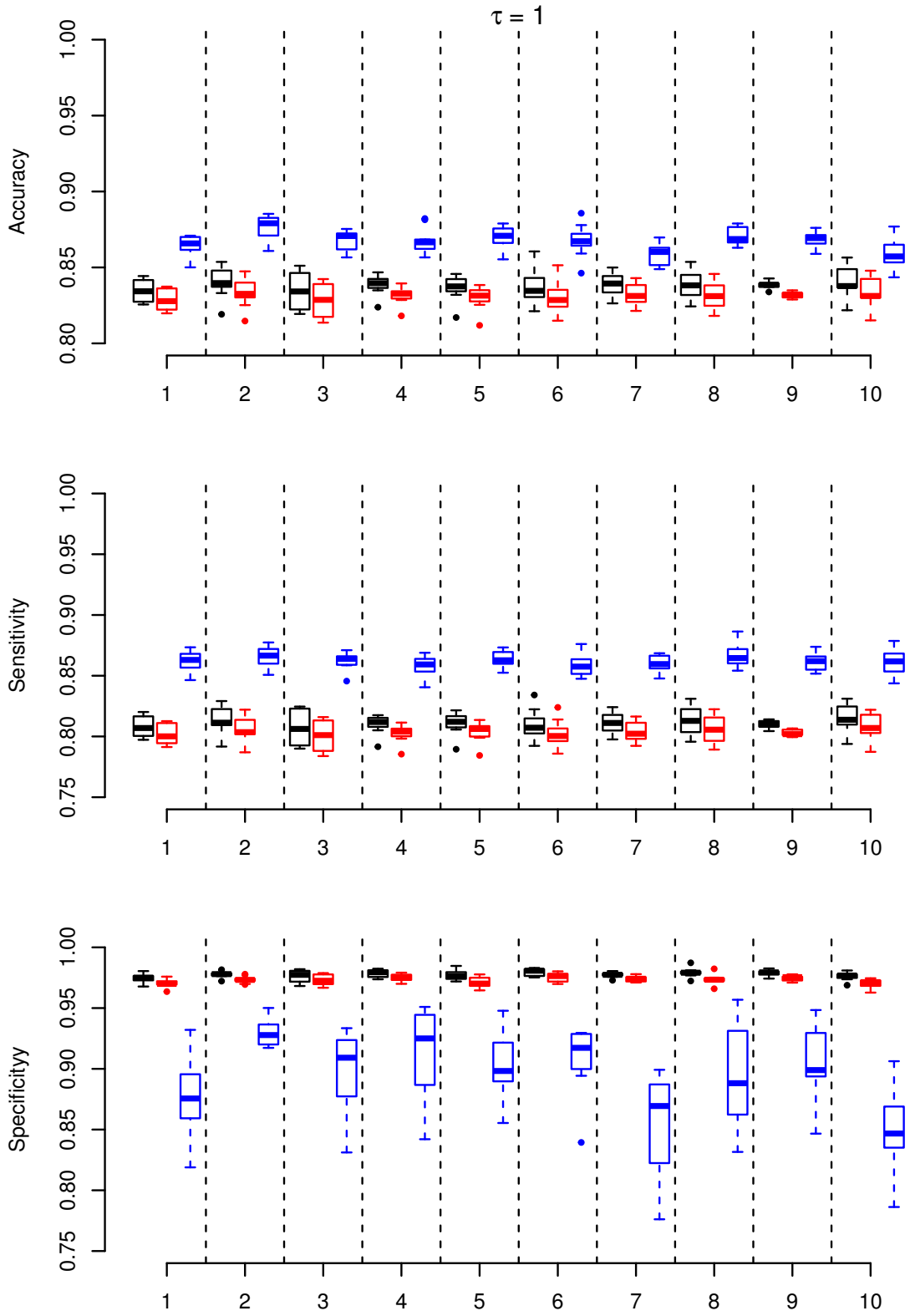

Figure 6.12: (Simulation $\tau = 1$ and large $\boldsymbol{\beta}$) Distribution of the accuracy (top panel), sensitivity (middle panel) and specificity (bottom panel) of BayMAP 2.0 (black box plots for each different simulation for ζ), BayMAP 1.0 (red), BayMAP 2.0 prior odds (blue) considering ten draws of $\boldsymbol{\zeta}$ for $\tau = 1$ with $\beta_0 = 0.5$, $\beta_{3'\text{UTR}} = 1.85$, $\beta_{\text{CDS}} = 1.15$ and $\beta_{5'\text{UTR}} = 0.75$.

It seems, thus, that the ordinary BayMAP 2.0 has a small advantage over BayMAP 1.0 and that the prior odds only have reliable results in some circumstances, in particular when $\boldsymbol{\beta}$ has no effect and/or when the variance τ^2 is large enough. In the other settings, the prior odds fall behind in terms of specificity with differences up to 0.3 (see Figure C.22). The specificity is here of special interest, as a high specificity ensures that the detected crosslinked T-to-C substitution positions are really method-induced. Some method-induced substitution positions may be missed depending on the sensitivity. Nevertheless, with a high specificity the number of detected positions can be highly reduced, so that for these positions further experiments can be conducted.

All in all, the ordinary BayMAP 2.0 seems to perform well, in particular when regarding the specificity, that is in the range of 0.97 to 0.98 in all settings. The values for the sensitivity are not as large as for the specificity with the lowest values between 0.65 and 0.7 in the setting with the small $\boldsymbol{\beta}$ and $\tau = 0.5$ (see Figure C.16) and the largest values around 0.9 in the settings with $\mu_{\text{exp}} = 0.2$ (see Figures C.22 - C.24).

These smaller values in sensitivity are somehow expected, since T-to-C substitution positions that do not have artificially changed substitutions but that lie on a binding site are here considered as crosslinked positions as explained in Section 6.2.1. Moreover, it is not necessary to detect every T-to-C substitution position on one binding site as method-induced, but it is sufficient to identify at least one of them.

If only one position is identified as crosslinked, this would nevertheless mean, that the whole cluster is declared as binding site. This will be seen in more detail in the next section, where BayMAP 2.0 is compared to wavClusteR and BMix. Because of the more stable results, only the ordinary BayMAP 2.0 is considered in the next section.

6.2.4 Comparison of BayMAP 2.0 to other methods

Position based comparisons

In the previous section, it was shown that BayMAP 2.0 leads to stable results and better values than BayMAP 1.0 in the here considered simulation settings with very high values of specificity. In Figure 6.13, BayMAP 2.0 is compared to wavClusteR and BMix for the simulation settings with the large $\boldsymbol{\beta}$, the beta distributed μ_{exp} and $\tau = 1$.

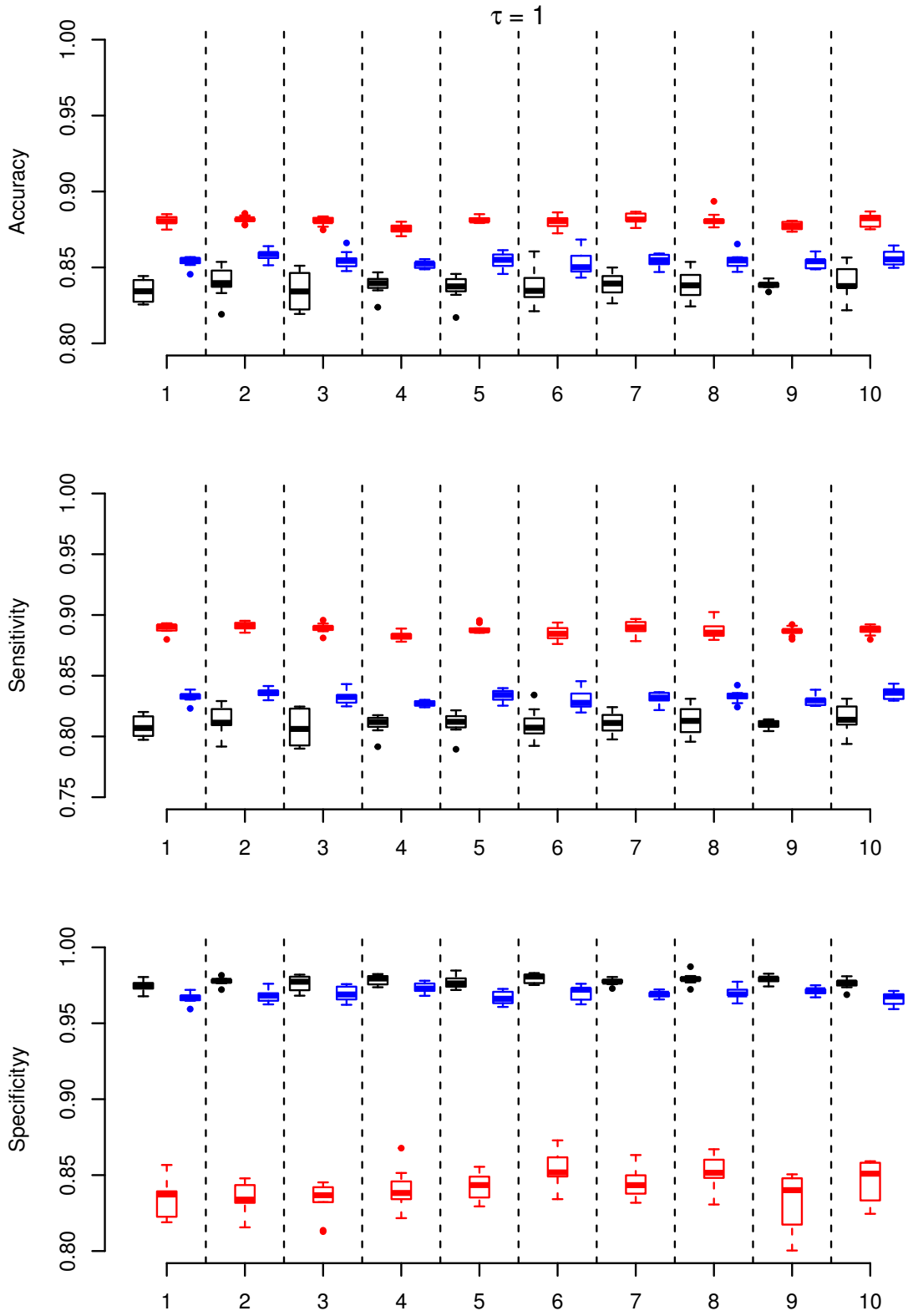

Figure 6.13: (Simulation $\tau = 1$ and large $\boldsymbol{\beta}$) Distribution of the accuracy (top panel), sensitivity (middle panel) and specificity (bottom panel) of BayMAP 2.0 (black box plots for each different simulation for ζ), wavClusteR (red), and BMix (blue) considering ten draws of $\boldsymbol{\zeta}$ for $\tau = 1$ with $\beta_0 = 0.5$, $\beta_{3'\text{UTR}} = 1.85$, $\beta_{\text{CDS}} = 1.15$ and $\beta_{5'\text{UTR}} = 0.75$.

It is first noticeable, that wavClusteR outperforms the other methods in terms of sensitivity but only achieves relatively small values in terms of specificity. This is even much more striking in the settings, in which $\boldsymbol{\beta}$ is small (see Figures C.27 - C.29) and in which $\boldsymbol{\beta}$ has no effect (see Figures C.30 - C.32). The values of sensitivity for wavClusteR are here mostly very close to one, whereas the values of specificity are mostly very close to zero, so that they are not even shown in the figures for a better graphical representation. Only in the settings, in which μ_{exp} is set to 0.2, the values for the specificity of wavClusteR are comparable to those of BayMAP 2.0 and BMix (see Figures C.30 and C.31). However, even in these settings, wavClusteR falls behind in terms of specificity, when τ is too large, i.e. $\tau = 2$ (see Figure C.30). In the settings, in which μ_{exp} is beta distributed with the large $\boldsymbol{\beta}$, it can also be observed, that wavClusteR's specificity diminishes with a larger τ (compare Figures C.25 and C.26).

In Figure 6.13 it seems that wavClusteR performs nevertheless better than BayMAP 2.0 and BMix when regarding the accuracy. However, these values are misleading, as the accuracy is here mostly determined by the sensitivity because of the large $\boldsymbol{\beta}$ that leads to much more crosslinked substitution positions than non-crosslinked ones. In Figures C.30 and C.31 wavClusteR's accuracy is even comparable to the other methods despite the very small values of specificity only because nearly all of the positions are estimated to have method-induced substitutions.

Thus, wavClusteR only seems to perform comparable to the other methods, if a distinction of method-induced substitutions to non-method-induced ones is not so difficult because of a fixed μ_{exp}, a large $\boldsymbol{\beta}$ and a small variance τ. In most settings, the results of wavClusteR are here however not reliable, since most positions are chosen to be method-induced, which leads to a high sensitivity but a very small specificity. This means that most of the T-to-C substitutions are detected as crosslinked when using wavClusteR with a high number of false positives. The aim of reducing the number of detected binding sites to only those binding sites that are very likely true binding sites, can therefore not be reached in a reliable way when using wavClusteR.

The results of BayMAP 2.0 and BMix are much more comparable as they are close to each other in all settings. It is remarkable that BMix outmatches BayMAP 2.0 in al-

most all settings in terms of sensitivity whereas BayMAP 2.0 wins over BMix in terms of specificity in all settings (see Figure 6.13 and Figures C.25 - C.35). This leads to slightly larger values of accuracy for BMix in nearly all settings because of the larger sensitivity.

Binding site based comparisons

However, as explained in the previous section, it is not necessary for detecting a binding site to detect all of the crosslinked positions of one binding site, but sufficient to detect as least one. Therefore, the methods are also compared binding site based instead of position based in Figure 6.14 as explained in the following. For BayMAP 2.0, wavClusteR and BMix, first for every T-to-C substitution position it is decided if this position is crosslinked or not. For the position based decisions, binding sites are then declared for regions with at least one position that is declared as experimentally induced. Lets suppose that all methods identify correctly the cluster boundaries. The number of correctly identified binding sites and correctly identified non-binding sites over all potential binding sites can then be compared for the different methods instead of the number of correctly identified positions over all positions.

In Figure 6.14 the values for accuracy are compared for the settings in which $\boldsymbol{\beta}$ is large and μ_{exp} beta distributed. It is obvious, that on a cluster based comparison, wavClusteR does not reach the same accuracy as the other methods, especially when τ is high. BayMAP 2.0 performs better than the other methods with values of accuracies around 0.95. The difference in accuracy is more visible for larger τ. In the settings with small $\boldsymbol{\beta}$ (see Figure C.36) results are similar with a larger gap to wavClusteR.

The fact that BayMAP 2.0 has better values in terms of accuracy compared to wavClusteR and BMix when the comparison is binding site based, leads to the conclusion that BayMAP's smaller position based sensitivity has an advantage over the other methods. This is due to the fact that only one detected crosslinked position is enough to declare a binding site, so that it is easy to falsely identify binding sites. With BayMAP's high specificity, it is less likely to falsely detect a binding site. On the other hand, BayMAP's sensitivity is large enough, that probably at least one position is identified correctly, so that the binding site based sensitivity is as good as the one for BMix and wavClusteR.

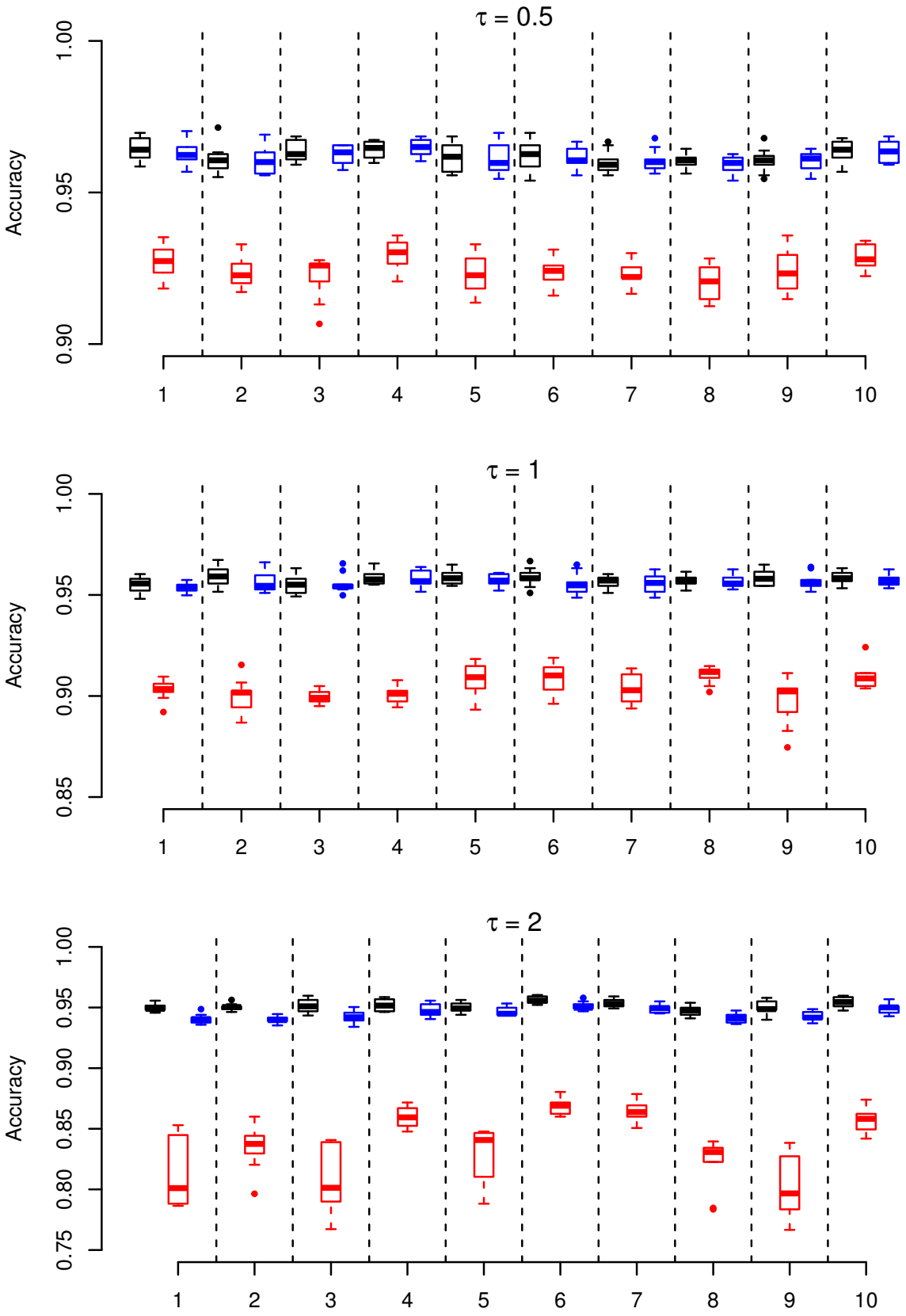

Figure 6.14: (Simulation binding site based with $\tau = 1$ and large $\boldsymbol{\beta}$) Distribution of the accuracy binding site based of BayMAP 2.0 (black box plots for each different simulation for ζ), wavClusteR (red), and BMix (blue) considering ten draws of $\boldsymbol{\zeta}$ for each value of τ with $\beta_0 = 0.5$, $\beta_{3\text{'UTR}} = 1.85$, $\beta_{\text{CDS}} = 1.15$ and $\beta_{5\text{'UTR}} = 0.75$.

Simulations with no binding site effect

It is also of interest to investigate the performance of BayMAP 2.0, when there is no difference in potential binding sites, i.e. the variance of $\tau^2 = 0$. The results of BayMAP 2.0 are added to the setting of Section 6.1 in which μ_{exp} is drawn from a Beta(2,10) distribution. Clusters to which the positions belong, are drawn from the one inflated zero truncated Poisson distribution as presented in Section 6.2.1. Figure C.37 is basically the same figure as Figure C.7 with the only difference, that results for BayMAP 2.0 are added.

In terms of sensitivity (see Figure C.37), BayMAP 2.0 is not performing as good as the other methods, whereas the specificity is larger than for the other methods. The values between the accuracy and the sensitivity do not differ much, since most of the positions are chosen as true experimentally induced substitution positions. The results show that even when no cluster dependencies are present, BayMAP 2.0 reaches still a high sensitivity close to 0.9 and a very high specificity around 0.95.

Comparisons dependent on part of crosslinked positions per binding site

In the simulation settings in this section, first it is chosen if cluster j is a binding site with probability p_j. For a chosen binding site, it is then drawn randomly for each position with probability p_j, if this position has method-induced substitutions or not. This means, in particular, that only a part, i.e. between 0 and 100%, of the positions of one binding site have PAR-CLIP induced substitutions. It could therefore also be of interest to analyze the performance depending on the fraction of method-induced substitution positions of a binding site to the aim of learning how many positions should have method-induced substitutions to predict in a reliable way if a position is crosslinked.

In the simulation study in this section, ten different draws for $\boldsymbol{\zeta}$ were generated for each setting with then ten simulated data sets for each draw of $\boldsymbol{\zeta}$. In none of the settings, an important difference between the ten draws for $\boldsymbol{\zeta}$ could be detected. This leads to the conclusion that there is not much loss of information, if analyzing the simulated data sets for the ten draws of $\boldsymbol{\zeta}$ altogether for one setting. As one setting consists of ten draws for $\boldsymbol{\zeta}$ with ten simulated data sets for each $\boldsymbol{\zeta}$, this combination results in one hundred

simulated data sets for each setting, that can be analyzed together. If one is interested in the performance depending on the fraction of crosslinked substitution positions of a binding site, the here described combination of the data sets ensures that there are observations even for fractions that are not very frequent.

In Figures 6.15 and C.38 for the large and the small $\boldsymbol{\beta}$, the sensitivity is displayed depending on the part of experimentally induced T-to-C substitution positions over all T-to-C substitution positions for a binding site. This range of zero to one for the fraction is divided into twenty intervals of equal length, so that for example all binding sites having less than five percent of crosslinked T-to-C substitution positions, i.e. in the interval $[0.00, 0.05)$, are grouped together.

For every binding site it is regarded if at least one of the T-to-C substitution positions is declared as method-induced. For each of the twenty groups, the true positive rate, i.e. sensitivity, is calculated. Note that in some intervals only binding sites with the same rate of experimental induced T-to-C substitutions belong to the group, e.g., the interval [0, 0.05) only consists of binding sites with zero experimental induced T-to-C substitutions and the interval [0.95, 1.00] only consists of experimentally induced T-to-C substitution positions. Even though only true binding sites are considered here, it is nevertheless preferable, when the method does not detect the true binding site if none of the T-to-C substitution positions have experimentally induced substitutions, since the data does not indicate the presence of a binding site in this case.

For BayMAP 2.0 and BMix the true positive rate for this group with zero PAR-CLIP induced substitutions is very low with values under 0.1, where BayMAP 2.0 reaches slightly smaller values than BMix. Even when no method induced substitutions are present, wavClusteR has high true positive rates, especially in the setting with the smaller $\boldsymbol{\beta}$ (see Figure C.38), where nearly all positions are chosen as method induced. This again indicates that wavClusteR has a high false positive rate and therefore a small specificity when nearly all of the non-changed T-to-C substitution positions are declared as method induced.

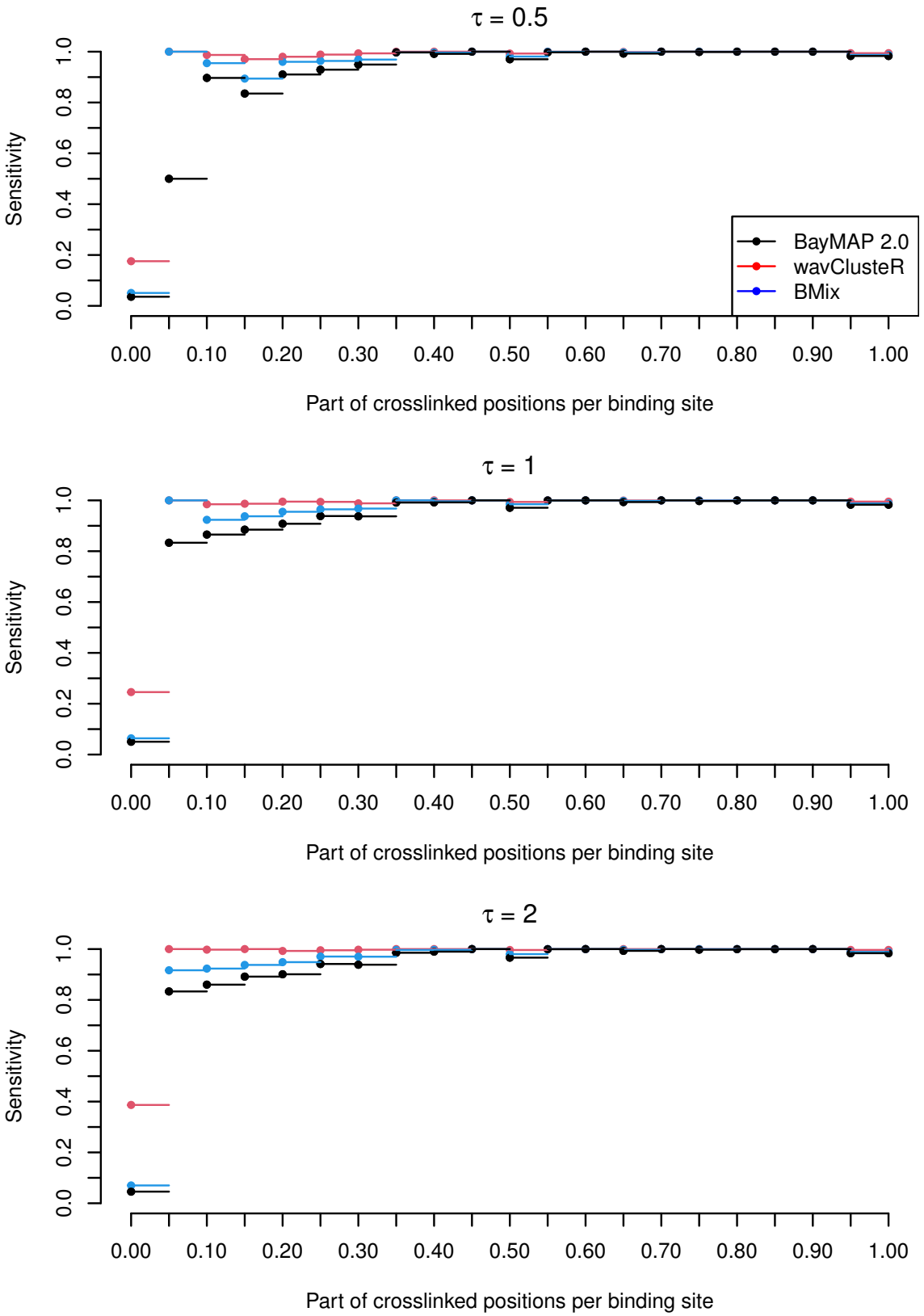

Figure 6.15: Distribution of the sensitivity of BayMAP 2.0 (black), wavClusteR (red), and BMix (blue) depending on the part of crosslinked positions per binding site for each value of τ with $\beta_0 = 0.5$, $\beta_{3'\text{UTR}} = 1.85$, $\beta_{\text{CDS}} = 1.15$ and $\beta_{5'\text{UTR}} = 0.75$. The x-axis is divided into intervals each of the length 0.05 (displayed by lines), where the left boundary of each interval is included (displayed by the point).

If at least one position has method-induced substitutions, all three methods perform very well with true positive rates over 0.8 with one exception for BayMAP 2.0 for the interval [0.05, 0.10) for $\tau = 0.5$ and the larger $\boldsymbol{\beta}$. However, there are only two binding sites belonging to this group, so that this outlier can be explained.

In general the true positive rate gets the closer to one, the higher the rate of PAR-CLIP induced substitution positions is. This result is somehow expected, since it should get easier to detect a binding site the more method-induced substitution positions the binding site has. However, some artifacts are noticeable, for example slightly smaller true positive rates for the intervals [0.50, 0.55) and [0.95, 1.00] in comparison to their neighbor intervals. This can be explained by the fact that many of the binding sites with all of the positions having method induced substitution positions only consist of one T-to-C substitution position and many of those with 50% only consist of two substitution positions. Because of this smaller amount of T-to-C substitution positions with only one position that has actually method-induced substitution, the detection is more difficult than for the neighbor intervals with more T-to-C substitution positions.

BayMAP 2.0 has always a slightly smaller true positive rate than the other methods but nevertheless very high true positive rates when at least one of the positions has method-induced substitutions and a very small true positive rate when no substitutions are PAR-CLIP induced. This again proves BayMAP's high capacity of not falsely detecting binding sites.

6.3 BayMAP combining several PAR-CLIP data sets

Sometimes, several PAR-CLIP data sets exist and one wish to combine the results in order to receive more reliable predictions. In the previous sections of this chapter, it was already verified via an extensive simulation study that BayMAP 1.0 and BayMAP 2.0 perform well in comparison to other methods. In this section, the improvement of BayMAP when combining the results of several PAR-CLIP data sets should be shown via simulation.

First the set up of the main simulation study is described (see Section 6.3.1). The bias does not have to be analyzed for the simulation study of the combined results, as this

method takes as input the results of several separate PAR-CLIP experiments obtained by BayMAP 1.0 or BayMAP 2.0, where the bias is already analyzed in the respective sections. For the simulation study with the combined results it is analyzed if the combination of several BayMAP results leads to an improvement in Section 6.3.2. Results are not compared to wavClusteR and BMix, as they are already compared in the respective sections.

6.3.1 Set up of simulation study

The data is generated in a similar way than in the previous Section 6.2.1. As BayMAP's general ability for detecting method-induced substitutions has already been proven, and as the aim is now only to investigate if there is still room for improvement when the results of several data sets are combined, here, only one setting with $\tau = 1$, the small $\boldsymbol{\beta}$ and the beta distributed μ_{exp} is considered.

First, clusters to which the positions belong, are drawn by the one inflated zero truncated Poisson distribution first presented in this thesis (see (6.1)). Then, it is drawn randomly with probability $p_j = \Phi(\boldsymbol{x_j}^\top \boldsymbol{\beta} + \alpha_j) = \Phi(\zeta_j)$ for every cluster j if the cluster is a binding site or not. Afterwards, based on this global data set, five different data sets are created. For each of the five data sets, only 80% of the clusters are selected randomly, so that not every cluster is represented in every data set. For the remaining binding sites it is then drawn randomly for every T-to-C substitution position of the binding site and for each of the five data sets separately if the position has method-induced substitutions and therefore artificially changed substitutions.

This means, that for a pair of those data sets, the probability is equal to 0.64 that both data sets contain cluster j. If both data sets contain cluster j, this cluster is either in both data sets a true binding site or in none of them. If cluster j is a binding site, the positions that have artificially induced substitutions could be different, as not all positions have to be crosslinked. Even if a position is chosen in both data sets to have experimentally induced substitutions, the number of substituted reads is probably different for the first and the second data set.

The above explained procedure is replicated twenty times, so that twenty times five

data sets are created. The total number of iterations for the application of BayMAP 2.0 and BayMAP 1.0 is here 45,000. The first 15,000 iterations of these 45,000 are discarded as burn-in. From the remaining iterations, every sixth iteration is used to the aim of autocorrelation reduction. This leads to 5,000 iterations that are kept for further analyses. Convergence is checked by trace plots. The chains appear to have converged, as the mean and the variance of the chains seem to be stable with jumps large enough that can traverse the whole space. Results are obtained by R.

6.3.2 Analysis of simulated results

In order to verify if the detection of binding site positions is getting better with a higher amount of PAR-CLIP data sets, first the accuracy, sensitivity and specificity are calculated for each of the five data sets separately. These values are then compared to the results of the combined posterior odds, where between two to five results could be combined.

The accuracy, the sensitivity and the specificity depending on the number of combined results are represented in Figure 6.16. Note, that for the combined results for two or more data sets, not all possible combinations are considered. This means, that for example for calculating the accuracy in the case of four data sets, only positions are considered, that are present in exactly four data sets.

In Figure 6.16, there is an improvement in terms of accuracy visible for each additional number of combined data sets. When having a closer look to the sensitivity and specificity, it is first remarkable, that the specificity seems not to get better with a higher amount of considered data sets. The mean value of specificity is in all five cases 0.98.

On the other hand, there are big differences in the sensitivity. The mean value of sensitivity for five combined data sets is almost equal to 0.9, whereas the mean value of sensitivity without combination is here only almost equal to 0.7. The highest jump can be observed from no combination to two combined data sets, where the mean value of sensitivity is almost equal to 0.8. Similar results are obtained when BayMAP 1.0 is applied to the data sets (see Figure C.39).

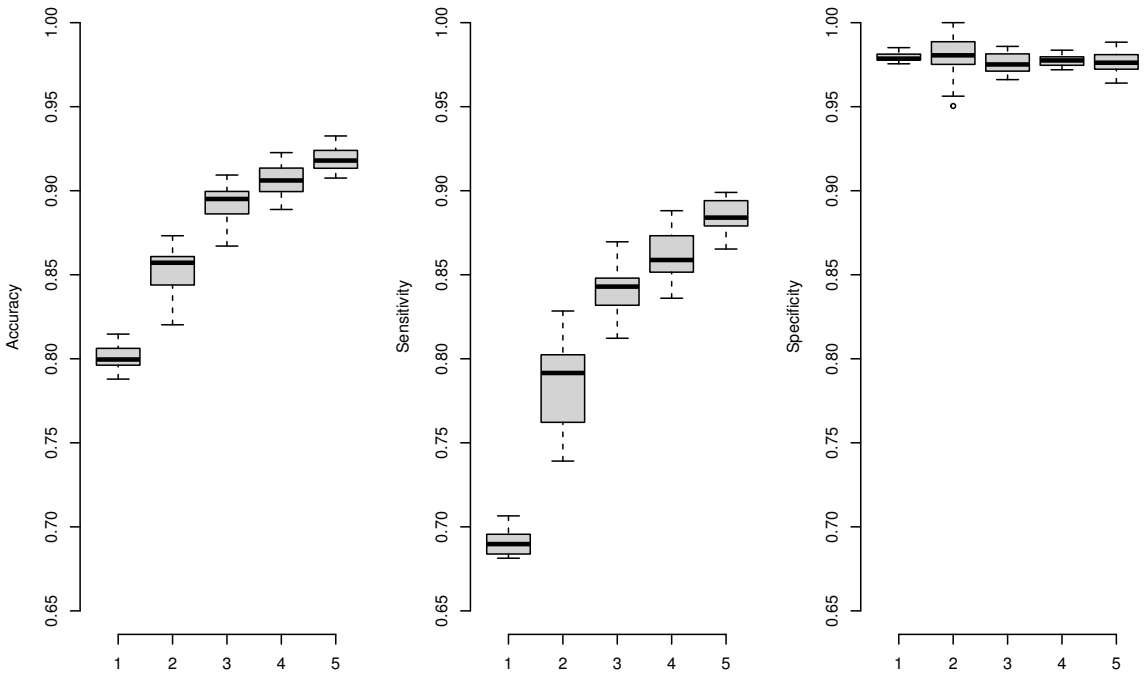

Figure 6.16: The accuracy (left panel), sensitivity (middle panel) and specificity (right panel) for the combined post odds of BayMAP 2.0 with one to five combined results, where one combined result means that only the results on one data set without combination are regarded.

As shown here as well as in the previous simulations and applications to PAR-CLIP data sets, BayMAP's strength lies in the high specificity, i.e. in not falsely declaring a position as binding site position. However, a high specificity can go along with a smaller sensitivity and therefore a higher amount of undetected binding site positions (false negatives). When having several PAR-CLIP data sets, it seems that BayMAP's sensitivity can be highly improved without loosing in specificity.

Chapter 7

Application to PAR-CLIP data sets

Five publicly available PAR-CLIP data sets are analyzed in this section to validate the performance of BayMAP with real PAR-CLIP data sets. The preprocessing of the data sets as well as a descriptive analysis are given in Section 2.2.2 and Section 2.2.3.

This section is divided into two parts. Section 7.1 focuses on the application of the different presented versions of BayMAP and the results on its own. In the second part in Section 7.2, the results of BayMAP 1.0 and BayMAP 2.0 are analyzed in terms of detected T-to-C substitution positions and compared to other methods for the analysis of PAR-CLIP data.

7.1 Application of BayMAP

In this section, the application will be described and parameter estimates analyzed first for BayMAP 1.0 (see Section 7.1.1) and then for BayMAP 2.0 (see Section 7.1.2). In addition, application results of the alternative for BayMAP 2.0, the intrinsic CAR model, that is represented in Section 5.3, will be shown (see Section 7.1.3). For those of the five data sets for which a replicate exists, the here presented method for combining the results (see Section 5.4) is then applied and analyzed in Section 7.1.4.

7.1.1 Application of BayMAP 1.0

The represented results here are obtained by applying BayMAP 1.0 employed in R [47] combined with WinBUGS [50]. The corresponding R package will be available online

[26]. The results in this Section 7.1.1 are the same results as have been published in Huessler et al. [27].

The total number of iterations for the application of BayMAP 1.0 is here 75,000. The first 7,500 iterations of these 75,000 are removed as burn-in, since the MCMC chains first have to be converged. To the aim of reducing autocorrelation, only every 15-th iteration is used of the remaining ones. This means that 4,500 iterations are kept for further analyses. Convergence of the chains and therefore of the estimation of the posterior distribution is checked by trace plots (see Appendix B.1). The chains appear to have converged, as the mean and the variance of the chains seem to be stable with jumps large enough that can traverse the whole space.

For the analysis of additional information modeled by the probit model, three indicator variables are considered for the type of the mRNA region, namely the 3'UTR, the CDS and the 5'UTR. Thus, the value of the indicator variable for, e.g., the 3'UTR is given by

$$x_{i,\,3\text{'UTR}} = \begin{cases} 1, & \text{if } i \text{ is a T-to-C substitution position on the 3'UTR} \\ 0, & \text{otherwise} \end{cases} . \qquad (7.1)$$

The corresponding effect is defined as $\beta_{3\text{'UTR}}$. $x_{i,\,\text{CDS}}$ and $x_{i,\,5\text{'UTR}}$ are defined in the same manner. In the two data sets of Gottwein et al. [22] only five (Gottwein A) and six (Gottwein B) positions on the 5'UTR are observed, so that this variable is not considered in further analyses for these two data sets.

Estimates for the conditional densities are represented in Table 7.1. For each parameter and each data set the median is shown as well as the 95% credible interval in square brackets.

Table 7.1: Estimated parameters of BayMAP 1.0 for the Ago PAR-CLIP data sets, where the parameter estimate is the median of the values of the Markov chain as presented in Huessler et al. [27]. The 95% credible interval is shown in brackets.

Parameter	Kishore A	Kishore B	Memczak	Gottwein A	Gottwein B
μ_{exp}	0.2655 [0.2651, 0.2659]	0.3033 [0.3016, 0.3051]	0.1941 [0.1931, 0.1951]	0.6282 [0.6088, 0.6454]	0.6426 [0.6261, 0.6592]
μ_{mm}	0.0052 [0.0052, 0.0052]	0.0068 [0.0067, 0.0068]	0.0031 [0.0031, 0.0031]	0.0061 [0.0059, 0.0062]	0.0070 [0.0068, 0.0072]
μ_{SNP}	0.9845 [0.9845, 0.9845]	0.9797 [0.9795, 0.9799]	0.9908 [0.9907, 0.9908]	0.9818 [0.9813, 0.9822]	0.9791 [0.9785, 0.9796]
q	0.9314 [0.9300, 0.9327]	0.7846 [0.7785, 0.7907]	0.6476 [0.6424, 0.6532]	0.1358 [0.1272, 0.1444]	0.1308 [0.1227, 0.1394]
β_0	-0.5531 [-0.5694, -0.5363]	-0.8044 [-0.8496, -0.7592]	-0.2587 [-0.2953, -0.2228]	-1.1210 [-1.2460,-1.0030]	-1.1230 [-1.2490, -1.0040]
$\beta_{3\text{'UTR}}$	-0.0028 [-0.0261, 0.0176]	-0.1312 [-0.1941, -0.0667]	0.6819 [0.6322, 0.7311]	0.5387 [0.2379, 0.8374]	0.5309 [0.2474, 0.8123]
β_{CDS}	0.1259 [0.0942, 0.1572]	-0.0036 [-0.1257, 0.1271]	0.4271 [0.3452, 0.5101]	0.4221 [-0.0012, 0.8326]	0.3587 [-0.1132, 0.8026]
$\beta_{5\text{'UTR}}$	0.0695 [-0.0154, 0.1556]	0.5859 [0.2831, 0.8927]	0.2036 [-0.0097, 0.4278]	-	-
p_0	0.2901 [0.2845, 0.2959]	0.2106 [0.1978, 0.2239]	0.3979 [0.3839, 0.4118]	0.1311 [0.1064, 0.1579]	0.1307 [0.1058, 0.1577]
$p_{3\text{'UTR}}$	0.2892 [0.2852, 0.2934]	0.1746 [0.1630, 0.1870]	0.6637 [0.6514, 0.6761]	0.2800 [0.1939, 0.3768]	0.2767 [0.1972, 0.3666]
p_{CDS}	0.3346 [0.3243, 0.3457]	0.2096 [0.1758, 0.2471]	0.5670 [0.5356, 0.5981]	0.2417 [0.1286, 0.3845]	0.2228 [0.1087, 0.3674]
$p_{5\text{'UTR}}$	0.3143 [0.2848, 0.3454]	0.4141 [0.3021, 0.5321]	0.4781 [0.3949, 0.5680]	-	-

As discussed in Section 2.2.3, the parameter μ_{mm} for the probability of observing by error a substitution, such as T-to-C, is assumed to be very close to zero. This assumption is justified by the high amount of substitution rates very close to zero in all data sets (see Section 2.2.3). In all data sets the median of its estimated density is very small with values between 0.0031 and 0.0070 (see Table 7.1). The 95% credible intervals are also very close to the estimated median with at most 0.0002 of difference. This means, that μ_{mm} is very narrowly distributed around its estimated median. Since all estimated values for μ_{mm} are very close to zero, the estimates seem to be reasonable. As μ_{SNP} is calculated as $1-3\mu_{\text{mm}}$, these values are as expected very close to 1 and hence also reasonable.

The probability that a read of a binding site position has the specific T-to-C substitution is μ_{exp}. In contrast, for μ_{mm} and μ_{SNP} it is not directly known which estimates are expected. The probability μ_{exp} could even vary among PAR-CLIP experiments, as each experiment is different. This diversity can be seen in the median estimates. In the data sets of Kishore et al. [30] and Memczak et al. [41], μ_{exp} varies around 0.2 and 0.3 whereas the estimates are around 0.63 or 0.64 in the two data sets of Gottwein et al. [22].

An explanation for this phenomenon could be that there is a high number of substitution rates close to one in the data sets from Gottwein et al. [22] (see Section 2.2.3). The relative high number of substitution rates close to one in comparison to substitution rates close to zero are partly due to the chosen threshold of a minimum coverage per position equal to five instead of twenty. However, when applying BayMAP 1.0 to the same data sets but with a minimum coverage of twenty, the median estimates for μ_{exp} are even slightly larger than before , so that this argument does not hold. The differences are therefore probably due to the differences in the PAR-CLIP experiments.

The parameter q represents the probability that a non-experimentally induced substitution position is a mismatch position. The median estimate for q vary between 0.13 (Gottwein B) and 0.93 (Kishore A). The estimates for q are very small for the data sets from Gottwein et al. [22] in comparison to the estimates for the other data sets. This is partly due to the smaller minimum coverage threshold. The estimates were around

Table 7.2: Estimated values for q for the AGO PAR-CLIP data sets using the naive approach as well as BayMAP 1.0

	Kishore A	Kishore B	Memczak	Gottwein A	Gottwein B
Naive approach	0.96	0.74	0.64	0.10	0.09
BayMAP 1.0	0.93	0.78	0.65	0.14	0.13

0.25 if a threshold of twenty was applied.

In order to verify if these values are reasonable, a naive approach can be used to estimate q as explained in the following. The parameter q is only relevant for non-experimentally induced substitution positions. It is supposed that substitution positions that are not T-to-C substitutions do not contain experimentally induced substitutions. These non-T-to-C substitution positions can, hence, be considered for the naive approach. If the number of mismatch and SNP positions were known, one could divide the number of mismatch positions by the sum of mismatch and SNP positions to estimate q. Since these numbers are unknown, they can be estimated. As discussed in Section 5.1.1, it is assumed that $\mu_{\text{SNP}} = 1 - 3\mu_{\text{mm}}$. It is also assumed, that $\mu_{\text{mm}} < \mu_{\text{SNP}}$. Under these restrictions μ_{mm} is smaller than 0.25 and μ_{SNP} is larger than 0.25. This value can be used as a threshold, that positions with a substitution rate smaller than 0.25 are assigned to be mismatch positions, whereas all other positions are assigned to be SNP positions. Then, the number of mismatch positions can be divided by the sum of the amount of mismatch and SNP positions.

These naive estimates for q are represented in comparison to the median estimate using BayMAP 1.0 in Table 7.2. The estimates for the naive approach for q and those obtained by BayMAP 1.0 have at most a difference of 0.04. This shows that the BayMAP 1.0 estimates for q seem to be reasonable.

Of more interest than q is the prior probability that position i has method-induced substitutions with $p_i = \Phi\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right)$, as one is interested here to distinguish crosslinked positions from non-crosslinked ones. The probability p_i is here supposed to be different depending on additional variables. As BayMAP is the first method for analyzing PAR-CLIP data allowing for the incorporation of additional information, the estimation

of $\boldsymbol{\beta}$ is of special interest.

First of all, it is remarkable that the estimated parameter distributions for the entries of $\boldsymbol{\beta}$ spread wider around the estimated median than it was the case for the other parameters, where the 95% intervals were very narrow. This means, that there is more uncertainty about the values of $\boldsymbol{\beta}$ than for the other parameters and it is more difficult to estimate it.

It was assumed that the effects of the different mRNA regions are positive with the highest effect for the 3'UTR, than for the CDS and the smallest positive impact for the 5'UTR. These expected positive effects cannot be seen in the data sets from Kishore et al. [30]. In data set B, even a small negative impact for the 3'UTR is supported with $\hat{\beta}_{3'\text{UTR}} = -0.1312$. In data set A the estimated effect is also negative with $\hat{\beta}_{3'\text{UTR}} = -0.0028$. However, these negative values are very close to zero. Moreover, the 95% credible interval includes zero, so that this effect does not seem to be important. The highest impact is estimated in data set A for the CDS with $\hat{\beta}_{\text{CDS}} = 0.1259$ and in data set B for the 5'UTR with $\hat{\beta}_{5'\text{UTR}} = 0.5859$. In contrast, the parameter estimates are as expected for the other data sets with the highest values for the impact of the 3'UTR than for the CDS and the smallest but positive impact for the 5'UTR.

As discussed in Section 2.2.3, the noise level in the data is probably the lowest in the data sets from Kishore et al. [30], where the expected impact of $\boldsymbol{\beta}$ could not be seen. The additional information could therefore be particularly important in PAR-CLIP data sets with a lot of noise, where it is more difficult to distinguish between method-induced and non-method-induced T-to-C substitution positions.

Estimates for p_i are also represented in Table 7.1. Since p_i is calculated as $p_i = \Phi\left(\boldsymbol{x_i}^\top \boldsymbol{\beta}\right)$ (see (5.8)), the probability $p_{3'\text{UTR}}$ is therefore equal to $\Phi\left(\beta_0 + \beta_{3'\text{UTR}}\right)$. It is noticeable that these estimates are mainly smaller than 0.5. This means that without knowing any data (i.e. number ob substitutions), the probability for a position having experimentally induced T-to-C substitutions is estimated being smaller than 50%. Only the estimates for the data set from Memczak et al. [41] are an exception, where the probability having method-induced substitutions for a position on the 3'UTR is even estimated to be around 66.37%.

One would assume that the T-to-C substitution positions do not contain a higher amount of non-method-induced substitutions than the highest number of substitutions for one of the other types of substitutions, e.g., T-to-A (see also Section 2.2.3). For the data sets from Kishore et al. [30] this would mean that at least 82.07% (Kishore A) and 74.50% (Kishore B) of the T-to-C substitution positions would be method-induced. For the data set from Memczak et al. [41] it would be at least 69.85% and for the data sets from Gottwein et al. [22] it would be at least 16.13% (Gottwein A) and 14.68%. The estimates for p_i seem therefore somehow reasonable for the data sets from Memczak et al. [41] as well as from Gottwein et al. [22]. However, the prior probability of having experimentally induced substitutions for a T-to-C substitution position would be expected to be much higher for the data sets from Kishore et al. [30]. This could be a sign that BayMAP 1.0 do not detect all method-induced T-to-C substitution positions. On the other hand, this could also be a sign, that the number of false positives could be very small, since only positions could be chosen that are likely crosslinked.

7.1.2 Application of BayMAP 2.0

Neighborhood information can be important to detect potential binding sites, as substitution positions close to each other belong probably to the same binding site. Consequently, the original BayMAP published in Huessler et al. [27] has been developed allowing in BayMAP 2.0 for the inclusion of read cluster or potential binding sites by random effects (see Section 5.3). For this purpose, read clusters are built and added to the data set prior to the application of BayMAP 2.0 as presented in Section 5.2. The application of BayMAP 2.0 is only meaningful, if the potential binding sites really contain more than one T-to-C substitution position, so that the information of the neighbor positions can be taken into account. Thus, the number of clusters and therefore of potential binding sites for each data set as well as the number of T-to-C substitution positions and the mean number of substitution positions per cluster are presented in Table 7.3.

The highest mean number of T-to-C substitution positions per cluster can be observed for the first data set from Kishore et al. [30] with a mean of 5.8. The mean number of T-to-C substitution positions is very low for the data sets from Gottwein et al. [22]

Table 7.3: Number of clusters and number of T-to-C substitution positions for the AGO PAR-CLIP data sets as well as the mean number of T-to-C substitution positions per cluster.

	Kishore A	Kishore B	Memczak	Gottwein A	Gottwein B
No. of clusters	14,685	2,318	4,675	767	777
No. of T-to-C positions	85,234	8,368	11,847	930	954
Mean of positions per cluster	5.80	3.61	2.53	1.21	1.23

with a mean of around 1.2. This implies that most of the clusters only contain one T-to-C substitution position. Again, this could be a sign for the high level of noise in the data sets from Gottwein et al. [22]. Consequently, the application of BayMAP 2.0 considering read cluster is more interesting for the other three data sets, where more T-to-C substitution positions are present per cluster on average.

The results here are obtained by applying BayMAP 2.0 employed in R [47]. The corresponding R package will be available online [26]. The total number of iterations for the application of BayMAP 2.0 is again 75,000, where the first 7,500 are removed as burn-in and only every 15-th iteration is used of the remaining ones. This means that 4,500 iterations are kept for further analyses. Convergence of the chains and therefore of the estimation of the posterior distribution is checked by trace plots (see Appendix B.2). The chains appear to have converged, as the mean and the variance of the chain seem to be stable and with jumps large enough that can traverse the whole space.

As random effects are now added to the model, a separate effect for each read cluster is estimated. One is here, however, not interested in the estimates of every single effect for the potential binding site, but in the variation of these effects. Therefore, the standard deviation parameter τ for the random effects is added to the table with estimates for the conditional posterior densities for every parameter for the five considered data sets (Table 7.4).

The parameter estimates for μ_{exp}, μ_{mm}, μ_{SNP} and q only differ very slightly in comparison to the results without consideration of read clusters. More different are the results of the estimations of $\boldsymbol{\beta}$.

Table 7.4: Estimated parameters of BayMAP 2.0 with random effects for the Ago PAR-CLIP data sets, where the parameter estimate is the median of the values of the Markov chain. The 95% credible interval is shown in brackets.

Parameter	Kishore A	Kishore B	Memczak	Gottwein A	Gottwein B
μ_{exp}	0.2778 [0.2776, 0.2779]	0.3030 [0.3026, 0.3034]	0.1941 [0.1938, 0.1944]	0.6148 [0.6123, 0.6173]	0.6377 [0.6347, 0.6403]
μ_{mm}	0.0052 [0.0052, 0.0052]	0.0068 [0.0067, 0.0068]	0.0031 [0.0031, 0.0031]	0.0061 [0.0060, 0.0061]	0.0070 [0.0070, 0.0070]
μ_{SNP}	0.9844 [0.9844, 0.9844]	0.9797 [0.9797, 0.9798]	0.9908 [0.9907, 0.9908]	0.9818 [0.9817, 0.9819]	0.9791 [0.9790, 0.9791]
q	0.9321 [0.9308, 0.9335]	0.7844 [0.7784, 0.7904]	0.6473 [0.6418, 0.6528]	0.1347 [0.1263, 0.1436]	0.1305 [0.1221, 0.1387]
τ	0.6416 [0.6226, 0.6614]	0.6335 [0.5635, 0.7079]	1.0770 [1.0040, 1.1549]	1.8432 [1.1191, 2.9177]	0.9111 [0.4882, 1.4919]
β_0	-0.7045 [-0.7329, -0.6775]	-0.9941 [-1.0722, -0.9211]	-0.5174 [-0.5898, -0.4465]	-2.4286 [-3.5984,-1.6931]	-1.5622 [-2.1311, -1.2313]
$\beta_{3'\text{UTR}}$	0.0728 [0.0398, 0.1057]	-0.0884 [-0.1827, 0.0059]	1.1176 [1.0168, 1.2207]	1.0778 [0.2870, 2.0778]	0.7562 [0.3201, 1.3295]
β_{CDS}	0.1843 [0.1349, 0.2352]	0.0282 [-0.1696, 0.2112]	0.6480 [0.4939, 0.8064]	0.6542 [-0.3815, 1.7929]	0.3965 [-0.2490, 1.0972]
$\beta_{5'\text{UTR}}$	0.1025 [-0.0394, 0.2490]	0.6904 [0.2309, 1.1637]	0.3850 [-0.0424, 0.8123]	-	-

However, the directions of $\boldsymbol{\beta}$ are mainly the same, so that for the data sets from Memczak et al. [41] and Gottwein et al. [22] the expected impacts of the 3'UTR, the CDS and the 5'UTR can be observed. In the data sets from Kishore et al. [30] the 3'UTR still has the smallest impact. It is interesting to see that the estimates for $\boldsymbol{\beta}$ differ now much more between the two data sets of Gottwein et al. [22]. This could be due to the fact, that the standard deviation of the random effects is nearly twice as high for the first data set ($\hat{\tau} = 1.8432$ in Gottwein A versus $\hat{\tau} = 0.9111$ in Gottwein B). In the data sets from Kishore et al. [30] this standard deviation parameter is estimated the smallest around 0.64. Even a relatively small standard deviation of around 0.64 is large enough to easily equalize the effects of $\boldsymbol{\beta}$. The random effect can hence play an important role for the prior probability p_i if a T-to-C substitution position has method-induced substitutions. Estimates for the conditional posterior densities of p_i are here not presented, as p_i is now different for every single read cluster.

In total, the parameter estimates for BayMAP 2.0 seem also to be reasonable. A comparison of BayMAP 1.0 and BayMAP 2.0 concerning the ability of distinguishing method-induced substitution positions from non-method-induced ones will be shown in Section 7.2.

7.1.3 Application of BayMAP with CAR

In Section 5.3 it is proposed to use random effects in order to take into account dependencies due to neighborhood. Not only the normal random effect model is presented there, that is called BayMAP 2.0, but also the intrinsic CAR model. Even though only results for the normal random effect model are shown in the chapters following Section 5.3, the intrinsic CAR model is also applied to a PAR-CLIP data set.

In order to apply the intrinsic CAR model, one has to decide which positions are defined to be neighbors and which are not. Here, only positions that lie on the same potential binding site are defined as neighbors. Two scenarios are regarded, one in which all positions on the same potential binding site are defined as neighbors and one in which no other T-to-C substitution positions lie in between two neighbor positions. The model is only applied to the data set of Memczak et al. [41] and only on the data

for chromosome 1, so that convergence can be reached in a shorter amount of time. Instead of the probit model, logistic regression is used, as it is already implemented in the R-package CARBayes. However, even after 46 million iterations, there does not seem to be convergence for the parameter α_i.

In Figures B.13 and B.14 trace plots are shown for the car model in which only T-to-C substitution positions on the same cluster are considered as neighbors, when no other T-to-C substitution positions lie in between these two positions. For the parameters q, τ and μ_{mm} convergence seems to be reached in Figure B.13. The chain of the parameter μ_{exp} does not seem to vary constantly around a certain value but that this value changes. However, it has to be noticed that this variation is still in a very small range with values between 0.1675 and 0.1700.

The parameters β_0 and $\beta_{3\text{'UTR}}$ seem to be correlated. When there are relatively small values for one parameter, there are relatively high values for the other parameter, so that at least for positions that lie on the 3'UTR, the effect is equalized. The estimation of $\boldsymbol{\beta}$ could be more difficult, as there is an additional parameter for every T-to-C substitution position, that is different for each position.

The estimated variance parameter τ^2 for this additional parameter α_i is estimated very close to zero with a value around 0.00004. As discussed in Section 5.3.1, α_i follows a normal distribution, where the mean parameter is equal to the mean of all α_s with $s \neq i$ that are in the neighborhood of position i, $i = 1, \ldots, N_{\text{TC}}$ and with variance τ^2 divided by the number of neighbor positions of position i. If the variance parameter is then very close to zero, it means that the parameter α_i is highly correlated to the parameters α_s from the same neighborhood. This high correlation leads to a high autocorrelation of the parameters of the parameter vector $\boldsymbol{\alpha}$ as can be seen in Figure B.14, for α_i with $i = 1043, 114, 448, 546, 26, 1014, 467, 586$, where the eight positions are drawn randomly out of all possible positions. For none of the eight examples, convergence seems to be reached and a high autocorrelation is visible even after 46 million iterations.

When having a look at the trace plots for the CAR model in which all T-to-C substitution positions of the same potential binding site are considered as neighborhood in Figures

B.15 and B.16, results are similar. Convergence seem to be reached for the parameters μ_{mm}, q and τ^2. Values for μ_{exp} seem to get smaller over time, even when the differences are very small, since the values lie in a range between 0.1665 and 0.1695. Again, there seem to be a correlation between $\beta_{\text{Intercept}}$ and $\beta_{3\text{'UTR}}$. Convergence cannot be reached for the eight positions for α_i.

Because of the high autocorrelation and therefore the high number of iterations that are needed to reach convergence, these car models do not seem to be a practicable solution. The solution, that is implemented in BayMAP 2.0, where each cluster has a separate random effect, appear to be the better answer to the problem.

7.1.4 Combining several PAR-CLIP data sets

In Section 5.4, a method is presented that enables the combination of the results of several PAR-CLIP data sets. The posterior odds given the data of all data sets can be estimated by taking the product of all Bayes factors and by multiplying this product by an estimation of the combined prior odds. The combined post odds are here applied to the two data sets from Kishore et al. [30] and to the two data sets from Gottwein et al. [22] with the results from BayMAP 2.0 (see previous section), since replicates are available.

The Kishore A data set consists of 85,234 T-to-C substitution positions, whereas the Kishore B data set only consists of 8,368 positions, which is nearly 10% of the T-to-C substitution positions of the first data set. 6,763 T-to-C substitution positions are present in both data sets, that is around 8% of the positions of the first data set and more than 80% of the second data set.

The number of T-to-C substitutions in the two Gottwein data sets are much more well-balanced. The first data set consists of 930 positions and the second of 954. 622 positions are T-to-C substitution positions in both data sets, that is around two third of the T-to-C substitution positions of each data set.

Since there are 6,763 T-to-C substitution positions, that are present in both data sets from Kishore et al. [30], there are 6,763 positions for which the combined post odds can

be calculated. Out of these 6,763 positions, there are 1,010 positions that are not assigned to the same group (method-induced or non-method-induced) when analyzing separately, that is 15% of the 6,763 positions. When applying the combined post odds, 83% of these 1,010 positions are assigned to the group the positions are assigned in the Kishore A data set. This can be explained by the fact that the Kishore A data set has a higher read depth, which results in a higher number of substitution positions. The higher read depth implies less incertitude so that there are more Bayes factors that are very small or very large so that the Bayes factor of the data set A outweighs the Bayes factor of data set B more often.

For the data sets from Gottwein et al. [22], there are 622 positions that are present in both data sets and thus 622 positions for which the combined post odds can be calculated. Out of these 622 positions, there are 55 positions not assigned to the same group when analyzing separately, that is nearly 9%. When applying the combined post odds, 42% ot these 622 positions are assigned to the same group as in the Gottwein A data set.

There is even one position that is assumed to have non-method-induced substitutions when analyzing the two data sets separately. When regarding the combined post odds, however, the position is assumed to have method-induced substitutions. This artefact is due to the fact, that in both data sets the Bayes factor is larger than 1 but the prior odds smaller than one and outweighing here the Bayes factor. When the combined post odds are regarded, the Bayes factor, and therefore the number of substitutions, gets more powerful, since the product of all Bayes factors is calculated, whereas only the mean of the prior odds is considered.

In Figure 7.1 the values of the logarithm of the post odds calculated for the Kishore A data set are ploted against the values of the logarithm of the Kishore B data set. For a better graphical representation only values in between the range of -200 and 200 are plotted for the Kishore B data set. As discussed above, the post odds of the Kishore A data set have a higher variation because of the larger read depth. It is noticeable that there is a linear dependence between the logarithm of the post odds of the first data set and of the second data set, i.e. the higher the post odds of the first data set are, the

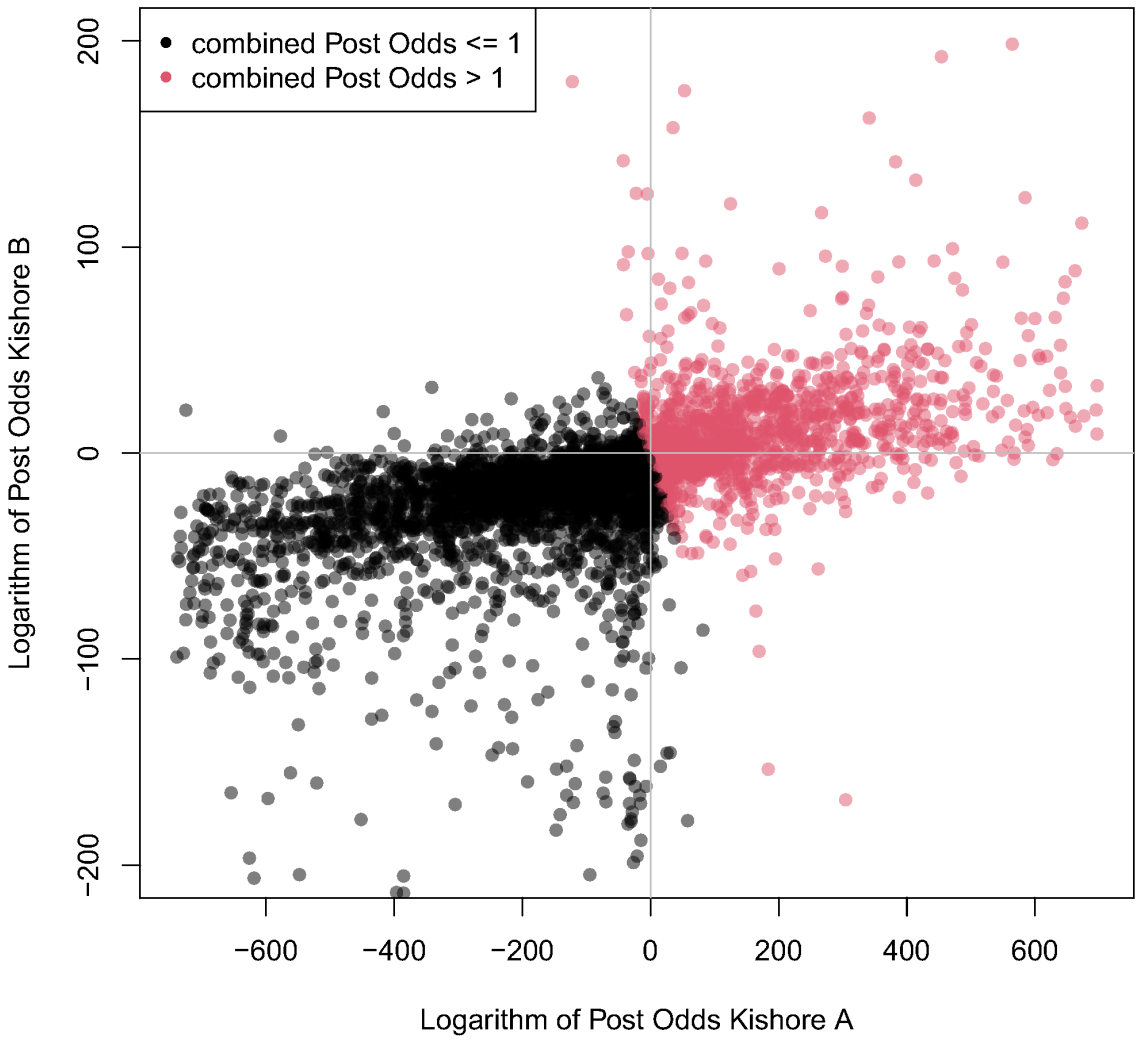

Figure 7.1: Scatter plot of the logarithm of post odds in the Kishore data sets.

higher they should be in the second data set. Red points in the figure represent the positions that are assumed to have method-induced substitutions based on the combined post odds, i.e. when the post odds are larger than one. Post odds larger than one are equivalent to the logarithm of post odds larger than zero. If a position is declared as method-induced or not is here mostly determined by the fact, if the logarithm of the post odds of the Kishore A data set are larger than zero due to the higher variation. There are, however, also positions that are determined by the post odds of the Kishore B data set (black points in the bottom right corner and red points in the upper left corner).

In Figure 7.2 the values of the logarithm of the post odds of the two Gottwein data sets are plotted against each other. Here, only points in between the range of −20 to 20 are

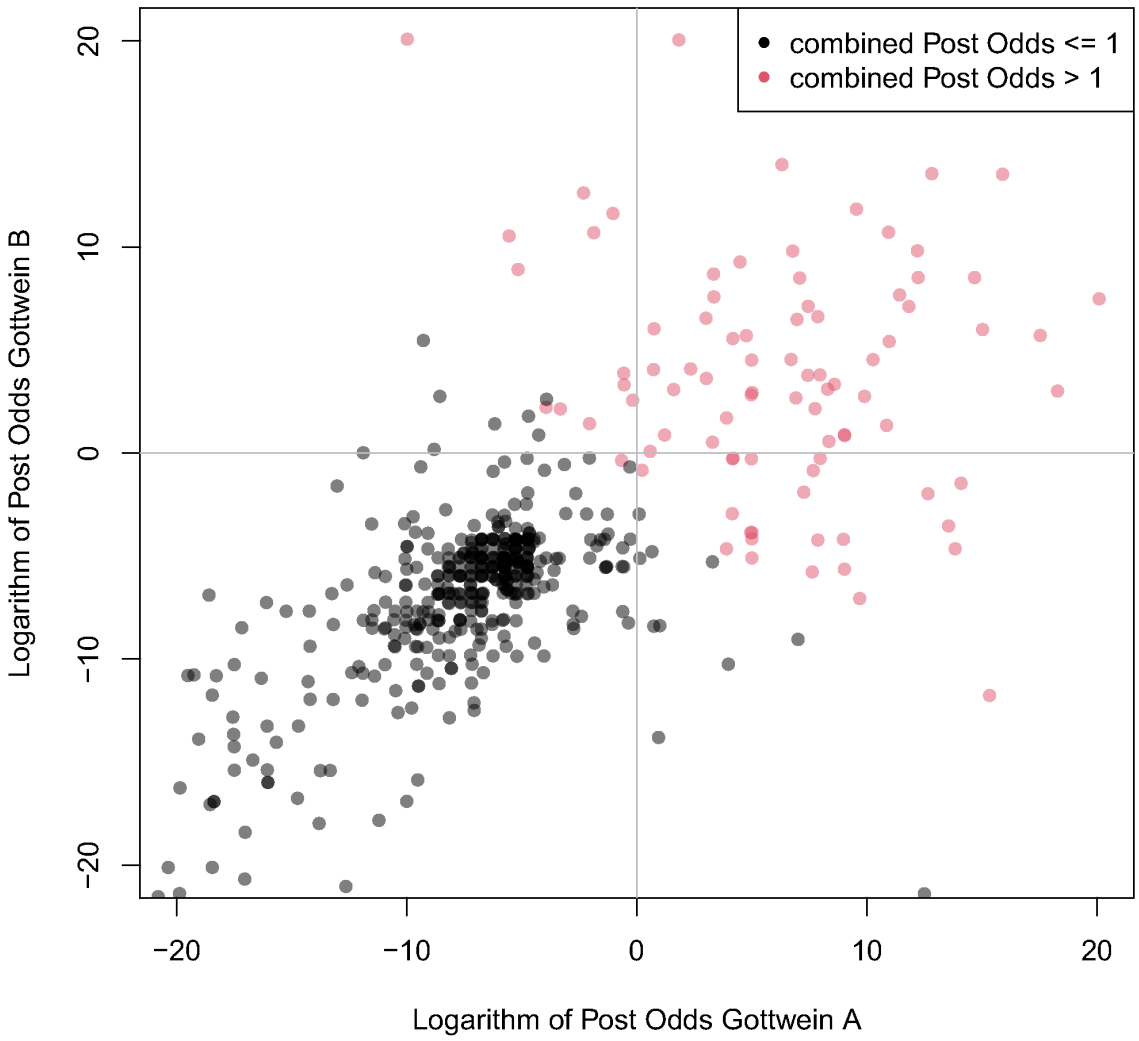

Figure 7.2: Scatter plot of the logarithm of post odds in the Gottwein data sets.

represented whereas there also exist very few values up to −600 or 600. There seems to be linear dependence with a slope of one, i.e. the logarithm of the post odds of the Gottwein B data set is expected to be more or less as high as the logarithm of the post odds of the Gottwein A data set.

The combined post odds are therefore particularly useful for positions for which the separate analyzes are ambiguous. Moreover, they even can change the results for one position, although the separate analyzes are unambiguous for this position as shown with the data sets from Gottwein et al. [22]. This is due to the larger impact of the data in comparison to the prior odds when combining results.

7.2 Comparison to other Methods

For the comparison of the obtained application results to other methods, first, in Section 7.2.1, the general set up is presented, how the methods are compared. Then, in Section 7.2.2, the results of BayMAP 1.0 and BayMAP 2.0 are analyzed and compared to BMix and wavClusteR. Finally, the detected T-to-C substitution positions of the different methods are compared to canonical and conserved targets of TargetScan [2] in Section 7.2.3 in order to further validate PAR-CLIP and BayMAP.

7.2.1 Set up of comparison

The main interest of BayMAP is the detection of method-induced T-to-C substitution positions. Thus, BayMAP 1.0 and BayMAP 2.0 are here analyzed in terms of detection of crosslinked and non-crosslinked positions, and compared to the methods on which BayMAP is based, namely wavClusteR and BMix. These methods predict if a T-to-C substitution position is method-induced or not. In contrast, PARalyzer takes as input aligned reads and predicts directly binding sites, i.e. regions of positions which makes a position-based comparison impossible. BayMAP 1.0 and BayMAP 2.0 are therefore compared to the position-based methods wavClusteR and BMix. Results for BayMAP 1.0, wavClusteR and BMix in this section were already published in Huessler et al. [27].

For each of the methods, i.e. BayMAP 1.0, BayMAP 2.0, wavCluster and BMix, it has to be specified, when a T-to-C substitution position can be declared as crosslinked. Here, the same probability cutoff of 0.5 is used for all methods [27]. This means that a position is declared as method-induced for a specific method, e.g., wavClusteR, if the probability of being crosslinked is estimated to be larger than 0.5 using this method.

Not only the number of as method-induced detected T-to-C substitution positions is of interest but also the number of falsely as method-induced detected ones. Obviously, these false positives cannot be drawn from the T-to-C substitution positions, as it is not known which of the positions have experimentally induced substitutions and which do not. However, other substitution types than T-to-C are not caused by the PAR-CLIP experiment, so that these positions can be taken to estimate the false positive rate as

proposed by Torkler [51].

The idea is to use the estimated model parameters to also predict for non-T-to-C substitution positions if the substitutions are crosslinked or not. Since it is assumed that the non-T-to-C substitution positions are not crosslinked, the number of as method-induced detected non-T-to-C substitution positions can be divided by the total number of all non-T-to-C substitution positions to the aim of estimating the false positive rate. BMix does not provide probabilities that a position has method-induced substitutions for positions other than T-to-C. However, these scores are here calculated separately by taking BMix' estimated model parameters for the considered data set.

BayMAP 1.0 is applied in WinBUGS, where the storage of all values of the latent variable $Z_i^{(h)}$ is not possible (see Section 5.1.4). The posterior odds (see Section 5.1.4) are therefore taken for the estimation if a position has method-induced substitution positions. Since BayMAP 2.0 is applied in R, the more naive approach of counting the number of times $Z_i^{(h)}$ = "exp" (see (5.39)) could be taken. However, this approach is only applicable for the T-to-C substitution positions as all other positions are never chosen to have experimentally induced substitutions in the MCMC chain. The posterior odds are therefore also taken for BayMAP 2.0.

As discussed in Section 5.1.4, the posterior odds can be calculated by multiplying the Bayes factor and the prior odds. The calculation of the prior odds depends in BayMAP 2.0 on the estimations of the random effects, so that prior odds are different for each cluster. When counting the number of as method-induced detected non-T-to-C substittuion positions, the random effects should be taken into account for a correct declaration. However, there are two main problems by the consideration of the random effects for substitution positions that are not T-to-C.

First, if the non-T-to-C substitution position is not close enough to a T-to-C substitution position, the position will not belong to a cluster, so that no random effects are estimated for these positions. Second, if the non-T-to-C substitution position belongs to a real binding site, this affiliation will ideally be reflected in a high score of the prior odds. Consequently, it will be more likely to predict a non-T-to-C substitution posi-

tion as method-induced when the probability is high for surrounding T-to-C substitution positions that they are method-induced. This dependency is volitional and should therefore not be used for the estimation of the false positive rate. For BayMAP 2.0, the random effects are therefore not taken into account, i.e. set to zero, when predicting if a non-T-to-C substitution position has method-induced substitutions.

7.2.2 Analysis of BayMAP and comparison to other methods

For each of the data sets and each of the methods the number of as method-induced detected T-to-C substitution positions as well as the number of as method-induced detected non-T-to-C substitution positions are presented in Table 7.5. The number of as method-induced detected non-T-to-C substitution positions in BayMAP 2.0 are as described above. Combined posterior odds for BayMAP 2.0 are also calculated, when possible. In addition to the position based decisions of BayMAP 2.0 with the posterior odds, the prior odds (see (5.67)) are used for deciding whether all positions of a cluster are binding site positions or not. As discussed in Section 5.3.4, they could be used here for a direct identification of binding sites. Additionally, the total number of all T-to-C substitution positions as well as the total number of all other substitution positions are given for all data sets. The percentages that are shown in brackets are the number of detected T-to-C (or non-T-to-C) substitution positions divided by the total number of T-to-C (or non-T-to-C) substitution positions.

Comparing BayMAP 1.0 and BayMAP 2.0 with Table 7.5, the numbers of detected T-to-C substitution positions do not differ a lot between the two versions. For all of the data sets except for Kishore A, BayMAP 2.0 detects more T-to-C substitution positions as crosslinked. There are only small differences between the estimated false positive rates of BayMAP 1.0 and BayMAP 2.0. It is, however, remarkable, that BayMAP 2.0 has smaller estimated false positive rates for all data sets even if more T-to-C substitution positions are declared as method-induced for most of the data sets. This leads to the conclusion that the two versions of BayMAP do not differ importantly but that BayMAP 2.0 has a better detection rate with less false positives than BayMAP 1.0.

Table 7.5: Number of detected experimentally induced substitution positions.

Method	Positions	Kishore A	Kishore B	Memczak	Gottwein A	Gottwein B
	All T-to-C	85,234	8,368	11,847	930	954
	All non-T-to-C	73,728	11,416	25,273	5,136	5,423
BayMAP 1.0	Detected T-to-C	24,528 (28.8%)	1,602 (19.1%)	6,505 (54.9%)	132 (14.2%)	143 (15.0%)
	Detected non-T-to-C	1,801 (2.4%)	519 (4.5%)	2,241 (8.9%)	399 (7.8%)	402 (7.4%)
BayMAP 2.0	Detected T-to-C	24,219 (28.5%)	1,613 (19.3%)	6,520 (55.1%)	150 (16.1%)	145 (15.2%)
	Detected non-T-to-C	1,743 (2.4%)	513 (4.5%)	2,225 (8.8%)	308 (6.0%)	376 (6.9%)
	Combined detected T-to-C	23,880 (28.0%)	1605 (19.2%)	-	152 (16.3%)	162 (17.0%)
	Prior odds detected T-to-C	6,865 (8.1%)	44 (0.5%)	6933 (58.5%)	106 (11.4%)	10 (1.0%)
wavClusteR	Detected T-to-C	79,300 (93.0%)	7,379 (88.2%)	9,197 (77.6%)	178 (19.1%)	181 (19.0%)
	Detected non-T-to-C	33,529 (45.5%)	6,227 (54.5%)	7,622 (30.2%)	493 (9.6%)	497 (9.2%)
BMix	Detected T-to-C	29,654 (34.8%)	2,118 (25.3%)	6,955 (58.7%)	206 (22.2%)	211 (22.1%)
	Detected non-T-to-C	2581 (3.5%)	714 (6.3%)	2528 (10.0%)	489 (9.5%)	506 (9.3%)

As Table 7.5 reveals, the number of T-to-C substitution positions declared as method-induced by wavClusteR and BMix is always higher for the five data sets than by BayMAP 1.0 or BayMAP 2.0. wavClusteR even declares around 90% of the T-to-C substitution positions as PAR-CLIP induced for the two data sets from Kishore et al. [30], which is about three times more as declared by BayMAP 1.0 or BayBAP 2.0. For the data set from Memczak et al. [41] all methods find more than 50% of method-induced substitutions. Again, wavClusteR declares more than the other methods with 78%. The percentages of detected T-to-C substitution positions are closer to each other for the data sets from Gottwein et al. [22], where wavCluster and BMix find around 20% and BayMAP around 15%.

The false positive rate on the other hand, that is estimated by the percentage of detected non-T-to-C substitution positions among all non-T-to-C substitution positions is the smallest for BayMAP for all data sets. Especially wavClusteR shows here a very high rate of false positives. The estimated false positive rate for the data sets from Kishore et al. [30] is around 50% for wavClusteR whereas for BayMAP 1.0 as well as BayMAP 2.0, it is only between 2.4 to 4.5%.

BayMAP is hence able to greatly reduce the number of false findings. This is important, since functional validation of miRNA target sites in the laboratory is cumbersome and cost intensive. However, some of the true PAR-CLIP induced T-to-C substitution positions may be missed out.

Regarding the number of detected T-to-C substitution positions, when the combined post odds are considered as presented in Section 5.4 and Section 7.1.4, there are less detected T-to-C substitution positions for the two Kishore data sets (0.5% less positions for Kishore A and 0.1% less positions for Kishore B). On the other hand, more positions are seletected for the two Gottwein data sets (0.2% more positions for Gottwein A and 1.8% more positions for Gottwein B).

When not using a position based decision but the prior odds for BayMAP 2.0 instead, only for the data set from Memczak et al. [41] more positions are chosen. For Gottwein A, the number of detected positions is at least comparable to the position based

number (106 instead of 150). For all other data sets, the number of detected T-to-C substitution positions is far behind the position based number. These results stand in contrast to the expectation, since the positions of a potential binding site are chosen all together as method-induced or not. Interesting is, that the standard deviation of the effects for the potential binding sites τ is estimated the highest for the Gottwein A and the Memczak data set. It seems therefore that the prior odds as a decision tool are only reasonable, if the estimated differences in the potential binding sites are large enough. However, the results indicate, that the position based decision, at least for the considered data sets, delivers the more preferable outcome.

The top panel of Figure 7.3 shows the probability that a position has method-induced substitutions as a function of the T-to-C substitution rate for the three methods BayMAP 1.0 (black points), BMix (red points) and wavClusteR (blue line) for the first data set of Kishore et al. [30] (for other data sets see Figures D.1 to D.4). Results from BayMAP 2.0 are not presented here, as they are very similar to the results from BayMAP 1.0. In wavClusteR this probability is derived from the average of posterior distributions that are calculated separately for each substitution position (see Section 4.2). This means that the estimated probability does only depend on the substitution rate, but not for example on the number of reads for a specific position. Notably, the probability in function of the substitution rate can be drawn as a line. On the contrary, in BayMAP and BMix the number of reads is a factor that is taken into account for calculating the probability that a position is method-induced. In BayMAP even the covariates regarded in the probit model, e.g., the type of mRNA region, affect the estimation of the probability.

Comparing the two histograms of non-T-to-C substitution rates (middle panel) and T-to-C substitution rates (bottom panel) in Figure 7.3, it is remarkable that very small substitution rates close to zero are frequent in both histograms, but that small but not very small substitution rates (around 0.05 to 0.3) only occur more frequent for T-to-C substitution rates. A crucial step is therefore to distinguish between mismatch substitution positions and method-induced substitution positions for small substitution rates.

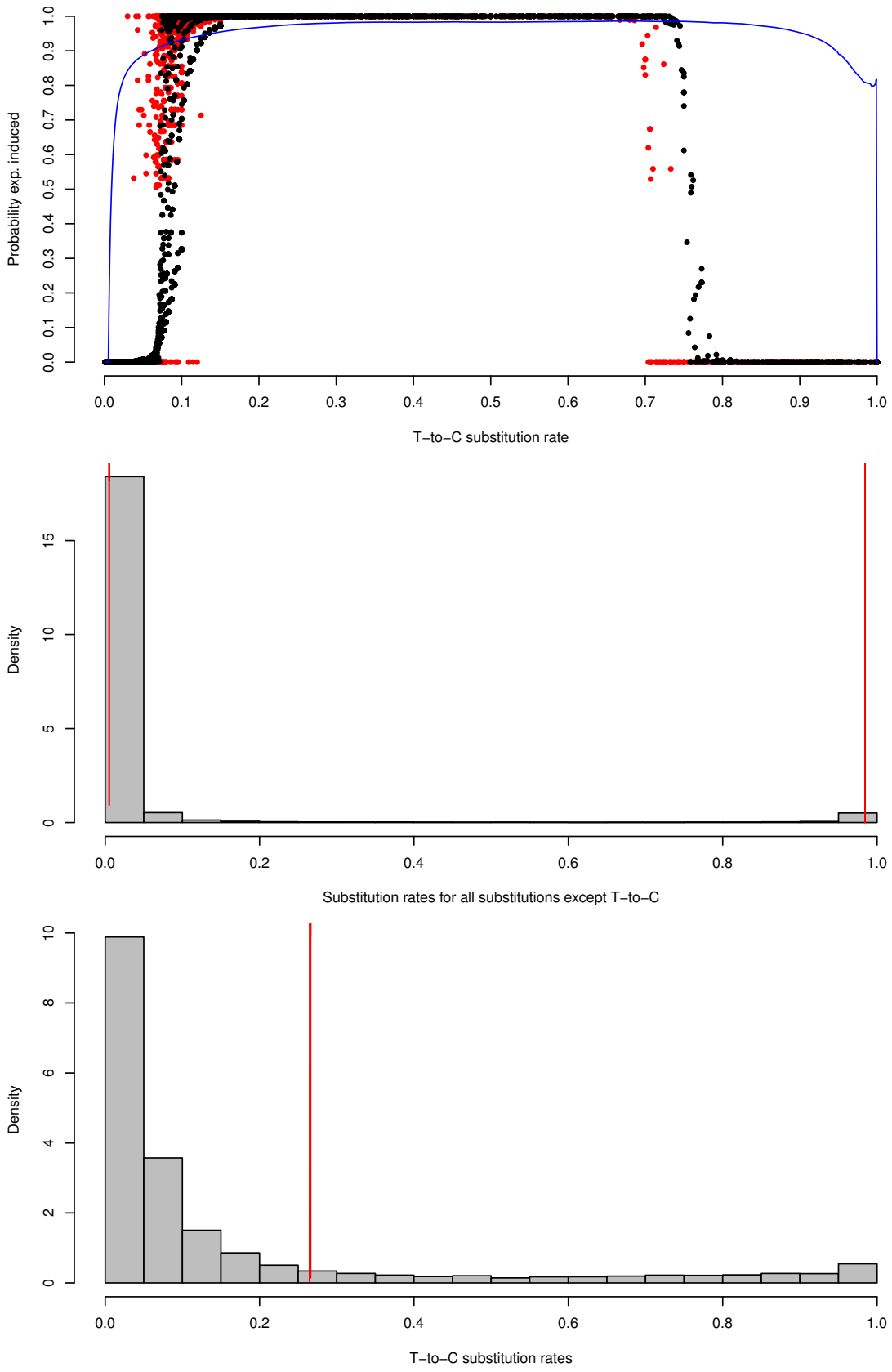

Figure 7.3: Top panel: T-to-C substitution rate in comparison to the probability that the position is experimentally induced for BayMAP 1.0 (black), BMix (red) and wavClusteR (blue) at 5,000 randomly chosen T-to-C substitution positions in the Kishore A data set. Middle/bottom panel: Histograms for the substitution rates for all substitutions except T-to-C with BayMAP 1.0 estimations for μ_{mm} and μ_{SNP} indicated by red lines (middle) and only for T-to-C substitutions with BayMAP 1.0 estimation for μ_{exp}. This graphic is published in Huessler et al. [27].

wavClusteR estimates already that positions with a very small substitution rate have experimentally induced substitutions (see top panel). This change point for wavClusteR is equal to 0.009, which means that a position with 100 observed reads with only one T-to-C substitution would be identified as method-induced substitution. In BayMAP and BMix this change point is higher (at a rate of about 0.08 for BayMAP and about 0.06 for BMix). These results also indicate that the number of false positives for wavClusteR is probably very high in comparison to BayMAP and BMix.

7.2.3 Comparison to TargetScan

As discussed in Section 2.2.1, there exist tools for target prediction such as TargetScan [2] that predict hundreds of potential binding sites for only one miRNA. These tools often only predict canonical binding sites whereas a miRNA could also have non-canonical ones. By contrast, canonical and non-canonical binding sites can be detected by PAR-CLIP. As seen in Table 7.5, BayMAP has a small false positve rate in comparison to wavClusteR and BMix. The number of potential binding sites, that can be considered for experimental validation, can therefore be highly reduced by BayMAP. It is, however, interesting to which extent the results of the methods for analyzing PAR-CLIP data are comparable to the results of target prediction tools such as TargetScan to the aim of further validation of PAR-CLIP and BayMAP. In Table 7.6 predicted canonical and conserved targets of conserved miRNA families derived by TargetScan are compared to all T-to-C substitution positions (± 10 nt) as well as the T-to-C substitution positions (± 10 nt) that are identified by BayMAP 1.0, BayMAP 2.0, wavClusteR or BMix as method-induced substitution positions.

BayMAP-identified crosslinked T-to-C substitution positions seem to slightly enrich for canonical miRNA target sites compared to all T-to-C substitution positions, since percentages of T-to-C substitution positions found in TargetScan are higher when using BayMAP. However, the degree of enrichment differs between data sets. The percentage of non-overlapping T-to-C substitution positions is still very high, which leads to the conclusion, that still a lot of binding sites that are non-canonical or non-conserved can be identified by using the combination of PAR-CLIP with BayMAP analysis.

Table 7.6: Overlap between TargetScan and the PAR-CLIP data. The total overlap describes the percentage of all T-to-C substitution positions that overlap with a canonical and conserved TargetScan site (+/- 10 nt). The percentage for the different methods is the proportion of detected experimentally induced T-to-C substitution positions that overlap with a canonical and conserved TargetScan site.

	Total overlap	BayMAP 1.0	BayMAP 2.0	Combined Post Odds	Prior Odds	wavClusteR	BMix
Kishore A	29.3%	29.7%	29.6%	29.6%	29.0%	29.4%	29.3%
Kishore B	31.8%	34.0%	33.9%	40.6%	9.1%	34.7%	34.1%
Memczak	29.3%	34.7%	34.7%	-	35.5%	33.3%	34.7%
Gottwein A	3.7%	8.6%	7.3%	7.2%	7.5%	10.1%	7.8%
Gottwein B	3.8%	9.1%	9.0%	8.0%	20.0%	7.2%	7.6%

When regarding the overlap for the combined post odds, it is interesting that the overlap does not change much in comparison to BayMAP 2.0 except for the Kishore B data set. Instead of an overlap of 33.9%, there is an overlap of 40.6%. This means, that at least for the Kishore B data set more of the detected T-to-C substitution positions are TargetScan targets when using the combined posterior odds and therefore also the Kishore A data.

The overlap is also shown for the prior odds when employing BayMAP 2.0. Except for the Memczak data set and the Gottwein A data set, the percentage is not reliable because of the small number of detected positions. In the Memczak data set, however, the highest percentage of overlap is reached for the prior odds. It is interesting, that BayMAP 2.0 has always a slightly smaller overlap with TargetScan than BayMAP 1.0. BayMAP 2.0 detects therefore a higher percentage of non-canonical or non-conserved binding sites.

Chapter 8

Discussion

BayMAP, a three component mixture model, is presented in this thesis for the detection of experimentally induced T-to-C substitution positions in PAR-CLIP data and therefore for the detection of miRNA targets on the mRNA. BayMAP is set into a fully Bayesian hierarchical framework, so that it is possible to include additional prior information that might be relevant for the probability that a position is a binding site. To the best of the author's knowledge, BayMAP is the first method for the analysis of PAR-CLIP data, that enables this inclusion. BayMAP 1.0 has already been already published in Huessler et al. [27]. BayMAP 2.0, firstly presented in this thesis, now allows the inclusion of neighborhood information in addition to other covariates, so that positions on the same read cluster are no longer assumed to be independent.

In an extensive simulation study, it is shown, that BayMAP 1.0 estimated very precisely, i.e. almost without bias the parameters, that is the true substitution probabilities and the regression parameters. For BayMAP 2.0, a small underestimation of β_0 could be detected. However, the other parameters of $\boldsymbol{\beta}$ seem to be unbiased, i.e. the part of $\boldsymbol{\beta}$ that is analyzed for interpretation. Notably, BayMAP can not only be used for distinguishing experimentally induced substitution positions from non-exerimentally induced ones, and therefore binding positions from noise, but also for the analysis and interpretation of additional variables that could be relevant for the probability that a position is a binding position

In this work, the analysis of additional variables is mainly focused on the type of the mRNA region. However, other factors that might influence the prior probability for

having a position with PAR-CLIP induced substitutions could be taken into account. Such other factors could be, for example, the number of PAR-CLIP reads relative to the mRNA abundance, or local AU-rich content near to the site.

Besides the precise estimation of the parameters, the simulation study reveals that BayMAP 1.0 and BayMAP 2.0 both outperform wavClusteR and BMix in terms of accuracy and specificity. In case of difficult decision making, i.e. when μ_{exp} is either very low or very high, the high performance of BayMAP 1.0 is even more pronounced in comparison to other methods. In the simulation setting for BayMAP 2.0, it is noticeable that wavClusteR seems to often predict almost all of the T-to-C substitution positions as method-induced, so that not only the sensitivity is close to one, but also the rate of false positives, leading to a very small specificity. BayMAP 2.0 in contrast, not only achieves a high sensitivity but in particular a high specificity in comparison to the other methods.

In applications to experimental Ago PAR-CLIP data sets, this higher performance in terms of specificity was confirmed. On the one hand, BayMAP 1.0 and BayMAP 2.0 detected less T-to-C substitution positions in total, but on the other hand the estimated false positive rates were very low. BayMAP seems therefore, particularly useful for the true detection of binding sites, which means that almost all of the by BayMAP detected binding sites are true binding sites.

This is a big advantage in comparison to the other methods, since one of the reasons for the implementation of the PAR-CLIP experiment is the very high number of target mRNAs by target prediction tools such as TargetScan. Targets identified by BayMAP can then for example, be further analyzed for functional and biological relevance in experiments.

The parameter estimates in these applications seemed to be reasonable. However, the positive impact of the 3'UTR could not be detected in every PAR-CLIP data set. It seemed that the impact of the 3'UTR was particularly important and was estimated with a positive value when the data is noisy, e.g., as was the case for the data sets from Gottwein.

For BayMAP 2.0 it was also proposed to use the prior odds as a decision criterion if a position has method-induced substitutions, instead of, e.g., the posterior odds. The prior odds provide particularly good results, when the variance of the random effect is high. However, for smaller variances, the results of the prior odds seemed to be less reliable in applications to the simulated as well as real PAR-CLIP data sets. It is therefore recommended that the posterior probability that a position is crosslinked is estimated.

It is also recommended to use BayMAP on all substitution positions, i.e. not only T-to-C substitution positions, with the dependence $\mu_{\text{SNP}} = 1 - 3\mu_{\text{mm}}$ and a zero truncated binomial model. It is however also possible to use simpler versions that still perform better than other methods as shown in the simulation study.

The applications to PAR-CLIP data sets as well as the simulation study indicate that BayMAP 2.0 performs slightly better than BayMAP 1.0. In the applications to the PAR-CLIP data sets, BayMAP 2.0 not only identified more T-to-C substitution positions as being method-induced, but also had a smaller estimated rate of false positives, so that BayMAP 2.0 should be preferably employed. However, when the main purpose is not the detection of binding sites but the analysis of variables, such as the mRNA region, it is advisable to utilize BayMAP 1.0.

It is also possible to adapt the general model of BayMAP. For instance, μ_{exp} is not supposed to be identical for different crosslinked T-to-C substitution positions as explained in the previous sections. This could be represented in the model by including a new level for μ_{exp}, e.g., $\mu_{\text{exp}} \sim \text{Beta}(a,\ b)$, where $a > 0$ and $b > 0$ are the parameters of the beta distribution. However, this new level would complicate the model with more parameters to estimate.

Furthermore, a new method is introduced in this thesis for the combination of reads into a read cluster. In comparison to the wavelet-based peak calling [49] and the MiniRank Norm [11], the here presented method is intuitive and easy to understand, as it is not based on an underlying model. Instead, it combines overlapping reads of the same T-to-C substitution position to a read cluster. The here presented method is similar to the one developed by Golumbeanu et al. [20]. In the latter method, overlapping

clusters of different T-to-C substitution positions are always combined to one single cluster. In contrast, the method presented in this thesis is able to identify several peaks and therefore several read clusters even if reads are overlapping, depending on the extent of overlap of two clusters.

Moreover, BayMAP is, to the best of the author's knowledge, the first method specialized on PAR-CLIP data, that enables the analysis of several PAR-CLIP data sets at the same time. The big advantage is that the information of several PAR-CLIP replicates can be combined so that decisions are more reliable. The more replicates are available, the higher the impact of the data on the here considered combined posterior odds in comparison to the prior information. This seems to be reasonable, since the more data is available, the more the data, i.e. the numbers of substitutions, should influence the results. The combined posterior odds take the BayMAP results of the seperate PAR-CLIP data sets as input. The data sets are, hence, first analyzed separately and the results are then used for the combination.

Instead of incorporating the results to the combined posterior odds after running BayMAP 1.0 or BayMAP 2.0, it might be advisable, to adapt the BayMAP model in such a way that several data sets can be analyzed simultaneously. Instead of K_i, the number of substitutions at position i, K_{id} could be modeled as the number of substitutions at position i for data set d. The allocation variable Z_i would still be independent of data set d, since a binding site is supposed to be a binding site in all replicates. However, the direct imputation makes the model more complex, especially when being interested in the effects of additional variables. In this thesis, hence, the combined post odds are chosen due to their simplicity and their easy interpretation.

BayMAP permits the detection of canonical and non-canonical miRNA binding sites. It is not possible to predict directly which individual miRNA binds to the identified target site on the given mRNA. However, target prediction tools, such as TargetScan, can also be used the other way around, so that binding sites that are detected by BayMAP can be linked to miRNAs. In this work, it is shown that the application of BayMAP leads to a slightly higher overlap with canonical and conserved TargetScan sites than without applying BayMAP. These target prediction tools are capable of predicting hundreds of

mRNAs for one miRNA and mainly focusing on the prediction of canonical and conserved targets. Only around one third of the by BayMAP detected binding site positions were also predicted by TargetScan and can thus be linked to a specific miRNA. In order to also detect non-canonical interactions, motif search algorithms can be implemented, e.g, as proposed by Khorshid et al. [29]. A new Bayesian algorithm could also be developed for finding miRNAs for the given targets, allowing the incorporation of genetic information such as primary pairing rules and RNA secondary structure predictions, i.e. the accessibility of the target site.

Taken together, a deeper understanding of the underlying biology of miRNA regulation using PAR-CLIP experimental data in combination with the highly specific target site prediction algorithm presented in this thesis is expected.

Appendix A

Additional full conditionals

As explained in Section 5.1.2, it could also be assumed that the number of substitutions K_i for positions $i = 1, \ldots, N$ could, for simplicity reasons, follow a binomial distribution instead of a zero truncated binomial distribution. Moreover, the restriction that $\mu_{\text{SNP}} = 1 - 3\mu_{\text{mm}}$ could be omitted, so that the model gets easier. In this section, the full conditional distributions for the easier models are presented. For notations and definitions see Section 5.1.2.

A.1 Additional full conditional distributions for $\boldsymbol{\mu}$

The full conditional density for $\boldsymbol{\mu}$ when K_i follows a binomial distribution and when there is no restriction on μ_1 and μ_2, is given by

$$
\begin{aligned}
f(\boldsymbol{\mu} \mid \boldsymbol{z}, \boldsymbol{k}, \boldsymbol{n}) &\propto f(\boldsymbol{k} \mid \boldsymbol{\mu}, \boldsymbol{n}, \boldsymbol{z}) \cdot f(\boldsymbol{\mu}) \\
&= \prod_{i=1}^{N} \prod_{m=1}^{3} \left(\binom{n_i}{k_i} \mu_m^{k_i} (1-\mu_m)^{n_i - k_i} \right)^{z_{im}} \mathbb{1}_{[0,1]}(\mu_m) \\
&= \left(\prod_{i=1}^{N} \binom{n_i}{k_i} \right) \left(\prod_{m=1}^{3} \mu_m^{N_m \bar{k}_m} (1-\mu_m)^{n_m^{\text{total}} - N_m \bar{k}_m} \right) \mathbb{1}_{[0,1]}(\mu_m) \\
&\propto \prod_{m=1}^{3} \mu_m^{N_m \bar{k}_m} (1-\mu_m)^{n_m^{\text{total}} - N_m \bar{k}_m} \mathbb{1}_{[0,1]}(\mu_m) \\
&\propto \prod_{m=1}^{3} \text{Beta}\left(\mu_m \mid N_m \bar{k}_m + 1, n_m^{\text{total}} - N_m \bar{k}_m + 1\right). \qquad \text{(A.1)}
\end{aligned}
$$

The full conditional density for $\boldsymbol{\mu}$ when K_i follows a zero truncated binomial distribution and when there is no restriction on μ_1 and μ_2, is given by

$$
\begin{aligned}
f\left(\boldsymbol{\mu} \mid \boldsymbol{z}, \boldsymbol{k}, \boldsymbol{n}\right) &\propto f\left(\boldsymbol{k} \mid \boldsymbol{\mu}, \boldsymbol{n}, \boldsymbol{z}\right) \cdot f\left(\boldsymbol{\mu}\right) \\
&= \prod_{i=1}^{N} \prod_{m=1}^{3} \left(\frac{\binom{n_i}{k_i} \mu_m^{k_i} \left(1-\mu_m\right)^{n_i-k_i}}{1-\left(1-\mu_m\right)^{n_i}} \right)^{z_{im}} \mathbb{1}_{[0,1]}(\mu_m) \\
&= \prod_{m=1}^{3} \frac{\mu_m^{N_m \bar{k}_m} \left(1-\mu_m\right)^{n_m^{\text{total}} - N_m \bar{k}_m}}{\prod_{i=1}^{N} \frac{1}{\binom{n_i}{k_i}} \left(1-\left(1-\mu_m\right)^{n_i}\right)^{z_{im}}} \mathbb{1}_{[0,1]}(\mu_m) \, . \qquad \text{(A.2)}
\end{aligned}
$$

When K_i follows a binomial distribution and $\mu_2 = 1 - 3\mu_1$, then the density for $\boldsymbol{\mu}$ given all model parameters except for $\boldsymbol{\mu}$ can be written as

$$
\begin{aligned}
f\left(\boldsymbol{\mu} \mid \boldsymbol{z}, \boldsymbol{k}, \boldsymbol{n}\right) &\propto f\left(\boldsymbol{k} \mid \boldsymbol{\mu}, \boldsymbol{n}, \boldsymbol{z}\right) \cdot f\left(\boldsymbol{\mu}\right) \\
&\propto \prod_{i=1}^{N} \prod_{m=1}^{3} \left(\binom{n_i}{k_i} \mu_m^{k_i} \left(1-\mu_m\right)^{n_i-k_i} \right)^{z_{im}} \mathbb{1}_{\mu_2}(1-3\mu_1) \cdot \mathbb{1}_{[0.25,1]}(\mu_2) \cdot \mathbb{1}_{[0,1]}(\mu_3) \\
&= \prod_{i=1}^{N} \binom{n_i}{k_i} \left(\mu_1^{k_i} \left(1-\mu_1\right)^{n_i-k_i} \right)^{z_{i1}} \left(\left(1-3\mu_1\right)^{k_i} \left(3\mu_1\right)^{n_i-k_i} \right)^{z_{i2}} \left(\mu_3^{k_i} \left(1-\mu_3\right)^{n_i-k_i} \right)^{z_{i3}} \\
&\quad \mathbb{1}_{\mu_2}(1-3\mu_1) \cdot \mathbb{1}_{[0.25,1]}(\mu_2) \cdot \mathbb{1}_{[0,1]}(\mu_3) \\
&= \prod_{i=1}^{N} \binom{n_i}{k_i} \mu_1^{N_1 \bar{k}_1} \left(1-\mu_1\right)^{n_1^{\text{total}} - N_1 \bar{k}_1} \left(1-3\mu_1\right)^{N_2 \bar{k}_2} \left(3\mu_1\right)^{n_2^{\text{total}} - N_2 \bar{k}_2} \mu_3^{N_3 \bar{k}_3} \left(1-\mu_3\right)^{n_3^{\text{total}} - N_3 \bar{k}_3} \\
&\quad \mathbb{1}_{\mu_2}(1-3\mu_1) \cdot \mathbb{1}_{[0.25,1]}(\mu_2) \cdot \mathbb{1}_{[0,1]}(\mu_3) \\
&\propto \mu_1^{N_1 \bar{k}_1} \left(1-\mu_1\right)^{n_1^{\text{total}} - N_1 \bar{k}_1} \left(1-3\mu_1\right)^{N_2 \bar{k}_2} \left(3\mu_1\right)^{n_2^{\text{total}} - N_2 \bar{k}_2} \mu_3^{N_3 \bar{k}_3} \left(1-\mu_3\right)^{n_3^{\text{total}} - N_3 \bar{k}_3} \\
&\quad \mathbb{1}_{\mu_2}(1-3\mu_1) \cdot \mathbb{1}_{[0.25,1]}(\mu_2) \cdot \mathbb{1}_{[0,1]}(\mu_3) \qquad \text{(A.3)}
\end{aligned}
$$

If there is no restriction on μ_{SNP} and K_i is assumed to follow a binomial distribution, the Gibbs sampler can be implemented in order to sample from the conditional distribution of $\boldsymbol{\mu}$, as μ_m, $m = 1, \ldots, 3$ follows then a beta distribution. In the other cases, the Metropolis-Hastings algorithm can be used.

A.2 Additional full conditional distribution for $\boldsymbol{Z}$

If K_i follows a binomial distribution, the density can be written as

$$f\left(\boldsymbol{z} \mid \boldsymbol{p}, q, \boldsymbol{\mu}, \boldsymbol{k}, \boldsymbol{n}\right) \propto f\left(\boldsymbol{k} \mid \boldsymbol{\mu}, \boldsymbol{n}, \boldsymbol{z}\right) \cdot f\left(\boldsymbol{z} \mid \boldsymbol{p}, q\right) \tag{A.4}$$

$$= \prod_{i=1}^{N} \prod_{m=1}^{3} \left(\binom{n_i}{k_i} \mu_m^{k_i} (1-\mu_m)^{n_i - k_i} \right)^{z_{im}} \left((1-p_i)q\right)^{z_{i1}} \left((1-p_i)(1-q)\right)^{z_{i2}} p_i^{z_{i3}}$$

With the weight

$$w_{im}^{\text{bin}} := \left(\binom{n_i}{k_i} \mu_m^{k_i} (1-\mu_m)^{n_i - k_i} \right)$$

Equations (A.4) can be rewritten, so that Z_i follows, conditional on the other parameters, a categorical distribution with density

$$\begin{aligned} f\left(\boldsymbol{z} \mid \boldsymbol{p}, q, \boldsymbol{\mu}, \boldsymbol{k}, \boldsymbol{n}\right) &\propto \prod_{i=1}^{N} \left(w_{i1}^{\text{bin}}(1-p_i)q\right)^{z_{i1}} \left(w_{i2}^{\text{bin}}(1-p_i)(1-q)\right)^{z_{i2}} \left(w_{i3}^{\text{bin}} p_i\right)^{z_{i3}} \\ &\propto \prod_{i=1}^{N} \left(\frac{w_{i1}^{\text{bin}}(1-p_i)q)}{\widetilde{w}^{\text{bin}}}\right)^{z_{i1}} \left(\frac{w_{i2}^{\text{bin}}(1-p_i)(1-q)}{\widetilde{w}^{\text{bin}}}\right)^{z_{i2}} \left(\frac{w_{i3}^{\text{bin}} p_i}{\widetilde{w}^{\text{bin}}}\right)^{z_{i3}} \\ &\propto \prod_{i=1}^{N} \text{Cat}\left(\frac{w_{i1}^{\text{bin}}(1-p_i)q)}{\widetilde{w}^{\text{bin}}}, \frac{w_{i2}^{\text{bin}}(1-p_i)(1-q)}{\widetilde{w}^{\text{bin}}}, \frac{w_{i3}^{\text{bin}} p_i}{\widetilde{w}^{\text{bin}}}\right) \end{aligned} \tag{A.5}$$

where $\widetilde{w}^{\text{bin}} = w_{i1}^{\text{bin}}(1-p_i)q) + w_{i2}^{\text{bin}}(1-p_i)(1-q) + w_{i3}^{\text{bin}} p_i$. The latent variable $\boldsymbol{Z}$ can therefore be sampled using the Gibbs sampler.

Appendix B

Trace plots

B.1 BayMAP 1.0

Example trace plots for simulated data set

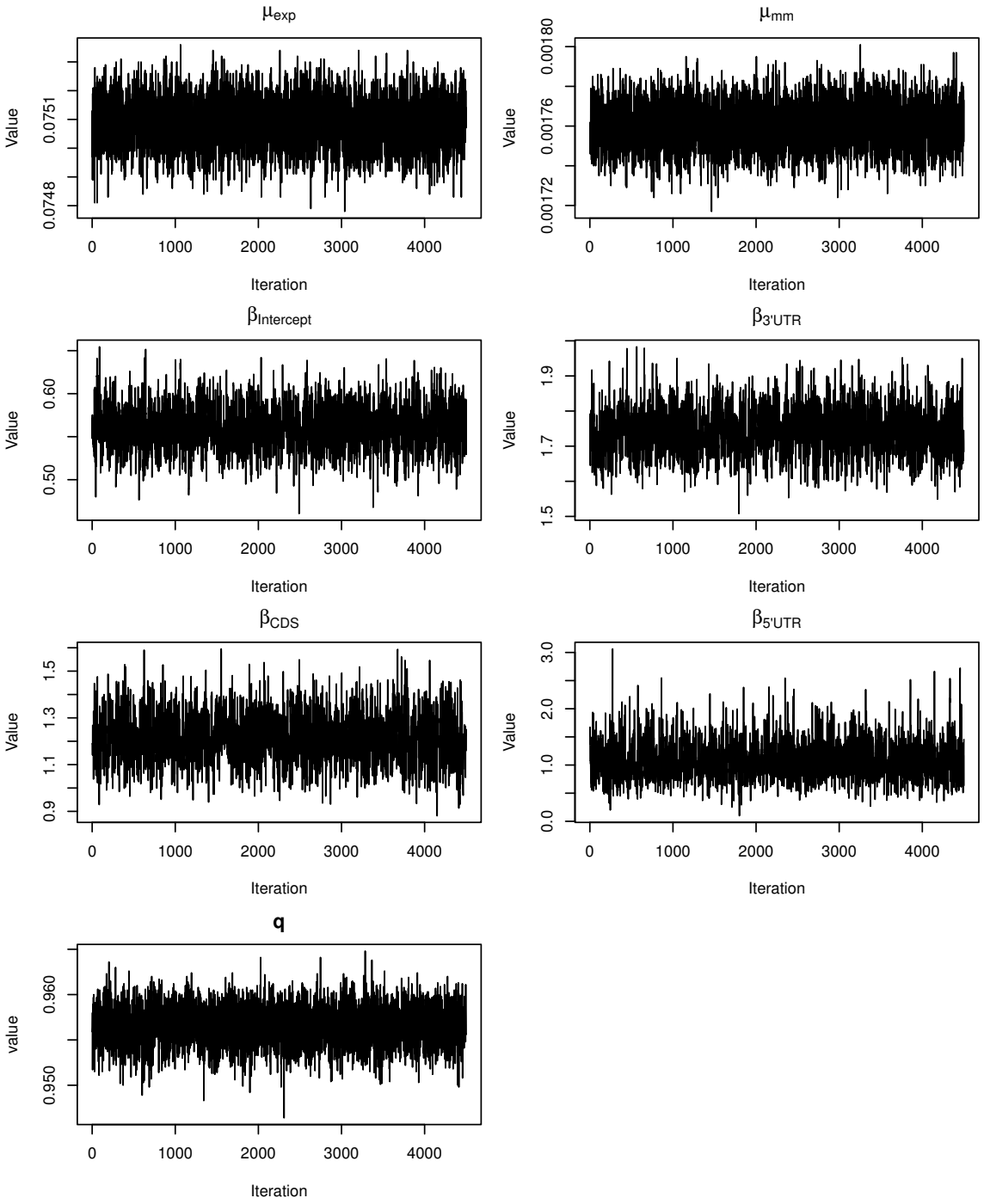

Figure B.1: Trace plots of the model parameters for BayMAP 1.0 applied to one simulated data set with $\mu_{\text{exp}} = 0.75$. 4,500 iterations out of 15,000 were used for parameter estimation using every third iteration after a burn in of 1,500.

Trace plots for Kishore A data set

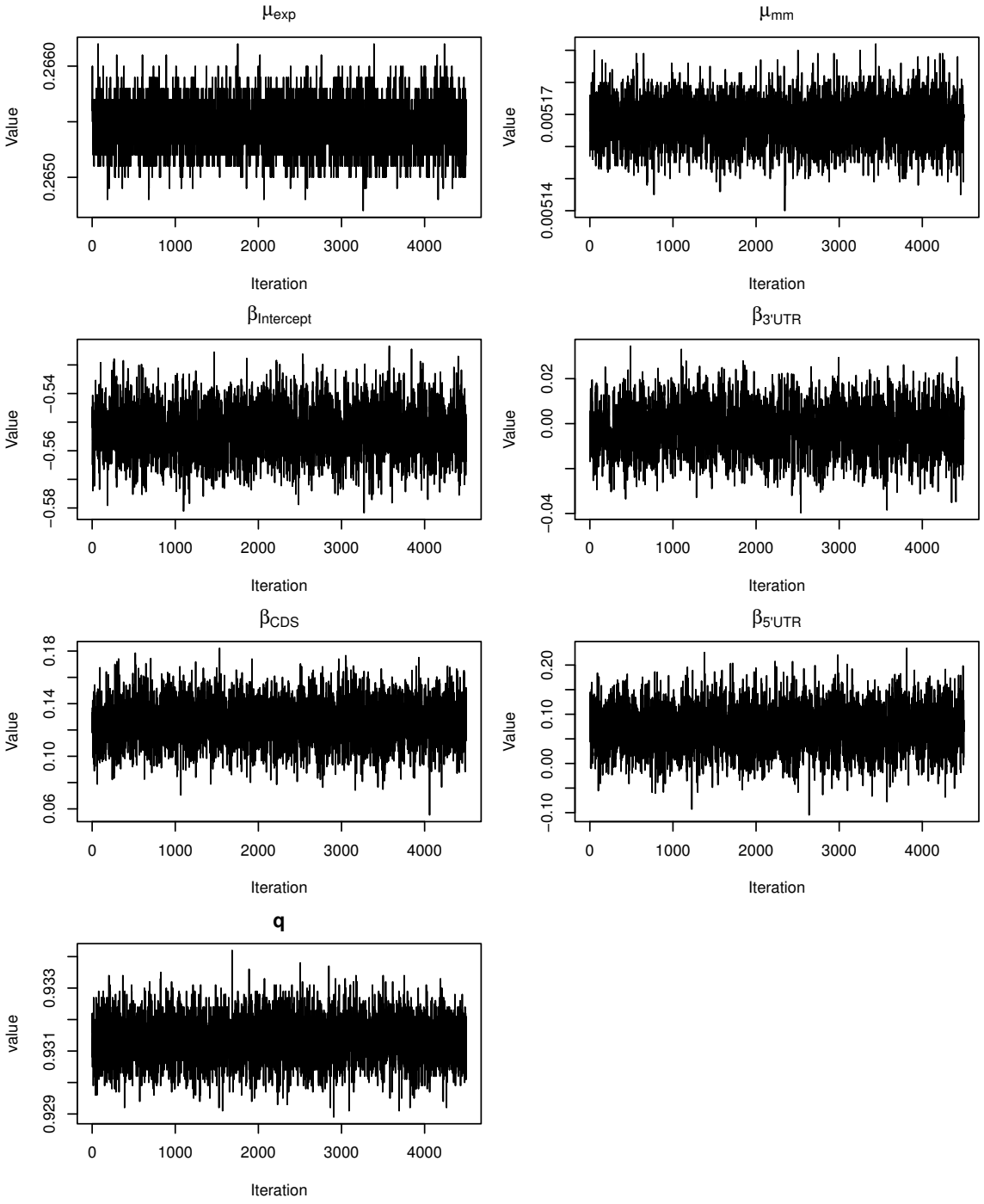

Figure B.2: Trace plots of the model parameters for BayMAP 1.0 applied to the Kishore A data. 4,500 iterations out of 75,000 were used for parameter estimation using every 15th iteration after a burn in of 7,500.

Trace plots for Kishore B data set

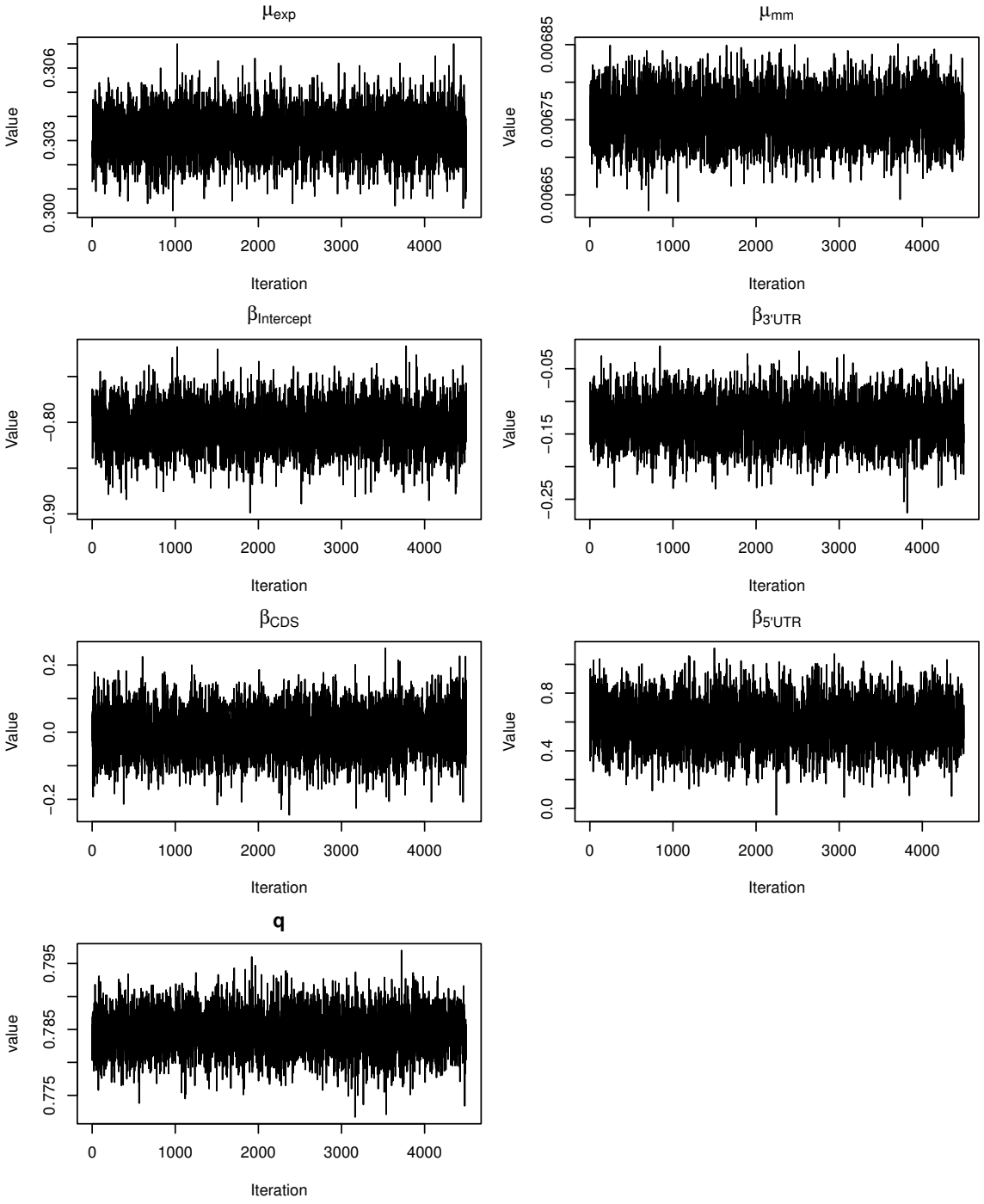

Figure B.3: Trace plots of the model parameters for BayMAP 1.0 applied to the Kishore B data. 4,500 iterations out of 75,000 were used for parameter estimation using every 15th iteration after a burn in of 7,500.

Trace plots for Memczak data set

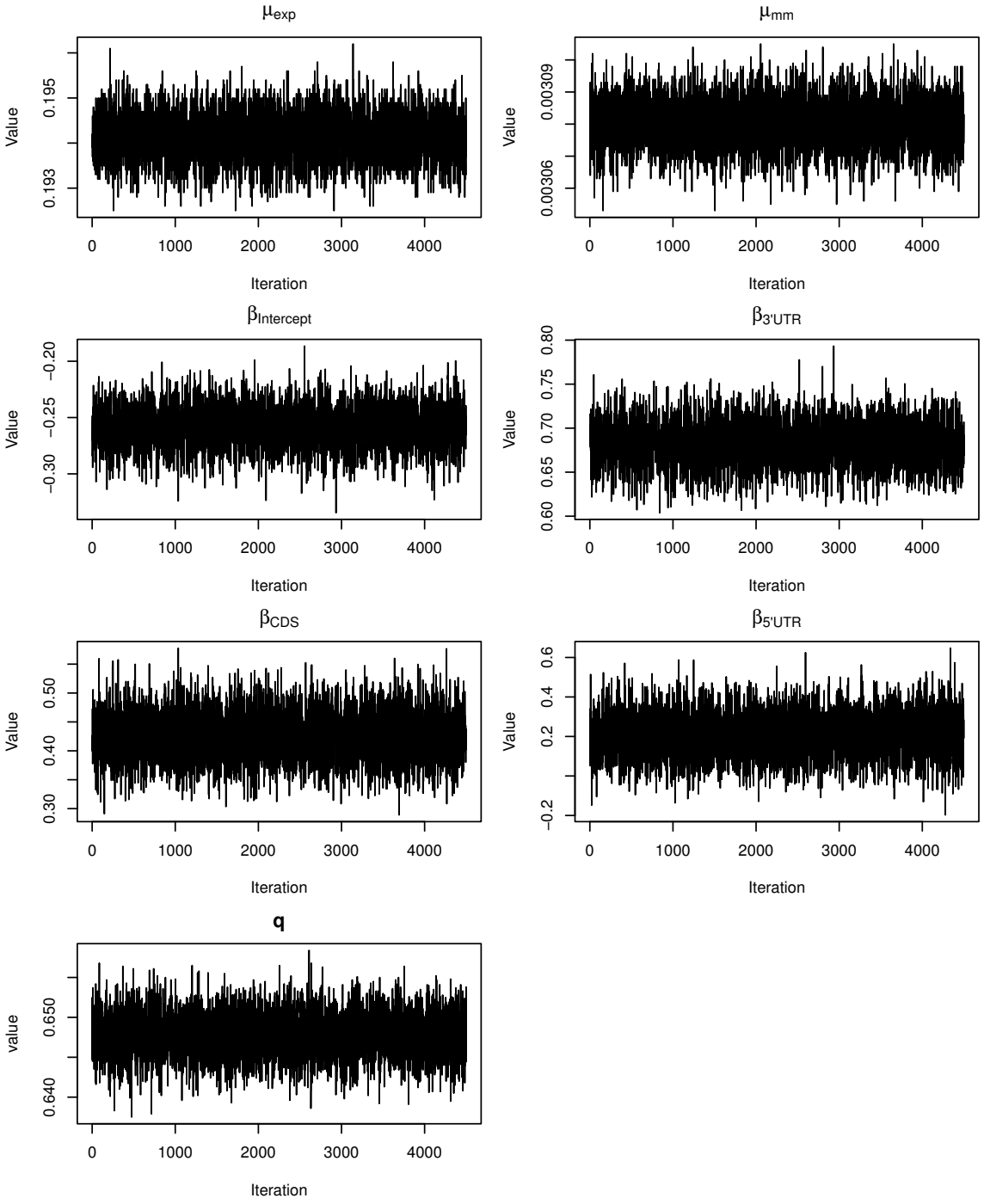

Figure B.4: Trace plots of the model parameters for BayMAP 1.0 applied to the Memczak data. 4,500 iterations out of 75,000 were used for parameter estimation using every 15th iteration after a burn in of 7,500.

Trace plots for Gottwein A data set

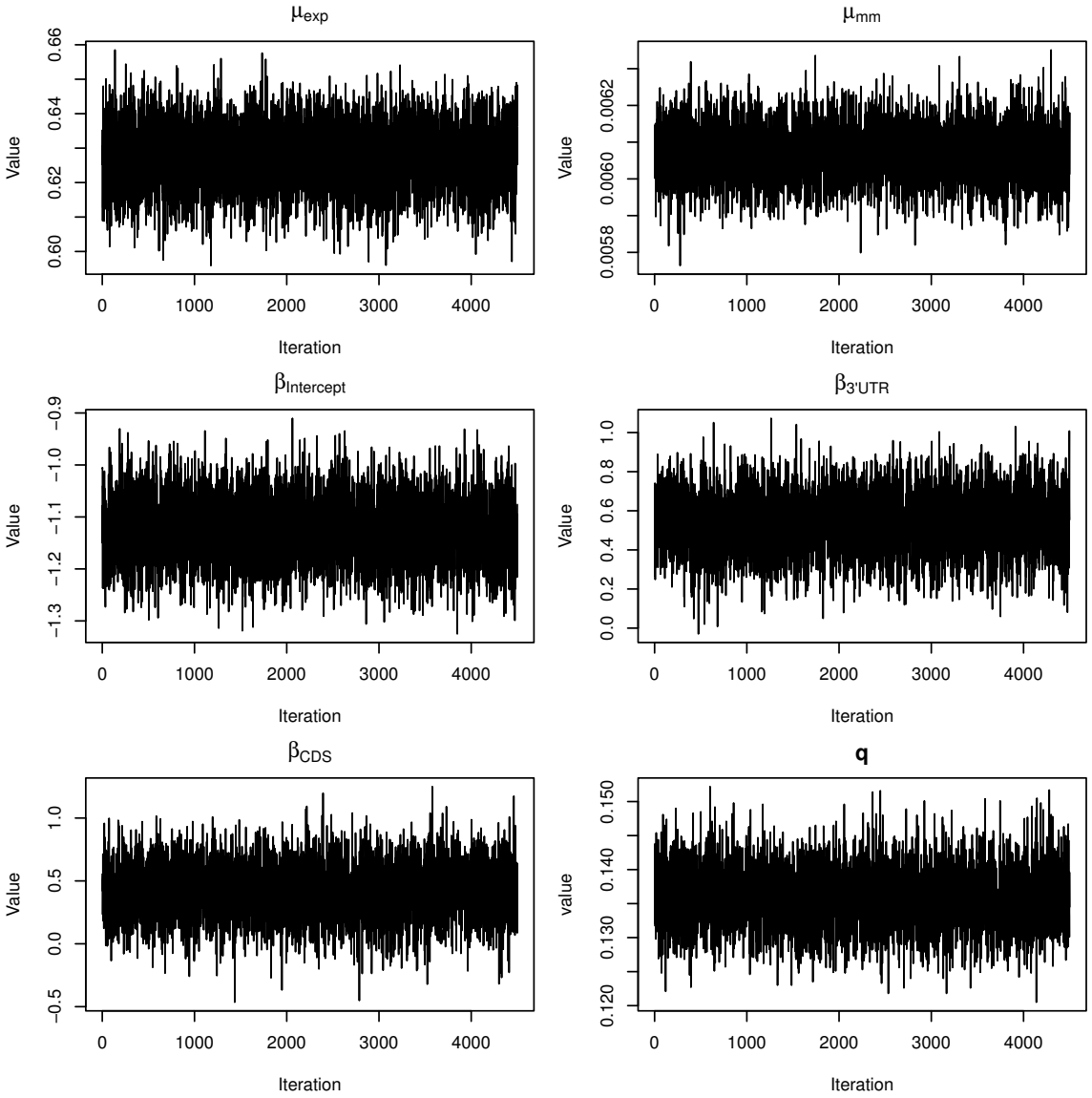

Figure B.5: Trace plots of the model parameters for BayMAP 1.0 applied to the Gottwein A data. 4,500 iterations out of 75,000 were used for parameter estimation using every 15th iteration after a burn in of 7,500.

Trace plots for Gottwein B data set

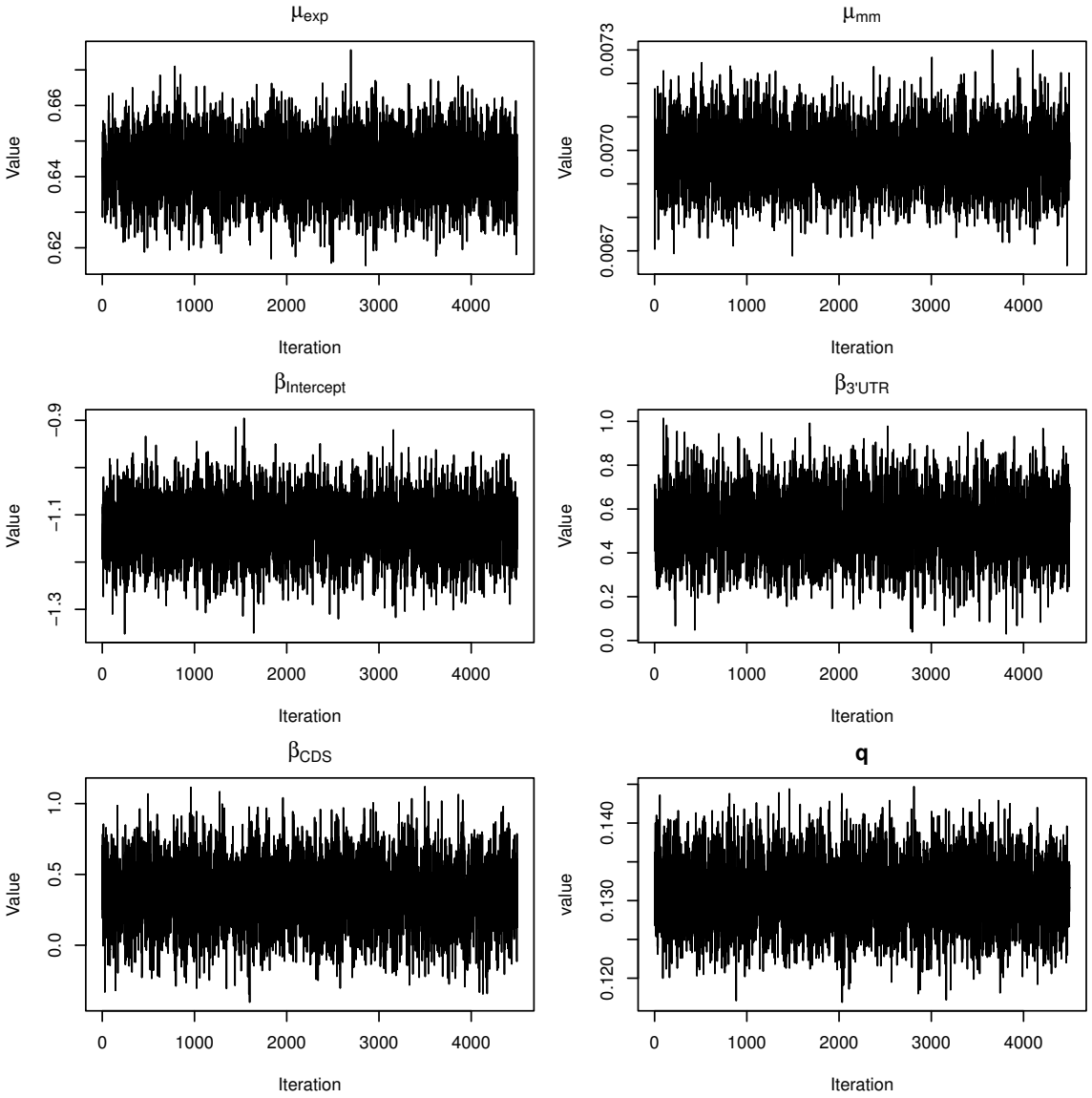

Figure B.6: Trace plots of the model parameters for BayMAP 1.0 applied to the Gottwein B data. 4,500 iterations out of 75,000 were used for parameter estimation using every 15th iteration after a burn in of 7,500.

B.2 BayMAP 2.0

Example trace plots for simulated data set

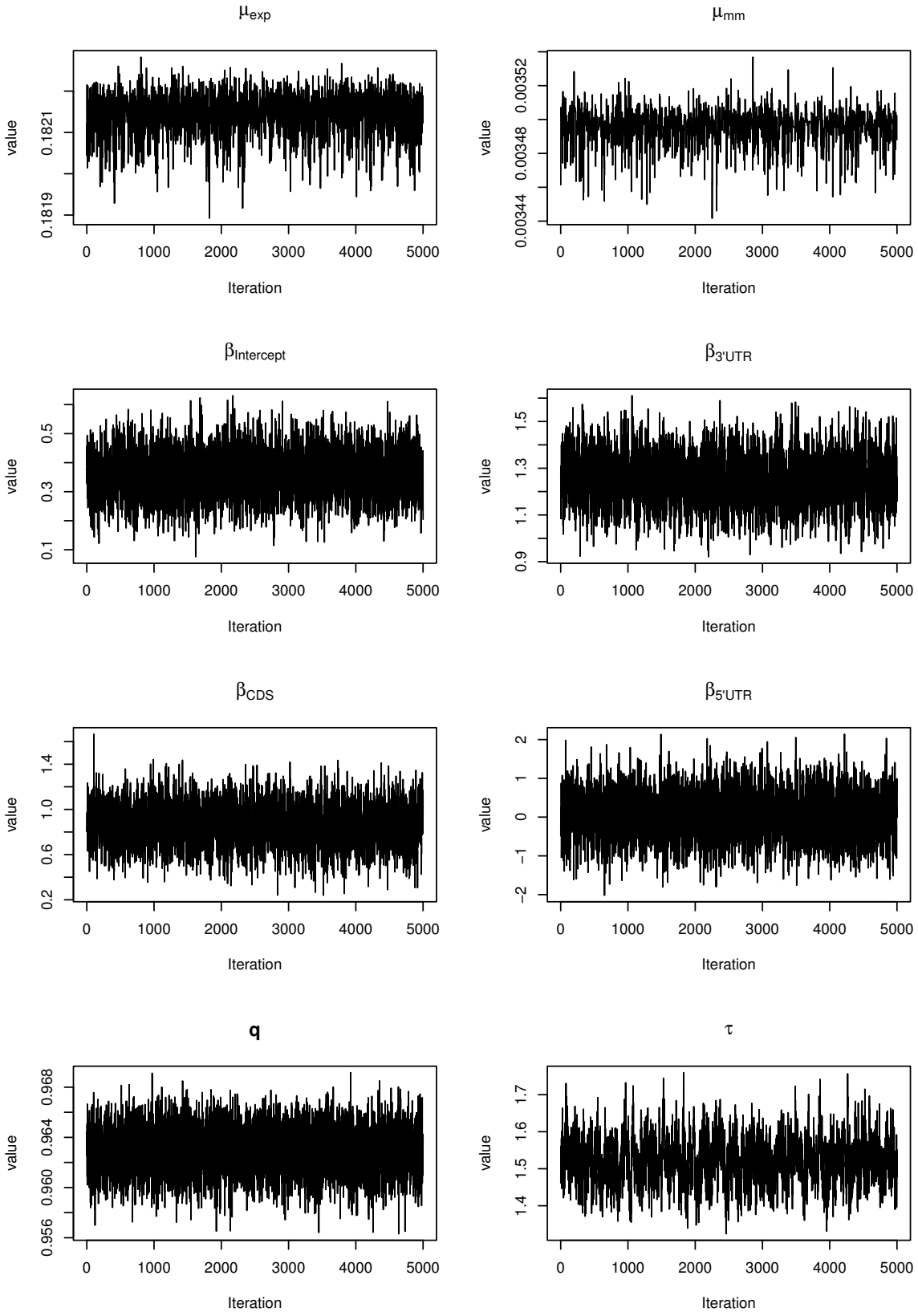

Figure B.7: Trace plots of the model parameters for BayMAP 2.0 applied to one simulated data set. 5,000 iterations out of 45,000 were used for parameter estimation using every sixth iteration after a burn in of 15,000.

Trace plots for Kishore A data set

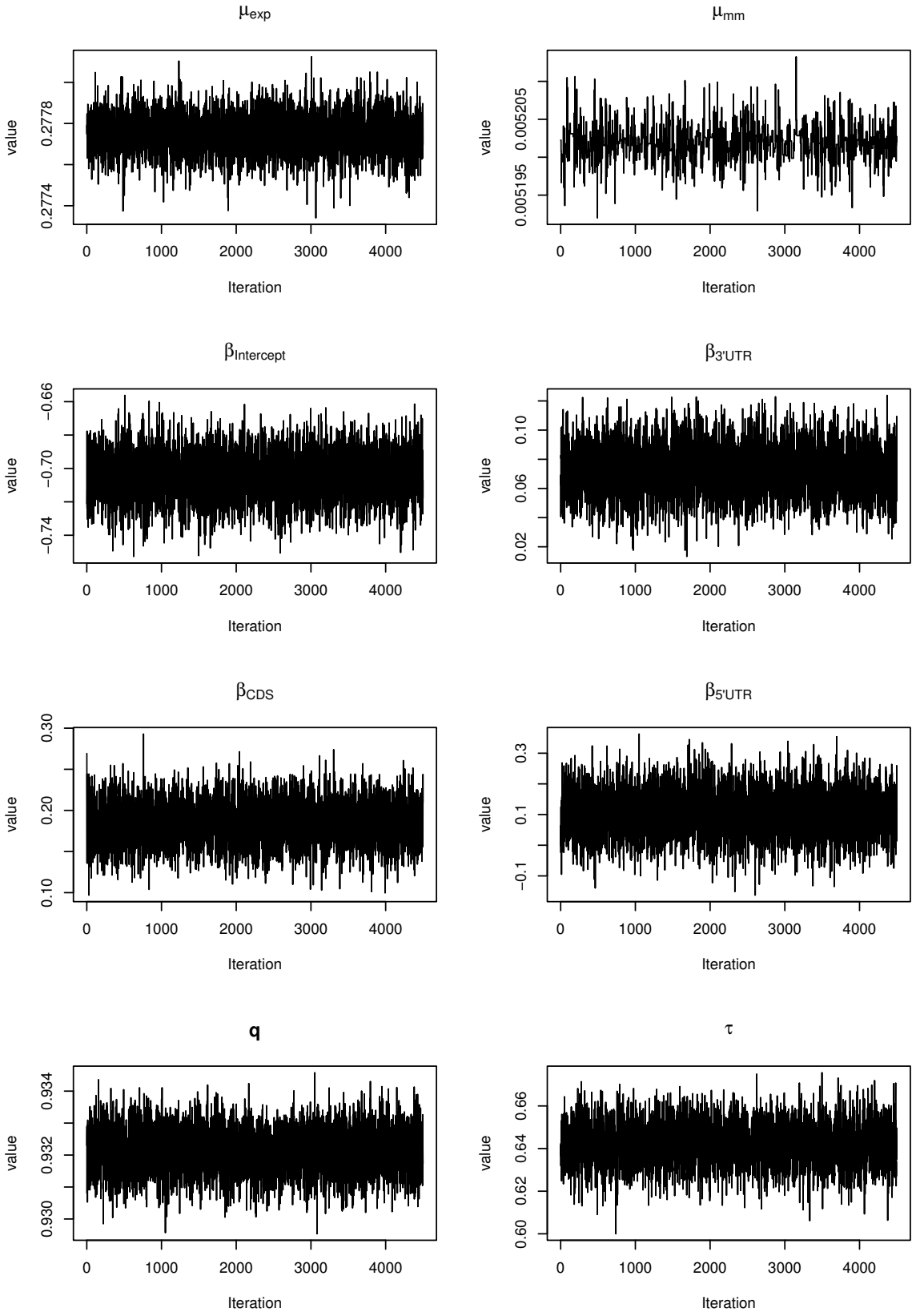

Figure B.8: Trace plots of the model parameters for BayMAP 2.0 applied to the Kishore A data. 4,500 iterations out of 75,000 were used for parameter estimation using every 15th iteration after a burn in of 7,500.

Trace plots for Kishore B data set

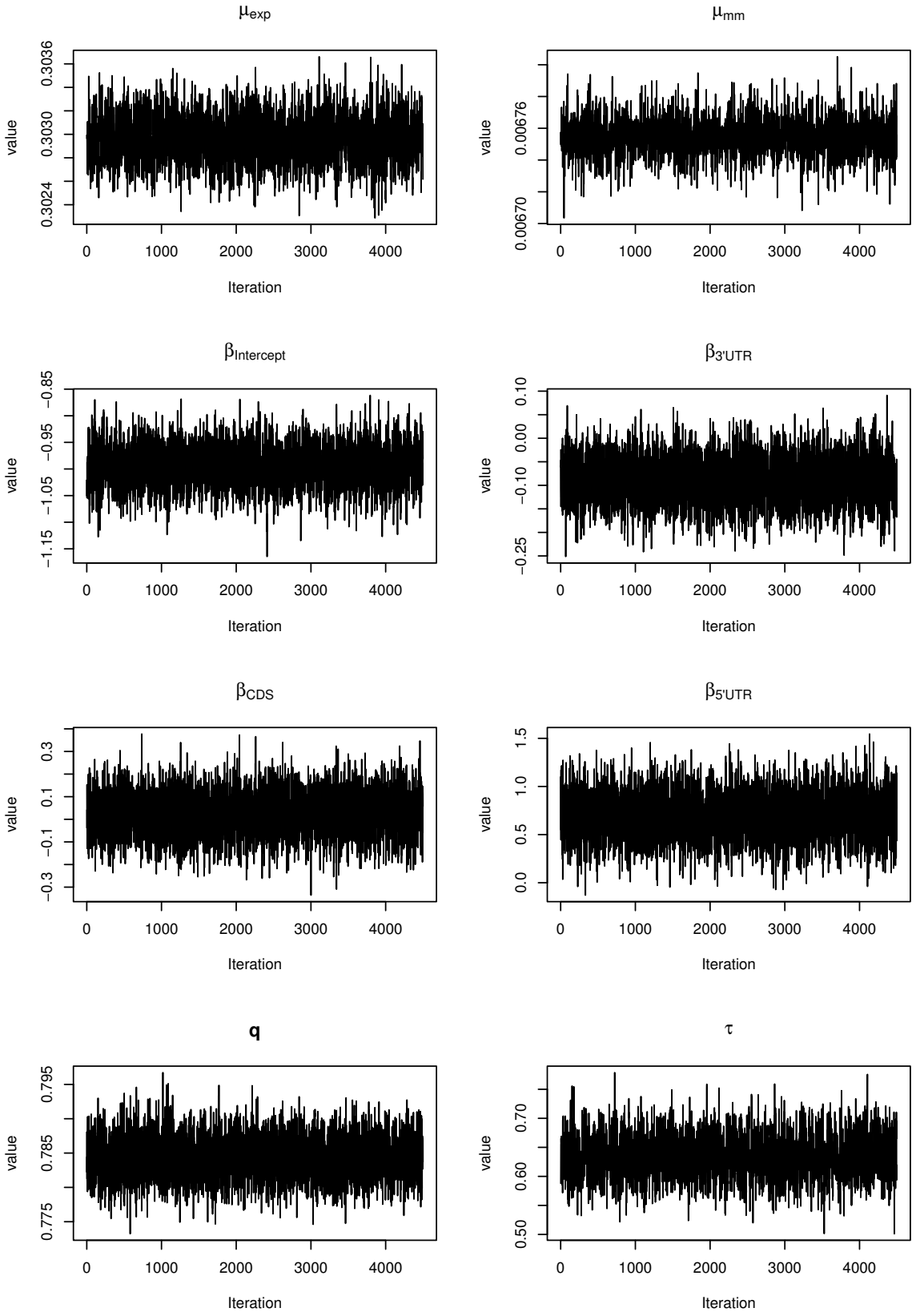

Figure B.9: Trace plots of the model parameters for BayMAP 2.0 applied to the Kishore B data. 4,500 iterations out of 75,000 were used for parameter estimation using every 15th iteration after a burn in of 7,500.

Trace plots for Memczak data set

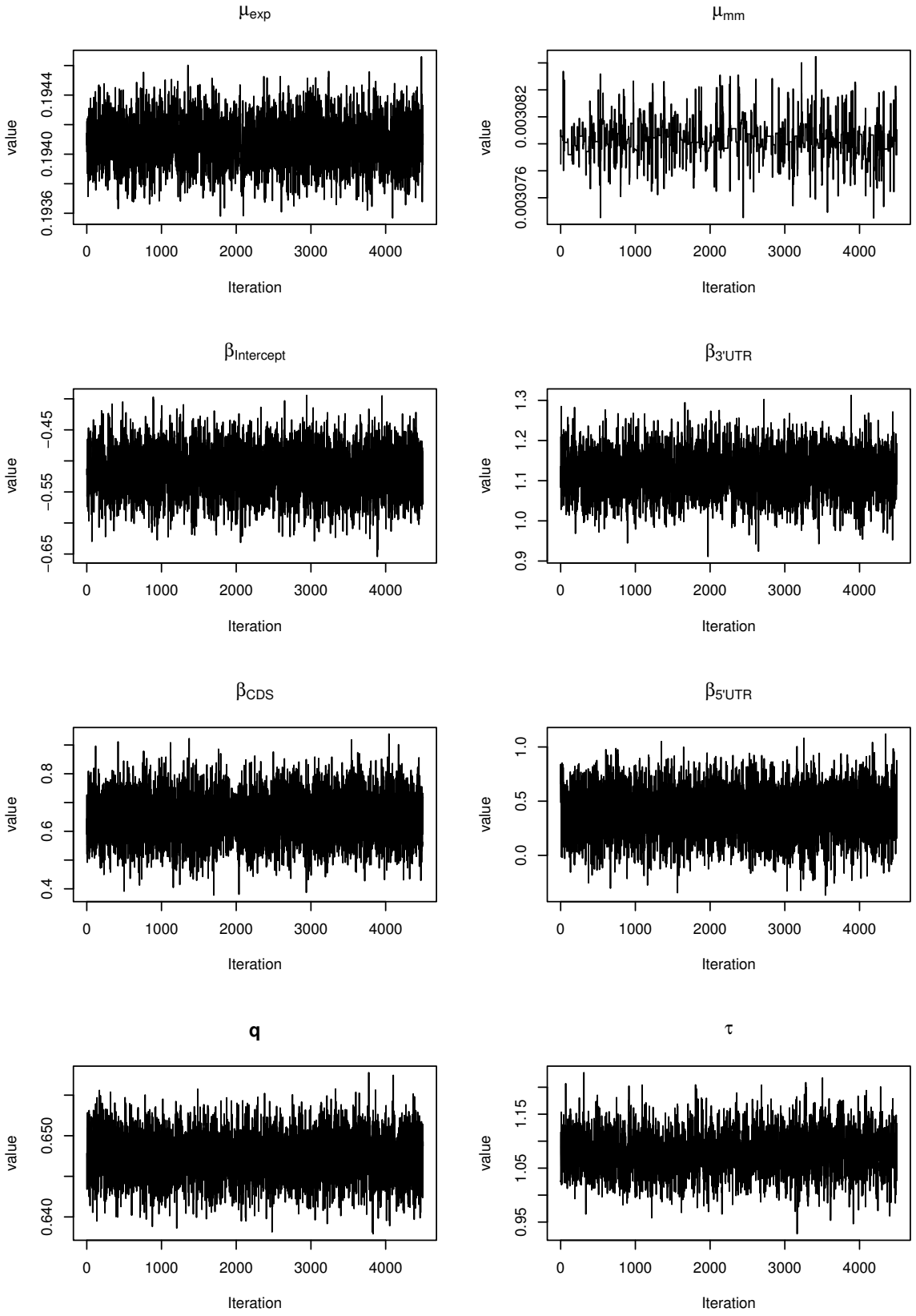

Figure B.10: Trace plots of the model parameters for BayMAP 2.0 applied to the Memczak data. 4,500 iterations out of 75,000 were used for parameter estimation using every 15th iteration after a burn in of 7,500.

Trace plots for Gottwein A data set

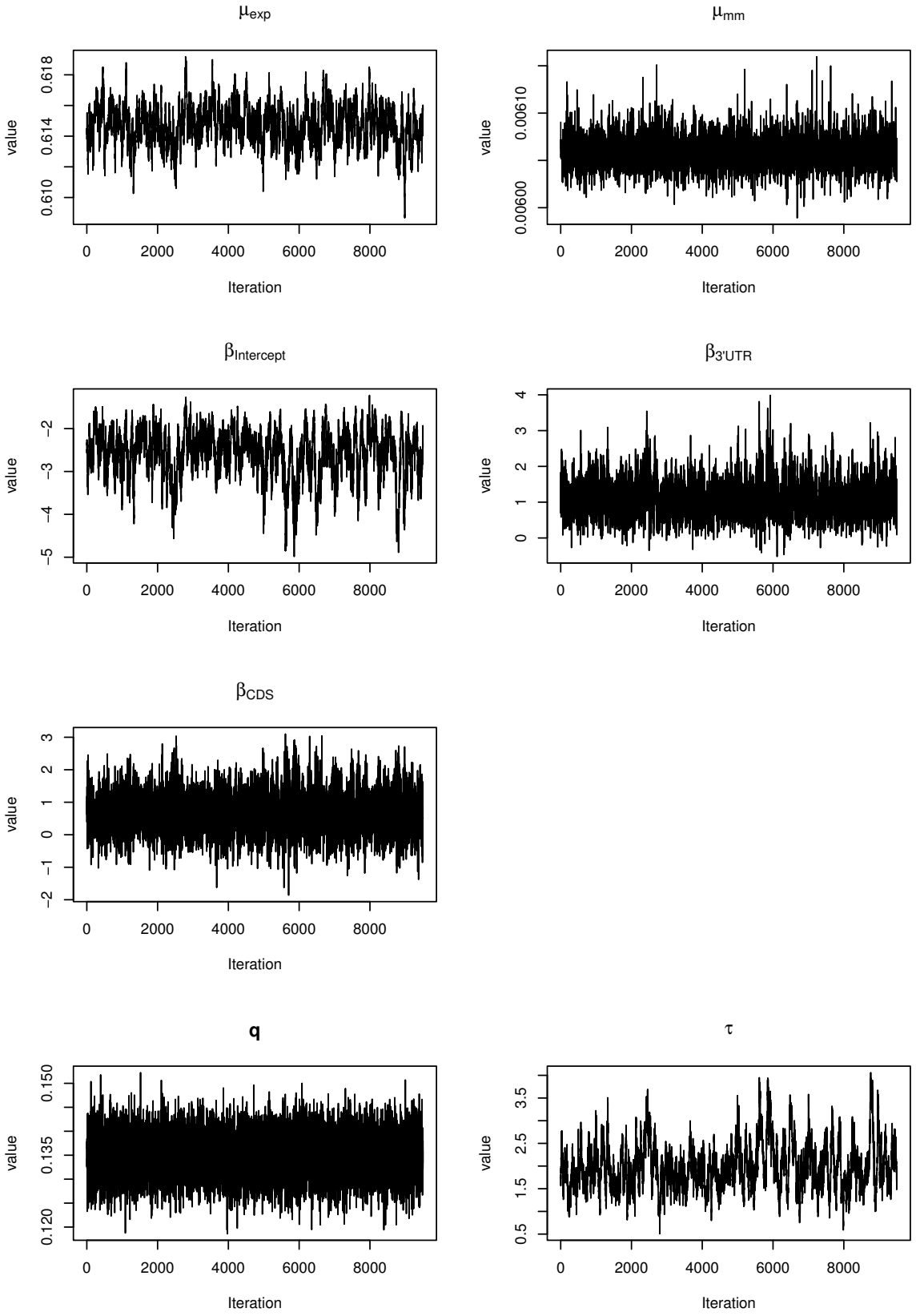

Figure B.11: Trace plots of the model parameters for BayMAP 2.0 applied to the Gottwein A data. 95,00 iterations are displayed out of 150,000 iterations here of which only the first 4,500 iterations were used for parameter estimation using every 15th iteration after a burn in of 7,500.

Trace plots for Gottwein B data set

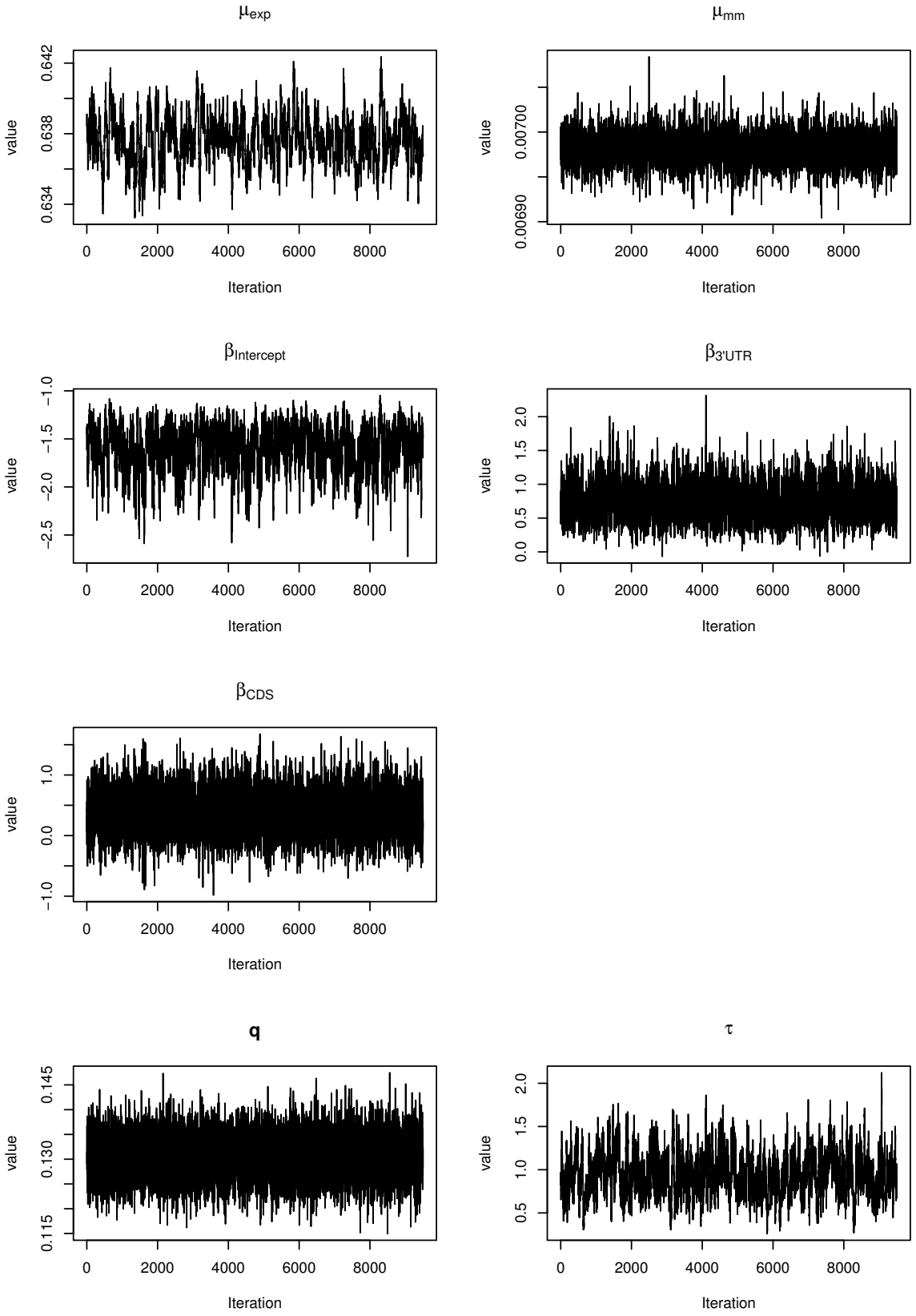

Figure B.12: Trace plots of the model parameters for BayMAP 2.0 applied to the Gottwein A data. 95,00 iterations are displayed out of 150,000 iterations here of which only the first 4,500 iterations were used for parameter estimation using every 15th iteration after a burn in of 7,500.

B.3 BayMAP with CAR

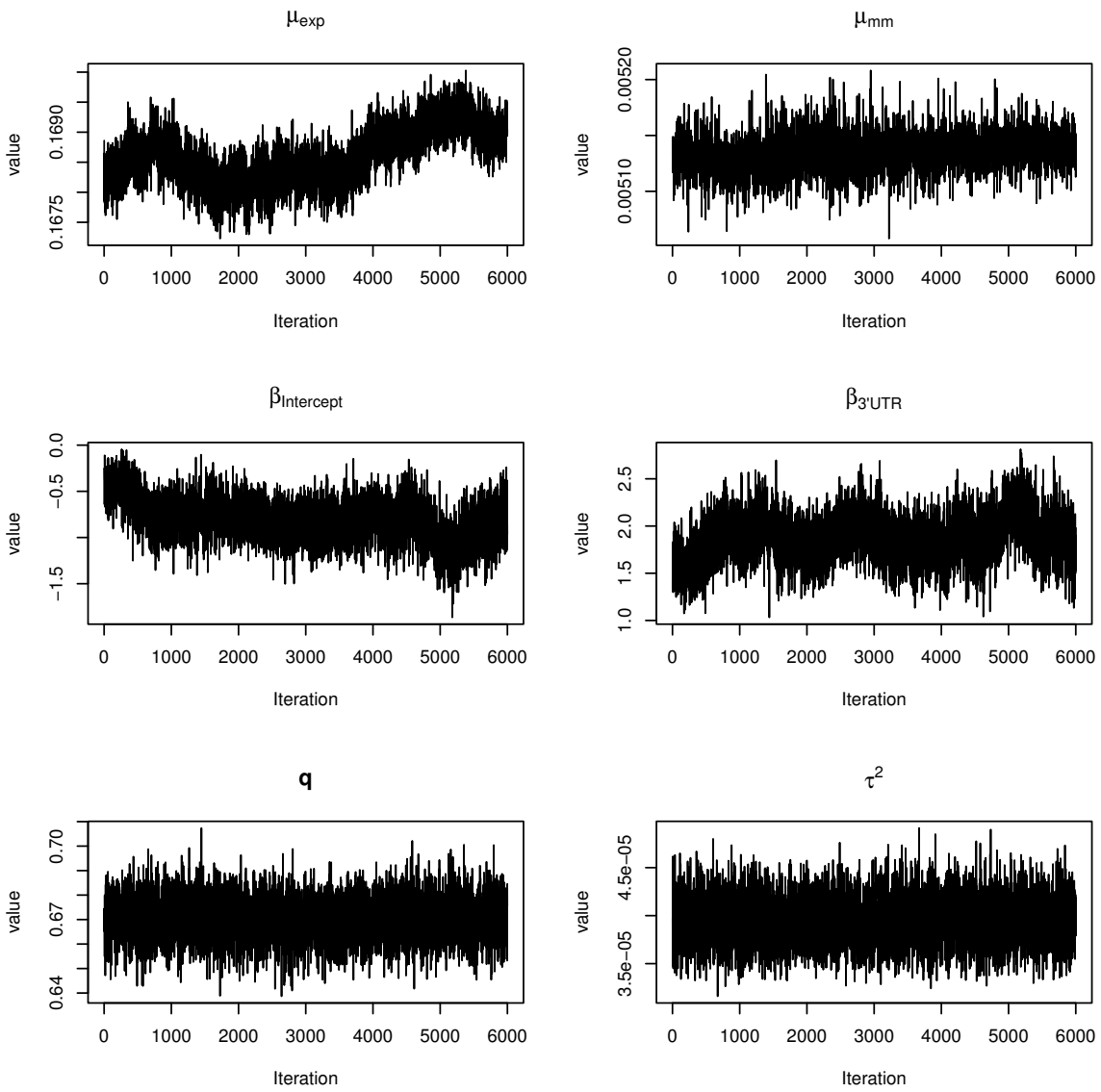

Figure B.13: Trace plots of the model parameters for the CAR model, where only T-to-C substitution positions are considered as neighbors when they are on the same binding site with no other T-to-C substitution positions in between. Out of the 46 million iterations, a burn-in of 16 million is used and only every 5,000th iteration is kept.

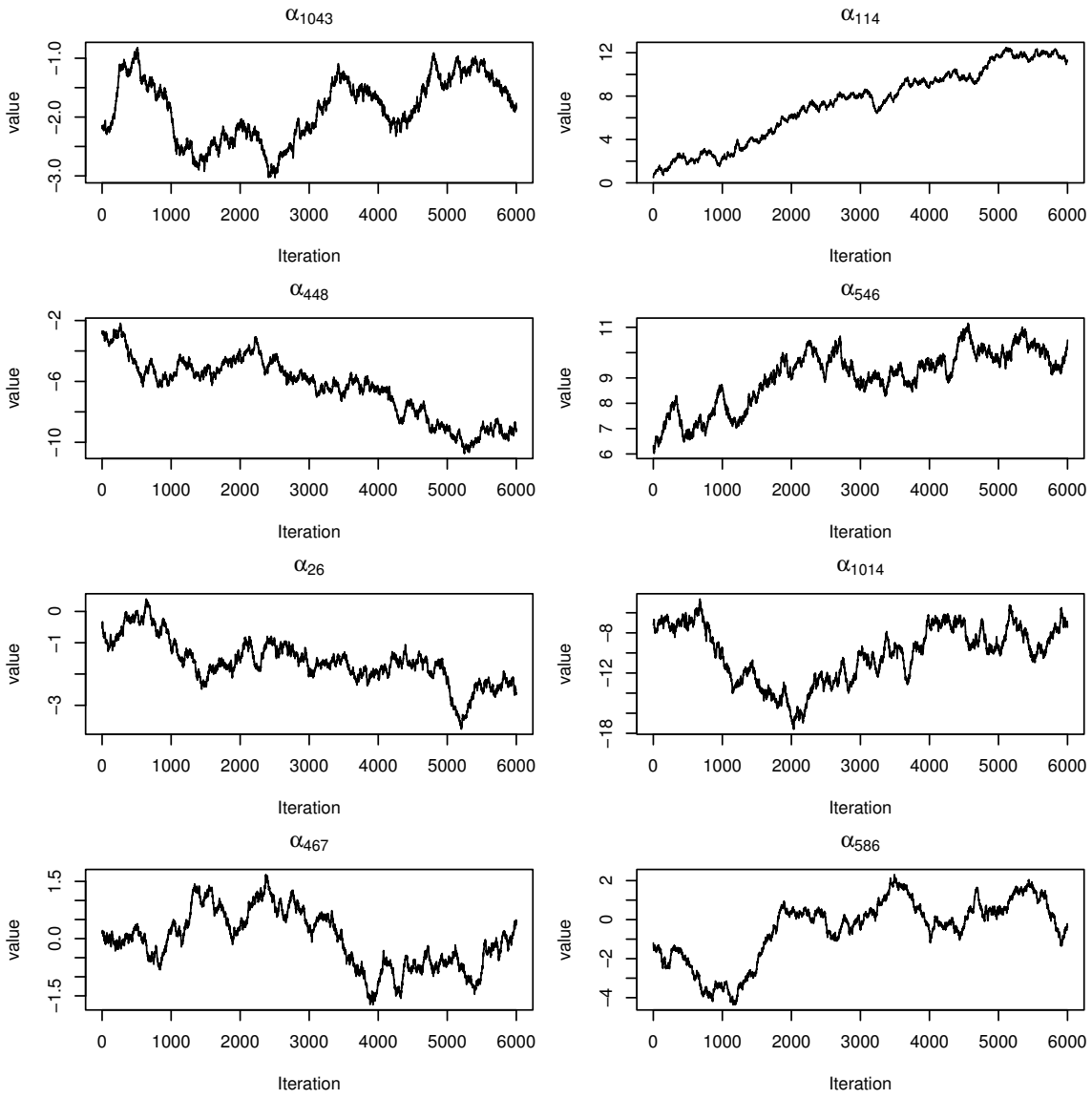

Figure B.14: Trace plots of eight randomly drawn $alpha_j$ for the CAR model, where only T-to-C substitution positions are considered as neighbors when they are on the same binding site with no other T-to-C substitution positions in between. Out of the 46 million iterations, a burn-in of 16 million is used and only every 5,000th iteration is kept.

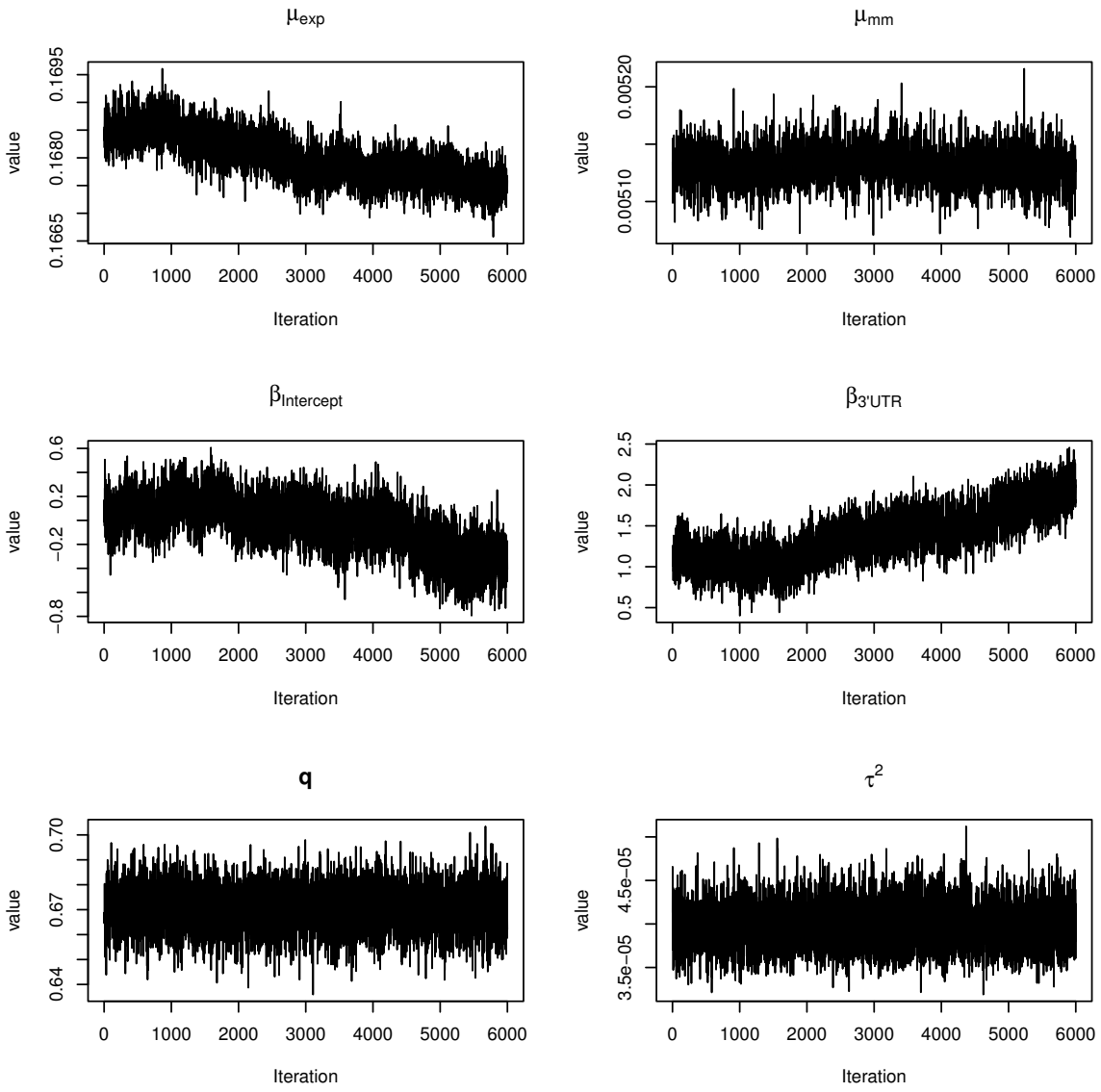

Figure B.15: Trace plots of the model parameters for the CAR model, where T-to-C substitution positions are considered as neighbors when they are on the same binding site. Out of the 46 million iterations, a burn-in of 16 million is used and only every 5,000th iteration is kept.

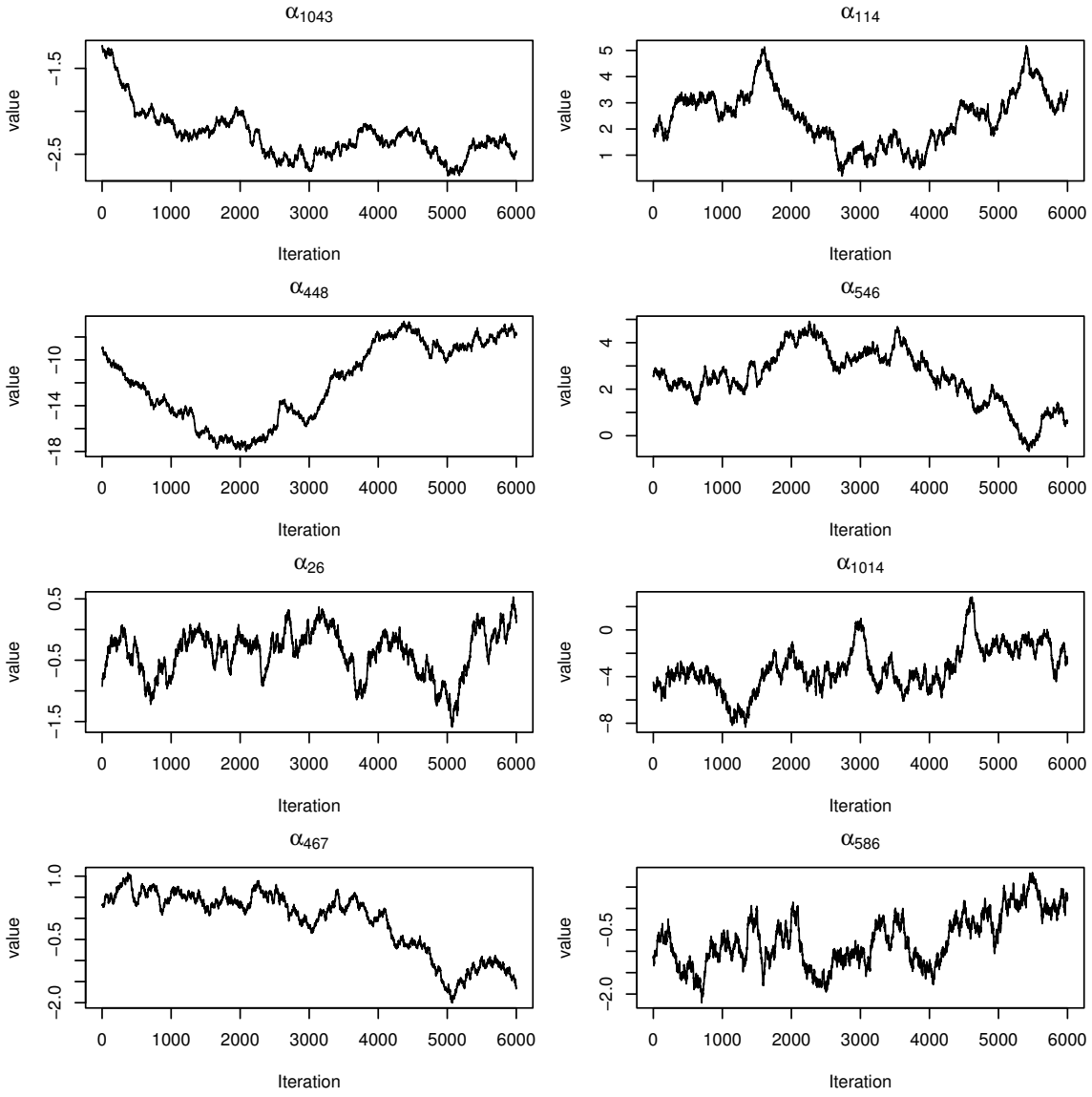

Figure B.16: Trace plots of eight randomly drawn α_j for the CAR model, where T-to-C substitution positions are considered as neighbors when they are on the same binding site. Out of the 46 million iterations, a burn-in of 16 million is used and only every 5,000th iteration is kept.

Appendix C

Additional simulation results

C.1 BayMAP 1.0

C.1.1 Bias in estimation

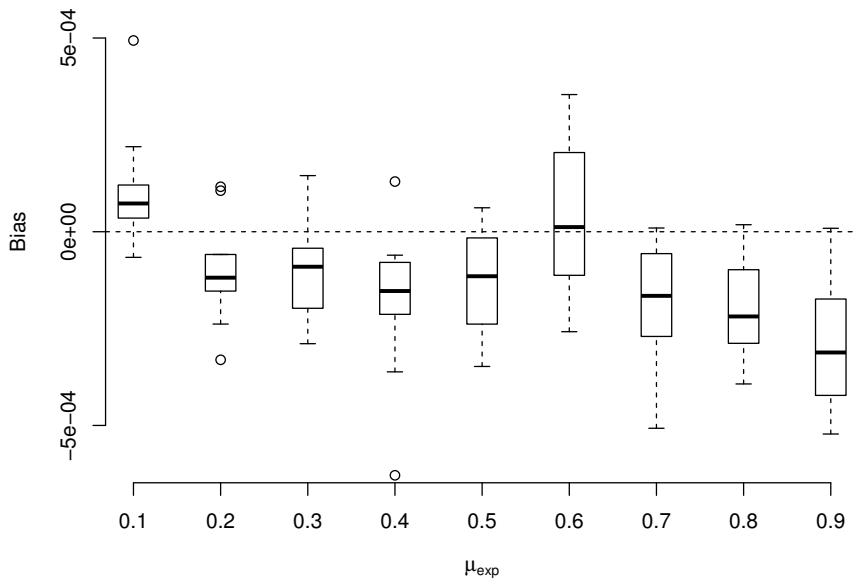

Figure C.1: Bias of the mean estimate for μ_{exp} for nine different values of μ_{exp} for BayMAP 1.0 for the simulation settings with $\beta_0 = -0.5$, $\beta_{3\text{'UTR}} = 1.5$, $\beta_{\text{CDS}} = 1.0$ and $\beta_{5\text{'UTR}} = 0.5$.

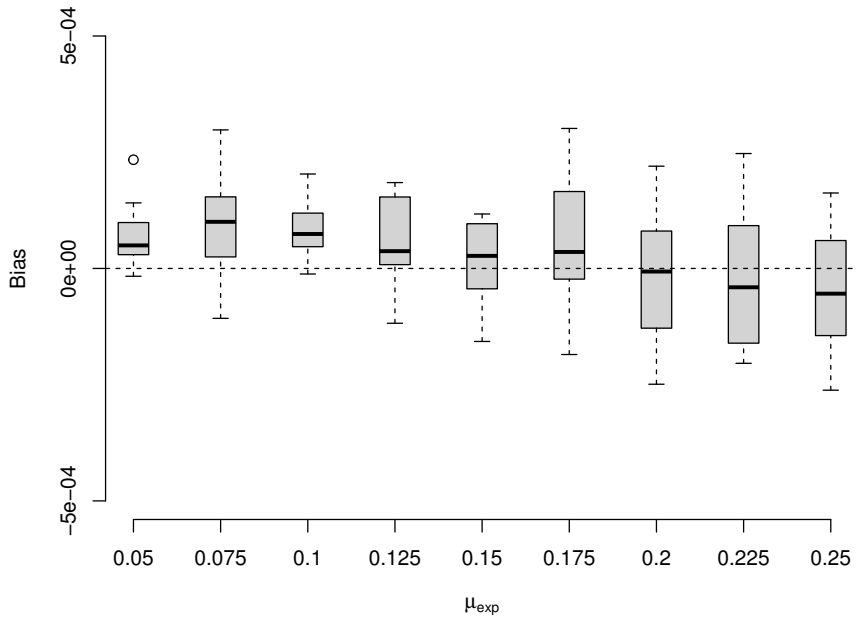

Figure C.2: Bias of the mean estimate for μ_{exp} for nine different values of μ_{exp} for BayMAP 1.0 for the simulation settings with $\beta_0 = 0.5$, $\beta_{3\text{'UTR}} = 1.85$, $\beta_{\text{CDS}} = 1.15$ and $\beta_{5\text{'UTR}} = 0.75$.

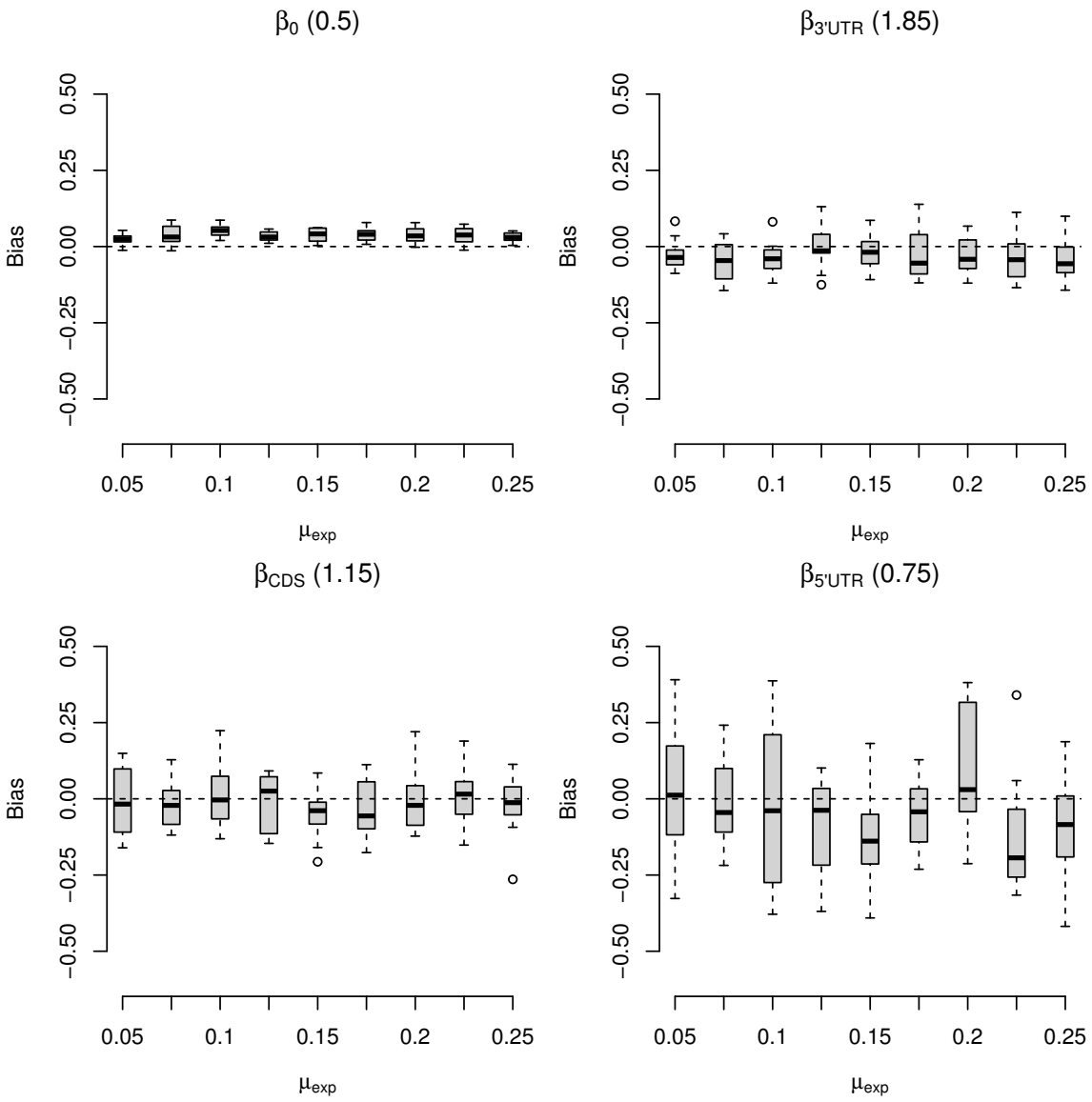

Figure C.3: Bias of the mean estimates for the regression parameters in the probit model in BayMAP in relation to the mRNA position (3'UTR, CDS, 5'UTR) with $\beta_0 = 0.5$, $\beta_{3'\text{UTR}} = 1.85$, $\beta_{\text{CDS}} = 1.15$ and $\beta_{5'\text{UTR}} = 0.75$. The true values of the parameters for β are shown in parentheses. For $\beta_{5'\text{UTR}}$ two outliers with values 0.70 and 0.76 for $\mu_{\text{exp}} = 0.4$ and $\mu_{\text{exp}} = 0.9$ are not displayed.

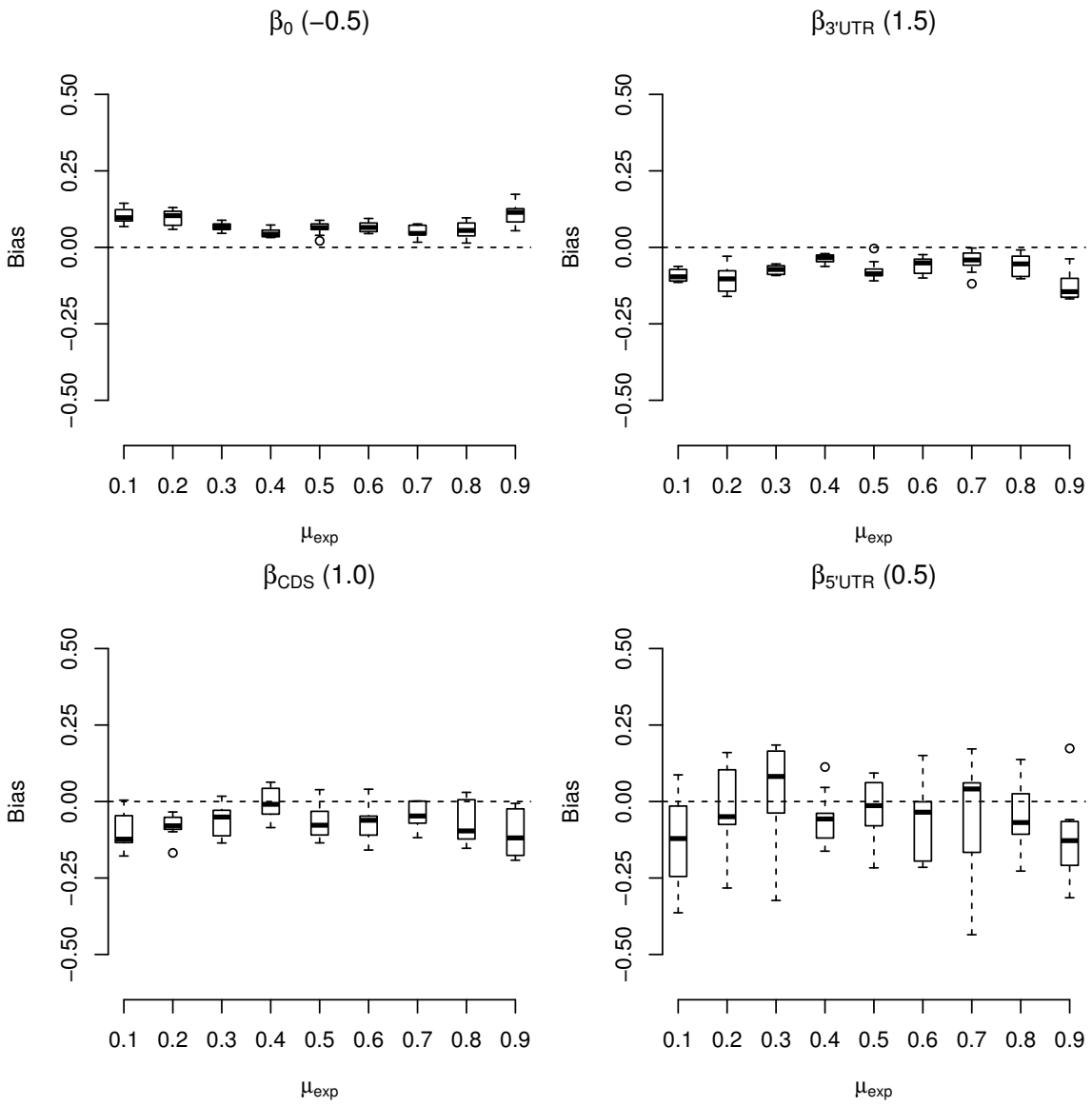

Figure C.4: Bias of the mean estimates for the regression parameters in the probit model in BayMAP in relation to the mRNA position (3'UTR, CDS, 5'UTR) with $\beta_0 = -0.5$, $\beta_{3'\text{UTR}} = 1.5$, $\beta_{\text{CDS}} = 1.0$ and $\beta_{5'\text{UTR}} = 0.5$. The true values of the parameters for β are shown in parentheses.

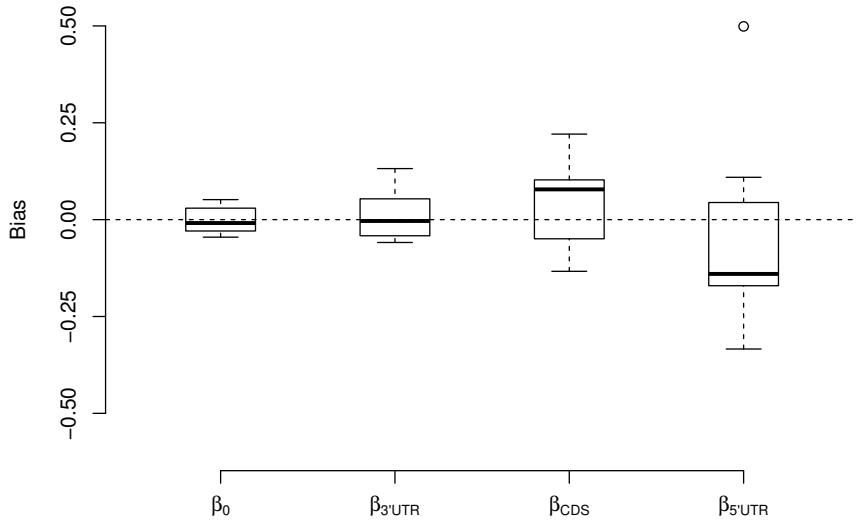

Figure C.5: Bias of the mean estimates for the regression parameters in the probit model in BayMAP in relation to the mRNA position (3'UTR, CDS, 5'UTR) with $\beta_0 = 0.5$, $\beta_{3'\text{UTR}} = 1.85$, $\beta_{\text{CDS}} = 1.15$ and $\beta_{5'\text{UTR}} = 0.75$ when μ_{exp} is drawn from the Beta$(2,10)$ distribution.

C.1.2 Comparison of BayMAP 1.0 to other methods

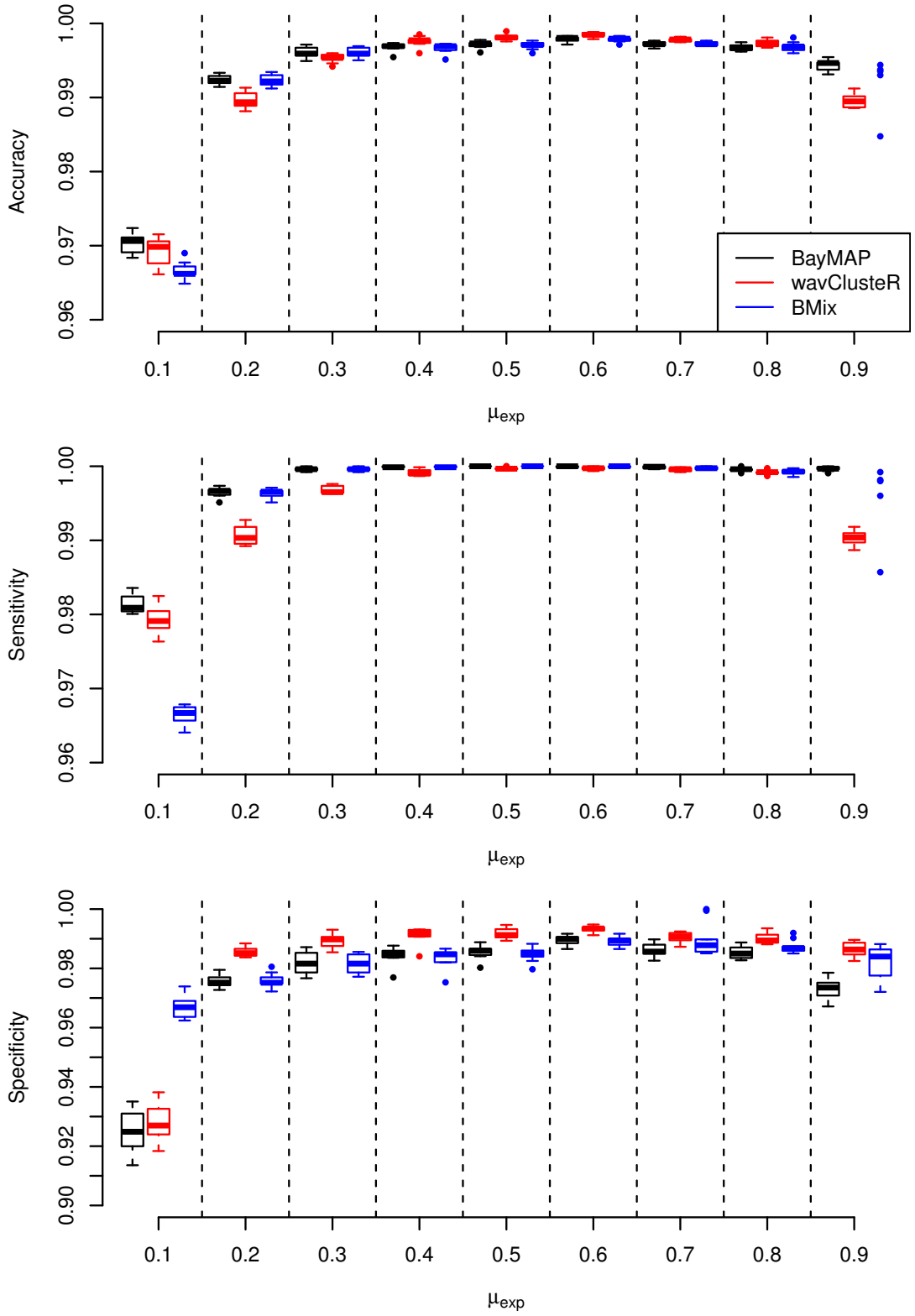

Figure C.6: (Simulation whole range μ_{exp} and no effect $\boldsymbol{\beta}$) Distribution of the accuracy (top panel), sensitivity (middle panel) and specificity (bottom panel) of BayMAP 1.0 (black box plots for each μ_{exp}), wavClusteR (red), and BMix (blue) considering probabilities μ_{exp} of T-to-C substitutions at an experimentally induced position between 0.1 and 0.9 with $p \approx 0.8$ for all positions in the simulated data sets and therefore without additional variables such as the mRNA. For a better graphical representation, very low values of sensitivity very close to zero and of accuracy of about 0.2 obtained in applications of BMix to data sets with $\mu_{exp} = 0.7$ or 0.9 are not shown. As thus the 25% quantiles for $\mu_{exp} = 0.9$ are about 0.2, the accuracies of BMix in this case are only displayed as points, but not as a box plot.

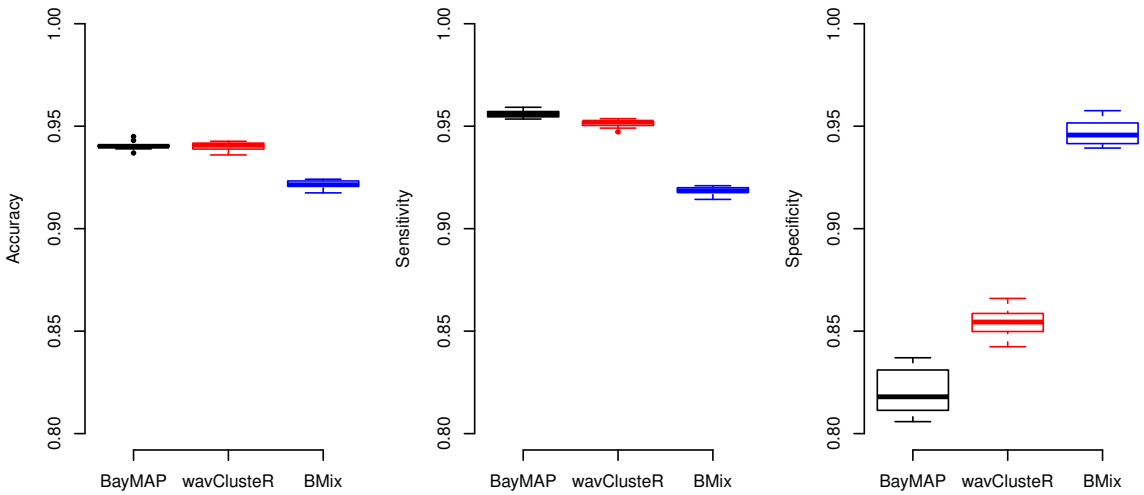

Figure C.7: (Simulation with beta distributed μ_{exp}) The accuracy (left panel), sensitivity (middle panel) and specificity (right panel) for BayMAP 1.0 (black box plots), wavClusteR (red box plots), and BMix (blue box plots) on ten simulated data sets for which μ_{exp} was drawn from a Beta(2,10) distribution for each position separately.

C.2 BayMAP 2.0

C.2.1 Bias in estimation

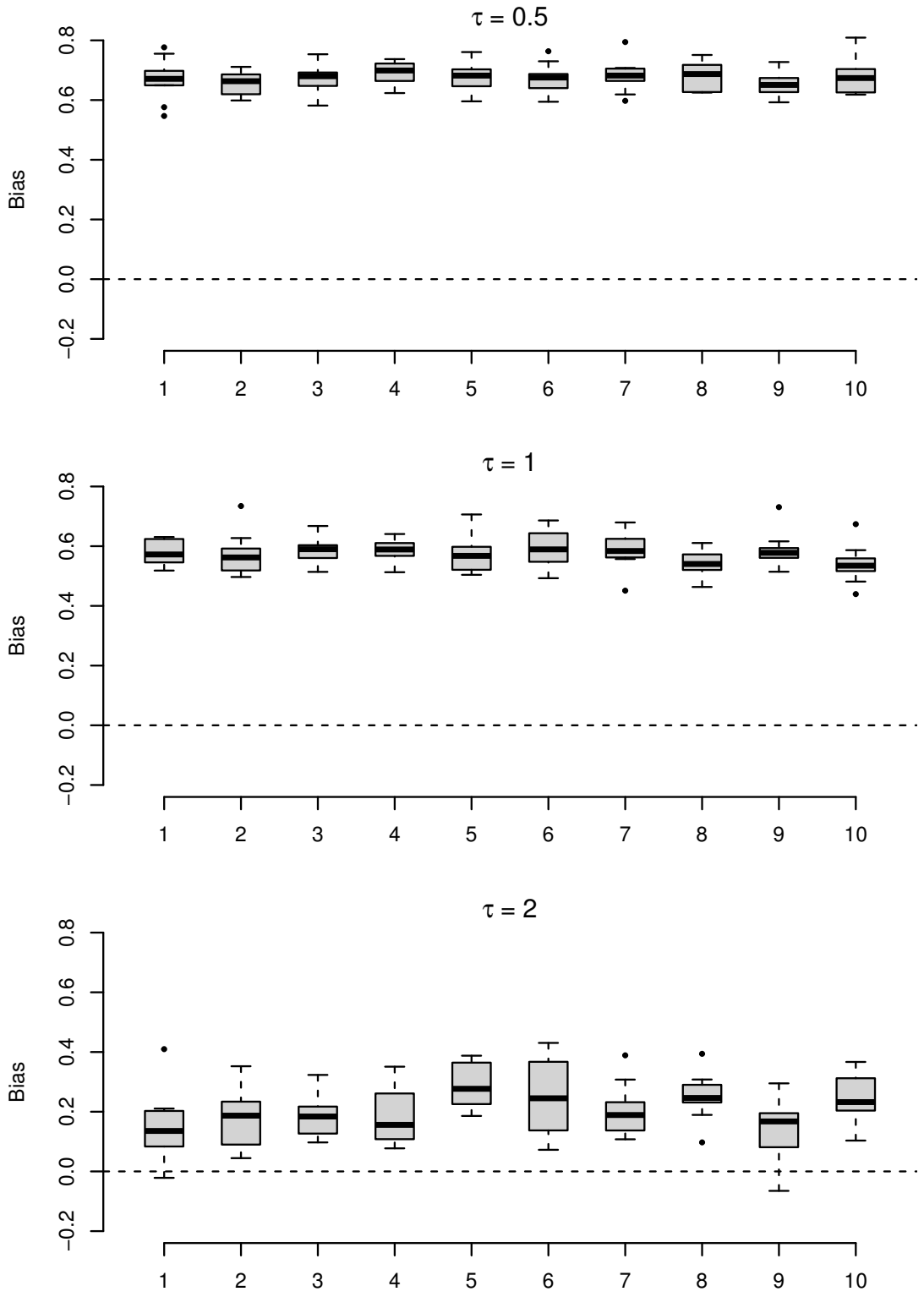

Figure C.8: Bias of the mean estimates for τ considering ten draws of $\boldsymbol{\zeta}$ for each value of τ with $\beta_0 = -0.5$, $\beta_{3'\text{UTR}} = 1.5$, $\beta_{\text{CDS}} = 1.0$ and $\beta_{5'\text{UTR}} = 0.5$.

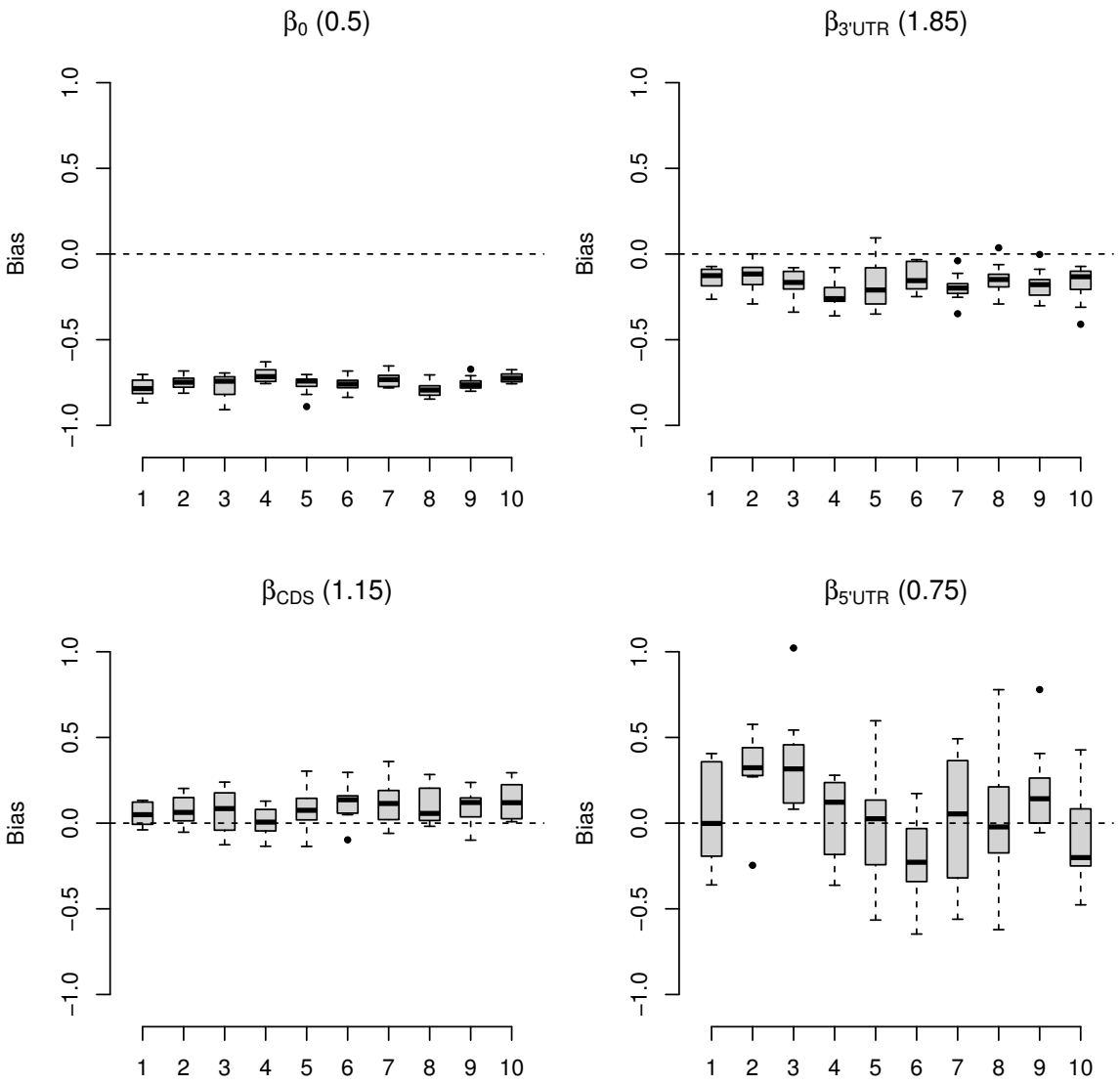

Figure C.9: Bias of the mean estimates for the regression parameters in the probit model in BayMAP 2.0 in relation to the mRNA position (3'UTR, CDS, 5'UTR) when $\tau = 0.5$. The true values of the parameters for β are shown in parentheses.

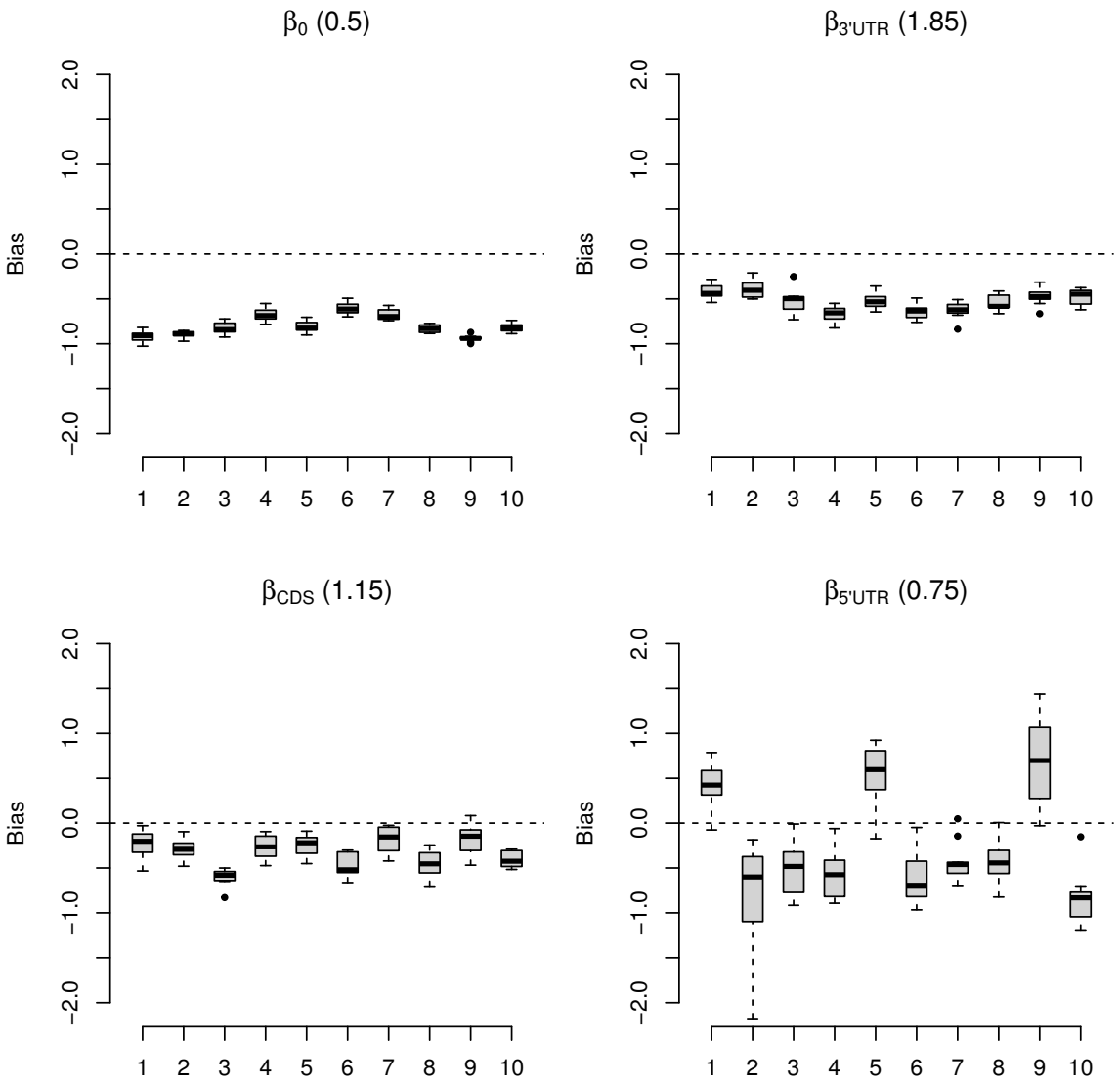

Figure C.10: Bias of the mean estimates for the regression parameters in the probit model in BayMAP 2.0 in relation to the mRNA position (3'UTR, CDS, 5'UTR) when $\tau = 2$. The true values of the parameters for β are shown in parentheses.

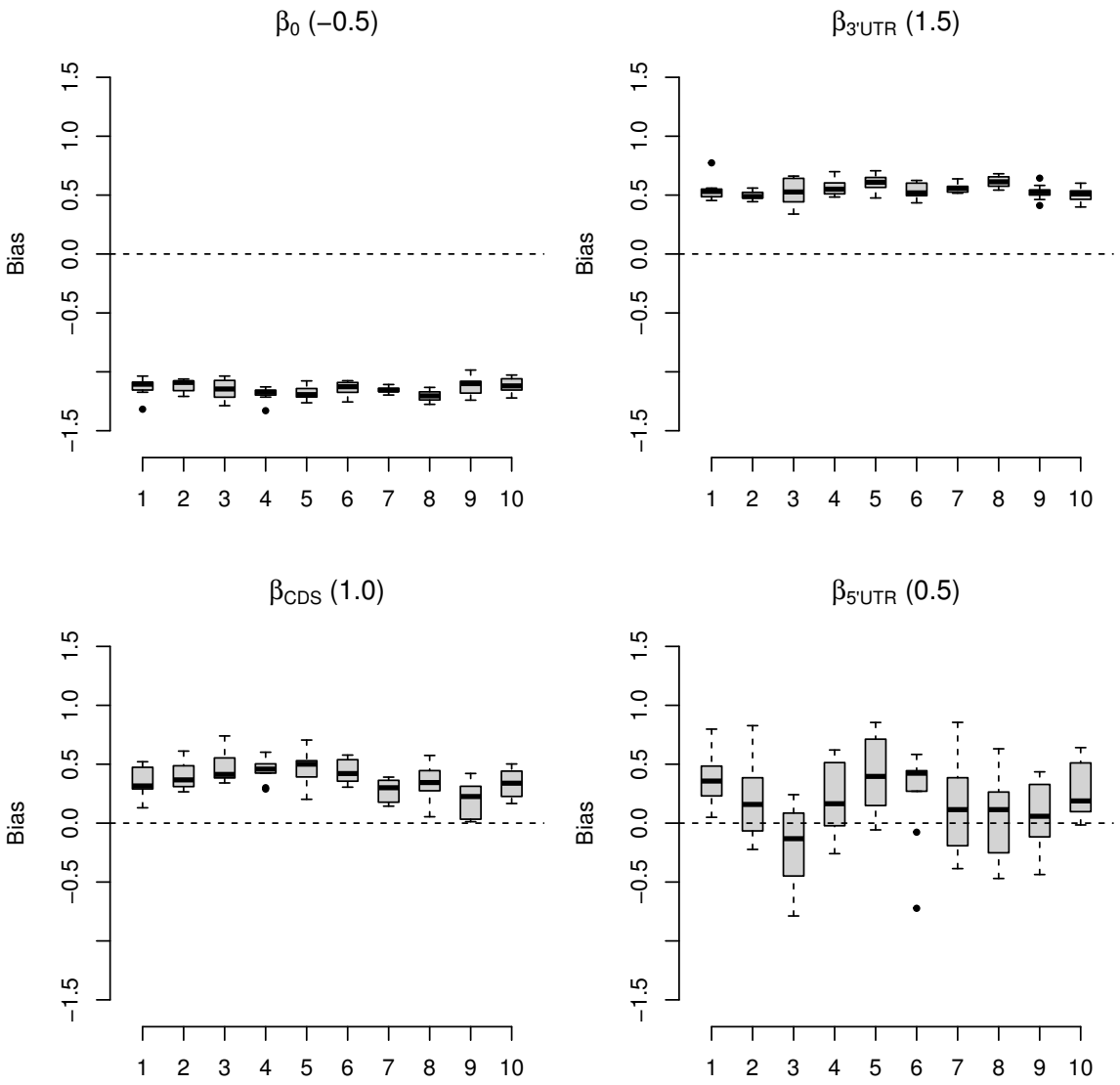

Figure C.11: Bias of the mean estimates for the regression parameters in the probit model in BayMAP 2.0 in relation to the mRNA position (3'UTR, CDS, 5'UTR) when $\tau = 0.5$. The true values of the parameters for β are shown in parentheses.

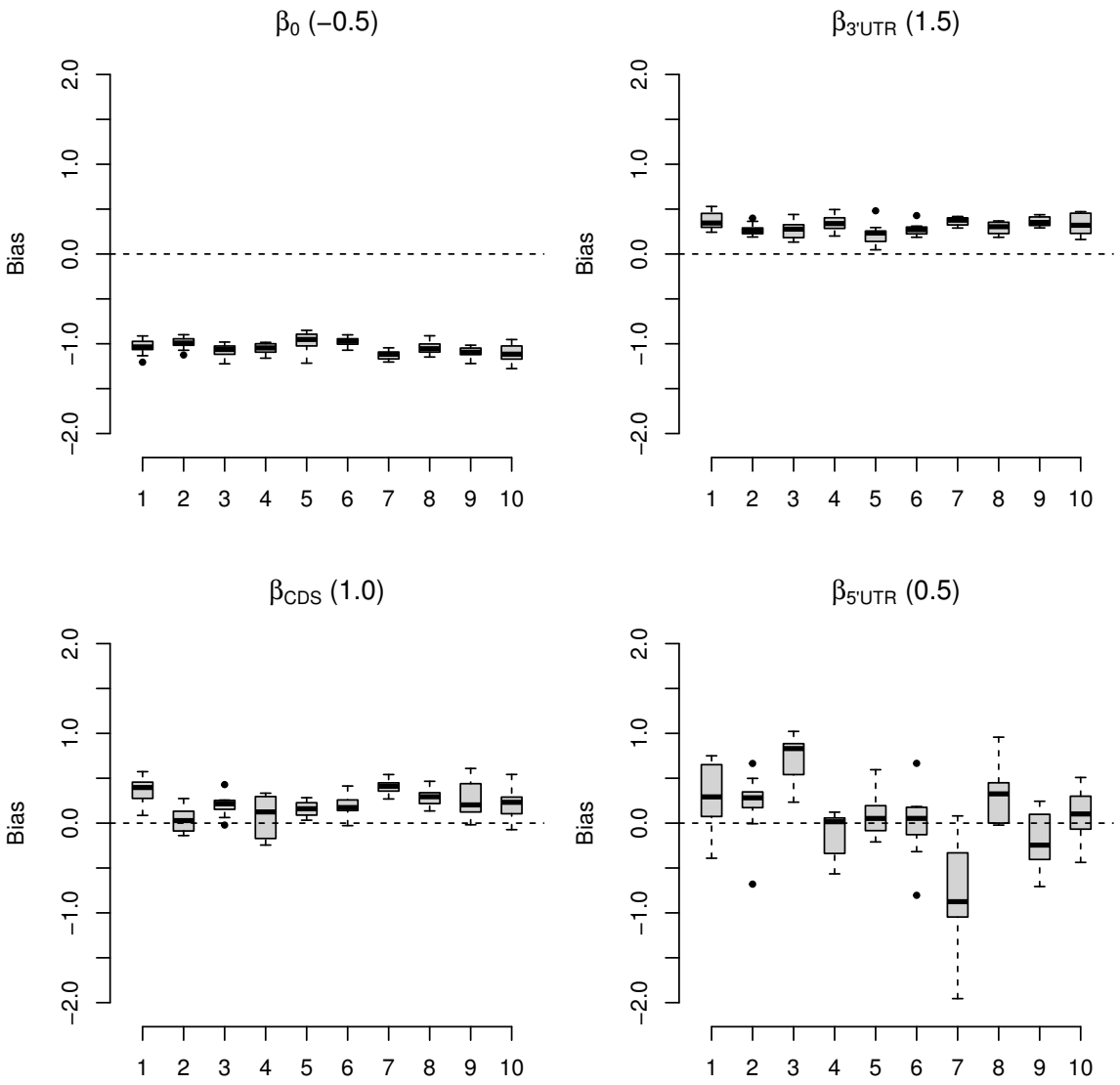

Figure C.12: Bias of the mean estimates for the regression parameters in the probit model in BayMAP 2.0 in relation to the mRNA position (3'UTR, CDS, 5'UTR) when $\tau = 1$. The true values of the parameters for β are shown in parentheses.

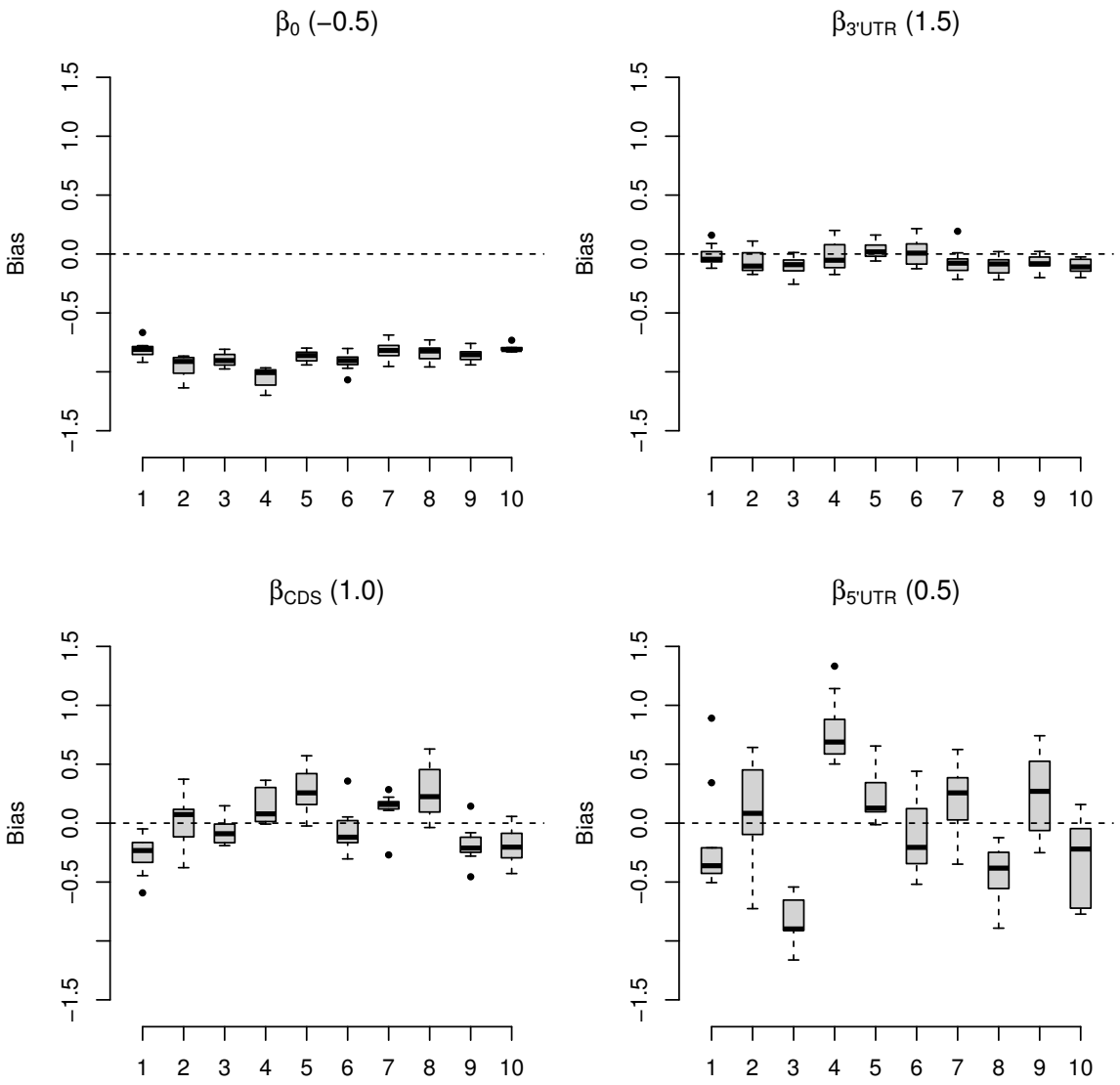

Figure C.13: Bias of the mean estimates for the regression parameters in the probit model in BayMAP 2.0 in relation to the mRNA position (3'UTR, CDS, 5'UTR) when $\tau = 2$. The true values of the parameters for β are shown in parentheses.

C.2.2 Comparison of BayMAP 2.0 to BayMAP 1.0

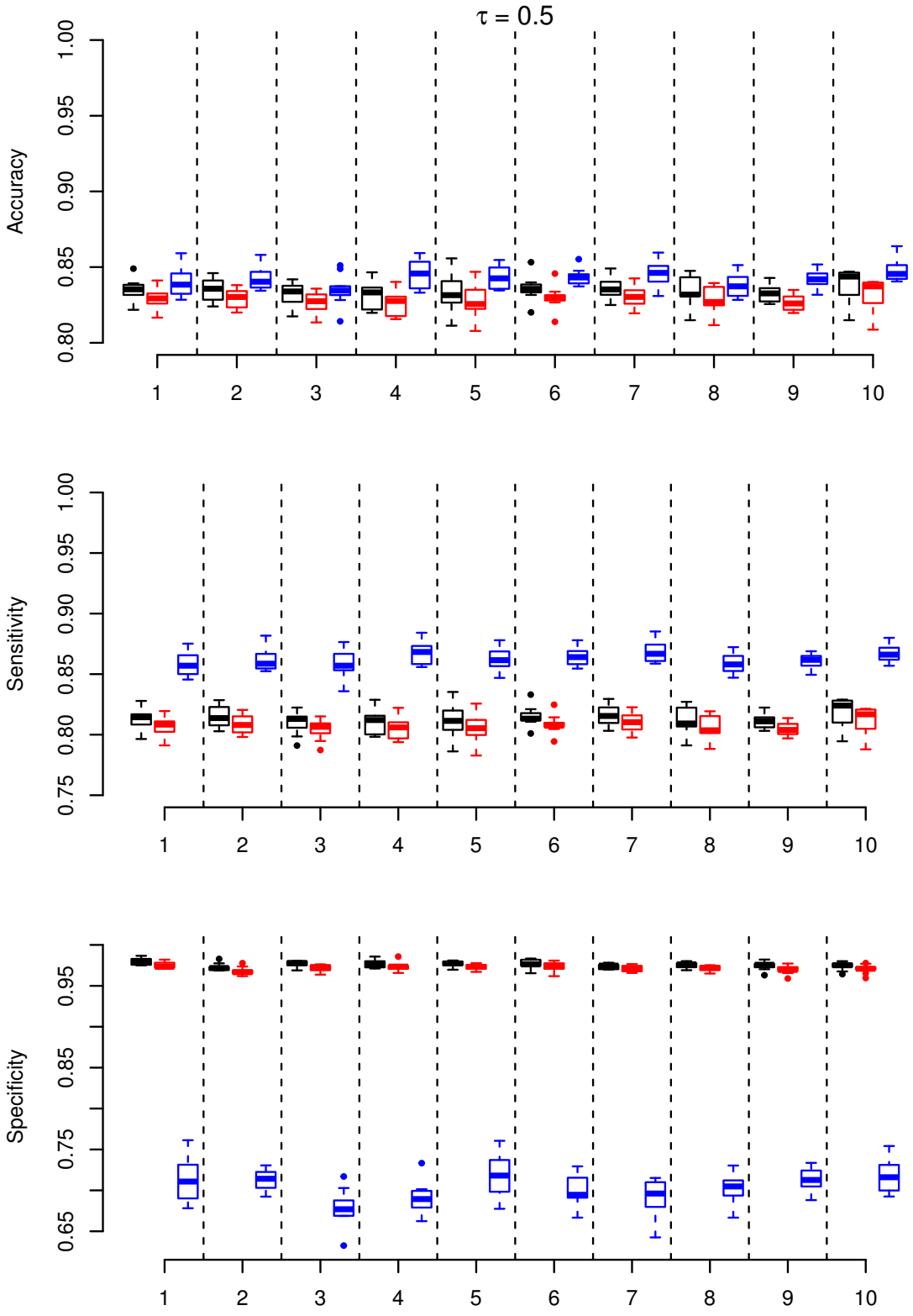

Figure C.14: (Simulation $\tau = 0.5$ with large $\boldsymbol{\beta}$) Distribution of the accuracy (top panel), sensitivity (middle panel) and specificity (bottom panel) of BayMAP 2.0 (black box plots for each different simulation for ζ), BayMAP 1.0 (red), BayMAP 2.0 prior odds (blue) considering ten draws of $\boldsymbol{\zeta}$ for $\tau = 0.5$ with $\beta_0 = 0.5$, $\beta_{3'\text{UTR}} = 1.85$, $\beta_{\text{CDS}} = 1.15$ and $\beta_{5'\text{UTR}} = 0.75$.

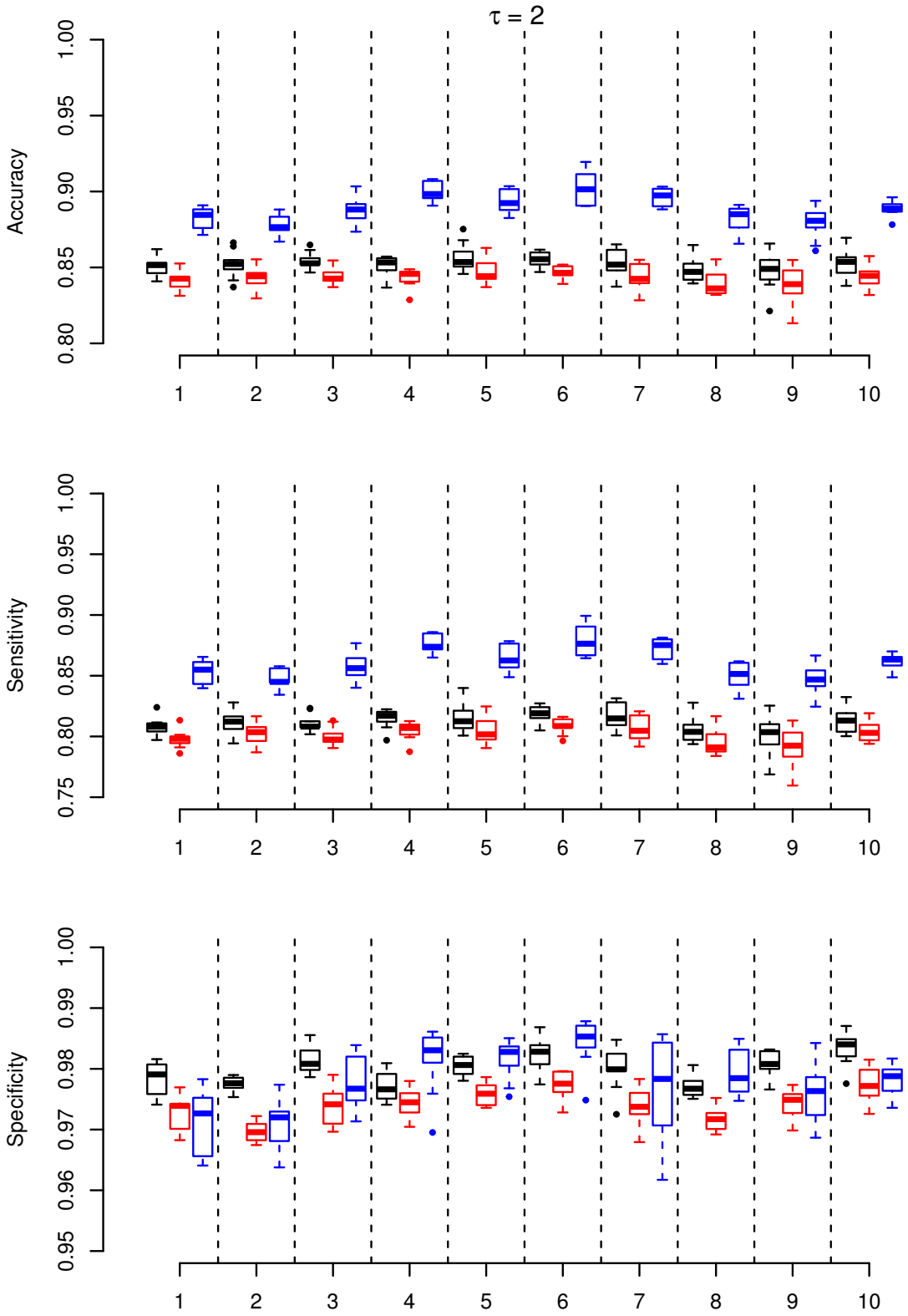

Figure C.15: (Simulation $\tau = 2$ with large $\boldsymbol{\beta}$)Distribution of the accuracy (top panel), sensitivity (middle panel) and specificity (bottom panel) of BayMAP 2.0 (black box plots for each different simulation for ζ), BayMAP 1.0 (red), BayMAP 2.0 prior odds (blue) considering ten draws of $\boldsymbol{\zeta}$ for $\tau = 2$ with $\beta_0 = 0.5$, $\beta_{3'\text{UTR}} = 1.85$, $\beta_{\text{CDS}} = 1.15$ and $\beta_{5'\text{UTR}} = 0.75$.

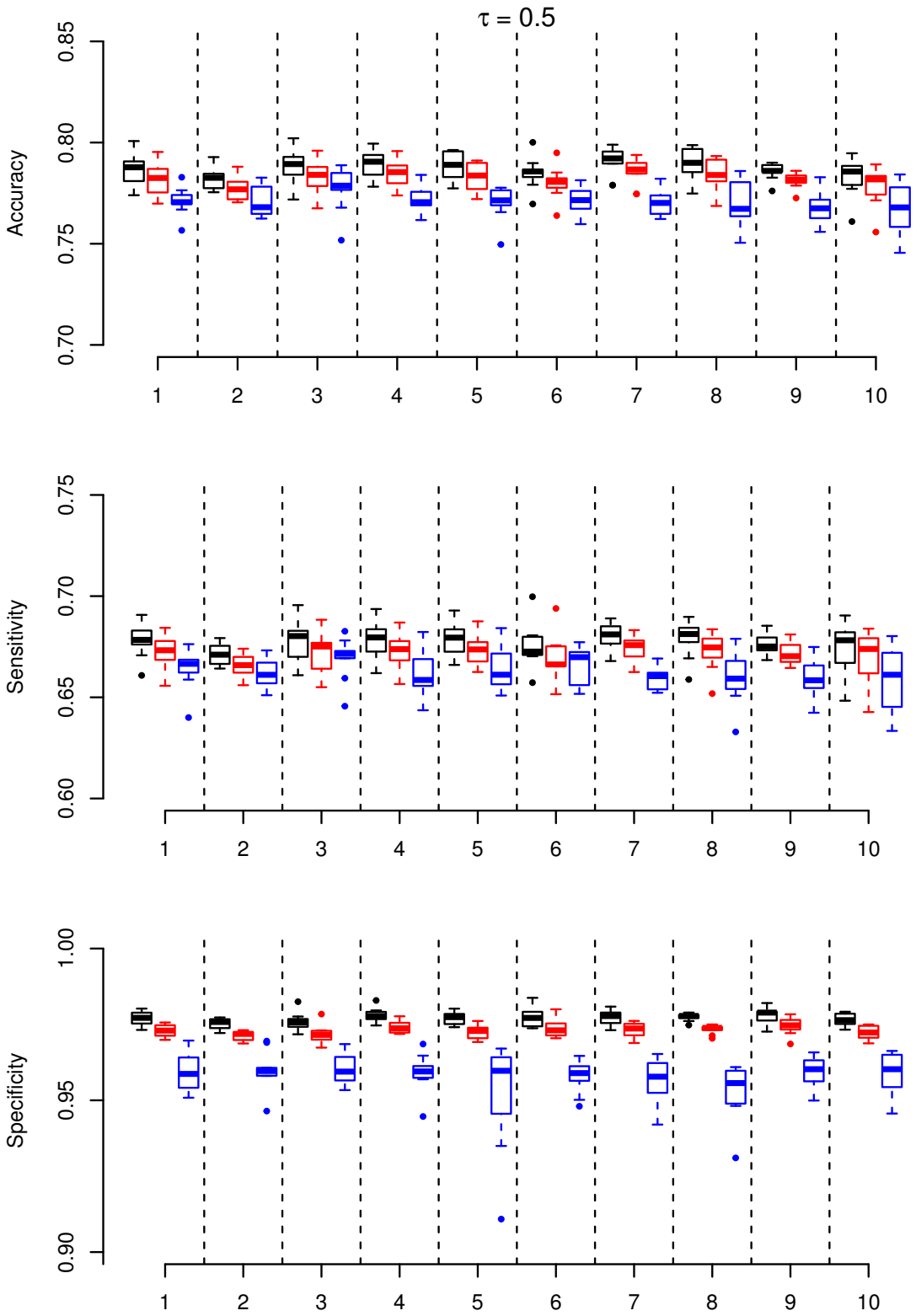

Figure C.16: (Simulation $\tau = 0.5$ with small $\boldsymbol{\beta}$) Distribution of the accuracy (top panel), sensitivity (middle panel) and specificity (bottom panel) of BayMAP 2.0 (black box plots for each different simulation for ζ), BayMAP 1.0 (red), BayMAP 2.0 prior odds (blue) considering ten draws of $\boldsymbol{\zeta}$ for $\tau = 0.5$ with $\beta_0 = -0.5$, $\beta_{3'\text{UTR}} = 1.5$, $\beta_{\text{CDS}} = 1.0$ and $\beta_{5'\text{UTR}} = 0.5$.

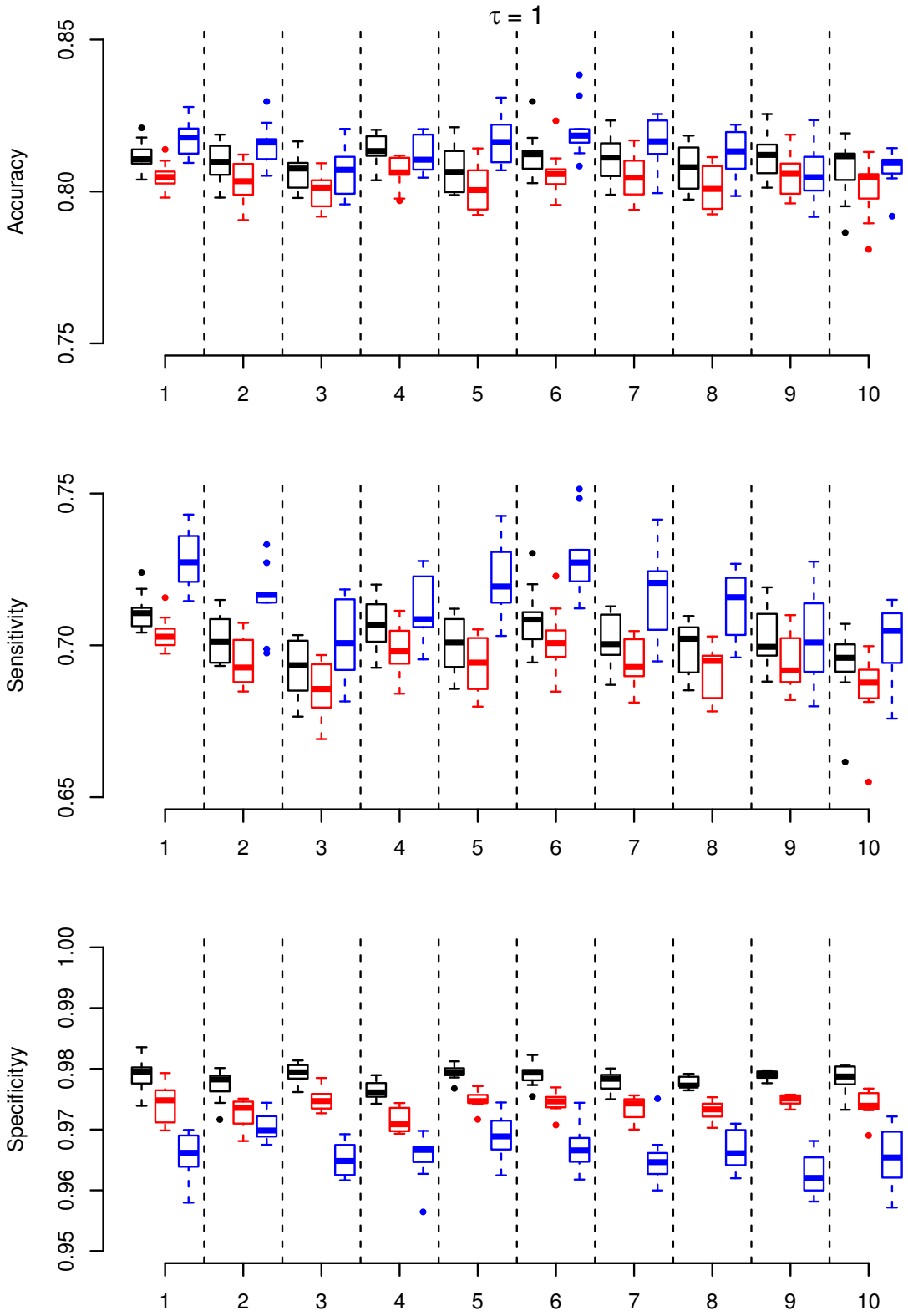

Figure C.17: (Simulation $\tau = 1$ with small $\boldsymbol{\beta}$) Distribution of the accuracy (top panel), sensitivity (middle panel) and specificity (bottom panel) of BayMAP 2.0 (black box plots for each different simulation for ζ), BayMAP 1.0 (red), BayMAP 2.0 prior odds (blue) considering ten draws of $\boldsymbol{\zeta}$ for $\tau = 1$ with $\beta_0 = -0.5$, $\beta_{3'\text{UTR}} = 1.5$, $\beta_{\text{CDS}} = 1.0$ and $\beta_{5'\text{UTR}} = 0.5$.

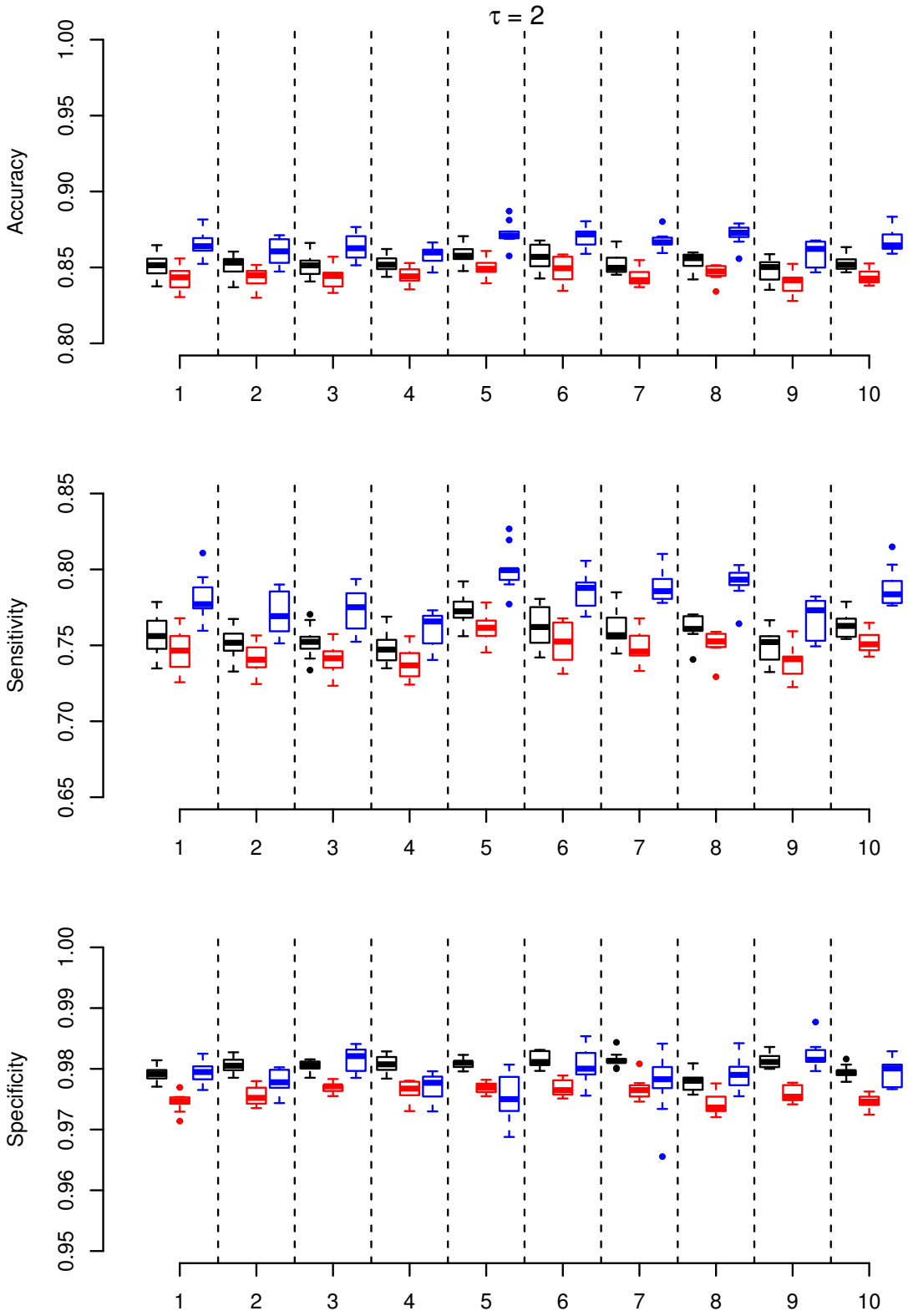

Figure C.18: (Simulation $\tau = 2$ with small $\boldsymbol{\beta}$) Distribution of the accuracy (top panel), sensitivity (middle panel) and specificity (bottom panel) of BayMAP 2.0 (black box plots for each different simulation for ζ), BayMAP 1.0 (red), BayMAP 2.0 prior odds (blue) considering ten draws of $\boldsymbol{\zeta}$ for $\tau = 2$ with $\beta_0 = -0.5$, $\beta_{3'\text{UTR}} = 1.5$, $\beta_{\text{CDS}} = 1.0$ and $\beta_{5'\text{UTR}} = 0.5$.

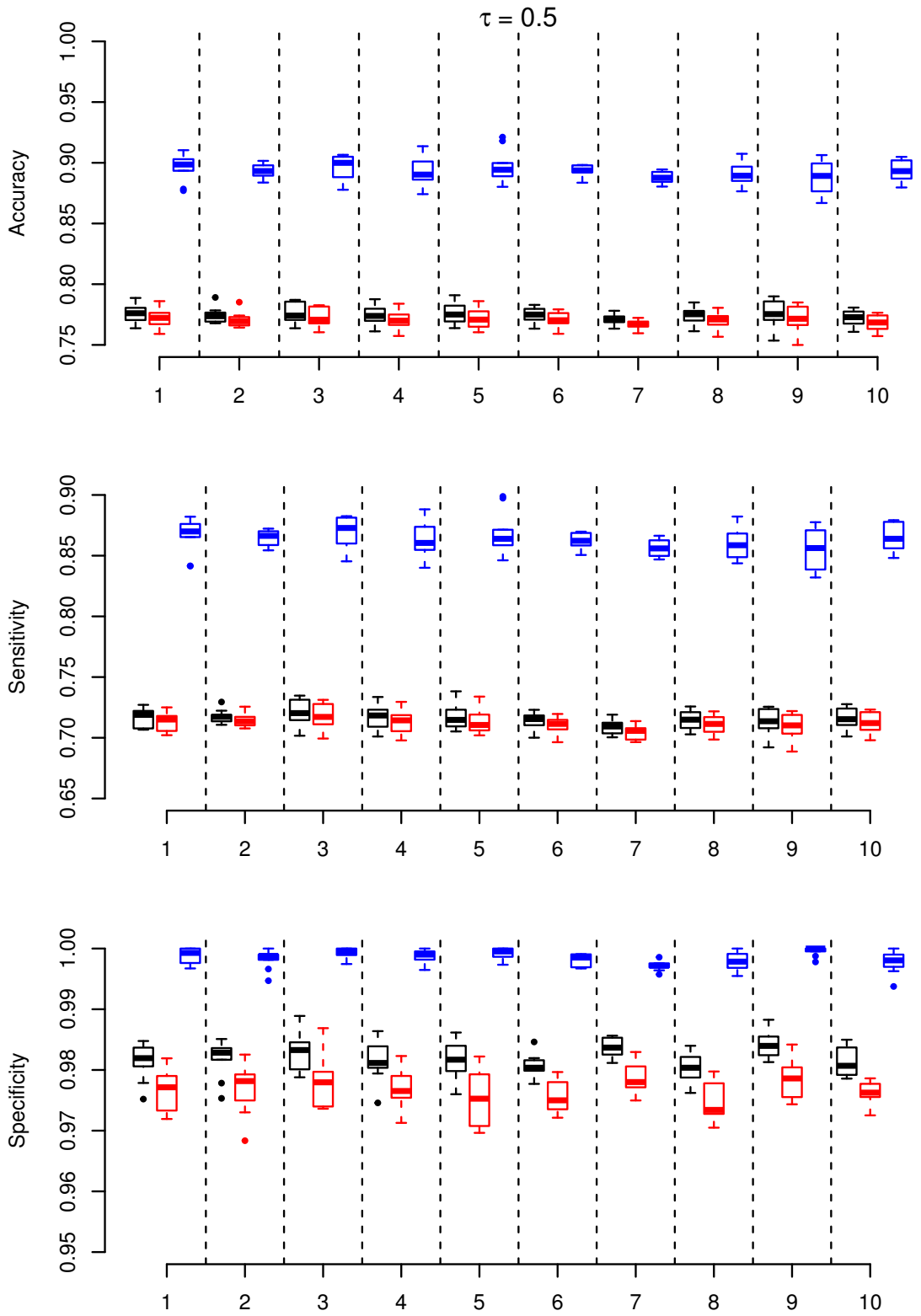

Figure C.19: (Simulation $\tau = 0.5$ with no effect $\boldsymbol{\beta}$) Distribution of the accuracy (top panel), sensitivity (middle panel) and specificity (bottom panel) of BayMAP 2.0 (black box plots for each different simulation for ζ), BayMAP 1.0 (red), BayMAP 2.0 prior odds (blue) considering ten draws of $\boldsymbol{\zeta}$ for $\tau = 0.5$ with $\beta_0 = 0.85$ and $\beta_{3'\text{UTR}} = \beta_{\text{CDS}} = \beta_{5'\text{UTR}} = 0$.

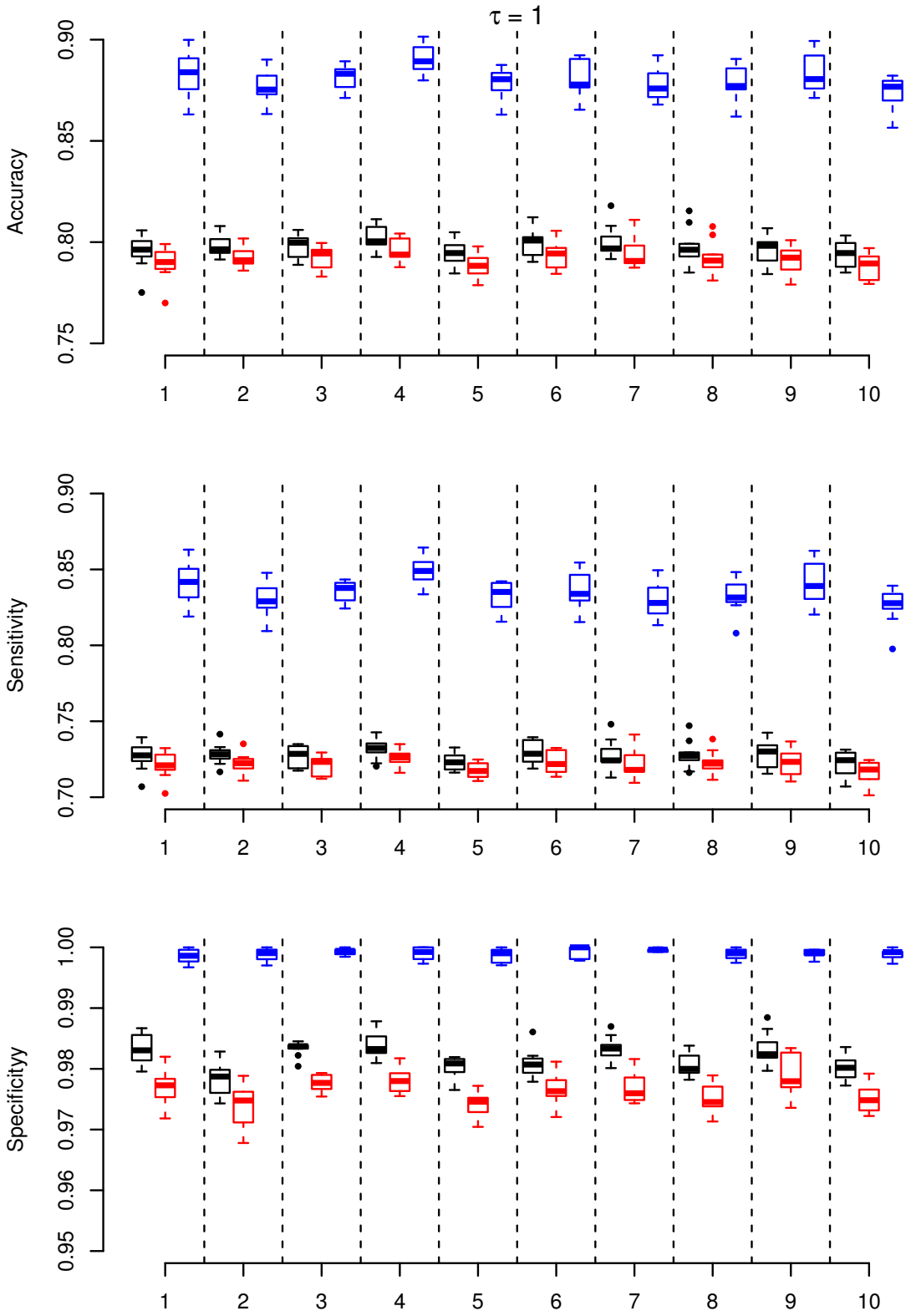

Figure C.20: (Simulation $\tau = 1$ with no effect $\boldsymbol{\beta}$) Distribution of the accuracy (top panel), sensitivity (middle panel) and specificity (bottom panel) of BayMAP 2.0 (black box plots for each different simulation for ζ), BayMAP 1.0 (red), BayMAP 2.0 prior odds (blue) considering ten draws of $\boldsymbol{\zeta}$ for $\tau = 1$ with $\beta_0 = 0.85$ and $\beta_{3'\text{UTR}} = \beta_{\text{CDS}} = \beta_{5'\text{UTR}} = 0$.

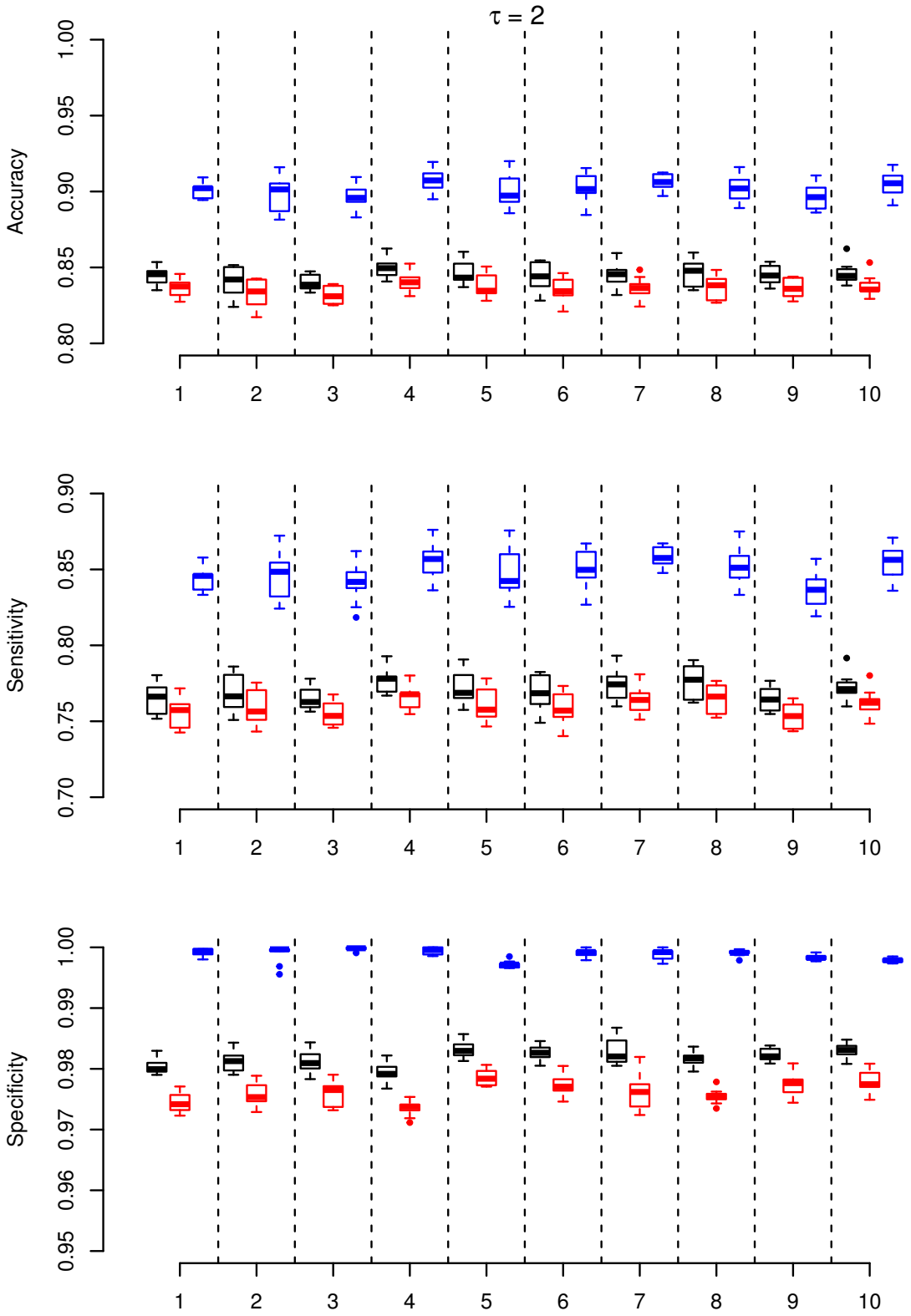

Figure C.21: (Simulation $\tau = 2$ with no effect $\boldsymbol{\beta}$) Distribution of the accuracy (top panel), sensitivity (middle panel) and specificity (bottom panel) of BayMAP 2.0 (black box plots for each different simulation for ζ), BayMAP 1.0 (red), BayMAP 2.0 prior odds (blue) considering ten draws of $\boldsymbol{\zeta}$ for $\tau = 2$ with $\beta_0 = 0.85$ and $\beta_{3'\text{UTR}} = \beta_{\text{CDS}} = \beta_{5'\text{UTR}} = 0$.

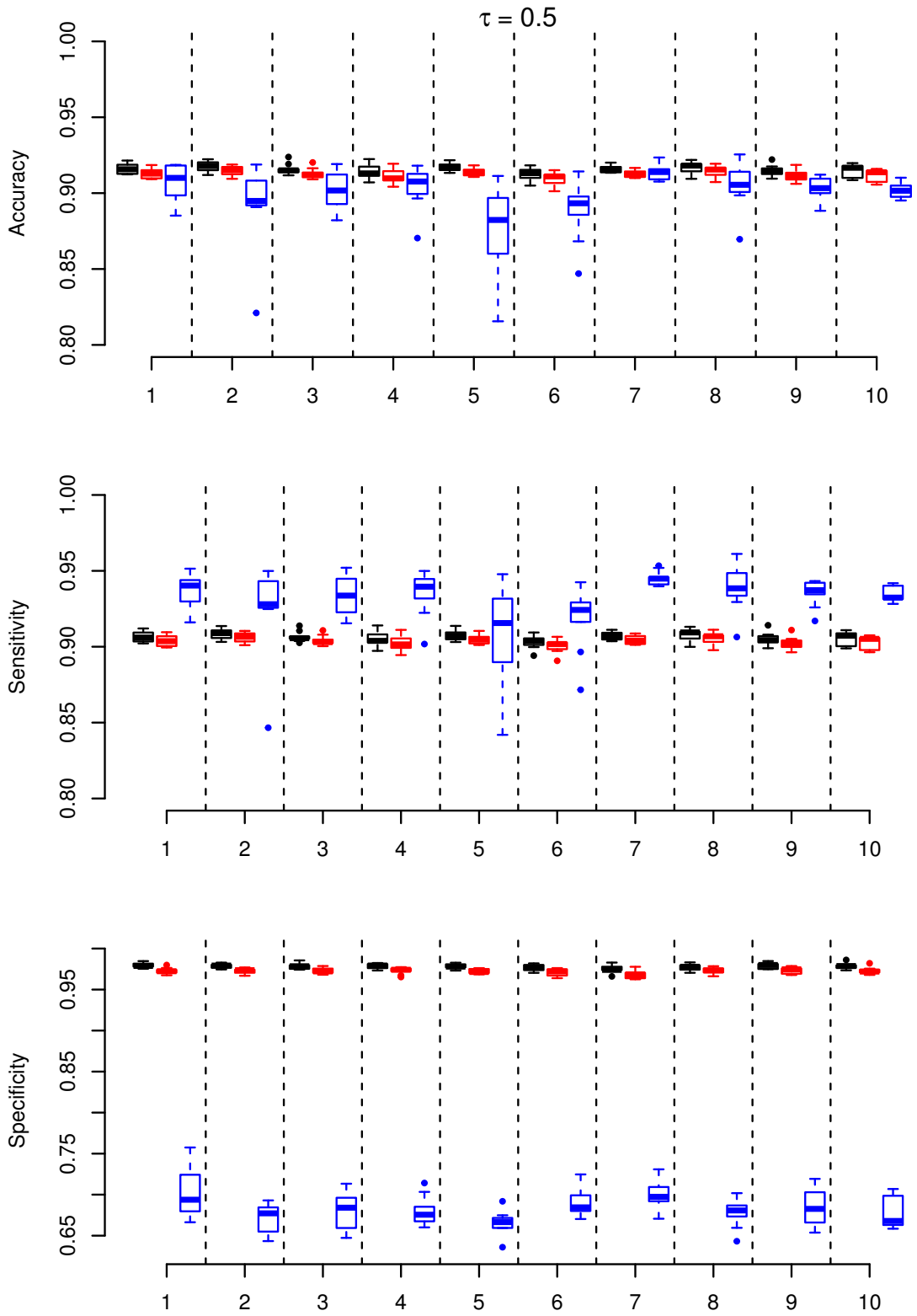

Figure C.22: (Simulation $\tau = 0.5$ with large $\boldsymbol{\beta}$ and $\mu_{\text{exp}} = 0.2$) Distribution of the accuracy (top panel), sensitivity (middle panel) and specificity (bottom panel) of BayMAP 2.0 (black box plots for each different simulation for $\boldsymbol{\zeta}$), BayMAP 1.0 (red), BayMAP 2.0 prior odds (blue) considering ten draws of $\boldsymbol{\zeta}$ for $\tau = 0.5$ with $\mu_{\text{exp}} = 0.2$ and $\beta_0 = 0.5$, $\beta_{3\text{'UTR}} = 1.85$, $\beta_{\text{CDS}} = 1.15$ and $\beta_{5\text{'UTR}} = 0.75$.

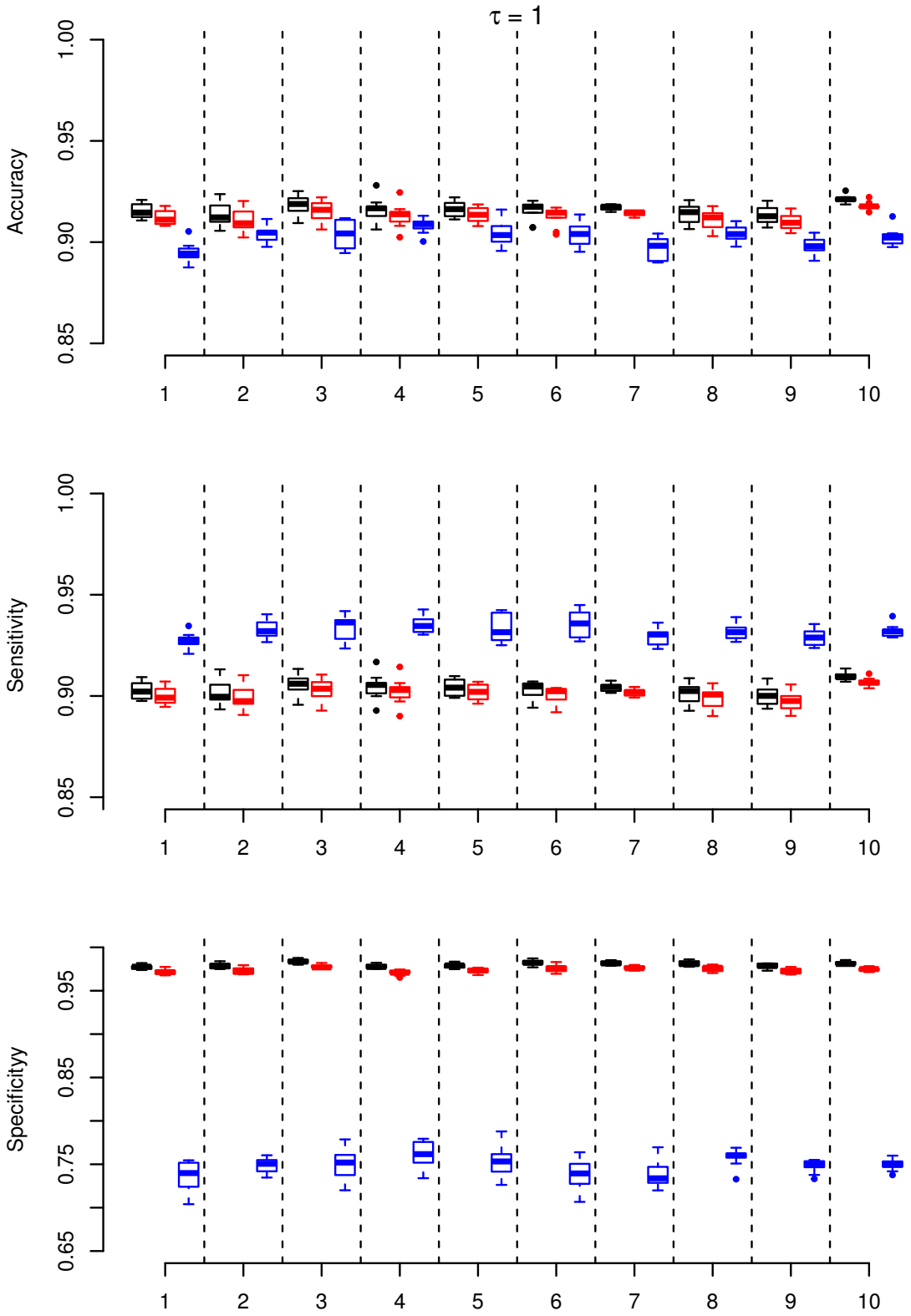

Figure C.23: (Simulation $\tau = 1$ with large $\boldsymbol{\beta}$ and $\mu_{\text{exp}} = 0.2$) Distribution of the accuracy (top panel), sensitivity (middle panel) and specificity (bottom panel) of BayMAP 2.0 (black box plots for each different simulation for $\boldsymbol{\zeta}$), BayMAP 1.0 (red), BayMAP 2.0 prior odds (blue) considering ten draws of $\boldsymbol{\zeta}$ for $\tau = 1$ with $\mu_{\text{exp}} = 0.2$ and $\beta_0 = 0.5$, $\beta_{3\text{'UTR}} = 1.85$, $\beta_{\text{CDS}} = 1.15$ and $\beta_{5\text{'UTR}} = 0.75$.

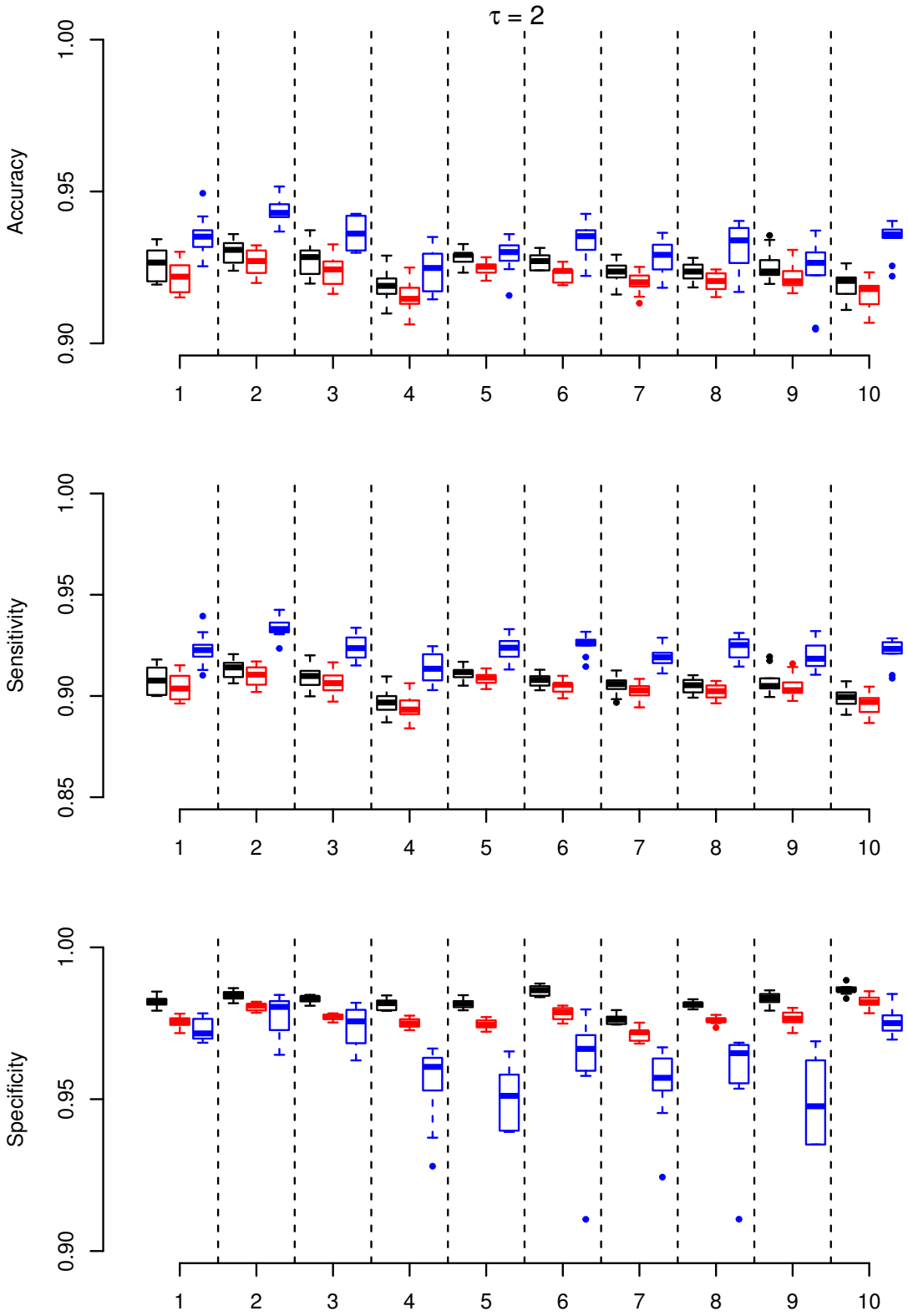

Figure C.24: (Simulation $\tau = 2$ with large $\boldsymbol{\beta}$ and $\mu_{\text{exp}} = 0.2$) Distribution of the accuracy (top panel), sensitivity (middle panel) and specificity (bottom panel) of BayMAP 2.0 (black box plots for each different simulation for $\boldsymbol{\zeta}$), BayMAP 1.0 (red), BayMAP 2.0 prior odds (blue) considering ten draws of $\boldsymbol{\zeta}$ for $\tau = 2$ with $\mu_{\text{exp}} = 0.2$ and $\beta_0 = 0.5$, $\beta_{3\text{'UTR}} = 1.85$, $\beta_{\text{CDS}} = 1.15$ and $\beta_{5\text{'UTR}} = 0.75$.

C.2.3 Comparison of BayMAP 2.0 to other methods

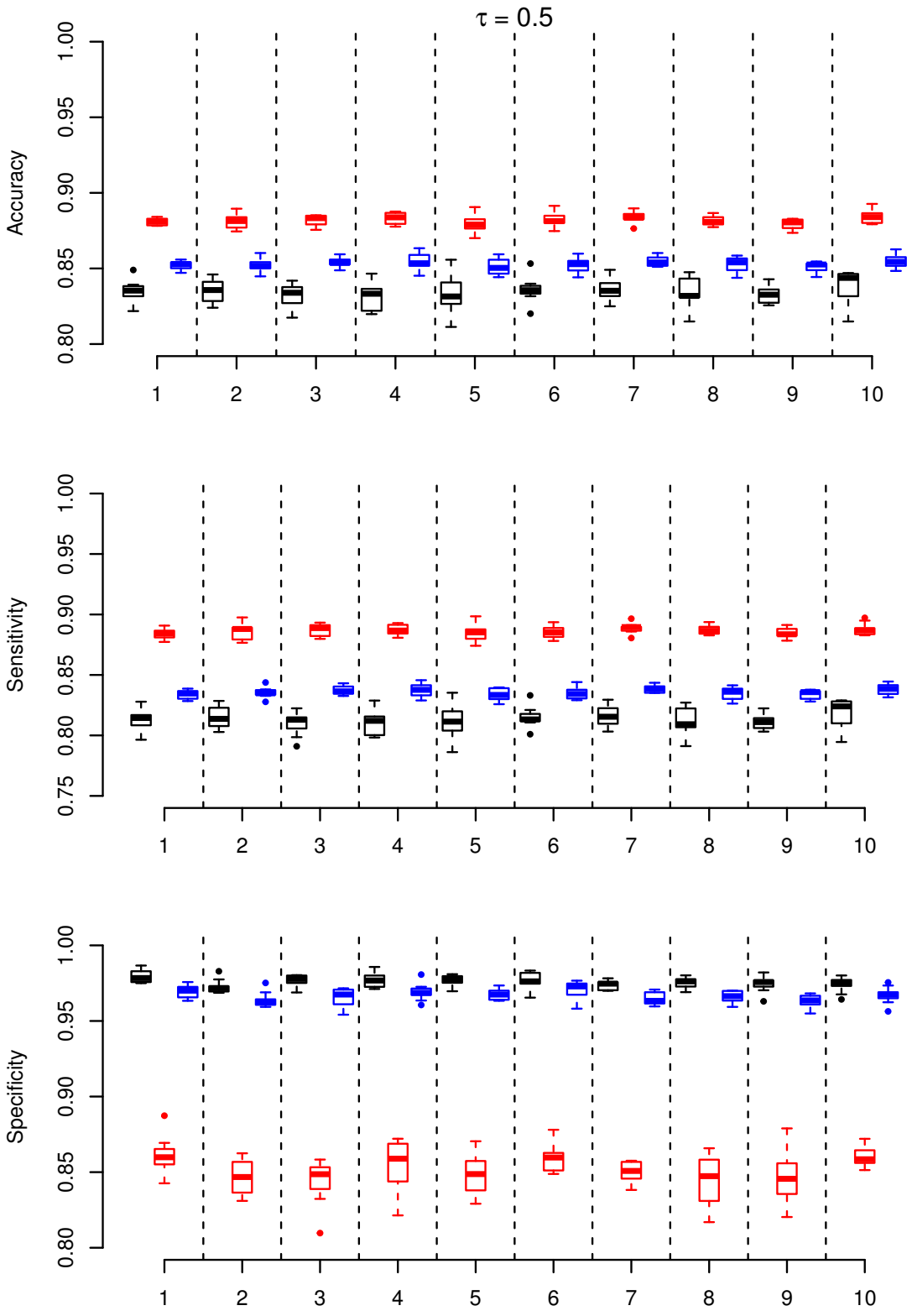

Figure C.25: (Simulation $\tau = 0.5$ with large $\boldsymbol{\beta}$) Distribution of the accuracy (top panel), sensitivity (middle panel) and specificity (bottom panel) of BayMAP 2.0 (black box plots for each different simulation for ζ), wavClusteR (red), and BMix (blue) considering ten draws of $\boldsymbol{\zeta}$ for $\tau = 0.5$ with $\beta_0 = 0.5$, $\beta_{3'\text{UTR}} = 1.85$, $\beta_{\text{CDS}} = 1.15$ and $\beta_{5'\text{UTR}} = 0.75$.

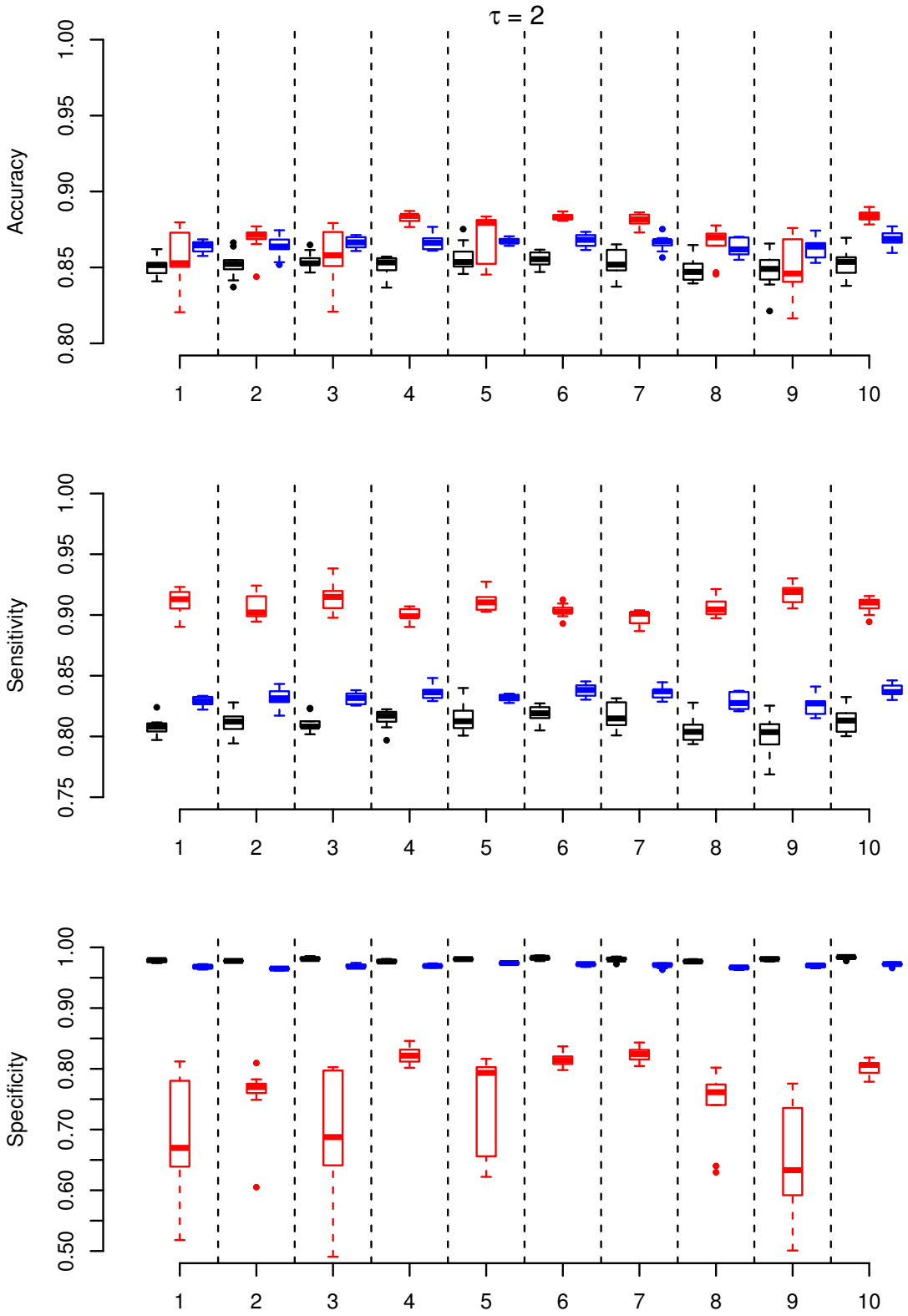

Figure C.26: (Simulation $\tau = 2$ with large $\boldsymbol{\beta}$) Distribution of the accuracy (top panel), sensitivity (middle panel) and specificity (bottom panel) of BayMAP 2.0 (black box plots for each different simulation for ζ), wavClusteR (red), and BMix (blue) considering ten draws of $\boldsymbol{\zeta}$ for $\tau = 2$ with $\beta_0 = 0.5$, $\beta_{3'\text{UTR}} = 1.85$, $\beta_{\text{CDS}} = 1.15$ and $\beta_{5'\text{UTR}} = 0.75$.

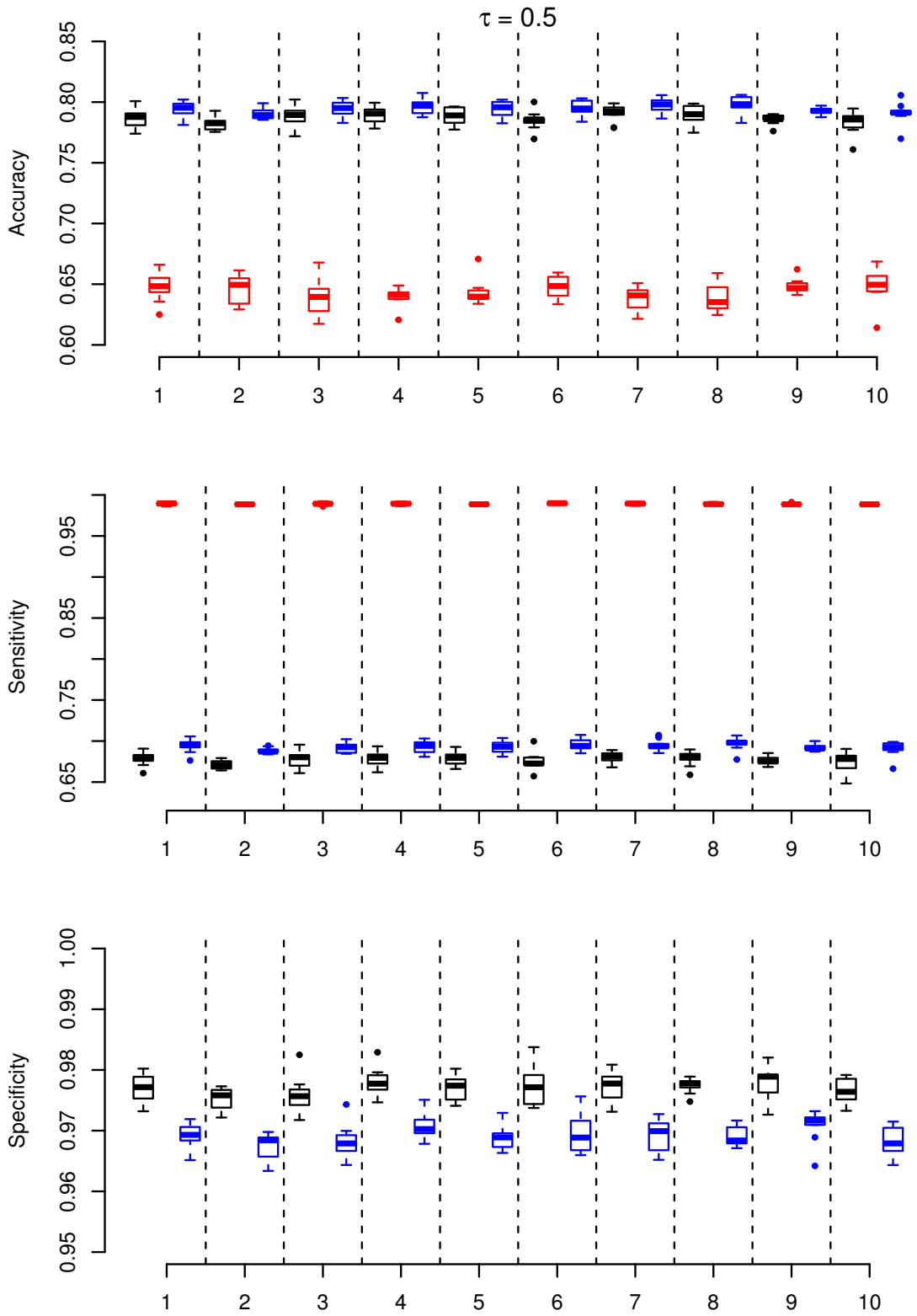

Figure C.27: (Simulation $\tau = 0.5$ with small $\boldsymbol{\beta}$) Distribution of the accuracy (top panel), sensitivity (middle panel) and specificity (bottom panel) of BayMAP 2.0 (black box plots for each different simulation for ζ), wavClusteR (red), and BMix (blue) considering ten draws of $\boldsymbol{\zeta}$ for $\tau = 0.5$ with $\beta_0 = -0.5$, $\beta_{3'\text{UTR}} = 1.5$, $\beta_{\text{CDS}} = 1.0$ and $\beta_{5'\text{UTR}} = 0.5$. For a better graphical representation, very small values of specificity for wavClusteR between 0.04 and 0.06 are not shown.

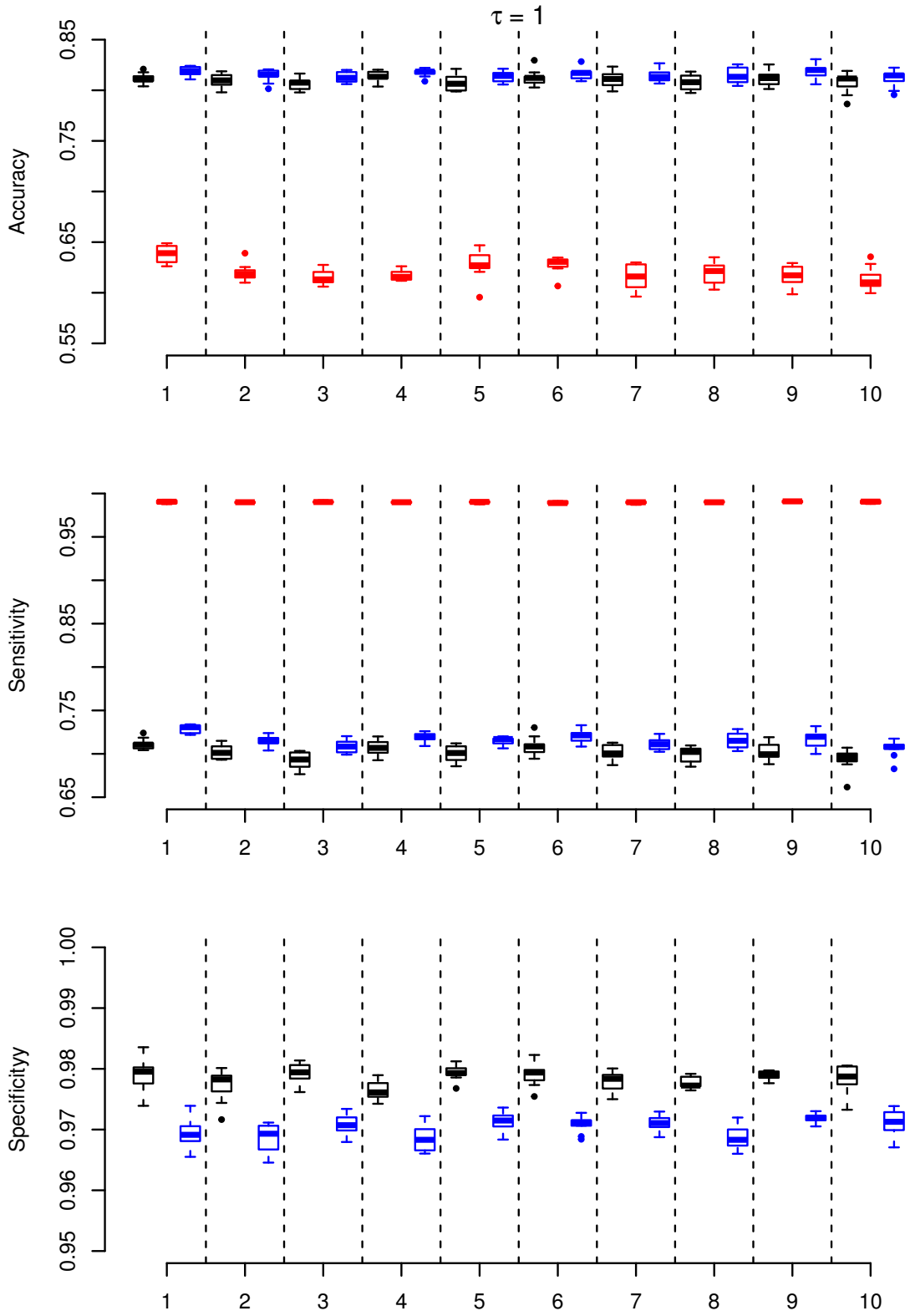

Figure C.28: (Simulation $\tau = 1$ with small $\boldsymbol{\beta}$) Distribution of the accuracy (top panel), sensitivity (middle panel) and specificity (bottom panel) of BayMAP 2.0 (black box plots for each different simulation for ζ), wavClusteR (red), and BMix (blue) considering ten draws of $\boldsymbol{\zeta}$ for $\tau = 1$ with $\beta_0 = -0.5$, $\beta_{3'\text{UTR}} = 1.5$, $\beta_{\text{CDS}} = 1.0$ and $\beta_{5'\text{UTR}} = 0.5$. For a better graphical representation, very small values of specificity for wavClusteR between 0.04 and 0.06 are not shown.

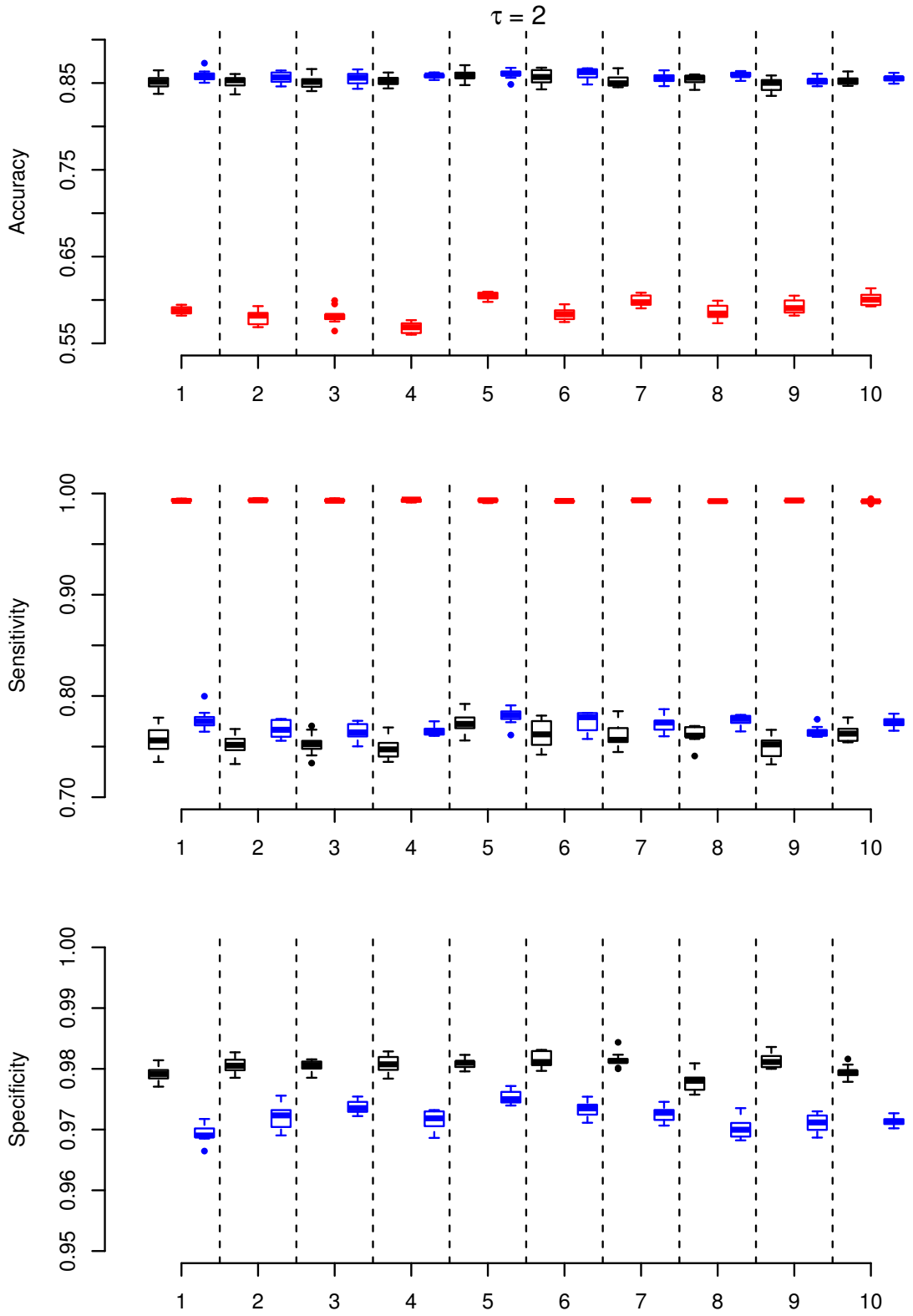

Figure C.29: (Simulation $\tau = 2$ with small $\boldsymbol{\beta}$) Distribution of the accuracy (top panel), sensitivity (middle panel) and specificity (bottom panel) of BayMAP 2.0 (black box plots for each different simulation for ζ), wavClusteR (red), and BMix (blue) considering ten draws of $\boldsymbol{\zeta}$ for $\tau = 2$ with $\beta_0 = -0.5$, $\beta_{3'\text{UTR}} = 1.5$, $\beta_{\text{CDS}} = 1.0$ and $\beta_{5'\text{UTR}} = 0.5$. For a better graphical representation, very small values of specificity for wavClusteR between 0.04 and 0.06 are not shown.

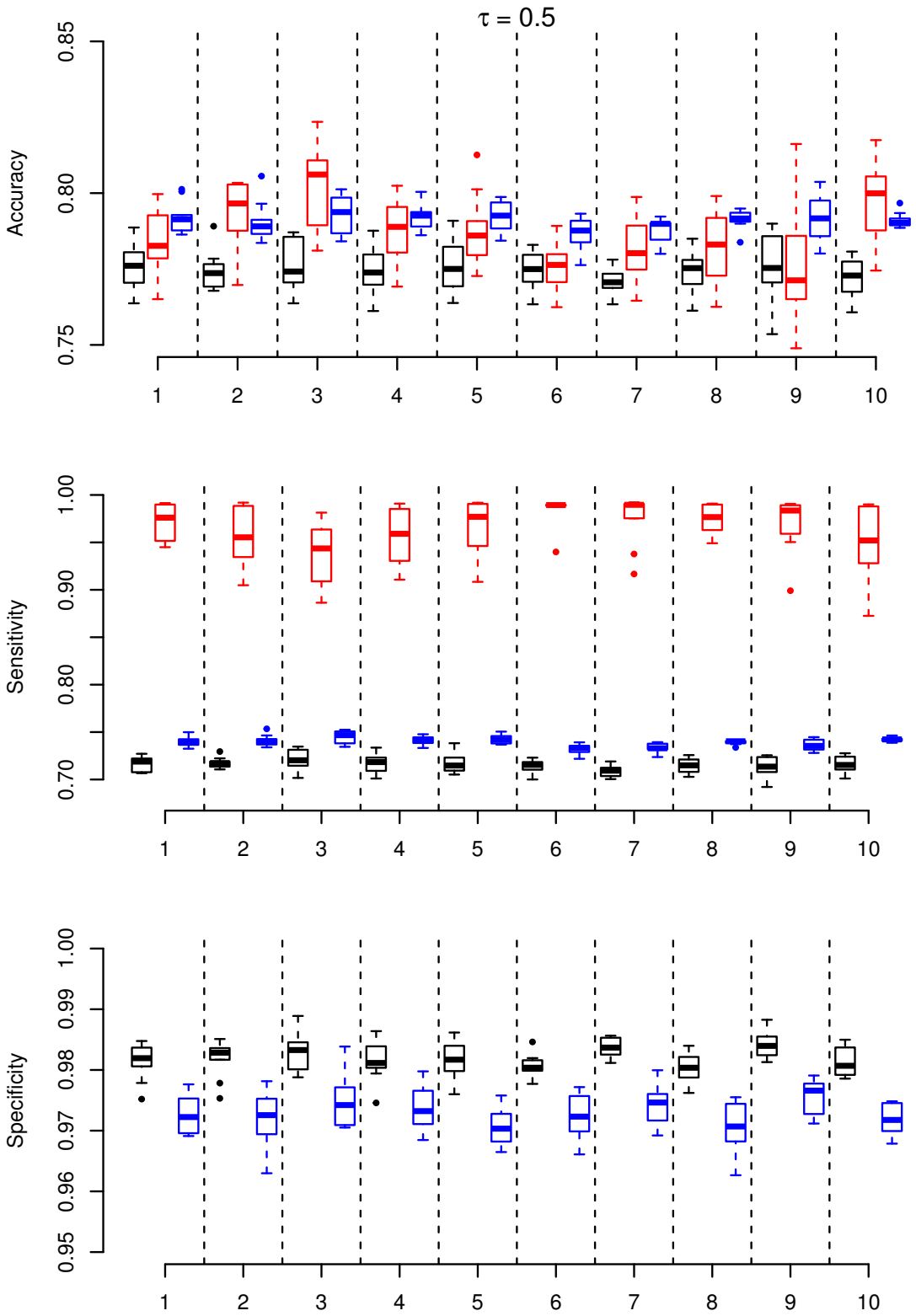

Figure C.30: (Simulation $\tau = 0.5$ with no effect $\boldsymbol{\beta}$) Distribution of the accuracy (top panel), sensitivity (middle panel) and specificity (bottom panel) of BayMAP 2.0 (black box plots for each different simulation for ζ), wavClusteR (red), and BMix (blue) considering ten draws of $\boldsymbol{\zeta}$ for $\tau = 0.5$ with $\beta_0 = 0.85$ and $\beta_{3\text{'UTR}} = \beta_{\text{CDS}} = \beta_{5\text{'UTR}} = 0$. For a better graphical representation, very small values of specificity for wavClusteR between 0.04 and 0.61 are not shown.

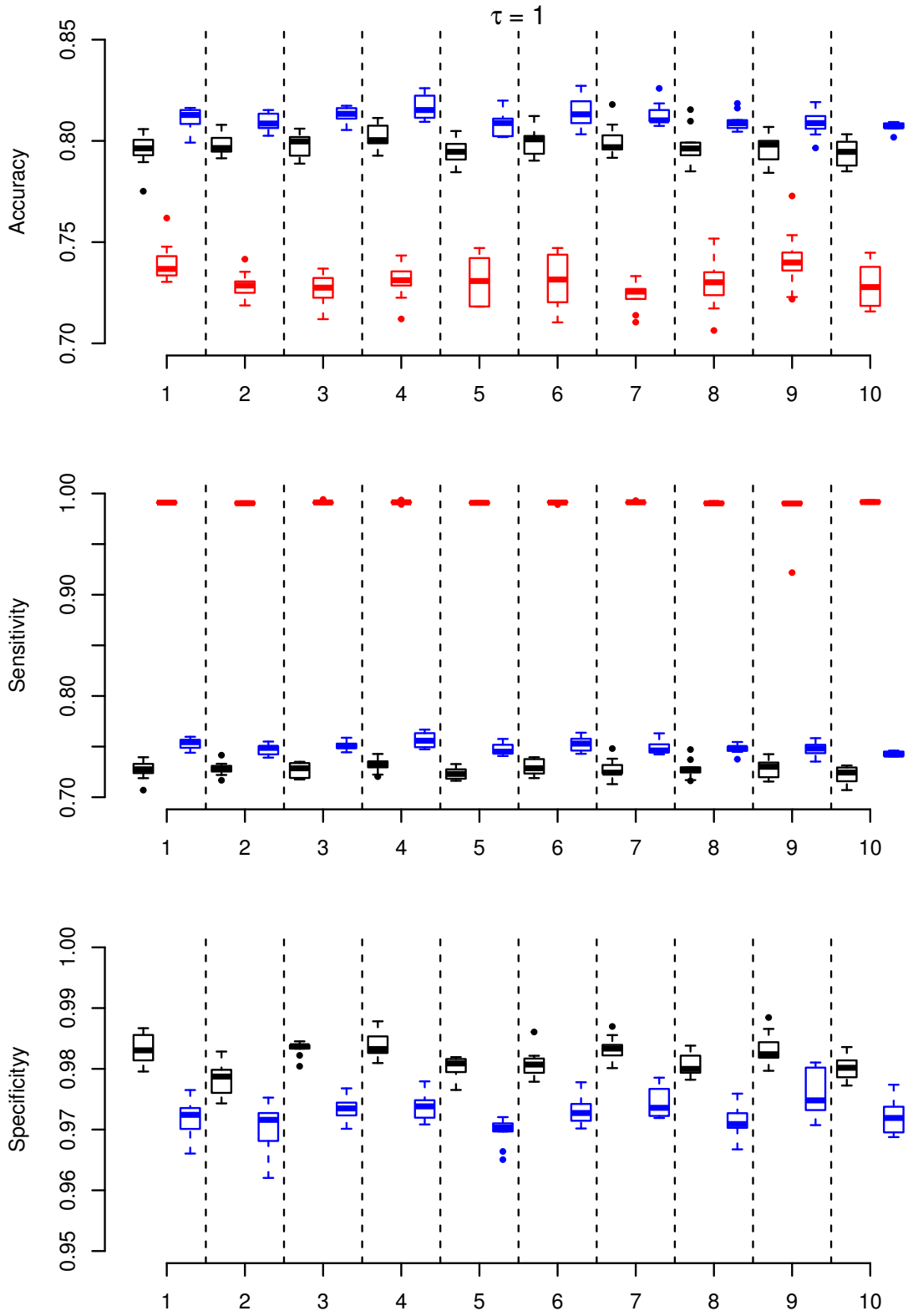

Figure C.31: (Simulation $\tau = 1$ with no effect $\boldsymbol{\beta}$) Distribution of the accuracy (top panel), sensitivity (middle panel) and specificity (bottom panel) of BayMAP 2.0 (black box plots for each different simulation for ζ), wavClusteR (red), and BMix (blue) considering ten draws of $\boldsymbol{\zeta}$ for $\tau = 1$ with $\beta_0 = 0.85$ and $\beta_{3'\text{UTR}} = \beta_{\text{CDS}} = \beta_{5'\text{UTR}} = 0$. For a better graphical representation, very small values of specificity for wavClusteR between 0.04 and 0.34 are not shown.

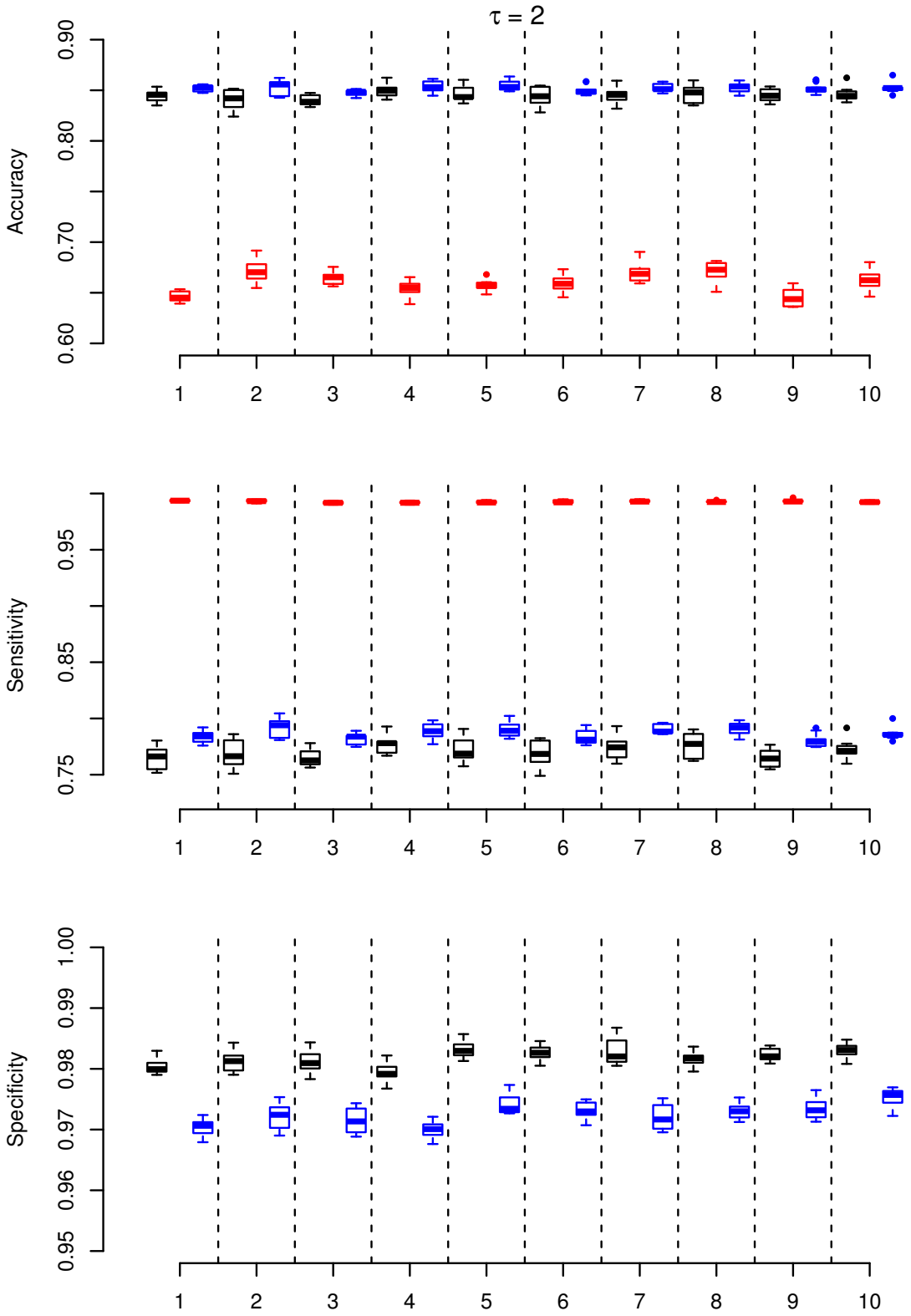

Figure C.32: (Simulation $\tau = 2$ with no effect $\boldsymbol{\beta}$) Distribution of the accuracy (top panel), sensitivity (middle panel) and specificity (bottom panel) of BayMAP 2.0 (black box plots for each different simulation for ζ), wavClusteR (red), and BMix (blue) considering ten draws of $\boldsymbol{\zeta}$ for $\tau = 2$ with $\beta_0 = 0.85$ and $\beta_{3'\text{UTR}} = \beta_{\text{CDS}} = \beta_{5'\text{UTR}} = 0$. For a better graphical representation, very small values of specificity for wavClusteR between 0.04 and 0.05 are not shown.

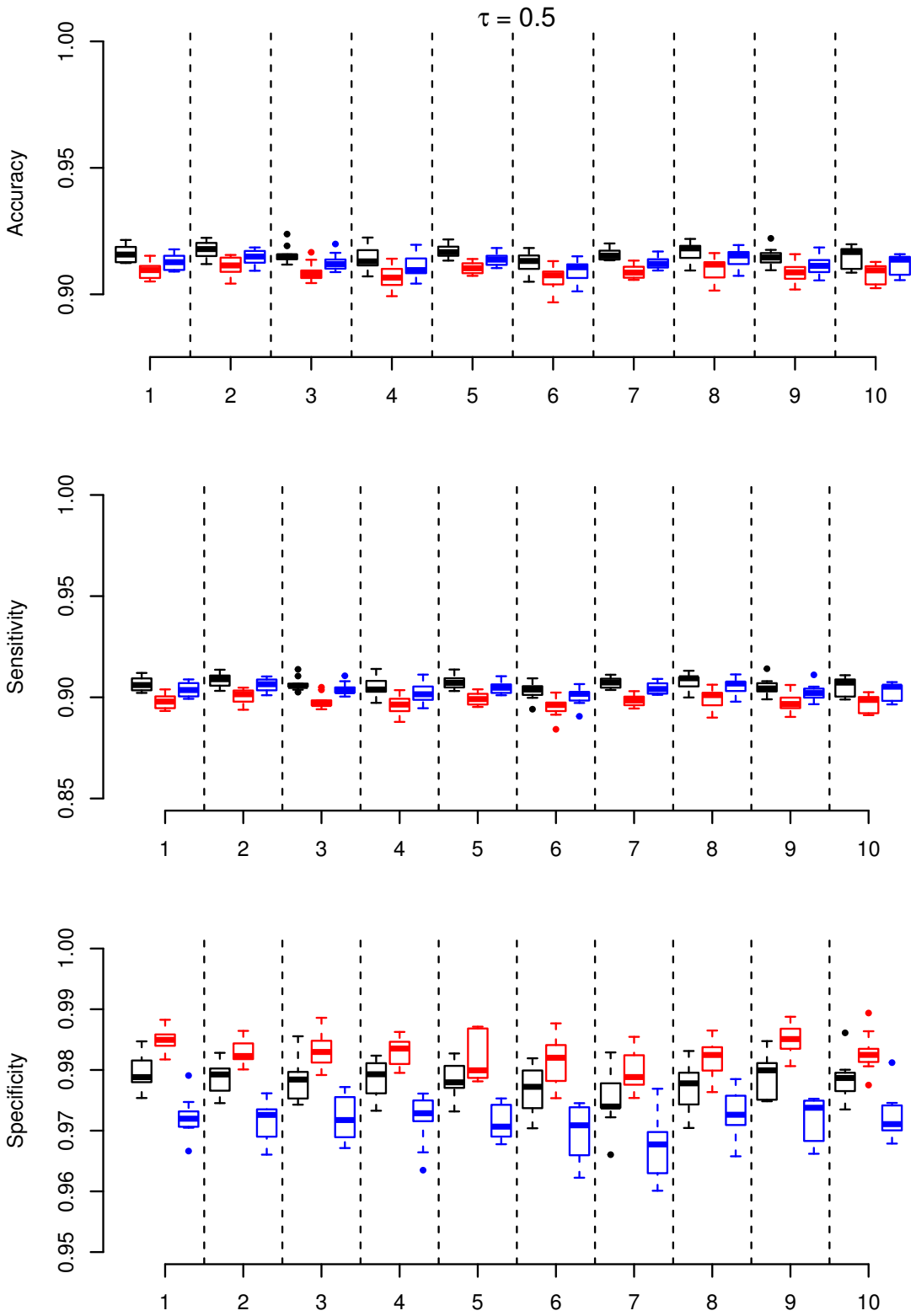

Figure C.33: (Simulation $\tau = 0.5$ with large $\boldsymbol{\beta}$ and $\mu_{\text{exp}} = 0.2$) Distribution of the accuracy (top panel), sensitivity (middle panel) and specificity (bottom panel) of BayMAP 2.0 (black box plots for each different simulation for ζ), wavClusteR (red), and BMix (blue) considering ten draws of $\boldsymbol{\zeta}$ for $\tau = 0.5$ with $\mu_{\text{exp}} = 0.2$ and $\beta_0 = 0.5$, $\beta_{3\text{'UTR}} = 1.85$, $\beta_{\text{CDS}} = 1.15$ and $\beta_{5\text{'UTR}} = 0.75$.

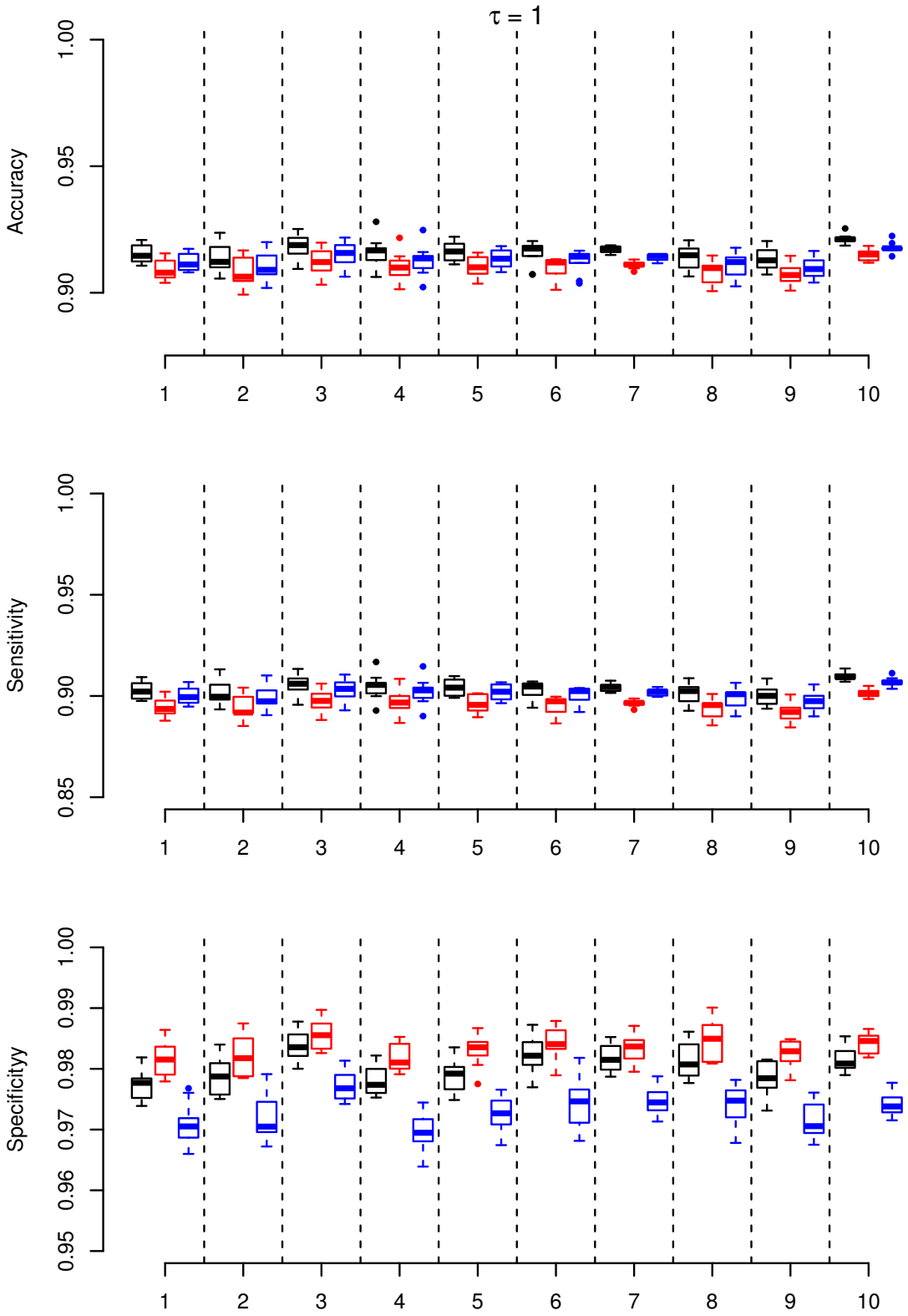

Figure C.34: (Simulation $\tau = 1$ with large $\boldsymbol{\beta}$ and $\mu_{\text{exp}} = 0.2$) Distribution of the accuracy (top panel), sensitivity (middle panel) and specificity (bottom panel) of BayMAP 2.0 (black box plots for each different simulation for ζ), wavClusteR (red), and BMix (blue) considering ten draws of $\boldsymbol{\zeta}$ for $\tau = 1$ with $\mu_{\text{exp}} = 0.2$ and $\beta_0 = 0.5$, $\beta_{3\text{'UTR}} = 1.85$, $\beta_{\text{CDS}} = 1.15$ and $\beta_{5\text{'UTR}} = 0.75$.

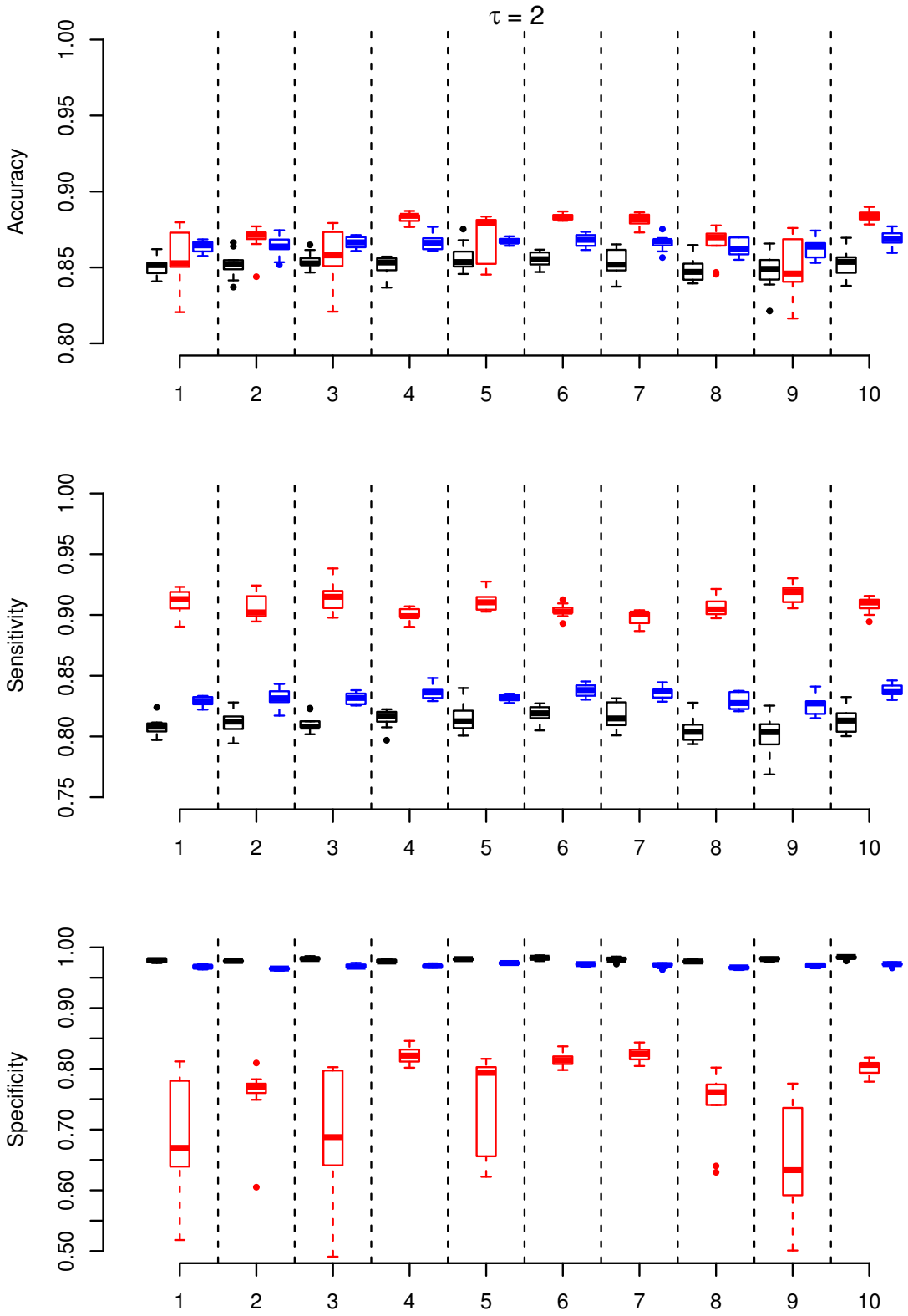

Figure C.35: (Simulation $\tau = 2$ with large $\boldsymbol{\beta}$ and $\mu_{\text{exp}} = 0.2$) Distribution of the accuracy (top panel), sensitivity (middle panel) and specificity (bottom panel) of BayMAP 2.0 (black box plots for each different simulation for ζ), wavClusteR (red), and BMix (blue) considering ten draws of $\boldsymbol{\zeta}$ for $\tau = 2$ with $\mu_{\text{exp}} = 0.2$ and $\beta_0 = 0.5$, $\beta_{\text{3'UTR}} = 1.85$, $\beta_{\text{CDS}} = 1.15$ and $\beta_{\text{5'UTR}} = 0.75$.

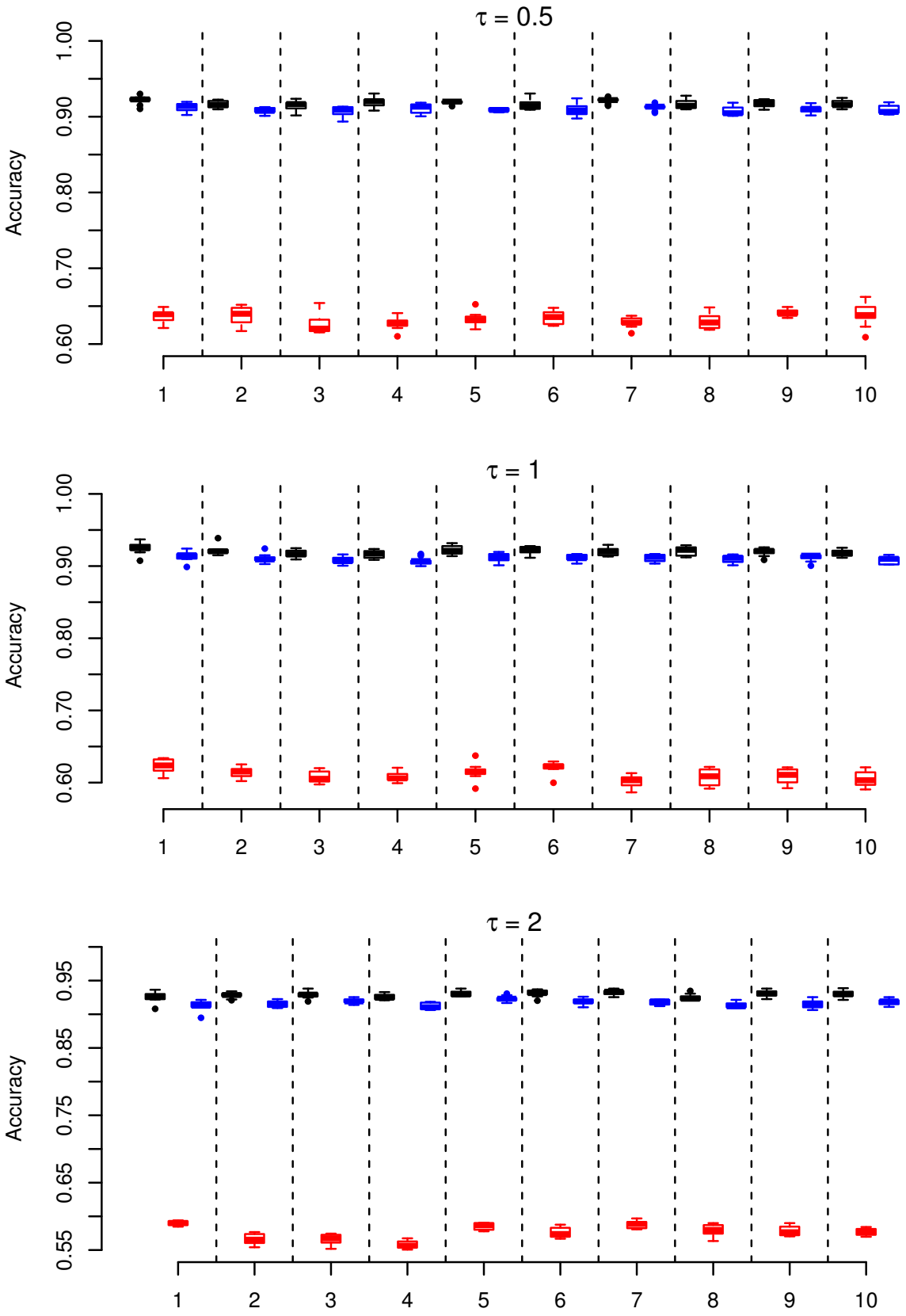

Figure C.36: (Simulation binding site based with small $\boldsymbol{\beta}$) Distribution of the accuracy binding site based of BayMAP 2.0 (black box plots for each different simulation for ζ), wavClusteR (red), and BMix (blue) considering ten draws of $\boldsymbol{\zeta}$ for each value of τ with $\beta_0 = -0.5$, $\beta_{3'\text{UTR}} = 1.5$, $\beta_{\text{CDS}} = 1.0$ and $\beta_{5'\text{UTR}} = 0.5$.

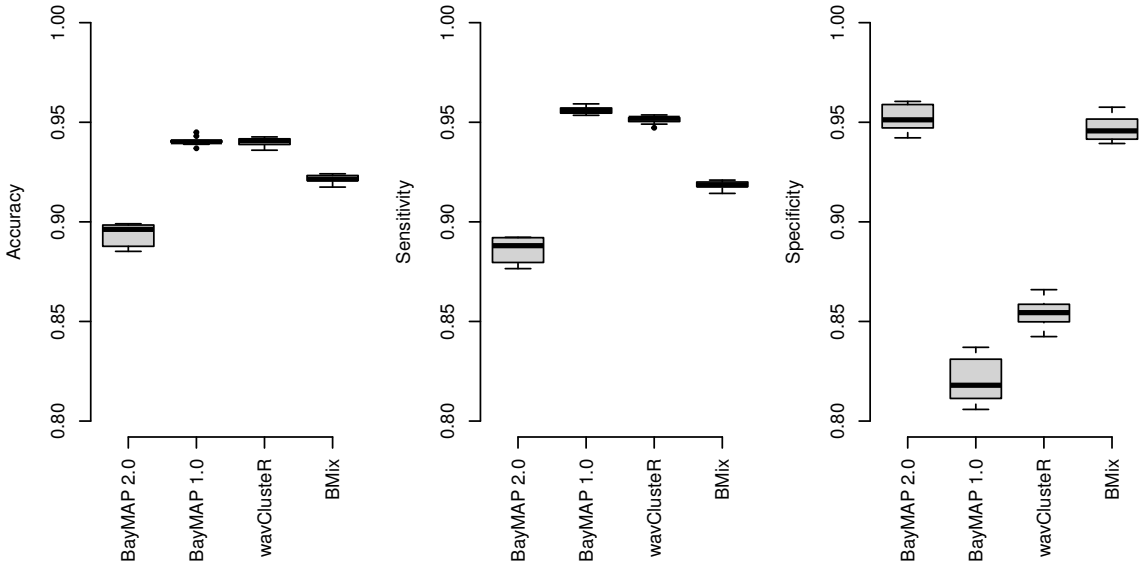

Figure C.37: (Simulation with beta distributed $\boldsymbol{\beta}$) The accuracy (left panel), sensitivity (middle panel) and specificity (right panel) for BayMAP 2.0, BayMAP 1.0, wavClusteR, and BMix on ten simulated data sets for which μ_{exp} was drawn from a Beta$(2,10)$ distribution for each position separately.

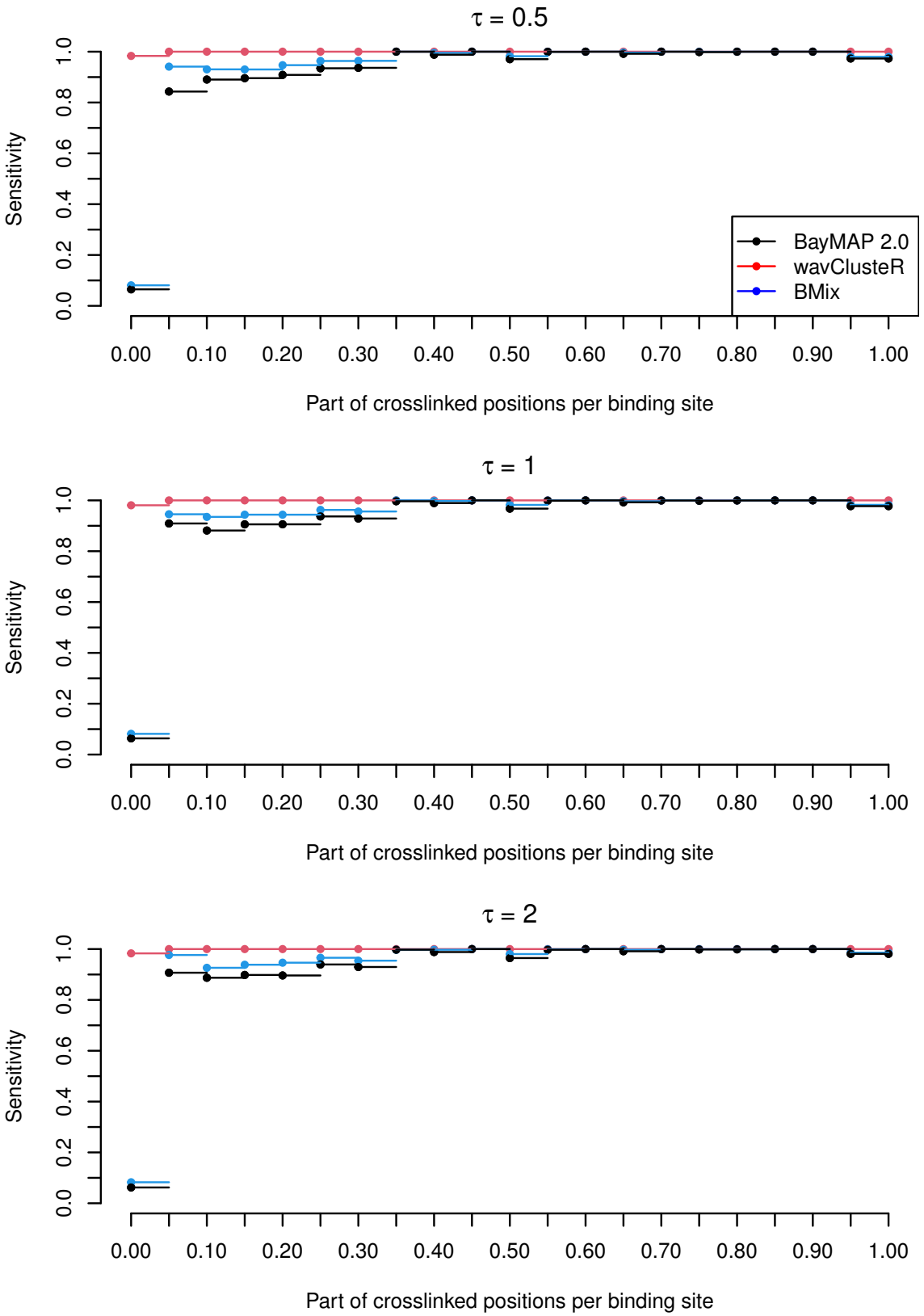

Figure C.38: Distribution of the sensitivity of BayMAP 2.0 (black), wavClusteR (red), and BMix (blue) depending on the part of crosslinked positions per binding site for each value of τ with $\beta_0 = -0.5$, $\beta_{3'\text{UTR}} = 1.5$, $\beta_{\text{CDS}} = 1.0$ and $\beta_{5'\text{UTR}} = 0.5$. The x-axis is divided into intervals each of the length 0.05 (displayed by lines), where the left boundary of each interval is included (displayed by the point).

C.3 BayMAP combining several PAR-CLIP data sets

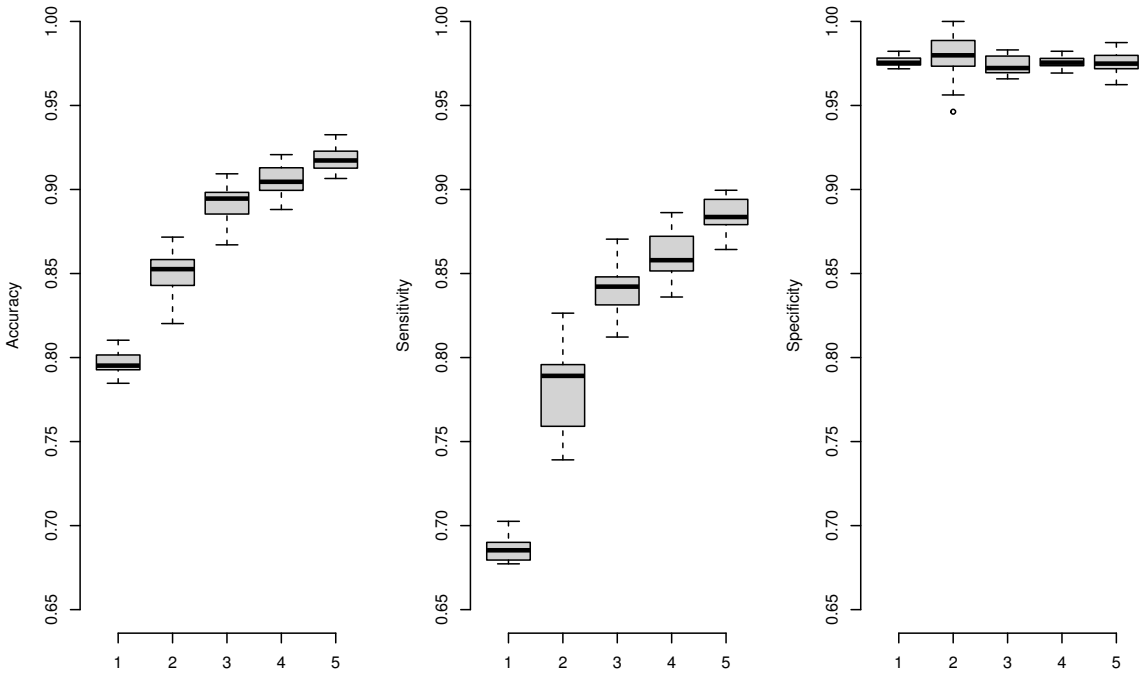

Figure C.39: The accuracy (left panel), sensitivity (middle panel) and specificity (right panel) for the combined post odds of BayMAP 1.0 with one to five combined results, where one combined result means that only the results on one data set without combination are regarded.

Appendix D

Additional application results

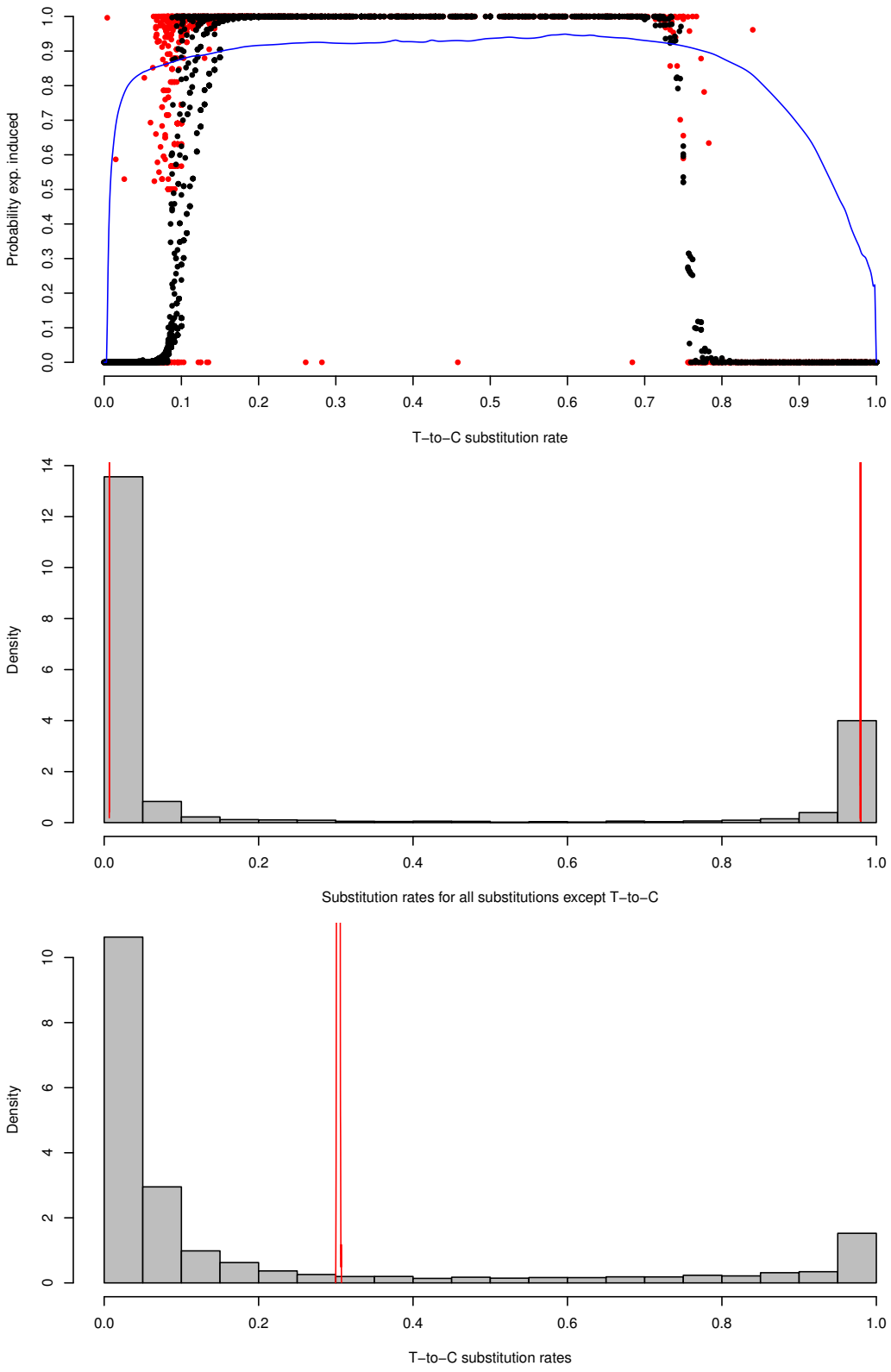

Figure D.1: (Kishore B) Top panel: T-to-C substitution rate in comparison to the probability that the position is experimentally induced for BayMAP (black), BMix (red) and wavClusteR (blue) at 5,000 randomly chosen T-to-C substitution positions in the Kishore B data set. Middle/bottom panel: Histograms for the substitution rates for all substitutions except T-to-C with BayMAP 1.0 estimations for μ_{mm} and μ_{SNP} indicated by red lines (middle) and only for T-to-C substitutions with BayMAP 1.0 estimation for μ_{exp}.

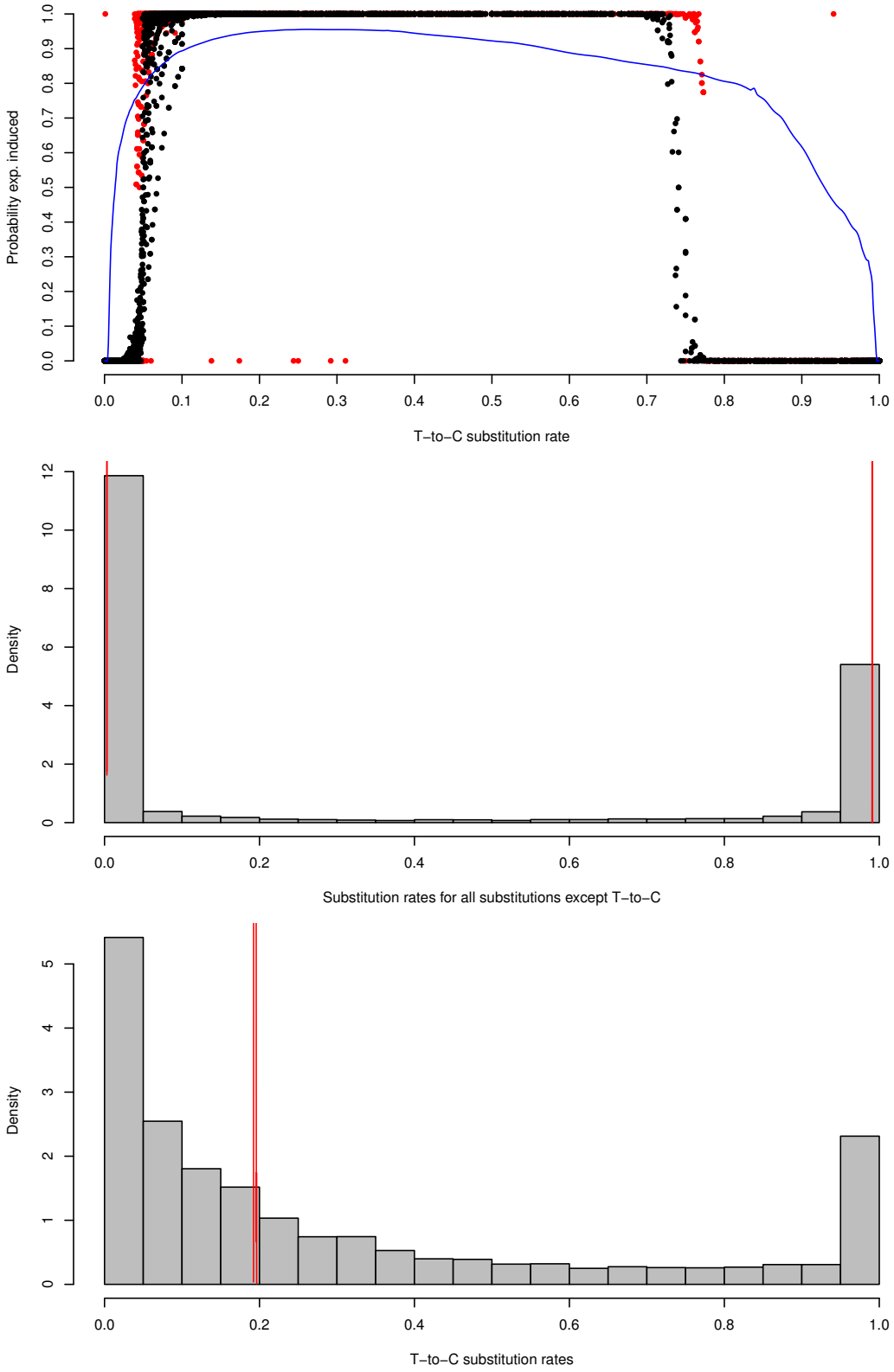

Figure D.2: (Memczak) Top panel: T-to-C substitution rate in comparison to the probability that the position is experimentally induced for BayMAP (black), BMix (red) and wavClusteR (blue) at 5,000 randomly chosen T-to-C substitution positions in the Memczak data set. Middle/bottom panel: Histograms for the substitution rates for all substitutions except T-to-C with BayMAP 1.0 estimations for μ_{mm} and μ_{SNP} indicated by red lines (middle) and only for T-to-C substitutions with BayMAP 1.0 estimation for μ_{exp}.

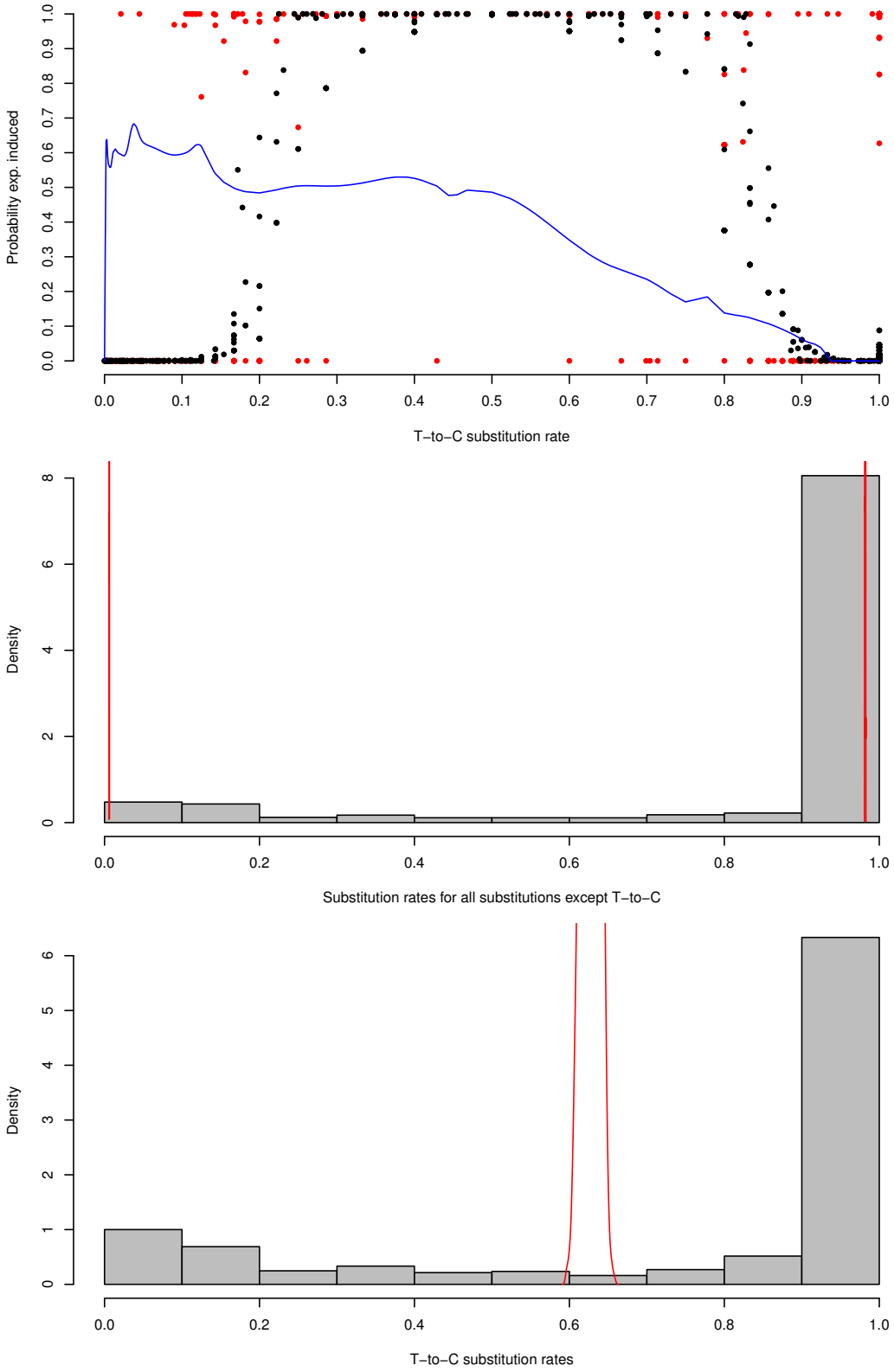

Figure D.3: (Gottwein A) Top panel: T-to-C substitution rate in comparison to the probability that the position is experimentally induced for BayMAP (black), BMix (red) and wavClusteR (blue) at 5,000 randomly chosen T-to-C substitution positions in the Gottwein A data set. Middle/bottom panel: Histograms for the substitution rates for all substitutions except T-to-C with BayMAP 1.0 estimations for μ_{mm} and μ_{SNP} indicated by red lines (middle) and only for T-to-C substitutions with BayMAP 1.0 estimation for μ_{exp}.

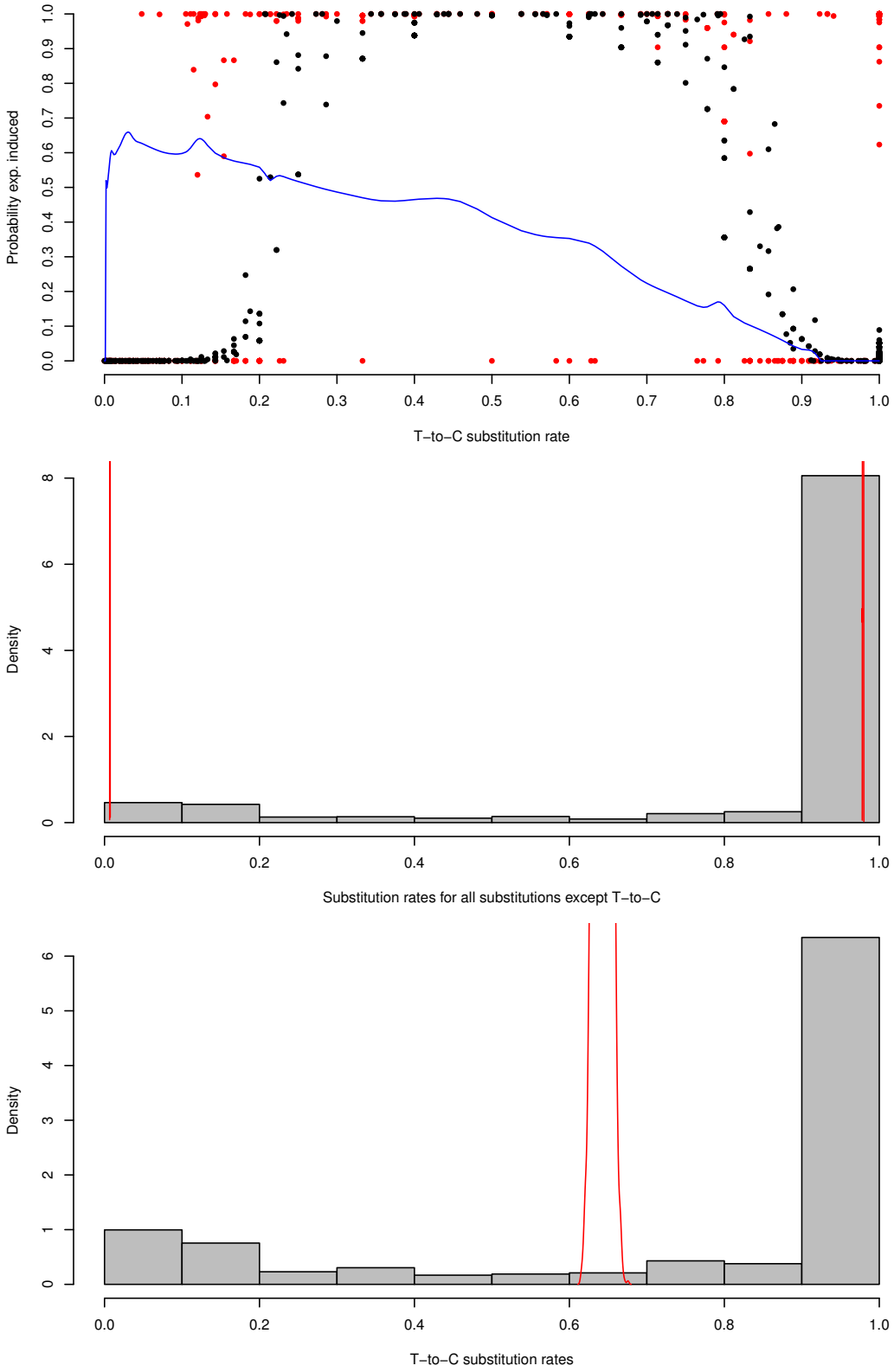

Figure D.4: (Gottwein B) Top panel: T-to-C substitution rate in comparison to the probability that the position is experimentally induced for BayMAP (black), BMix (red) and wavClusteR (blue) at 5,000 randomly chosen T-to-C substitution positions in the Gottwein B data set. Middle/bottom panel: Histograms for the substitution rates for all substitutions except T-to-C with BayMAP 1.0 estimations for μ_{mm} and μ_{SNP} indicated by red lines (middle) and only for T-to-C substitutions with BayMAP 1.0 estimation for μ_{exp}.

Appendix E

Software

E.1 R Documentation BayMAP 1.0

Description

The baymap function runs **WinBUGS** on PAR-CLIP data.

Usage

```
baymap(data, count = "count", coverage = "coverage",
       mutation = "mutation", mutation.type = "TC",
       covariates = NULL, dist = c("truncated", "binomial"),
       dep = TRUE, n.chains = 1, n.iter = 1500, n.thin = 1,
       n.burnin = 500 * n.thin, inits.Z = NULL, inits.q = NULL,
       inits.u = NULL, inits.b = NULL, ...)

## S3 method for class 'baymap'
predict(object, data, count = "count", coverage = "coverage",
       covariates = NULL, dist = c("truncated", "binomial"),
       print.i = NULL, add.thin = NULL, ...)
```

Arguments

`data`	a data frame with at least the count for mutations per genomic position, the number of reads/coverage and the mutation type (e.g., T-to-C).
`count`	the name of the variable that counts the number of mutations.
`coverage`	the name of the variable that contains the number of reads.
`mutation`	the name of the variable that contains the different types of mutations.
`mutation.type`	the name of the mutation type that is induced by the PAR-CLIP method.
`covariates`	a vector containing the names for the covariates for the regression model, e.g., c("tpUTR", "cds", "fpUTR"). Intercept is automatically added as first variable.
`dist`	the distribution for the number of mutations. Possible entries are "truncated" (default) and "binomial.
`dep`	a logical value for defining if dependencies between mismatches and SNPs are considered (default) or not.
`n.chains`	number of Markov chains (default: 1)
`n.iter`	number of total iterations per chain (including burn in; default 4500).
`n.thin`	thinning rate. Must be a positive integer. Set n.thin > 1 to save memory and computation time if n.iter is large.
`n.burnin`	length of burn in, i.e. number of iterations to discard at the beginning.
`inits.Z`	a vector containing as inits an allocation for each position, where 1 stands for an experimental induced substitution position, 2 for a SNP and 3 for a mismatch.

inits.q	a numerical value between 0 and 1 containing as init the conditional probability for a mismatch position given the subsitions are not experimentally induced.
inits.u	a numerical vector containing as inits three values between 0 and 1 for the substitution probability due to the PAR-CLIP method, due to SNPs and due to mismatches. If dep = TRUE the second entry of this vector should be set to NA.
inits.b	a numerical vector containing as inits the parameter vector for the covariates. Only necessary if the vector covariates is not NULL.
...	Additional arguments to be passed to the baymap function (see bugs).
object	a baymap object obtained by baymap().
print.i	a positive integer indicating if every ith iteration step should be printed.
add.thin	a positive integer containing an additional thinning rate that should be applied on the baymap() outcome.

Value

The returned object is a result of the bugs function of the **R2WinBUGS** package.

Author(s)

Eva-Maria Huessler, eva-maria.huessler@uni-duesseldorf.de

References

Huessler, E. M., Schäfer, M., Schwender, H., Landgraf, P. (2019). BayMAP: a Bayesian hierarchical model for the analysis of PAR-CLIP data. Bioinformatics, 35(12), 1992-2000.

See Also

bugs in the **R2WinBUGS** package

Examples

```
## Not run:
data(data_test)
res <- baymap(data = data_test)
data_new <- predict(res, data_test)

## End(Not run)
```

E.2 R Documentation BayMAP 2.0

---

baymap *A Bayesian hierarchical model for the analysis of PAR-CLIP data*

Description

The baymap function runs BayMAP on PAR-CLIP data to detect PAR-CLIP induced T-to-C substitution positions on binding sites.

Usage

```
baymap(data, count = "count", coverage = "coverage",
    mutation = "mutation", mutation.type = "TC",
    covariates = NULL, dist = c("truncated", "binomial"),
    dep = TRUE, n.chains = 1, n.iter = 1500, thin = 1,
    sd.mu = c(1e-04, 1e-04, 1e-04), inits.z = NULL,
    inits.q = NULL, inits.mu = NULL, inits.beta = NULL,
    ran = FALSE, cluster = "cluster", inits.tau = NULL,
    print.i = NULL, save_log = FALSE,  save_file = "./results_tmp.RData")
```

Arguments

`data`	a data frame with at least the count for mutations per genomic position, the number of reads/coverage and the mutation type (e.g., T-to-C).
`count`	the name of the variable that counts the number of mutations.
`coverage`	the name of the variable that contains the number of reads.
`mutation`	the name of the variable that contains the different types of mutations.
`mutation.type`	the name of the mutation type that is induced by the PAR-CLIP method.
`covariates`	a vector containing the names for the covariates for the regression model, e.g., c("tpUTR", "cds", "fpUTR"). Intercept is automatically added as first variable.
`dist`	the distribution for the number of mutations. Possible entries are "truncated" (default) and "binomial.
`dep`	a logical value for defining if dependencies between mismatches and SNPs are considered (default) or not.
`n.chains`	number of Markov chains (default: 1)
`n.iter`	number of total iterations per chain (including burn in; default 4500).
`n.thin`	thinning rate. Must be a positive integer. Set n.thin > 1 to save memory and computation time if n.iter is large.
`sd.mu`	a vector containing three values of standard deviations for the sampling of `mu` with a normal jumping distribution.
`inits.Z`	a vector containing as inits an allocation for each position, where 1 stands for an experimental induced substitution position, 2 for a SNP and 3 for a mismatch.

`inits.q`	a numerical value between 0 and 1 containing as init the conditional probability for a mismatch position given the subsitions are not experimentally induced.
`inits.mu`	a numerical vector containing as inits three values between 0 and 1 for the substitution probability due to the PAR-CLIP method, due to SNPs and due to mismatches.
`inits.beta`	a numerical vector containing as inits the parameter vector for the covariates. Only necessary if the vector covariates is not NULL.
`ran`	a logical value indicating if neighborhood dependencies should be included via a random effect (default) or not.
`cluster`	the name of the varialbe indicating to which cluster a position belongs. Only necessary if `ran` is set to `TRUE`.
`inits.tau`	a numerical value containing as inits the standard deviation of the random effect if `ran` is set to `TRUE`.
`print.i`	a positive integer indicating if every ith iteration step should be printed.
`save_log`	a logical value indicating if temporary results should be saved or not (default).
`save_file`	file name where temporary results should be stored if `save_log` is set to `TRUE`.

Value

The returned object is a list with sampled MCMC chains for each parameter as entries and an entry with acceptance values for each sampled value for the parameter `mu`.

Author(s)

Eva-Maria Huessler, eva-maria.huessler@uni-duesseldorf.de

References

Huessler, E. M., Schäfer, M., Schwender, H., Landgraf, P. (2019). BayMAP: a Bayesian hierarchical model for the analysis of PAR-CLIP data. Bioinformatics, 35(12), 1992-2000.

See Also

`predict.baymap`

Examples

```
## Not run:
data(data_test)
res <- baymap(data = data_test,
inits.mu = c(0.05, 0.85, 0.2), n.iter = 4500)

## End(Not run)
```

`predict` *Prediction method for BayMAP results*

Description

Predictions if T-to-C substitution positions are PAR-CLIP induced substitutions. Results of several PAR-CLIP experiments can be combined.

Usage

```
predict.baymap(object, data, chr = "chr", pos = "pos",
          count = "count", coverage = "coverage",
          mutation = "mutation", mutation.type = "TC",
          covariates = NULL,
```

```
dist = c("truncated", "binomial"),
ran = FALSE, cluster.id = NULL,
print.i = 100, thin = NULL, burn.in = 0, ...)
```

Arguments

object	either a baymap object or a list of baymap objects if several PAR-CLIP experiments should be analyzed jointly. If a list with baymap objects is read in, the class "baymap" should be assigned to the list prior the analysis by class.
data	either a data frame with at least the count for mutations per genomic position, the number of reads/coverage and the mutation type (e.g., T-to-C) or a list of data frames.
chr	the name of the variable that contains the chromosome information.
pos	the name of the variable that contains the position information.
count	the name of the variable that counts the number of mutations.
coverage	the name of the variable that contains the number of reads.
mutation	the name of the variable that contains the different types of mutations.
mutation.type	the name of the mutation type that is induced by the PAR-CLIP method.
covariates	a vector containing the names for the covariates for the regression model, e.g., c("tpUTR", "cds", "fpUTR"). Intercept is automatically added as first variable.
dist	the distribution for the number of mutations. Possible entries are "truncated" (default) and "binomial.

`ran`	a logical value indicating if neighborhood dependencies should be included via a random effect (default) or not.
`cluster`	the name of the varialbe indicating to which cluster a position belongs. Only necessary if `ran` is set to `TRUE`.
`print.i`	a positive integer indicating if every ith iteration step should be printed.
`thin`	an additional thinning rate that should be applied on the `baymap` outcome. Must be a positive integer.
`burn.in`	length of burn in, i.e. number of iterations to discard at the beginning.
`...`	further arguments for `predict`.

Value

If a single PAR-CLIP data set is analyzed, the returned object is a data frame including the prior odds, the bayes factor and the posterior odds. If several PAR-CLIP data sets are analyzed, the returned object is a list combined predictions as well as separate predictions for each dara set. Posterior odds greater than one means that the probability of having a method-incuced substitution position given the data is larger than 0.5.

Note

Predictions of the separate analyzis are made for all entries of the included data set even for substitution types other than T-to-C.

Author(s)

Eva-Maria Huessler, eva-maria.huessler@uni-duesseldorf.de

References

Huessler, E. M., Schäfer, M., Schwender, H., Landgraf, P. (2019). BayMAP: a Bayesian

hierarchical model for the analysis of PAR-CLIP data. Bioinformatics, 35(12), 1992-2000.

See Also

baymap

Examples

```
## Not run:
data(data_test)
res <- baymap(data = data_test,
inits.mu = c(0.05, 0.85, 0.2), n.iter = 4500)
data_new <- predict(res, data_test, burn.in = 3000)

## End(Not run)
```

List of Abbreviations

A	Adenine
Ago	Argonaute protein
BayMAP	Bayesian hierarchical Model for the Analysis of PAR-CLIP data
C	Cytosine
cDNA	Complementary DNA
CDS	Coding Sequence
CWT	Continuous Wavelet Transforms
CLIP	Crosslinking and Immunoprecipitation
DNA	Deoxyribonucleic Acid
e.g.	Exampli Gratia (lat.), for example
G	Guanine
HEK	Human Embryonic Kidney
i.e.	Id Est (lat.), this means
MCMC	Marcov Chain Monte Carlo
MRN	Mini-Rank Norm
miRNA	MicroRNA
mRNA	Messenger RNA
nt	Nucleotide
PAR-CLIP	Photoactivatable-Ribonucleoside-Enhanced CLIP
PCR	Polymerase Chain Reaction
RISC	RNA-Induced Silencing Complex
RNA	Ribonucleic Acid

SNP	Single Nucleotide Polymorphism
T	Thymine
U	Uracil
UTR	Untranslated Region
UV	Ultraviolet Light
4-SU	4-thiouridine
6-SG	6-thioguanosine

List of Figures

List of Tables

Contribution to manuscripts

BayMAP: a Bayesian hierarchical model for the analysis of PAR-CLIP data

Eva-Maria Huessler[1], Martin Schäfer[1,2], Holger Schwender[1] and Pablo Landgraf[3]

[1]Mathematical Institute, Heinrich Heine University, 40225 Düsseldorf

[2]Epidemiology Unit, German Rheumatism Centre, 10117 Berlin

[3]Department of Pediatric Oncology and Hematology, Children's Hospital, University of Cologne, 50937 Cologne

Authorship:	first author
Contributed part:	85%
Contribution:	Development of the statistical method
	Implementation in software
	Creating simulated data sets
	Statistical data analysis
	Preparing figures and tables
	Interpretation of results
	Writing the paper
Journal:	Bioinformatics
Impact factor:	5.610
Date of publication:	09 November 2018
DOI:	10.1093/bioinformatics/bty904
PubMed-ID:	30418480

Bibliography

[1] Brian D Adams, Andrea L Kasinski, and Frank J Slack. Aberrant regulation and function of microRNAs in cancer. Current Biology, 24(16):R762–R776, 2014.

[2] Vikram Agarwal, George W Bell, Jin-Wu Nam, and David P Bartel. Predicting effective microRNA target sites in mammalian mRNAs. elife, 4:e05005, 2015.

[3] B. Alberts. Molecular Biology of the Cell. CRC Press, 2017. ISBN 9781317563747.

[4] David P Bartel. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell, 116 (2):281–297, 2004.

[5] David P Bartel. MicroRNAs: target recognition and regulatory functions. Cell, 136(2):215–233, 2009.

[6] A. Berk, C.A. Kaiser, H. Lodish, A. Amon, H. Ploegh, A. Bretscher, M. Krieger, and K.C. Martin. Molecular Cell Biology. Macmillan Learning, 2016. ISBN 9781464183393.

[7] Juan S Bonifacino, Esteban C Dell'Angelica, and Timothy A Springer. Immunoprecipitation. Current protocols in molecular biology, 48(1):10–16, 1999.

[8] Beibei Chen, Jonghyun Yun, Min S Kim, Joshua T Mendell, and Yang Xie. PIPE-CLIP: a comprehensive online tool for CLIP-seq data analysis. Genome Biol, 15(1):R18, 2014.

[9] Sung Wook Chi, Julie B Zang, Aldo Mele, and Robert B Darnell. Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps. Nature, 460(7254):479–486, 2009.

[10] Sung Wook Chi, Gregory J Hannon, and Robert B Darnell. An alternative mode of microRNA target recognition. Nature structural & molecular biology, 19(3):321, 2012.

[11] Federico Comoglio, Cem Sievers, and Renato Paro. Sensitive and highly resolved identification of RNA-protein interaction sites in PAR-CLIP data. BMC Bioinformatics, 16(1):32, 2015.

[12] David L Corcoran, Stoyan Georgiev, Neelanjan Mukherjee, Eva Gottwein, Rebecca L Skalsky, Jack D Keene, Uwe Ohler, et al. PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. Genome Biol, 12(8):R79, 2011.

[13] Svenja Daschkey, Silja Röttgers, Anamika Giri, Jutta Bradtke, Andrea Teigler-Schlegel, Gunter Meister, Arndt Borkhardt, and Pablo Landgraf. MicroRNAs distinguish cytogenetic subgroups in pediatric AML and contribute to complex regulatory networks in AML-relevant pathways. PlOS ONE, 8(2):e56334, 2013.

[14] Annette J Dobson and Adrian G Barnett. An introduction to generalized linear models. Chapman and Hall/CRC, 2008.

[15] Florian Erhard, Lars Dölken, Lukasz Jaskiewicz, and Ralf Zimmer. PARma: identification of microRNA target sites in AGO-PAR-CLIP data. Genome Biol, 14(7):R79, 2013.

[16] Thalia A Farazi, Jessica I Hoell, Pavel Morozov, and Thomas Tuschl. MicroRNAs in human cancer. In MicroRNA Cancer Regulation, pages 1–20. Springer, 2013.

[17] Alan E Gelfand and Adrian FM Smith. Sampling-based approaches to calculating marginal densities. Journal of the American statistical association, 85(410):398–409, 1990.

[18] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. Bayesian data analysis, volume 2. CRC press Boca Raton, FL, 2014.

[19] Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on pattern analysis and machine intelligence, (6):721–741, 1984.

[20] Monica Golumbeanu, Pejman Mohammadi, and Niko Beerenwinkel. BMix: probabilistic modeling of occurring substitutions in PAR-CLIP data. Bioinformatics, 32(7):976–983, 2015.

[21] Jane EA Gordon, Justin J-L Wong, and John EJ Rasko. MicroRNAs in myeloid malignancies. British journal of haematology, 162(2):162–176, 2013.

[22] Eva Gottwein, David L Corcoran, Neelanjan Mukherjee, Rebecca L Skalsky, Markus Hafner, Jeffrey D Nusbaum, Priscilla Shamulailatpam, Cassandra L Love, Sandeep S Dave, Thomas Tuschl, et al. Viral microRNA targetome of KSHV-infected primary effusion lymphoma cell lines. Cell host & microbe, 10(5):515–526, 2011.

[23] Frank L Graham, J Smiley, WC Russell, and R Nairn. Characteristics of a human cell line transformed by DNA from human adenovirus type 5. Journal of general virology, 36(1): 59–72, 1977.

[24] Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano, Anna-Carina Jungkamp, Mathias Munschauer, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. Cell, 141(1):129–141, 2010.

[25] W Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. 1970.

[26] Huessler. Baymap. http://stat.math.uni-duesseldorf.de/baymap.

[27] Eva-Maria Huessler, Martin Schäfer, Holger Schwender, and Pablo Landgraf. BayMAP: a Bayesian hierarchical model for the analysis of PAR-CLIP data. Bioinformatics, 35(12): 1992–2000, 2018.

[28] Lukasz Jaskiewicz, Biter Bilen, Jean Hausser, and Mihaela Zavolan. Argonaute CLIP–A method to identify in vivo targets of miRNAs. Methods, 58(2):106–112, 2012.

[29] Mohsen Khorshid, Jean Hausser, Mihaela Zavolan, and Erik Van Nimwegen. A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. Nature methods, 10(3):253, 2013.

[30] Shivendra Kishore, Lukasz Jaskiewicz, Lukas Burger, Jean Hausser, Mohsen Khorshid, and Mihaela Zavolan. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. Nature methods, 8(7):559–564, 2011.

[31] Andreas Kloetgen, Arndt Borkhardt, Jessica I Hoell, and Alice C McHardy. The PARA-suite: PAR-CLIP specific sequence read simulation and processing. PeerJ, 4:e2619, 2016.

[32] Julian Konig, Kathi Zarnack, Gregor Rot, Tomaz Curk, Melis Kayikci, Blaz Zupan, Daniel J Turner, Nicholas M Luscombe, and Jernej Ule. iCLIP-transcriptome-wide mapping of

protein-RNA interactions with individual nucleotide resolution. Journal of visualized experiments: JoVE, (50), 2011.

[33] Diane Lambert. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics, 34(1):1–14, 1992.

[34] Charles H Lawrie. MicroRNAs and lymphomagenesis: a functional review. British journal of haematology, 160(5):571–581, 2013.

[35] Duncan Lee. A comparison of conditional autoregressive models used in Bayesian disease mapping. Spatial and spatio-temporal epidemiology, 2(2):79–89, 2011.

[36] Benjamin P Lewis, Christopher B Burge, and David P Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell, 120(1):15–20, 2005.

[37] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, 25(14):1754–1760, 2009.

[38] Donny D Licatalosi, Aldo Mele, John J Fak, Jernej Ule, Melis Kayikci, Sung Wook Chi, Tyson A Clark, Anthony C Schweitzer, John E Blume, Xuning Wang, et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. Nature, 456(7221):464–469, 2008.

[39] Yao-Cheng Lin, Morgane Boone, Leander Meuris, Irma Lemmens, Nadine Van Roy, Arne Soete, Joke Reumers, Matthieu Moisse, Stéphane Plaisance, Radoje Drmanac, et al. Genome dynamics of the human embryonic kidney 293 lineage in response to cell biology manipulations. Nature communications, 5(1):1–12, 2014.

[40] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal, 17(1):pp–10, 2011.

[41] Sebastian Memczak, Marvin Jens, Antigoni Elefsinioti, Francesca Torti, Janna Krueger, Agnieszka Rybak, Luisa Maier, Sebastian D Mackowiak, Lea H Gregersen, Mathias Munschauer, et al. Circular RNAs are a large class of animal RNAs with regulatory potency. Nature, 495(7441):333, 2013.

[42] Ioannis Ntzoufras. Bayesian modeling using WinBUGS, volume 698. John Wiley & Sons, 2011.

[43] Swiss Institute of Bioinformatics. CLIPZ homepage. http://www.clipz.unibas.ch. Online; accessed July 22, 2015.

[44] DJ Patel, J-B Ma, Y-R Yuan, K Ye, Y Pei, V Kuryavyi, Lucy Malinina, G Meister, and T Tuschl. Structural biology of RNA silencing and its functional implications. In Cold Spring Harbor symposia on quantitative biology, volume 71, pages 81–93. Cold Spring Harbor Laboratory Press, 2006.

[45] Lasse Peters and Gunter Meister. Argonaute proteins: mediators of RNA silencing. Molecular cell, 26(5):611–623, 2007.

[46] Ashley J Pratt and Ian J MacRae. The RNA-induced silencing complex: a versatile gene-silencing machine. Journal of Biological Chemistry, 284(27):17897–17901, 2009.

[47] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

[48] Holger Schwender, Sylvia Rabstein, and Katja Ickstadt. Do You Speak Genomish? Chance, 19(3):3–8, 2006.

[49] Cem Sievers, Tommy Schlumpf, Ritwick Sawarkar, Federico Comoglio, and Renato Paro. Mixture models and wavelet transforms reveal high confidence RNA-protein interaction sites in MOV10 PAR-CLIP data. Nucleic Acids Research, 40(20):e160–e160, 2012.

[50] David Spiegelhalter, Andrew Thomas, Nicky Best, and Dave Lunn. WinBUGS user manual. http://www.politicalbubbles.org/bayes_beach/manual14.pdf, 2003. Online; accessed June 29, 2015.

[51] Phillipp Torkler. STAMMP - A statistical model and processing pipeline for PAR-CLIP data reveals transcriptome maps of mRNP biogenesis factors. Ludwig Maximilian University of Munich, 2015.

[52] Andrea Ventura and Tyler Jacks. MicroRNAs and cancer: short RNAs go a long way. Cell, 136(4):586–591, 2009.

[53] Tao Wang, Beibei Chen, MinSoo Kim, Yang Xie, and Guanghua Xiao. A model-based approach to identify binding sites in CLIP-seq data. PlOS ONE, 9(4):e93248, 2014.

[54] Julia Winter, Stephanie Jung, Sarina Keller, Richard I Gregory, and Sven Diederichs. Many roads to maturity: microRNA biogenesis pathways and their regulation. Nature cell biology, 11(3):228–234, 2009.

[55] Jonghyun Yun, Tao Wang, and Guanghua Xiao. Bayesian hidden Markov models to identify RNA–protein interaction sites in PAR-CLIP. Biometrics, 70(2):430–440, 2014.

Eidesstattliche Versicherung

Ich versichere an Eides statt, dass die Dissertation von mir selbstständig und ohne unzulässige fremde Hilfe unter Beachtung der Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf erstellt worden ist.

Eva-Maria Hüßler, Juli 2021, Düsseldorf