Chapter 5

Genome scale motif inference in *Saccharomyces cerevisiae*

The algorithmic work presented in the previous chapters, particularly the **metamatti** motif classification framework, was partly motivated by the needs of the sequence analysis projects in which I have been involved. One part of this has been involvement in collaborative projects, where I have analysed human and mouse noncoding sequence with computational regulatory motif inference, scanning and statistical testing tools [1], some of which I had developed for the purpose. The more substantial part has however been the genome scale *de novo* regulatory motif inference work with the *Saccharomyces cerevisiae* genome that will be discussed in this chapter.

5.1 Background

Budding yeast is an organism of great interest for regulatory genomics, given its small genome, amenability to genetic manipulation, and relatively simple regulatory mechanisms including a small total number of transcription factors (Goffeau et al., 1996). The DNA specificity of many of its TFs has been characterised in a combination of several high throughput *in vitro* studies (Badis et al., 2008; Zhu et al., 2009), providing a high quality reference set of regulatory motifs that

[1] Majority of this work is now published in (Lewis et al., 2009) and Murray *et al.* (in press)

are useful for comparison with *de novo* discoveries. Information on the genomic binding positions for many of its TFs are also known from large scale ChIP-chip based studies (Harbison et al., 2004; Lee et al., 2002). Gene expression studies comparing knock-out lines for nearly all of its known sequence specific TFs to the wild-type are available (Hu et al., 2007; Reimand et al., 2010). Furthermore, many of the target genes of these TFs are known, as a result of the above ChIP-chip and expression studies, and literally thousands of other primary publications that have been manually curated (Teixeira et al., 2006). The *in vivo* DNA specificity of many budding yeast TFs is yet to be studied in high resolution, but nevertheless, budding yeast currently offers the best available knowledge base of TFs, TF target genes and binding site specificity, of any eukaryotic genome.

These resources together allow us to assess the ability of *de novo* motif inference algorithms to find large collections of regulatory motifs on a genome scale. Information from this large scale study is valuable most importantly because it indicates which of the algorithms, if any, are sufficiently accurate for complex regulatory problems that are aplenty in large genomes of multicellular eukaryotes.

5.1.1 Genome scale motif inference

Motif inference studies have traditionally been made to infer one or more recurring signals from a sequence set – of dozens to at most a few hundred – of sequences assumed to be co-regulated or involved in the same biological process. The rapid expansion in the number of complete genomes and computational power has however made it possible to use motif inference for a more ambitious goal: genome scale inference of comprehensive motif collections or 'dictionaries' from a significant subset of promoter sequences of a genome. I will below review a selection of previous literature on genome scale motif inference – both *ab initio* methods [1], and methods which apply gene expression or sequence conservation as a guide. See Section 1.2 for a more general discussion of motif inference methods.

To my knowledge, the earliest motif discovery study which fits the above criteria of *de novo* genome-scale motif inference is that of Brazma et al. (1998), who

[1] Ab initio suggests in this context that no other information but the reference genome sequence and the predicted transcription start sites (putative promoter locations) are used as input for inferring the motifs.

predicted a series of regular expression like patterns from the *S. cerevisiae* genome using the SPEXS algorithm (Vilo, 1998), in an experiment where the algorithm was run 'blindly' with 6,000 upstream sequences. Assessing the significance of the found patterns, however, proved troublesome: top scoring regular expressions are matched to TRANSFAC binding site entries, but the authors attempted to draw few conclusions based on the found matches, except to note the surprise at being able to discover TFBS-like patterns with sequence information alone. Bussemaker et al. (2000) also presented a word enumeration based study where they found 11 known matching k-mers from a genome-wide study of *S. cerevisiae* promoters.

Several large, gene expression cluster-driven motif inference studies have been published. Among the earliest were Roth et al. (1998), who successfully recapitulated motifs of some of the key regulators of galactose response, heat shock and mating type regulatory systems in the *S. cerevisiae*, using the Gibbs sampling based AlignACE algorithm. Vilo et al. (2000) on the other hand used a word enumeration based method to find 62 clustered consensus strings reported to be match words in the SCPD database (Zhu and Zhang, 1999). Methods that go beyond clustering genes (and applying motif inference algorithms separately per cluster) have also been developed: Bussemaker et al. (2001) introduced a gene expression correlation based method REDUCE, which they apply to *S. cerevisiae* cell cycle regulation (Bussemaker et al., 2001). Elemento and Tavazoie (2005) use mutual information between gene expression patterns and the absence or presence of motifs as a means to infer *cis*-regulatory elements, in both mammalian, the yeast, and the *Plasmodium falciparum* genomes.

Whereas gene expression patterns are useful in inferring regulators which act in a certain state of the cell, use of sequence conservation has been used as a general 'cell state blind' informant for large scale motif inference. One of the earliest studies was Kellis et al. (2003) with a study of *S. cerevisiae*: a whole-genome multiple alignment of *S. cerevisiae* with *S. paradoxus*, *S. mikatae* and *S. bayanus*, which identified highly conserved consensus strings by clustering instances of shorter 'mini-motifs'. Amongst the 78 motifs found, 28 closely match known TFBS consensus strings. Comparative techniques were later used by the same authors and others (Elemento and Tavazoie, 2005; Ettwiller, 2005; Jones

and Pevzner, 2006; Xie et al., 2005, 2007).

In conclusion, different large scale approaches to inferring *cis*-regulatory elements have been proposed, and several of them have been applied to the *S. cerevisiae* genome. In contrast to these previous studies, my perspective to inferring motif dictionaries from the budding yeast is primarily to find out how different previously published algorithms perform at this task, rather than setting out to discover novel functional motifs. This assessment is now made possible due to the availability of regulatory motifs, and sets of target genes for many of the budding yeast TFs. This is important, because performance of *de novo* motif inference methods have not previously been systematically assessed on biologically relevant, realistic problems.

5.1.2 Performance inference method assessments

Publications describing regulatory motif inference algorithms typically contain a comparison of the algorithm introduced with at least some previously published ones. Standard assessment criteria or benchmark datasets have not surfaced, and new methods are often compared only with a small number of common existing methods, so it is not always clear how they compare with the state of the art. An objective assessment of the merits of the hundreds of different available algorithms is therefore difficult. To my knowledge, the most comprehensive *de novo* motif inference algorithm benchmark, involving 13 different methods and discussed in more detail below, has been conducted by Tompa et al. (2005). As more and more motif inference methods are published on top of the hundreds already available, being able to assess the performance of methods relative to each other becomes increasingly important.

Two types of approaches have been used in previous literature for ranking methods:

1. Finding TFBSs motifs from motifs from well studied collections of *cis*-regulatory elements (Ao et al., 2004; Liu et al., 2002; Roth et al., 1998; Thijs et al., 2002).

2. Finding TFBS motifs from synthetic sequence created by planting, or 'spiking' motifs into background sequence. The background is usually some neu-

tral sequence thought to be devoid of other motifs (e.g. intronic sequence). This approach is taken for instance by Down and Hubbard (2005); Pevzner and Sze (2000); Workman and Stormo (2000).

Measuring the performance of algorithms in either of the above cases is done most often by counting instances of motifs above some significance level, and comparing the overlap of the list of predicted motif instances to a reference binding site collection. The reference is either a set of known sites, if the assessment is made with real sequence, or a known set of planted instances of the target motif in the case of synthetic sequence. Some commonly used metrics derived from comparing binding site matches on nucleotide and binding site level are discussed below in Section 5.1.3. Testing a motif discovery algorithm in its capacity to find motifs from unmodified biological sequence would perhaps seem as the most intuitive approach. However, to date, performance assessment with unmodified biological sequence has been limited to small numbers of individual genomic regions because of our limited knowledge of regulatory regions. Perhaps for this reason, synthetic regulatory sequence is often used, and is also the primary type of sequence used in the Tompa et al. (2005) assessment, detailed below. Regardless of the sequence type, the above assessment criteria also make the assumption that a motif inference algorithm should be able to partition sequences into binding sites and background sequence. The appropriateness of this partitioning assumption is also discussed below.

5.1.3 The Tompa *et al.* (2005) assessment

Tompa et al. (2005) compared 13 different motif inference methods in their ability to predict motif binding sites from mostly synthetic promoter sequence sets. The authors assessed the algorithms with summary statistics derived from motif hit instances predicted in the sequences. A thorough review of the assessment is provided here, because it is the most comprehensive performance assessment of its kind, and has been influential for performance assessments presented in later publications. It also suffers from a number of self-professed flaws, some of which I intend to address in the present work.

The binding site sequences used in their assessment were retrieved from the

TRANSFAC database (Matys et al., 2006), and inserted into a mixture of the types of background sequences: 1) randomly chosen promoter sequences from the same genome, or 2) sequences generated from a 3^{rd} order Markov chain. Unmodified binding site sequences are used in a third type of benchmark dataset. In total, 52 datasets were created for different TFs of fly, human, mouse, rat and yeast (one dataset per TF), and four negative control sequence sets created from the Markov chain background were added to the set. The benefit of testing algorithms with synthetic sequences (types 2 and 3) is the controlled environment they provide: inserted binding site positions are known, and motif frequency or sequence length can be varied at will. This is the reason that a benchmark with synthetic sequences, consisting of sampled TFBS hits in intronic background sequence, is also used in my work in Chapter 3 to allow the known motif frequency (sequence length) to be varied in a predictable way. Making sure that synthetic benchmarking sequence sets are realistic is not possible, especially in a genome scale problem, because of our limited understanding of regulatory sequences. In this case the background sequence is sampled from a 3^{rd} order Markov chain (trained from genomic sequence) in the Tompa et al. (2005) assessment are almost certainly not closely related to real promoter sequence in their properties (nucleotide content in genomic sequence varies in discrete regions, as discussed in Section 1.3.3).

At the nucleotide level, four types of measurements were defined, to measure the overlap of real binding sites with those predicted:

- **nTP**: the number of nucleotide positions in both known sites and predicted sites.
- **nFN**: the number of nucleotide positions in known sites but not in predicted sites.
- **nFP**: the number of nucleotide positions not in known sites but in predicted sites.
- **nTN**: the number of nucleotide positions in neither known sites nor predicted sites.

Similar metrics were also defined for binding site overlap, with an arbitrarily chosen 25% overlap required between the nucleotides of the sites to be considered overlapping.

Tompa et al. (2005) then defined a number of further statistics based on nTP, nFN, nFP, nTN. Firstly, sensitivity nSn, specificity nSp, and positive predictive value $nPPV$:

$$nSn = nTP/nTP + nFN \tag{5.1}$$

$$nSp = nTN/(nTN + nFP) \tag{5.2}$$

$$nPPV = nTP/(nTP + nFP) \tag{5.3}$$

A nucleotide level performance coefficient nPC, intended to "in some sense average (some of) [the above] quantities", is also reported (Equation 5.4), following the work of Pevzner and Sze (2000).

$$nPC = nTP/(nTP + nFN + nFP) \tag{5.4}$$

Following Burset and Guigó (1996), the authors also report a nucleotide level Pearson product-moment correlation coefficient (Equation 5.5), and an average site performance $sASP$ (Equation 5.6).

$$nCC = \frac{nTP \times nTN - nFN \times nFP}{\sqrt{(nTP + nFN)(nTN + nFP)(nTP + nFP)(nTN + nFN)}} \tag{5.5}$$

$$sASP = (sSn + sPPV)/2 \tag{5.6}$$

The measures nSn, nSp, $nPPV$, nPC, nCC, $sASP$ are then summarised in three different ways per tool across the datasets: either as an average, as a Z-score, or a 'combined' weighted average score where all the measures are computed as if the real and predicted sites were part of one large dataset instead of 56 individual ones. Most of the chosen performance measures however present problems with the four negative control datasets with no motifs: nSn, nCC,

$sASP$ are not defined, and $nPPV$, nPC and $sPPV$ are uninformative. Most troubling however is that when a tool makes no prediction in datasets containing motif instances, $TP + FP = 0$, causing $nPPV$, nCC, $sPPV$ to be undefined and nSn, nPC and sSN to be uninformative. The 'combined' average score works around this to the extent where these predictions consisting of entirely false negative predictions do not contribute at all. The score does however still penalise methods which make a small number of false positive prediction against those which attempt to make no predictions whatsoever (also pointed out by the authors). The statistics used also leave no intuition for how any of the tools performed on any individual dataset, and no guidance is given by the authors for the interpretation or relative importance of the various different measures.

A further problem with the above performance measures is that if the binding site positions called either positive or negative for a predicted binding event are dramatically affected by the motif significance thresholds used (high significance cutoff increases the false positive rate). Indeed, given that different experts ran the experiments, it is possible that this assessment tested not only the ability to detect recurring motifs with different algorithms, but also the stringency and parameter choices involved in deciding which of the potential binding site matches to report based on the inferred motifs. The problem of inferring a motif, and finding its binding site matches are independent in the formulation used by many motif inference algorithms. Some Bayesian motif inference algorithms do not in fact report individual binding site matches as part of the motif inference process (Down and Hubbard, 2005). Furthermore, when the above binding site level measures are computed for real promoter sequence with experimentally determined TFBSs, the quality of binding site data affects all of the above-mentioned measures. For example, some of the false positives can in fact be true, unknown binding sites.

The authors cite several gene finding assessments (Burge and Karlin, 1997; Burset and Guigó, 1996; Reese et al., 2000) as the inspiration for their approach. In those studies protein coding gene models are inserted to large sets of vertebrate sequence. I question the analogy between gene finding and TFBS finding, and advocate the use of comparison of motifs, rather than comparison of individual motif matches, as the primary means to benchmark motif inference performance.

TFBSs are several orders of magnitude shorter and lower in information, transient and turned over during evolutionary time scale, tend to co-occur, and vary in frequency and stringency of matches, depending on the TFBS in ways that are not well understood (see Section 1.1). Furthermore, weak binding sites which can be very 'distant' matches to the motif, and therefore both difficult to find experimentally or by scanning computational motifs, can also contribute to regulatory responses (Gertz et al., 2009). A motif match alone does not determine if a genomic position binds a TF or not; other levels of information relevant for regulation is stored in genomes, including for instance tissue specific epigenetic marks and the DNA melting propensity. Making use of such additional sources of evidence substantially improves classification of sites as either binding or non-binding Ernst et al. (2010); Lähdesmäki et al. (2008); Ramsey et al. (2010). For many eukaryotic TFs, even a perfect motif inference algorithm cannot predict its binding sites accurately, in turn raising questions about the use of binding site or nucleotide level based methods for their performance assessment.

The authors required the experts applying prediction methods to report a single high confidence prediction. Especially when inferring motifs from real-world genomic sequence, one cannot be sure of the absence of unexpected 'real' sequence motifs, which a good computational motif prediction tools should in fact be able to report. Indeed, the authors also state that "no attempt was made to eliminate sequences that might contain additional transcription factor binding sites, since our ability to identify such sites accurately is limited." Therefore, methods which were (correctly) able to report additional motifs present in the sequence, but where the genomic matches of the correct motif was not submitted for analysis, can in fact be penalised for it heavily, perhaps explaining in part the reportedly bad prediction performance seen with the real sequences. Inferring motifs, and ranking them, should be considered independently. I would argue also that the algorithm assessment should be made with a collection of inferred motifs per method, instead of a single motif per method. Otherwise the assessment measures, in part, the correctness of post-processing and motif ranking steps which can be made by the experts – and were not detailed by the authors.

In conclusion, the design of the Tompa et al. (2005) study suffers from certain troubling assumptions and sources of potential bias. It is also inconclusive; the

authors do not offer direct advice or a ranking of methods based on the measures, and point out many of the study's shortcomings also themselves. To my surprise, I have been unable to find later performance assessments which would directly try to address these shortcomings, apart from Li and Tompa (2006); Sandve et al. (2007) who mostly confirm problems apparent in the Tompa et al. (2005) assessment, but do not offer a new thorough assessment. On the contrary, several motif inference method publications after this paper have used the same statistical measures or synthetic datasets provided by Tompa et al. (2005), as supporting evidence for the favourable performance of their computational tools to previous work (Chan et al., 2009; Fauteux et al., 2008; Gunewardena and Zhang, 2008; Hu et al., 2006; Klepper et al., 2008; Lu et al., 2008; Peng et al., 2006; Reddy et al., 2007; Robinson et al., 2006; Sandve et al., 2008; Wang and Zhang, 2006; Wijaya et al., 2008; Zare-Mirakabad et al., 2009).

5.2 Materials & Method

This project had two phases: running a number of DNA motif inference algorithms on a large series of genomic sequence, and then assessing the discovered motifs. The sections below firstly describe the sequence sets used in the project (Section 5.2.1), before giving an account of the tested motif inference algorithms (Section 5.2.2). The remaining sections then detail the methodology of the various analyses conducted on the predicted motif sets. Notably, the performance assessment of methods is made in a parameter free manner when possible. Motif scanning with a motif hit significance cutoff parameter is done primarily for exploration of the data, for instance to find subsets of potentially interesting motifs which do not match the reference motif sets (Section 5.3.6.4).

5.2.1 Sequence and annotation retrieval

The *S. cerevisiae* promoter sequence used in all motif inference runs consisted of 200 base long upstream sequences from 1,000 randomly chosen protein coding genes with 5-way orthologs between the hemiascomycetous yeast species *S. cerevisiae*, *Candida glabrata*, *Kluyveromyces lactis*, *Debaryomyces hansenii* and

Yarrowia lipolytica. These sequence sets were collated by Dr Thomas Down. Briefly, Ensembl Compara (Birney et al., 2004) formatted database schemas were created of the genomic sequence data retrieved from the hemiascomycete comparative genomics database Genolevures (Sherman et al., 2004). BLASTP (Altschul and Gish, 1996) and reciprocal matching was then used to assign orthology between genes. *S. cerevisiae* sequences for orthologous genes were then retrieved, and a randomly selected subset of 1,000 200 bases long promoters chosen from the subset (other organisms were only used for selecting candidate genes).

I fetched additional sequence sets from the Ensembl database Hubbard et al. (2009) for the purposes of assessing the motifs (e.g. positional bias in Section 5.2.7.1, or the conservation analysis in Section 5.3.6.1). Most sequence fetching tasks were done from the Ensembl database with tools which I created with Dr Thomas Downs help using the BioJava toolkit Holland et al. (2008). Sequences for the assessment originated from version 57 of the Ensembl Core database. An usage example for the **nmensemblseq** retrieval tool is provided below:

```
nmensemblseq \
-database saccharomyces_cerevisiae_core_57_1j \
-host ensembldb.ensembl.org \
-user anonymous \
-port 5306 -noRepeatMask \
-noExcludeTranslations \
-proteinCoding -known \
-fivePrimeUTR 500 0 -type protein_coding
```

The genomic coordinates for the sequence regions were also retrieved for the sequences, similarly using **nmensemblseq**, by adding the command line flag `-outputType gff`. A more thorough tutorial on using this utility, as well as some of the others included in the nmica-extra package I created during my project, are provided in Appendix B. The sequence retrieval tools were also integrated with the iMotifs sequence motif visualization and inference environment which I created during my project (Piipari et al., 2010b) (Figure 5.1A,B).

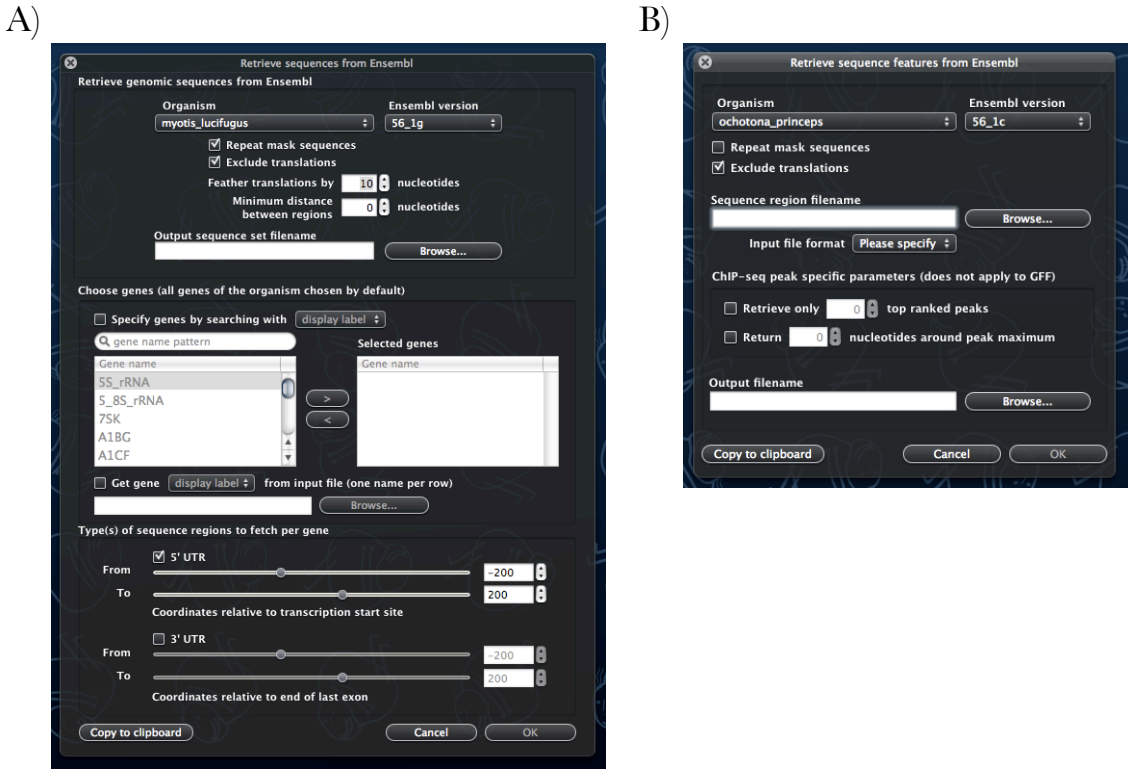


Figure 5.1: The sequence retrieval tools included in iMotifs. A) Configuration dialog for the 5 / 3 UTR sequence retrieval tool nmensemblseq. B) Configuration dialog for the GFF/BED sequence feature and ChIP-seq peak retrieval tools (**nmensemblfeat** and **nmensemblpeakseq**).

5.2.2 Motif inference

I tested predicting motifs with all of the thirteen motif inference algorithms from the Tompa et al. (2005) assessment, as well as SOMBRERO (Mahony et al., 2005b), PRIORITY (Narlikar et al., 2006), MoAn (Valen et al., 2009) and BayesMD (Tang et al., 2008). The Tompa et al. (2005) methods were chosen because it is perhaps the most comprehensive assessment to date, and the additional methods (NestedMICA, SOMBRERO, PRIORITY, BayesMD, MoAn) were tested because of their reported favourable performance in comparison to those tested in Tompa et al. (2005). The input parameters used for all of the successfully run algorithms are described in Appendix C. All inference experiments were made with the random orthologous promoter sequence set detailed in Section 5.2.1. If possible, each algorithm was made to predict 200 motifs. In case this was not possible, the largest motif set output by the tool was used for evaluation.

The PWMs output by each of the programs were converted to the XMS format used by the NestedMICA suite and iMotifs, with scripts that use the libxms Ruby bindings which I wrote Piipari et al. (2010b). Two of the algorithms which successfully returned results use a consensus string representation of their output (YMF and Oligoanalysis). These were converted to a PWM representation, applying a very small pseudo-count of 0.001 to the motifs.

I ran all of the motif inference programs myself after consulting the publications describing the algorithms, and other available documentation regarding each of them. This is in contrast with the Tompa et al. (2005) assessment, which was a large collaborative project where outside experts (the authors of the algorithms) created the motif predictions, which were assessed independently.

Conservation of noncoding sequence has been applied in some earlier studies as a means of selecting candidate sequences for motif inference (Elemento and Tavazoie, 2005; Hardison, 2000; Kellis et al., 2003; Xie et al., 2005). However, I decided not to choose or weight promoter sequences for my study according to conservation. There were several reasons for this decision. Firstly, leaving sequence conservation aside from the motif inference step allows it to be used as an independent way of assessing the motifs. Secondly, the traditional con-

servation scoring methods, such as the PhastCons (Siepel et al., 2005) used in the present study, assume an alignment between the sequences; given the small alphabet size of DNA, and repetitive nature of genomic sequence, alignment errors are inevitable. Thirdly, biologically active TFBSs are known to be turned over quickly, and some experience near to neutral mutation rates (Kunarso et al., 2010; Schmidt et al., 2010). Although success has been reported in studies using conservation as a criterion of choosing motifs amongst candidates (Xie et al., 2005), it does not always lead to detection of correct ones. For instance, Li et al. (2005) suggest that a simple conservation based significance score would lead to the selection of an incorrect TFBS motif in 28% of cases with yeast ChIP-chip data of Lee et al. (2002).

The rate of binding site turnover has been studied in high resolution with ChIP-seq assaying in the CEBPA and HNF4A transcription factors, which are strongly conserved across placental mammals (Schmidt et al., 2010). Less than 0.3% of binding events were shown to be conserved in all assayed species. A study by Kunarso et al. (2010) finds that in the case of Oct4 and Nanog, 2.0% of sites are conserved in sequence. The binding regions however are functionally conserved at a much higher rate of between 50% and 10% depending on the chosen stringency of statistical significance. The strength of binding was not seen to associate with conservation, suggesting that the wide binding site spectrum of TFs is important (Schmidt et al., 2010), and that weak binding sites can have a biological effect. Several studies of human (Kasowski et al., 2010; McDaniell et al., 2010) and yeast (Zheng et al., 2010) individuals and related yeast species (Borneman et al., 2007) have shown results pointing in the same direction: individual TFBS events undergo rapid divergence, but a weak conservation signal tends to be found from a collection of TFBSs. The excess conservation of motifs is considered here, in combination with other lines of evidence, as a potential sign of function for computationally predicted motifs.

5.2.2.1 Unsuccessfully run algorithms

Several motif inference programs which were assessed in the Tompa et al. (2005) assessment by the authors of each of the algorithms were unsuccessfully attempted

to be used in the assessment, due to various reasons. Firstly, ANN-SPEC (Workman and Stormo, 2000) and Improbizer (Ao et al., 2004) are not distributed in binary or source code form without request from their authors, and the web servers provided are not suitable for discovering motifs on a genome scale. MITRA (Eskin and Pevzner, 2002) was not available at the URL noted by the authors[1], and no suitable online prediction server was found. QuickScore (Egnier, 2004) is only available as an online prediction server, and it was found not to handle the large (200,000nt) input sequence size. CONSENSUS (Hertz and Stormo, 1999) failed to compile on either 32 or 64 bit Linux or Mac OS X with the available compiler versions (gcc 4.2 and 4.3), and I was unable to find a binary distribution, or an online CONSENSUS prediction server suitable for the large analysis task at hand.

MoAn (Valen et al., 2009), PRIORITY (Narlikar et al., 2006), and SeSiMCMC (Favorov et al., 2005) were each successfully run with example data sets, but each only allowed for a single motif to be estimated.

BioProspector (Liu et al., 2001) was attempted to be run (`BioProspector -i orthologs-sc-1000.fa -r 200 -f yeast_all.bg -n 100 -h 1`). The currently distributed version of the program[2] does not parse the FASTA files used in the assessment. The file did appear to conform to the required variant of the file format given in the program's example file, and all of the other attempted tools processed it without problems. Furthermore, the BioProspector web server (http://robotics.stanford.edu/ xsliu/BioProspector/) only allows reporting a maximum of ten motifs (and its documentation specifically warns against specifying too large an input sequence set), which made it inapplicable for this benchmark (the target is 200 motifs). The same reason also made it impossible to run MDscan from the same authors (Liu et al., 2002) [3].

The Bayesian motif inference method BayesMD, which reportedly performs better with long promoter sequence than NestedMICA (Tang et al., 2008), was also tested, but it failed to report any output motifs due to persistently running

[1] `http://www.cs.columbia.edu/compbio/mitra`

[2] 'BioProspector.2004.zip', downloaded 1st June, 2010 from `http://motif.stanford.edu/distributions/bioprospector/`)

[3] 'MDScan.2004.zip', download made 1st June, 2010 from http://motif.stanford.edu/distributions/mdscan/

out of runtime memory, even with cluster nodes with 15.5G of allocatable memory.

5.2.3 Motif comparison

The computationally inferred *S. cerevisiae* motifs were compared to two different, partially overlapping reference sets of regulatory motifs: the JASPAR 2010 database (Portales-Casamar et al., 2010), and the Zhu et al. (2009) PBM motifs (some of which are included in the JASPAR dataset). The discovered motifs were also compared against one another to measure the level of redundancy across the sets.

To study the capacity of each of the motif inference methods to detect motifs that resemble known regulatory motifs, I compared them to motifs in the JASPAR 2010 database (Portales-Casamar et al., 2010). The JASPAR fungal motif dataset was chosen as the primary gold standard comparison set because it covers the great majority of all *S. cerevisiae* transcription factor motifs (177 TFBS non-redundant motifs in the database). It is an open access database, and its curation appears to be of more uniform quality than its competitor TRANSFAC which suffers from infrequent missing annotations such as species or publication references. Furthermore, JASPAR 2010, unlike previous versions of the database, includes a high coverage, non-redundant [1] set of *S. cerevisiae* motifs. The dataset originates mostly from two large scale studies; The single largest set included, and one preferred by Portales-Casamar et al. (2010) in case of conflicts, is the set of motifs from a study by Badis et al. (2008). This study includes data for a total of 112 TFs (107 of which are included in the non-redundant dataset, see Figure 5.2) from a combination of universal protein binding microarray assays (Berger et al., 2006; Mintseris and Eisen, 2006), cognate site identifier (CSI) microarrays (Warren et al., 2006), and DIP-chip (Liu et al., 2005) assays. The second large dataset included in JASPAR 2010 is the PBM based study by Zhu et al. (2009) (89 motifs). The remaining motifs from two datasets containing primarily literature based motifs from the SCPD binding profile database and literature (Zhu and Zhang, 1999),and the ChIP-chip based SwissRegulon database (Pachkov et al.,

[1] In this context, non-redundant means that only one motif prediction is included in the set for each TF.

2007) as well as computationally inferred motif dataset from the genome-wide ChIP-chip study of *S. cerevisiae* by MacIsaac et al. (2006). The motif comparisons presented in this chapter rely on these original studies and the manual curation conducted for the JASPAR database.

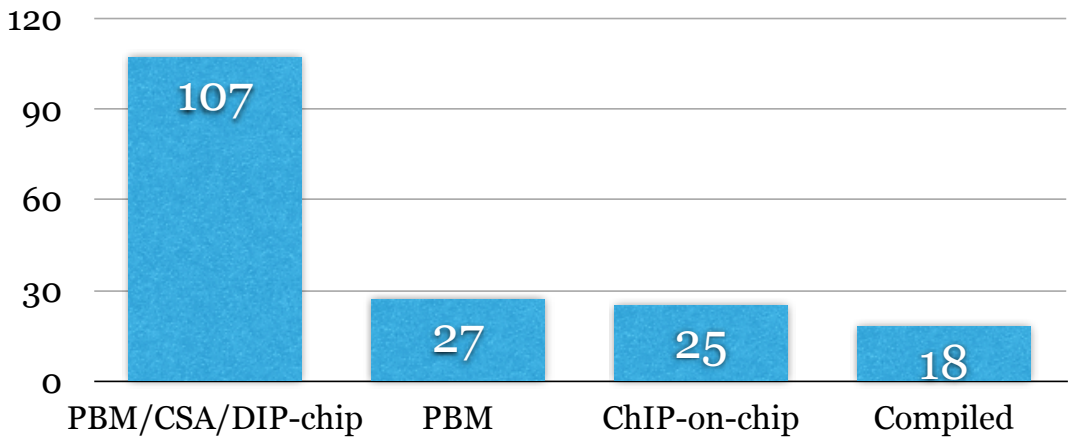


Figure 5.2: The number of motifs from different experimental sources in the JASPAR 2010 non-redundant fungal motif dataset. Note that some datasets contain motifs for TFs covered by other datasets. The PBM/CSA/Dip-chip dataset of Badis et al. (2008) for example contains in total 112 motifs, but only 107 of these are used in the non-redundant dataset by Portales-Casamar et al. (2010).

I also compared the inferred motifs to the Zhu et al. (2009) PBM motifs because they form the highest coverage regulatory motif dataset originating from a single type of experiment in the *S. cerevisiae*; the Badis et al. (2008) dataset with 112 motifs is in fact larger than the 89 motifs estimated by Zhu et al. (2009), but Badis et al. (2008) apply a combination of three different high-throughput methods, rather than one. A reference dataset additional to JASPAR was useful also because some of the JASPAR motifs could in fact originate from one of the tested algorithms (the 25 ChIP-on-chip and 18 'other' motifs in JASPAR are suspect). In contrast, the Zhu et al. (2009) motifs are all estimated from PBM data with the Seed-and-wobble algorithm (Berger et al., 2006), and these data should therefore not suffer from circularity in the comparison of the *de novo*

predictions to a reference.

5.2.3.1 Motif clustering with the SSD metric

The pairwise sum of squared differences (SSD) metric between PWMs, introduced by Down et al. (2007) (Equation 2.7 in Section 2.2.3), was computed systematically between all pairs of motifs. The distance matrix between all inferred motifs and JASPAR reference motifs were computed. Motif-to-motif distances allowed probing the redundancy of motifs within inferred sets with complete linkage clustering (Johnson, 1967). All motif sets were also clustered together with the JASPAR reference set, to summarise and visualise the trends in motif types found by each of the algorithms.

5.2.4 Motif scanning

After predicting sequence motifs with a selection of motif inference algorithms from the putative *S. cerevisiae* promoters, I scanned all putative promoter sequences of lengths 200bp, 500bp and 2000bp for the inferred motifs using the `nmscan` program included in the NestedMICA suite (Down and Hubbard, 2005). Sequences on the reverse strand with respect to the reference genome were reverse-complemented. The 200bp and 500bp sequence ends were aligned to the TSS. 2000bp sequences were centered on the TSS (i.e. they contain 1000bp upstream and 1000bp downstream sequence). The motif bit score function evaluated by nmscan for all PWMs W at positions p in sequence S is explained in Section 1.2.1 (Equation 1.1).

I also scanned all the sequences again to report the maximum bit score achieved in 200nt and 500nt upstream sequence regions of all *S. cerevisiae* genes (the `-maxPerSeq` mode in nmscan). Maximum bit scores were computed because they allow a parameter free comparison of score distributions between groups of promoters (genes). In Section 5.2.6 the maximum bit scores achieved by promoters are used to compare putative target genes of TFs to non-target genes (to see if the maximum bit scores discriminate TF targets from non-targets).

The match positions identified are dependent on the choice of the bit score threshold chosen for each of the motifs. Finding a meaningful statistical measure

of significance for motifs found from genomic DNA sequence itself is an active research problem. Approximate (Thijs et al., 2001) and even exact P-value calculation of PWM matches in DNA sequence (Zhang et al., 2007) is possible for PWMs given a sequence background with independent and identically distributed (i.i.d.) nucleotides, but i.i.d. is not a realistic model of background genomic DNA (Section 1.3.3). I therefore used a method for assigning the significance threshold of motif hits which can account for varying DNA dinucleotide content (Down et al., 2007).

In brief, the significance scores are computed with respect to a 1^{st} order mosaic sequence background model. I compare the score distribution of k-mers drawn from a 1^{st} order Mosaic sequence background model to the motif matches in each bin (both the expected and the observed score distribution are binned on 1 bit intervals). The benefit of this approach is that it allows a comparison to be made to a more representative background model of nucleotide sequence than what is commonly done (with a GC-content based background model). The drawback is that the computation is not exact, and the scores are reliant on the score bin sizes, and the total number of hits. This led to some difficulties, particularly with the motifs output by MEME, which are discussed below.

The total number of motif hits identified at different confidence thresholds varies dramatically. For instance, in the case of the 200 motifs predicted by NestedMICA, the total genomic hit count in 200 base upstream sequences ranges from $47,312$ with the 0.01 confidence threshold to $139,312$ hits with the 0.05 threshold. All analyses presented here were made with a stringent 0.01 cutoff.

5.2.5 Predicted binding site overlap

I computed the overlap between matches of different motifs within the inferred sets, and with the JASPAR database motifs, with a score similar to the one used by (Down et al., 2007) (Equation 5.7). In brief, the overlap score O, between binding sites B_1 of motif 1 and binding sites B_2 of motif 2, is the fraction of overlapping predicted sites which are hits for motif 1. O is 0 when the sets are disjunct, and 1 when a motif matches all of the other ones sites.

$$O = \frac{|B_1 \cap B_2|}{min(|B_1|, |B_2|)} \tag{5.7}$$

This allows the detection of similar motifs within the inferred motif sets, and also between the inferred and the experimentally validated JASPAR motifs. The overlap scores were considered for binding sites at the 0.01 significance cutoff (see Section 5.2.4 for discussion of determining motif hit significance). Overlapping motifs were analysed in an orientation independent manner, simply as chromosomal coordinate ranges with no strand information. This was done because all of the motif inference algorithms were run in a mode which allows for matches of a motif to occur in either orientation.

5.2.6 Association of motif hits to transcription factor target genes

A set of target genes is known for the great majority of *S. cerevisiae* regulatory TFs. For many of them, there is also an experimentally verified DNA motif in the JASPAR database. This makes it possible to judge if high-scoring matches of the predicted motifs distinguish target promoters of their likely TFs from non-target promoters. That is, for each computationally predicted motif with a closely related known TFBS motif, I test if the distribution of its maximum scoring occurrences differs between targets of the likely TF genes, and non-target genes.

I considered three different TF target gene datasets in this work. These datasets were:

1. **YEA**st **S**earch for **T**ranscriptional **R**egulators **A**nd **C**onsensus **T**racking database (Teixeira et al., 2006). Introduced in Section 5.2.6.1.

2. TF target calls from a reanalysis (Reimand et al., 2010) of a sequence specific TF knockout expression dataset Hu et al. (2007). Introduced in Section 5.2.6.2.

3. The Harbison et al. (2004) dataset of genome-wide location analysis by ChIP-chip (Iyer et al., 2001; Lieb et al., 2001). Introduced in Section 5.2.6.3.

For all of the target gene sets (introduced below), I extracted the curated TF–target dataset for all of the factors which also had a corresponding motif available in the JASPAR database. For each of these JASPAR motifs, I then calculated the closest motif from each predicted motif set (using the SSD distance metric by Down et al. (2007)). Maximum bit scores of the computationally predicted motifs were then compared in 500 base upstream regions of the *S. cerevisiae* genome using a two-sample single-tailed Kolmogorov-Smirnov (KS) test. The target genes of the TF, and the non-target genes, were the two different sets whose maximum bit score distributions were compared for each motif. In the KS test a low p-value indicates skewing of the bit score distribution of TF target promoters to the high bit-score end when compared to non-target genes. In addition to the two-sample KS-test, the rank-based two-sample Mann-Whitney (MW) test was computed for the maximum bit score distributions to see if the ranks of the maximum motif bit scores would be higher amongst the TF target genes. The non-parametric KS and MW tests were used due to the non-normal shape of the maximum bit score distribution.

5.2.6.1 YEASTRACT

YEAst **S**earch for **T**ranscriptional **R**egulators **A**nd **C**onsensus **T**racking database is a curated repository of transcriptional regulatory interactions in the *S. cerevisiae* genome (Teixeira et al., 2006). It currently collates a total of 12,346 TF–target associations for 149 TFs, each derived from one of a number of possible experimental sources, described in as many as 861 primary publications (download date 18/3/2010). The possible lines of evidence accepted as support of a target association in it are either:

1. change in the expression of the gene of interest owing to deletion or mutation of the TF gene (as measured by either gene by gene or genome-wide microarray).

2. binding of the transcription factor to the promoter region of the target gene, as supported by a band-shift assay (Fried and Crothers, 1981a), DNAse footprinting (Brenowitz et al., 1986), or ChIP assaying (Harbison et al., 2004).

In other words, the evidence sources in this dataset range from detailed individual genetic or physical interaction studies to high throughput ChIP-chip experiments.

5.2.6.2 Reimand *et al.* (2010) TF knockout and expression data based target set

Reimand et al. (2010) present a reanalysis of the sequence specific TF knockout expression dataset by Hu et al. (2007) of 269 sequence specific regulatory factors, including both general and specific TFs and factors involved in regulating chromatin state. The re-analysed dataset applied a series of corrections and processing steps to the expression data which were not made by original authors. These include a correction for non-specific background and print-tips (Huber et al., 2002), as well as correction for multiple-testing which was not made by false-discovery rate estimates (Reiner et al., 2003). TF target calls made by Reimand et al. (2010) were downloaded from the ArrayExpress database (Parkinson et al., 2009). Genes called as targets for a TF have a highly significant expression difference between the knock-out and the wild-type, with a 0.05 p-value cutoff. The problem of possible indirect targets being included amongst the predicted target genes is however not directly addressed by Reimand et al. (2010).

5.2.6.3 Harbison *et al.* (2004) ChIP-chip dataset

The Harbison et al. (2004) dataset of genomic occupancy of 203 TFs is a result of genome-wide location analysis by ChIP-chip (Iyer et al., 2001; Lieb et al., 2001). They made measurements in a number of growth conditions (1 to 12 conditions, depending on the TF). I use a re-analysis of the Harbison et al. (2004) dataset by MacIsaac et al. (2006). This dataset contains lists of ORFs likely to be regulated by the TFs, based on conservation in other related yeasts, and a significance cutoff of the signals identified close to the ORFs in the ChIP-chip measurements. The analysis I present was made with the most stringent dataset provided by MacIsaac et al. (2006): ChIP-chip signal significance $p < 0.001$, with the binding site conserved in at least 2 other yeast species.

5.2.6.4 Relationship between discovered motifs and inter-species sequence conservation

The relationship between discovered motifs and sequence conservation were studied with 7-way phastCons conservation scores (Nielsen, 2005; Siepel et al., 2005) derived of an alignment of the *S. cerevisiae* genome with genomes of six other *Saccharomyces* species (*S. paradoxus*, *S. kudriavzeii*, *S. bayanus*, *S. castelli*, and *S. kluyveri*). The phastCons scores were retrieved from the UCSC Genome Browser FTP server (sacCer2 conservation track, available at `ftp://hgdownload.cse.ucsc.edu/goldenPath/sacCer1/phastCons/`, downloaded on 12/02/2010).

The conservation scores of motif match positions at the stringent confidence cutoff of 0.01 were contrasted with phastCons scores of 10,000 randomly sampled intergenic regions of the same lengths (10,000 regions were sampled at all lengths between 6 and 20 nucleotides). The random intergenic regions were sampled and retrieved from Ensembl (Hubbard et al., 2009) with the help of tools I wrote as part of the project. See Appendix B for usage examples for some of the tools included in the nmica-extra toolkit. The difference in conservation score distributions of the motif matches and random intergenic sequences were measured with the single-tailed two-sample Kolmogorov-Smirnov test.

5.2.7 Relationship between discovered motifs and sequence variation in *cerevisiae* strains

The *S. cerevisiae* reference genome was the first eukaryotic genome to be published (Goffeau et al., 1996; Mewes et al., 1997). Because the budding yeast is so amenable for genomic study and manipulation, and because its association to human activity and migration, its genetic variation in and between its different populations has also been studied. Large genetic studies began from typing microsatellites of over 600 *S. cerevisiae* strains (Legras et al., 2007). In this work I however use the more recent whole genome sequencing data from 42 *S. cerevisiae* strains conducted by the *Saccharomyces* genome resequencing project (SGRP) (Liti et al., 2009). This study presents the 1x to 4x coverage whole-genome capillary sequencing of the *S. cerevisiae* strains. Genotypes reported by Liti et al. (2009) for individual positions in the multiply aligned strains were imputed using

ancestral recombination graphs (Minichiello and Durbin, 2006) and the sequencing traces, instead of 'trusting' the base calls alone. On top of the low coverage sequence, the PALAS alignment method built for assembling and aligning the low coverage sequences is not a principled, probabilistic method with predictable properties, but instead an ad hoc iterative algorithm. The common occurrence of binding sites with large numbers of mismatches in aligned binding site matches suggested that alignment errors were prevalent (Edmund Duesbury, personal communication), especially between the *S. cerevisie* and *paradoxus* strains. Because of the limitations of the low coverage data and the SNP calls derived from it, I resorted to a simple comparative study between the SNP rates in binding sites when compared to intergenic sequence, with the aim of detecting motifs with likely function (those which show lower SNP rate than intergenic sequence). Only the *S. cerevisiae* strains were considered (no S. paradoxus strains), with two or less SNPs per regions of interest, as well as filtering out SNPs with less than 1×10^{-6} error probability. Putative TFBS matches with more than two SNPs were rejected because they are most likely caused by misalignments.

I applied a simple bootstrapping based statistical test to assess the significance of the difference of SNP rates seen in motif matches and random intergenic regions of the matching length. This was done for each predicted motif by counting the number of SNPs in a randomly chosen sub-selection of binding sites of the same length as the motif, and repeating this 10,000 times. The number of binding sites in each of the 10,000 random intergenic region sets was matched to the number of motif hits above the significance cutoff of 0.01. The significance score was derived as the fraction of the 10,000 sets where the mean SNP rate was higher than that observed for the motif's binding sites. Higher coverage Solexa based resequencing data, which (at the time of writing) is expected soon, could allow a more detailed analysis, for instance using the mutation spectra of motifs.

5.2.7.1 Positional bias of motifs

Regulatory motifs often match positions close to transcription start sites. Many cases of characteristic positional biases have been described for TFs, especially for elements bound by the general TFs, such as TATA-box (at around -30) or the

B-recognition element (BRE) which is found immediately upstream from TATA (Lagrange et al., 1998). An inverse linear association between the distance of the binding site to the TSS and its effect on gene expression has been suggested based on an *in vivo* study of factors acting in the liver and the immune system (MacIsaac et al., 2010). An earlier *in vitro* study of differently spaced Gal4 activator sites upstream to Gal4 also suggest a simple inverse relation between the distance of binding site to the transcription start site and its gene expression activating effect (Ross et al., 2000). I therefore analysed the positional bias of the computationally discovered motifs as an indicator of potential function.

I counted the motif matches in all matches overlapping 100-base windows between -1000 to 1000 from the TSS of all known protein-coding genes in the *S. cerevisiae* genome, and tested for the enrichment of sites within the region -500–0 with respect to the TSS, compared to sequence regions outside this window. I used the exact one tailed binomial test with the null hypothesis success probability of 0.25 (the interval -500 to 0 covers a quarter of the 2000 base sequence length of interest). The interval was chosen because it is expected to contain the great majority of S. cerevisiae TFBSs (Venters and Pugh, 2008).

5.2.8 Classification of motifs with metamatti

Metamotifs were constructed from the JASPAR 2010 motif dataset similarly as described in chapter 4: motifs were labelled with their structural class, and clustered at cutoff 4.0 (complete linkage clustering) using the SSD metric from Down et al. (2007). However, in this classification exercise I did not use the structural classification terminology from the TRANSFAC database, but instead the binding structural mode taxonomy introduced by Luscombe et al. (2000), which is included for majority of motifs in JASPAR 2010. The Luscombe et al. (2000) classification terminology describes 'classes' and 'families' for TFs. Classes are defined by a manual, visual comparison of protein structures, and families by a computational clustering of the domain structures with the SSAP secondary structure alignment algorithm (Orengo and Taylor, 1996).

The JASPAR database was used for building a *S. cerevisiae* motif classifier because it contains the largest selection of high quality training data for the *S.*

cerevisiae genome; The emphasis in TRANSFAC is on vertebrate genomes, and as of version 12.2 its non-redundant coverage of the *S. cerevisiae* genome is only 43 as opposed to 177 motifs in JASPAR 2010. As described in Section 1.1.1, eukaryotic genomes have experienced lineage specific expansion of TF domains. Therefore for an accurate organism specific TFBS motif classifier it important to have a good coverage of the domains that are present in that genome. For example in the case of *S. cerevisiae* the largest domain class is that of zinc coordinated domains, especially the fungal specific zinc cluster (Macpherson et al., 2006) (47 of 99 *S. cerevisiae* zinc finger motifs belong to this family, and very few are present in TRANSFAC).

Metamotifs were trained from each of the motif clusters with `nmmetainfer` (minimum length 6, maximum length 15) and metamotif density features were then computed per training set motif as described in Section 4. Based on the classification labels and probabilities that the random forest classifier produces, I computed a precision-recall curve using the ROCR R package (Sing et al., 2005), and applied a probability cutoff to the classification decisions such to provide a high confidence labelling of motifs.

5.3 Results & Discussion

I apply eight motif inference tools in this work primarily as a genome scale performance benchmark. To my knowledge, these algorithms have not been judged before on problems involving the prediction of large motif collections from promoter sequence. The rationale in the assessment is simple: a well performing *de novo* motif discovery algorithm should find as many as possible motifs closely matching known TFBS motifs in the *S. cerevisiae* genome (Section 5.3.2).

5.3.1 Properties of inferred motifs

The motifs predicted by different computational methods were found to differ clearly by visual inspection. A selection of the top matches between the inferred motifs and motifs in the JASPAR database are shown in Figure 5.3. The closest matches identified vary considerably between different methods. The familial

patterns of motifs found by different methods is also apparent amongst the closest matches; MEME, in particular, shows clear preference towards discovering GC-rich fungal Zn cluster motifs, whereas SOMBRERO and NestedMICA show more variability amongst the closest matches.

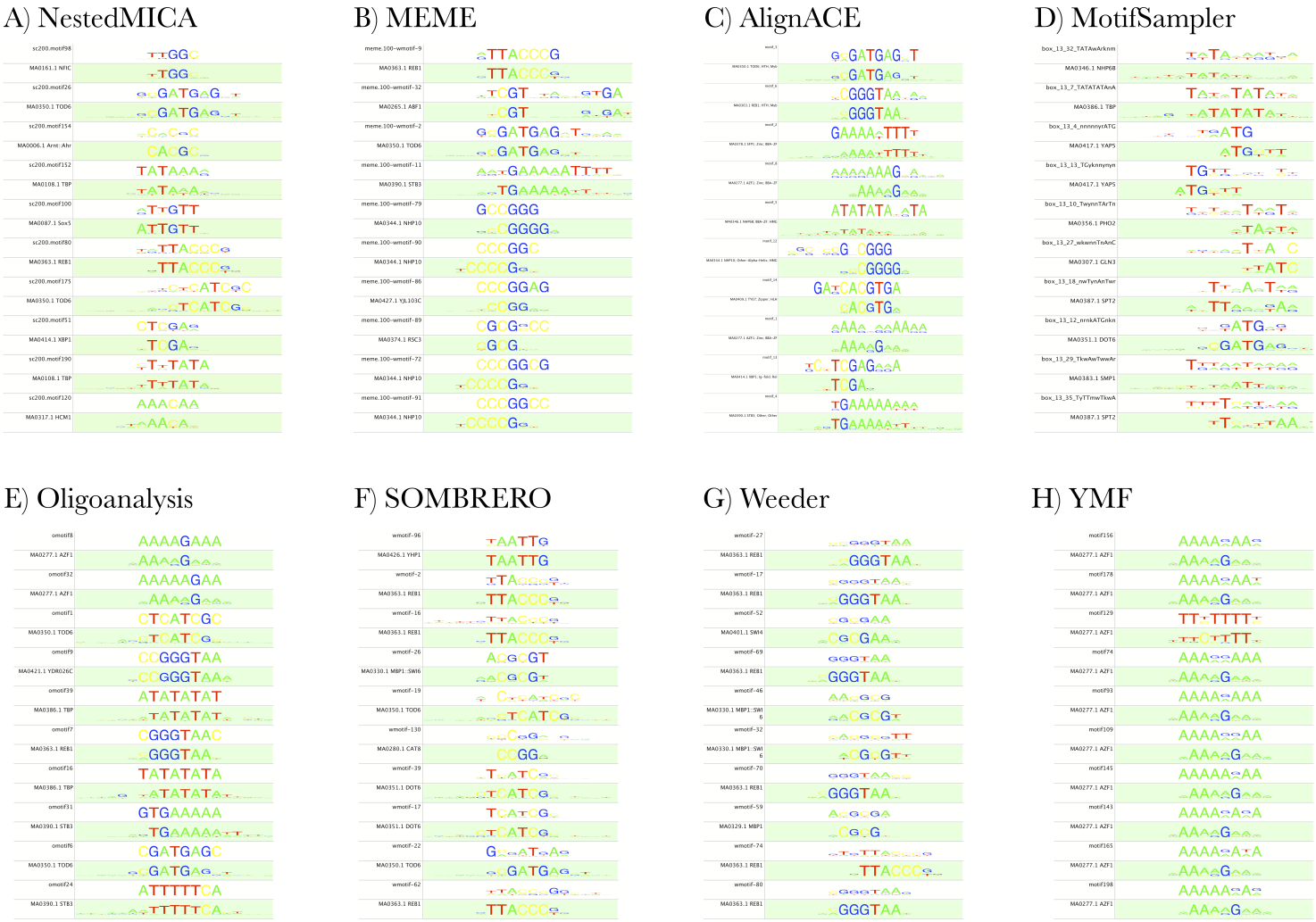


Figure 5.3: The ten closest matches between inferred motif sets, and JASPAR motifs. The JASPAR motifs are shown on green background.

The lengths, information contents and column wise average information contents are summarised for reference motifs and all inferred motif sets in Table 5.4. NestedMICA predicts the shortest motifs (6.6 columns), whereas Weeder has both the smallest information content (7.1 bits) and lowest per-column information content (0.9 bits per column). In contrast, MEME's motifs are almost twice as long as those of NestedMICA, at 12.6 columns, and they have the highest information content (over three times as high on average as motifs predicted by Weeder, at 21.7 bits). It should be noted that these motif set summary statistics and the relative performance measures reported in the following sections also depend on the chosen input parameters (Appendix C).

In terms of information content, the methods are divided to two groups: SOMBRERO, MotifSampler, NestedMICA and Weeder all predict motifs with smaller information content than their closest JASPAR matches, whereas AlignACE, Oligoanalysis, MEME and YMF have higher information content. The median per column information content is slightly higher with the JASPAR motifs with all but Weeder and MotifSampler. The combination of short motif lengths, with less information in total but with higher per-column information could be explained by the computational motifs lacking ends with low information columns, which are common in the experimentally verified motifs. The systematically low information content seen in the case of Weeder and MotifSampler is apparent already by visual inspection of the sequence logos: the columns tend to be less constrained than those in the reference set, or those output by the other methods.

Oligo-analysis and YMF results are included in this study for the sake of completeness: both are word enumeration based methods, and therefore not strictly comparable to the other methods which output a PWM, but they could be run also on my benchmarking dataset. Oligo-analysis motifs are in fact individual 8-mers (not IUPAC consensus strings, like those predicted by YMF). This inflates its information and per-column average information content measures shown in Table 5.4.

Motif set	Average length	Information content	Average column info content
NestedMICA (200 motifs)	6.6	9.5	1.5
AlignACE (16 motifs)	11.6	17.2	1.5
MEME (100 motifs)	12.6	21.7	1.7
MotifSampler (37 motifs)	10.0	10.0	1.0
Oligoanalysis (50 motifs)	8.0	16.0	2.0
SOMBRERO (200 motifs)	9.4	9.6	1.1
Weeder (200 motifs)	8.3	7.1	0.9
YMF (200 motifs)	8.6	14.2	1.6
JASPAR (177 motifs)	10.3	11.6	1.3
Zhu et al. (2009) PBM motifs (89 motifs)	9.6	11.7	1.3

Figure 5.4: Summary of the average lengths and information contents of the different inferred motifs, and the two reference datasets (JASPAR and Zhu et al. (2009) PBM motifs, shown on a grey background in the bottom).

5.3.2 Finding matches to known regulatory motifs amongst *de novo* motif discoveries

The number of JASPAR motifs with matches in each of the predicted motif sets ($p < 0.05$) are shown in Figure 5.5. Results appear to be rather consistent with two different reference databases (JASPAR in Figure 5.5A, and Zhu et al. (2009) PBM motifs in Figure 5.5B). The top performers, by a clear margin, are NestedMICA (54 matches to JASPAR amongst its 200 motifs, 44 matches with 100 motifs), MEME (39 matches) and SOMBRERO (38 matches). NestedMICA was tested with two different motif set sizes, in part to measure its robustness with differing motif count, and also to allow direct comparison with MEME which was incapable of predicting more than 100 motifs. AlignACE reports a mere 16 motifs, but surprisingly, these map to 31 JASPAR motifs; almost all of the motifs predicted by AlignACE are in fact contributing to the JASPAR matches (14 out 16 motifs). With the (Zhu et al., 2009) PBM motifs as a reference, NestedMICA is consistently the top performer, with SOMBRERO outperforming MEME.

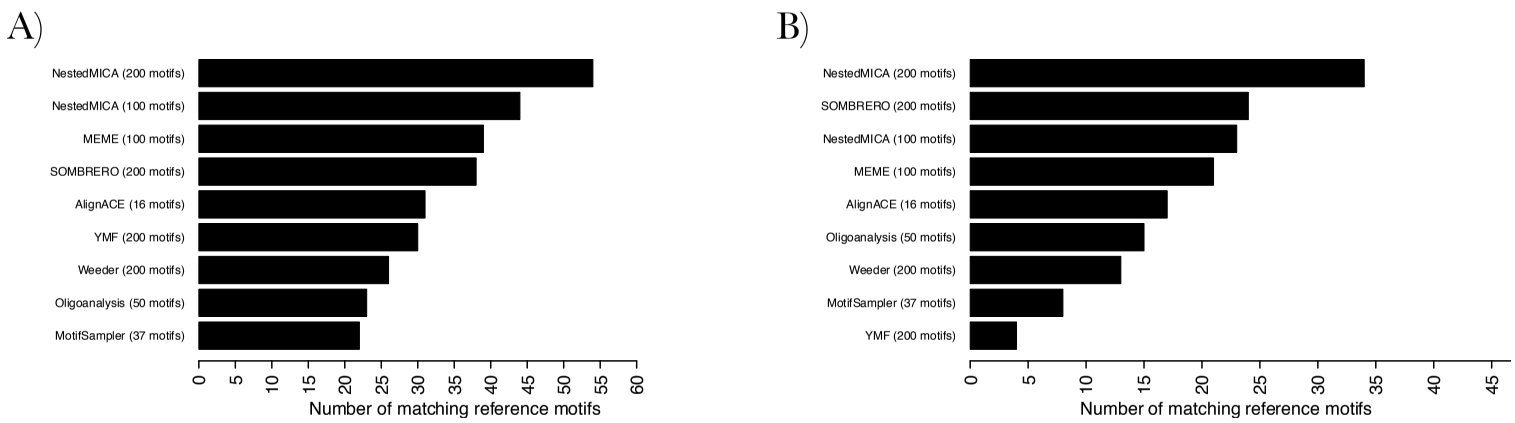


Figure 5.5: The number of statistically significant matches of the predicted motifs with A) JASPAR, and B) Zhu et al. (2009) PBM motifs.

Counting reciprocal matches between the predictions and the reference motifs is a more stringent way to assess motif relatedness (Figure 5.6). This measure penalises motif sets containing several closely related motifs. Some of the motifs amongst the reference motif sets are also highly similar to one another. NestedMICA also tops this ranking. With the JASPAR dataset of 177 motifs, it has 14 reciprocal matches, with SOMBRERO behind it, again with a clear margin (10 reciprocal matches) and MEME and AlignACE third (both with 6 reciprocal matches). Note again that the AlignACE program, which outputs a small motif set and has little redundancy in its predictions (Sections 5.3.4 and 5.3.5), is more likely to perform well by chance in this comparison than MEME with 100 motifs with several closely related motifs. Overall, the most likely reason for low numbers of reciprocal matches seen is due to the partial redundancy and large size of the experimental and inferred motif sets. NestedMICA however outperforms MEME and AlignACE also with a 100 motif count which matches that of MEME (9 reciprocal matches). There is little qualitative difference between the rankings with JASPAR or Zhu et al. (2009) PBM dataset as the reference.

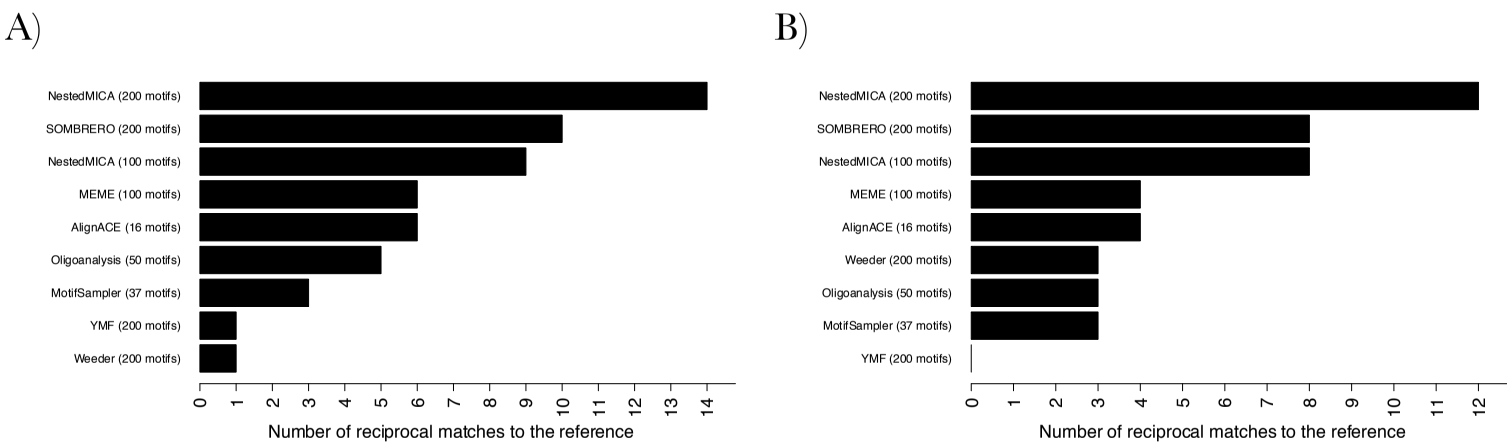


Figure 5.6: The number of reciprocal matches between the predicted motifs and A) JASPAR, and B) Zhu et al. (2009) PBM motifs.

The significant JASPAR and PBM motif matches suggest NestedMICA, SOMBRERO and MEME as the top performing methods. I also studied the overlap between the reference motifs covered by the different methods. I did this by computing the numbers of overlapping motifs between the top performers with the JASPAR motifs (Figure 5.7). NestedMICA has the highest overlap with the two other top performing methods (13 overlapped with SOMBRERO, and 9 with MEME). The number of motifs predicted by it and not covered by the other top performers (22 motifs) is also higher than either of MEME or SOMBRERO (14, and 9 motifs respectively), suggesting it covers more reference motifs than either of the other two top performers. Ten JASPAR motifs are found by all of SOMBRERO, NestedMICA, and MEME.

The number of statistically significant matches is informative of the extent to which the predictions cover the reference motif sets with detectably related motifs. The distribution of SSD distances between the inferred motifs, and their significant reference motif matches however also varies between algorithms (Figure 5.8). These results are consistent with above ranking in that NestedMICA also tends to have the shortest median distance, with SOMBRERO ranking the second. Once again the top performers are also consistent between the two different reference motif sets (JASPAR and the Zhu et al. (2009) PBM motifs).

The substantial disjunction of discoveries between the top-performing NestedMICA, MEME and SOMBRERO suggests that differences exist in the types of motifs that different algorithms are capable of finding. To study this further, I visualised the JASPAR dataset matches as a heatmap of matching or non-matching states, labelling the JASPAR motifs with its associated structural taxonomy of TFs, and clustering the motifs (Figure 5.9).

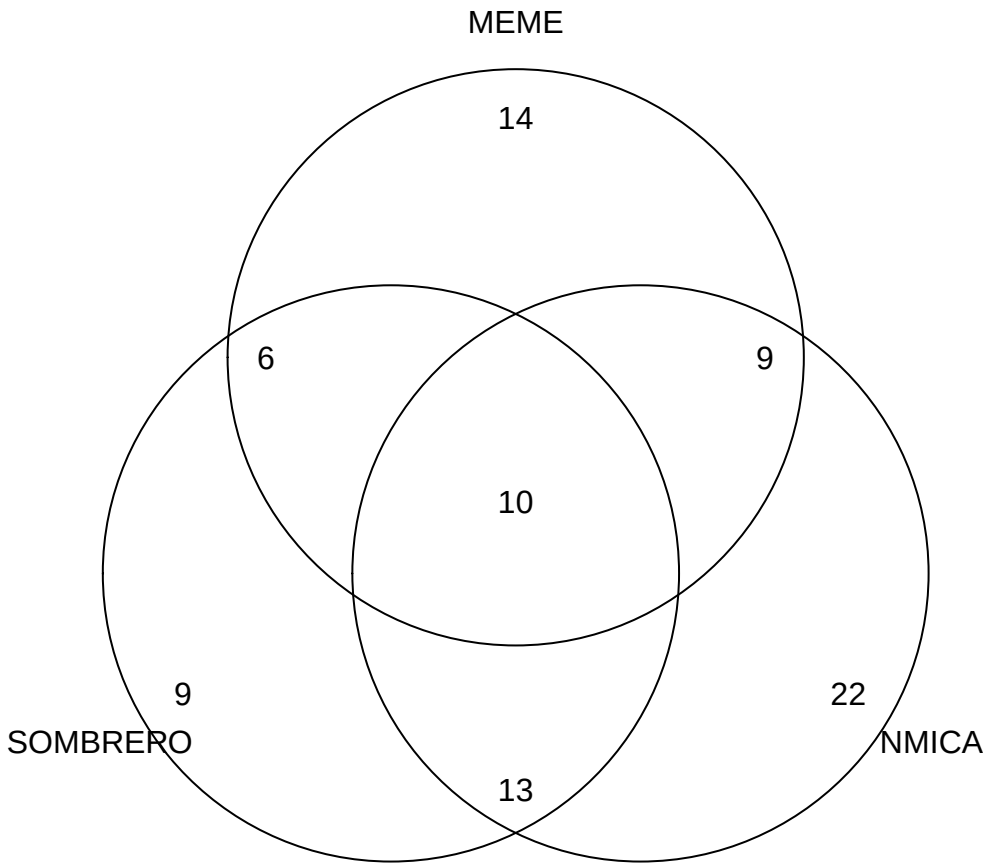


Figure 5.7: Overlap of significant matches to the JASPAR database between the three top performing motif prediction methods: NestedMICA, MEME and SOMBRERO.

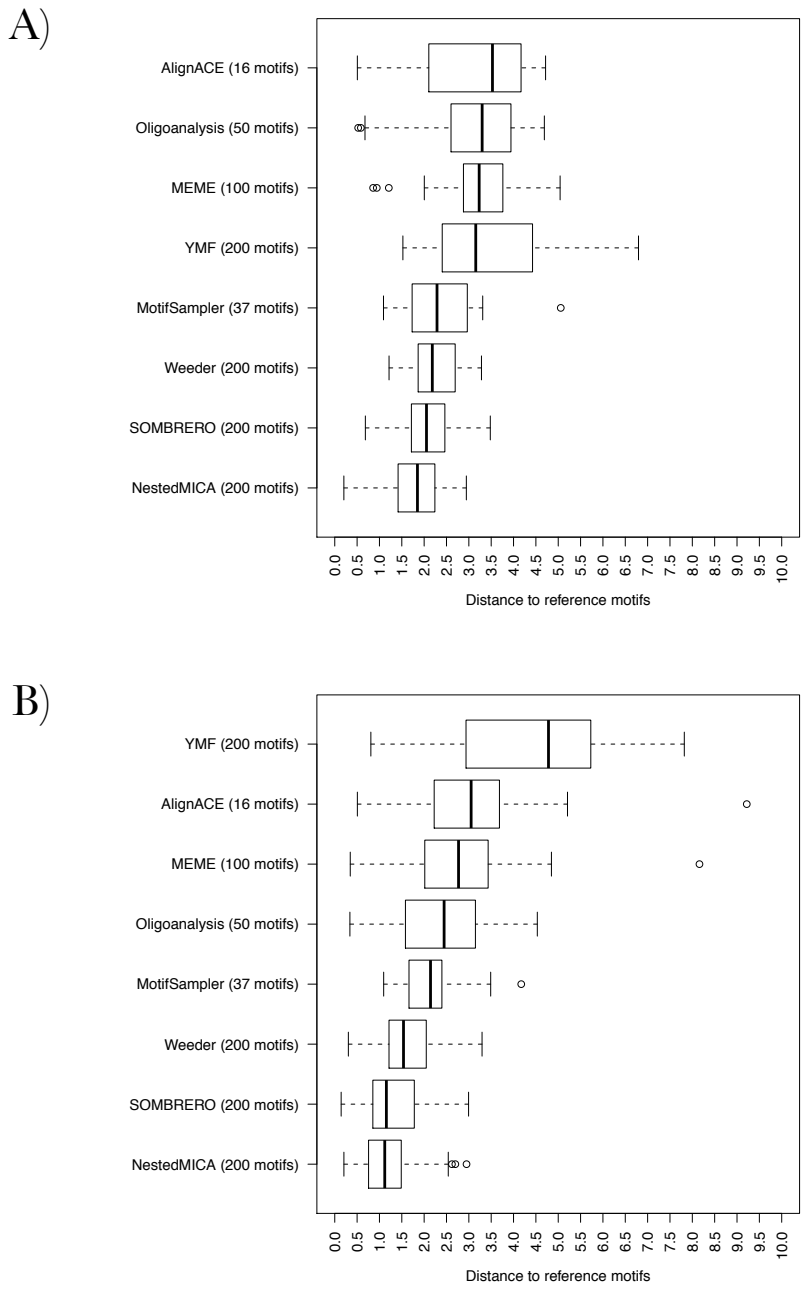


Figure 5.8: Distribution of SSD distances of predicted motifs to significant matches in the A) JASPAR and B) Zhu et al. (2009) PBM motif sets.

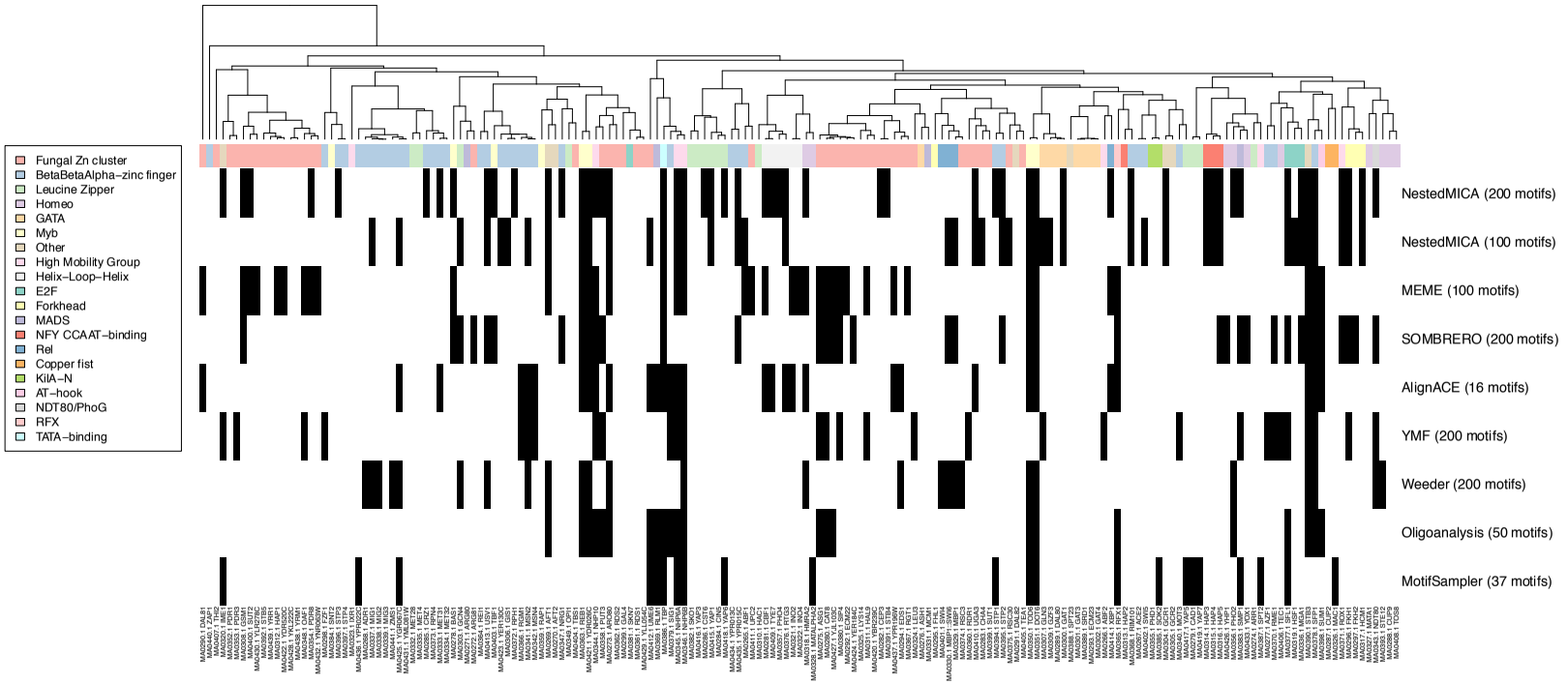


Figure 5.9: A heatmap showing the JASPAR motifs found or missed by each of the prediction methods ($p < 0.05$). Black cells in the matrix correspond to statistically significant matches ($p < 0.05$) between the JASPAR reference motifs (columns) and the computationally inferred motifs (rows). The reference motifs are ordered based on hierarchical clustering with the SSD distance. The inferred motif sets are ordered by their number of matches.

Some clustering of shared predictions by different computational methods is evident. Examples of JASPAR motifs predicted by different subsets of the methods are shown in Figure 5.10. Few clusters are covered by the majority of the algorithms, in fact only four such clusters appear. However, most JASPAR motifs in fact match by two or more methods, suggesting that consensus based predictions could perhaps be developed for more successful large scale motif inference, using combinations of different agreeing predictions. For example, SOMBRERO, and especially MEME, succeed with a large homogeneous cluster of 15 Zn cluster motifs (MEME identifies matches to 9, SOMBRERO to 5), to which NestedMICA predicts only two matches (CEP3, STB4). In contrast, NestedMICA shares motifs with SOMBRERO which match the FKH1 and FKH2 forkhead motifs, and the relatively closely related ROX1 motif, matches to which are not discovered by any of the other algorithms. All of the top performing methods also have motifs unique to them. Some examples of these motifs are also shown in Figure 5.10.

I studied the motif families predicted by the different methods, using the structural taxonomy provided by JASPAR. Some of these families, such as fungal Zn clusters, or $\beta\beta\alpha$-zinc fingers are present in high numbers in the yeast genome. I separated the JASPAR motifs to groups based on their structural family, and counted the numbers of matches to each of these families (Figure 5.11). Stratification of the matches by motif family provides another natural way of ranking the motif inference methods.

Most methods (MEME especially) appear to find several of the fungal Zn cluster motifs (the single most abundant TF domain family in the yeast (Wilson et al., 2008a)). The $\beta\beta\alpha$ zinc finger, Myb and HMG motifs are also covered with predictions by most methods. Substantial differences between methods do however exist. MEME, for instances, appears to be unable to find any instances of E2F, forkhead, MADS, or NFY CCAAT-binding domains, whereas it discovers motifs similar to the only AT-hook and RFX-like motifs present in the JASPAR motif set. NestedMICA and SOMBRERO find the most varied collection of motifs: 16 different structural families, whereas AlignACE only finds 12, and MEME 11.

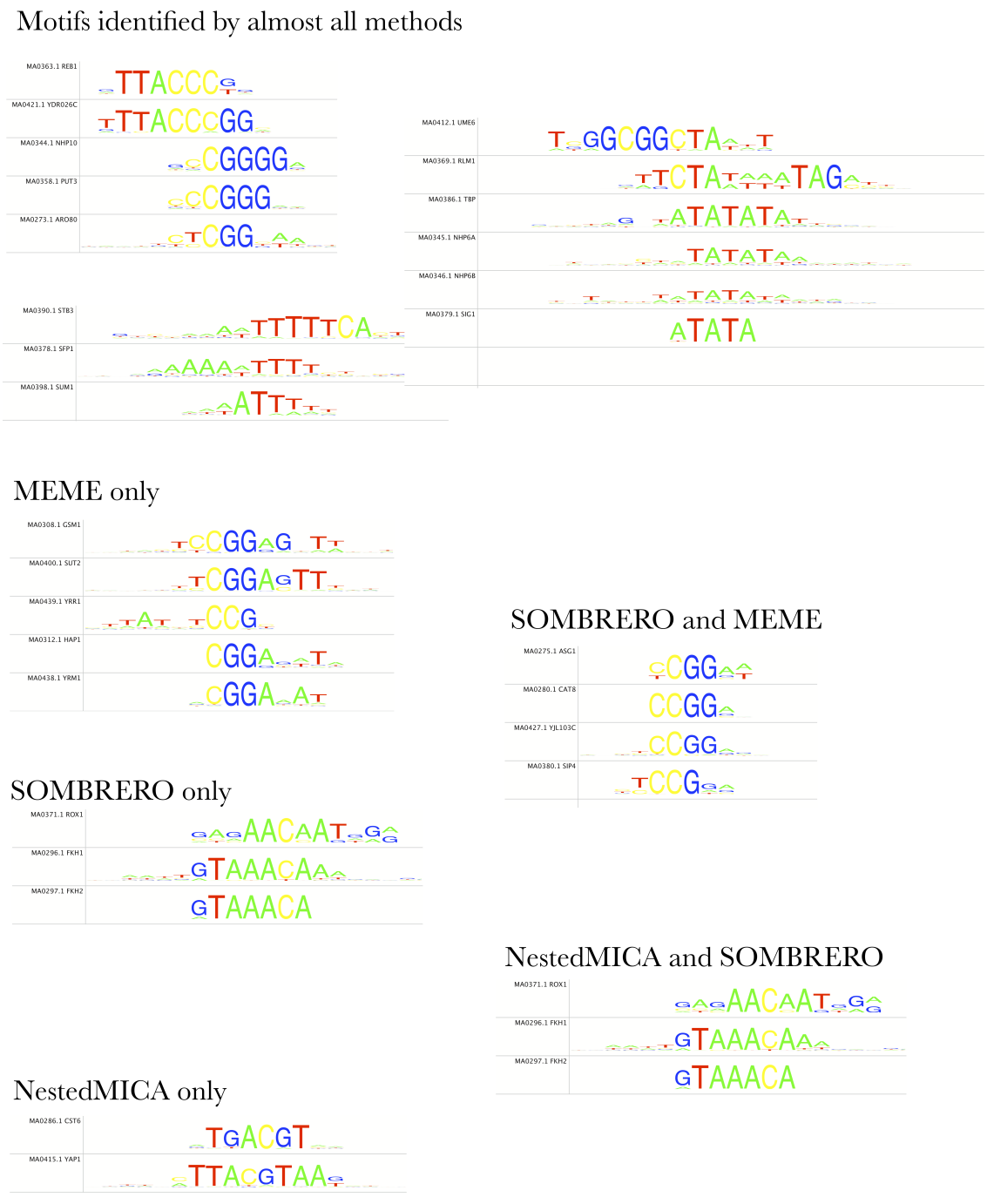


Figure 5.10: Different algorithms find matches to partially overlapping subsets of the JASPAR motif set. Example motif clusters found by different subsets of the algorithms are presented.

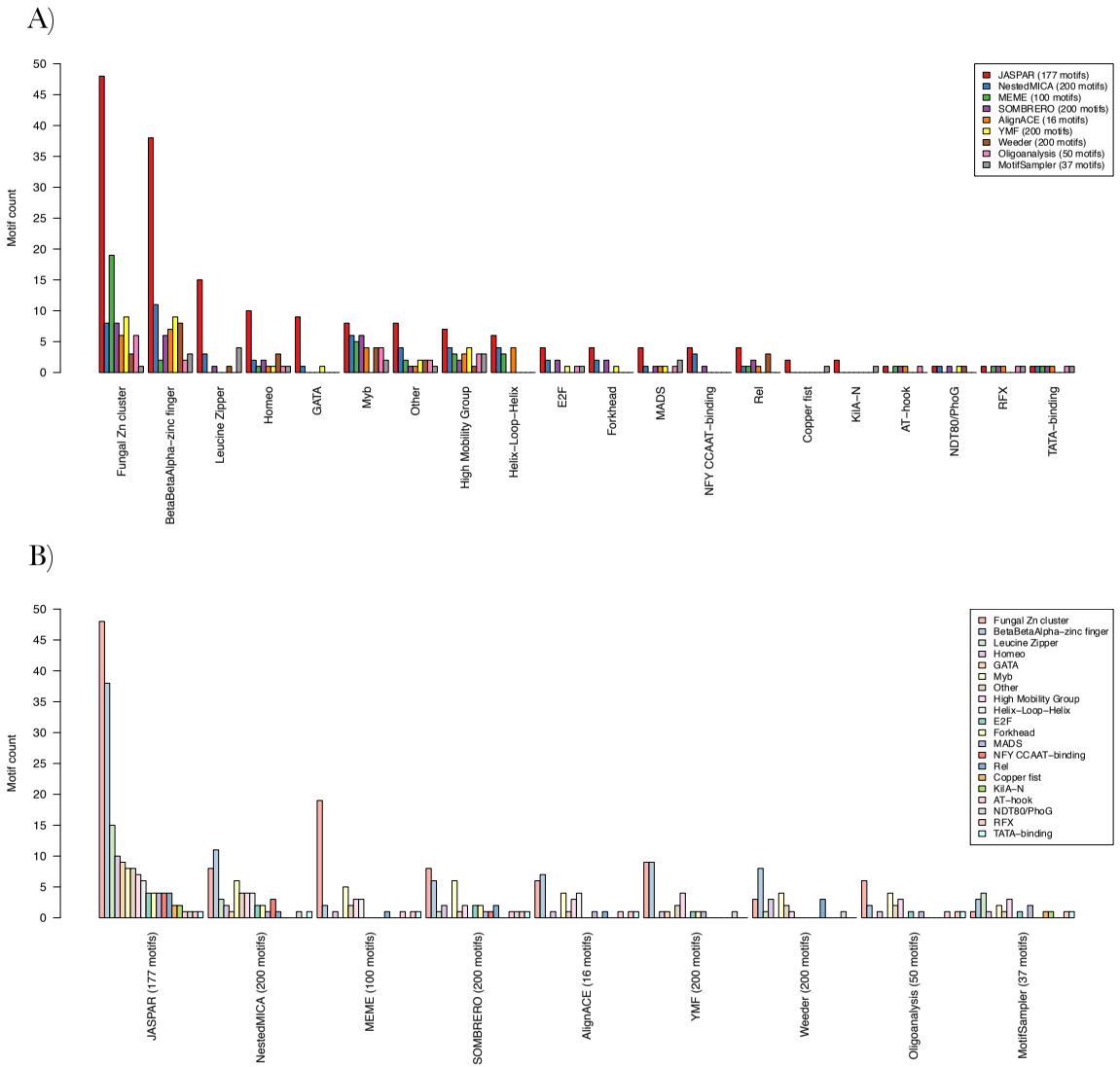


Figure 5.11: JASPAR motifs and computationally predicted motif, grouped according to their A) domain family and B) the motif set.

Figure 5.12 summarises the differences seen between the motif inferred by the eight different methods, and their closest, statistically significant reference motif matches. The properties shown are the motif lengths, information contents, and per-column information contents, similarly as shown above in Table 5.4. Once again, the analysis conducted with the JASPAR reference motif set is largely consistent with the Zhu et al. (2009) PBM motif set.

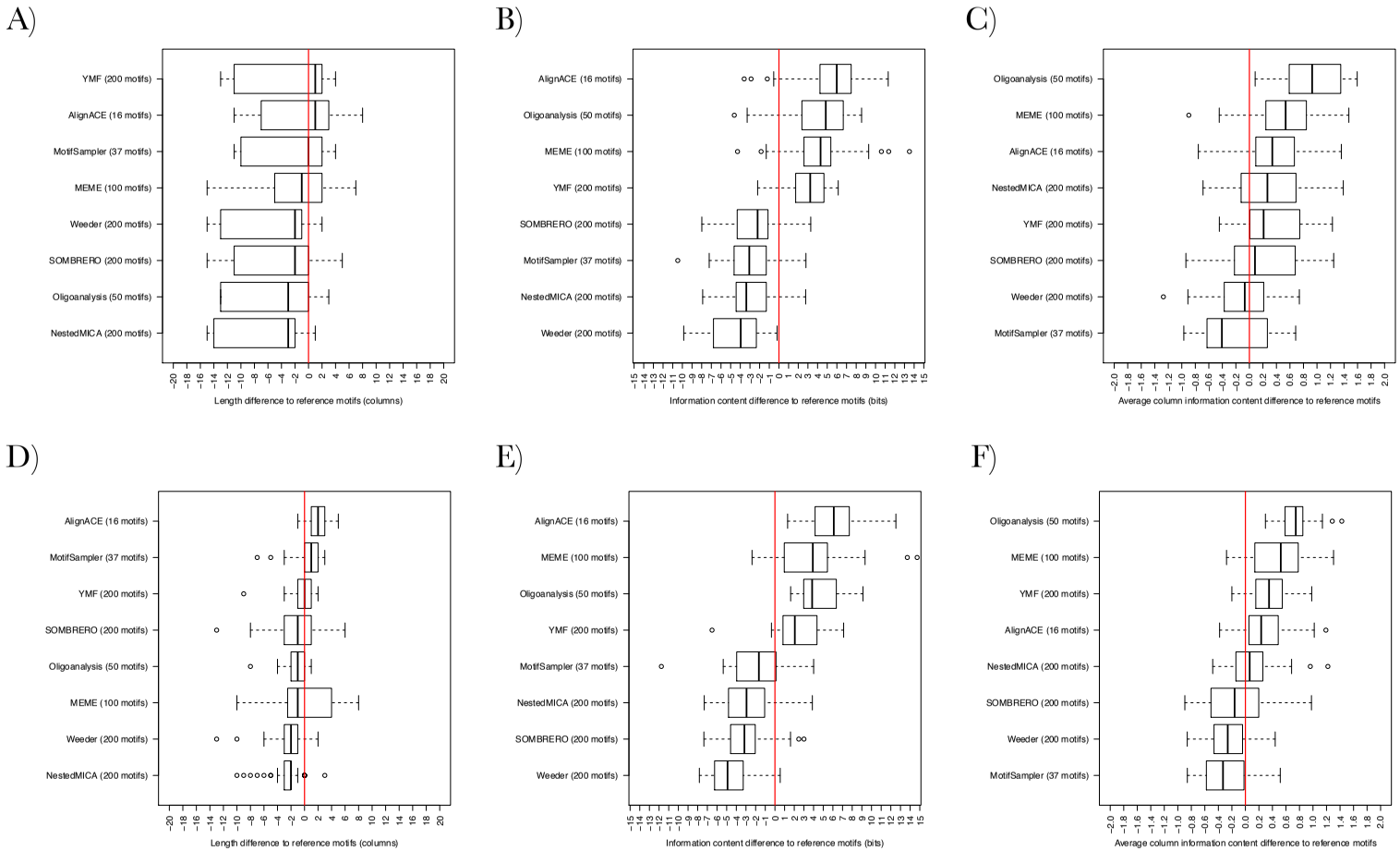


Figure 5.12: Differences in length, information content, and column-wise information content between the predicted and the JASPAR reference motifs. Panels A,B and C show comparisons of the predicted motifs with JASPAR motifs. Panels D,E and F are for comparisons of the predicted motifs with Zhu et al. (2009) PBM motifs. A & D: length difference; B & E: information content difference; C & F: average column-wise information content.

5.3.3 TF target gene associations of the discovered motifs

I tried to associate the genomic matches of inferred motifs with known target genes of TFs in the yeast genome (see Section 5.2.6 for details regarding the method). I did this with a parameter-free approach, assuming no significance threshold for the genomic matches of a motif. Each inferred motif was paired with its closest match in the non-redundant JASPAR database. With one exception (the MBP1:SWI6 complex), the 177 motifs in the JASPAR motif sets correspond to individual TFs, which in turn have associated target gene data available. The distribution of maximum bit scores are then compared with the non-targets to identify differences. Because there is no single authoritative source of TF–target gene pairings for the yeast genome, as discussed in Section 5.2.6, I therefore studied three alternative datasets. It is possible to rank methods based on the number of motifs identified by each, where a statistically significant difference is observed between the maximum bit score distribution of the target versus the non-target genes. Results with the three alternative datasets are shown in Figure 5.15.

An illustrative example of the maximum bit score distribution difference between target and non-target genes of a TF is shown in Figure 5.13, where motif58 from the NestedMICA 200 motif set is studied with the targets of the REB1 TF (the REB1 motif is the closest match to motif58). There is a highly significant difference between the maximum score distributions.

High scoring TFBS motifs are not expected to cleanly partition promoter sequences of the yeast to disjoint target and non-target gene sets. For instance, motif158 from NestedMICA's prediction set is found to be a close match to both the CBF1 and the PHO4 helix-loop-helix domain containing TFs (Figure 5.14). A statistically significant pattern is seen for the enrichment of motif158 with both CBF1 and PHO4. The DNA motifs of these two factors have been previously described as being closely similar, but they are known to act under different conditions and have partially different target gene sets; CBF1 acts under sulphur limitation, and PHO4 under phosphorus limitation Clements et al. (2007). High scoring motif matches of motif158 score highly for both of these only partially overlapping gene sets. One can therefore imagine that the motifs alone – espe-

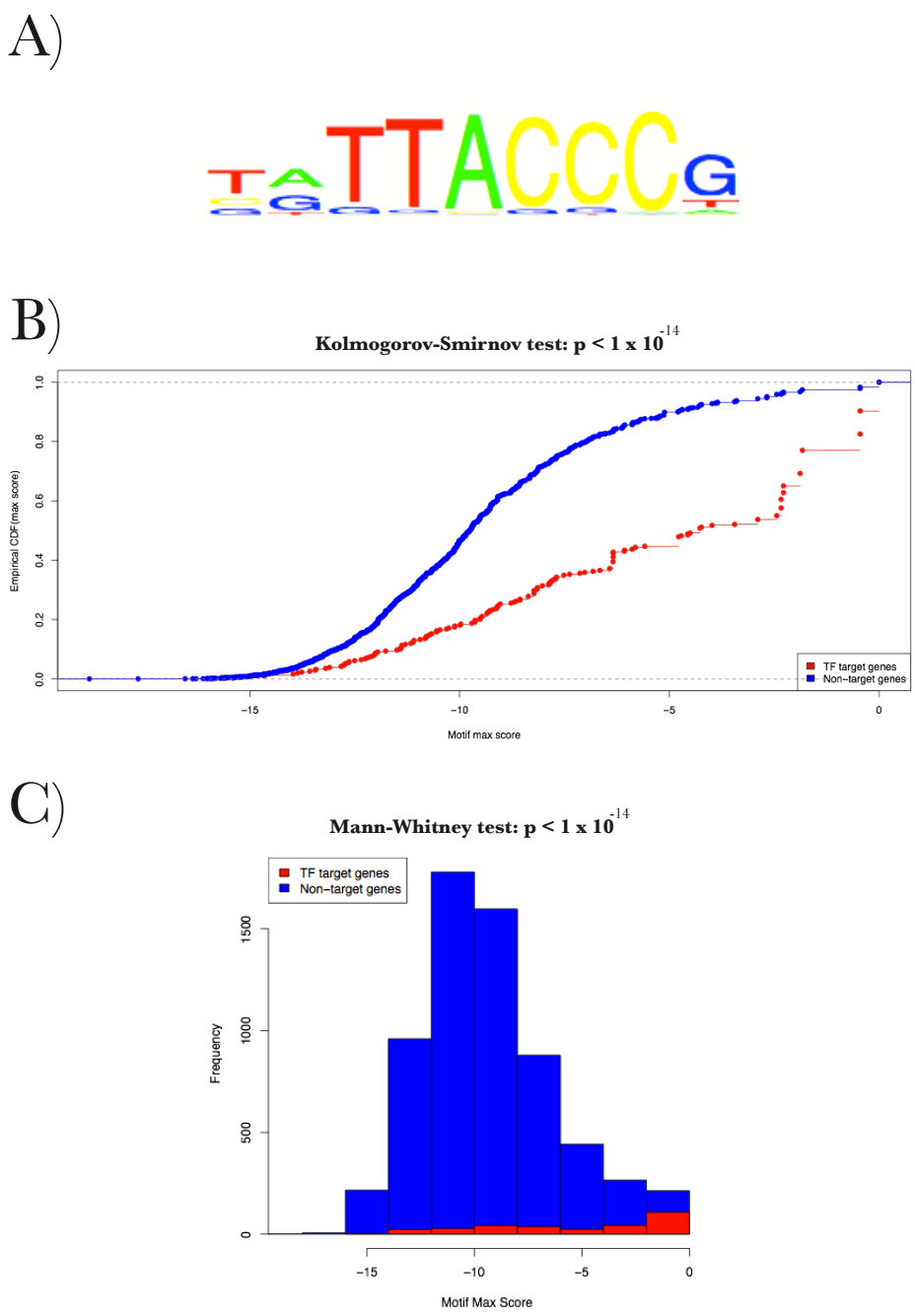


Figure 5.13: Some *de novo* inferred motifs are able to distinguish putative TF target genes from non-target genes by the maximum bit scores achieved by the gene promoter sequences (500bp upstream promoter sequences considered). A) Motif 83 predicted by NestedMICA is one such motif. B) The cumulative distribution of the maximum bit scores of non-targets (blue) and targets (red) as judged by the YEASTRACT database. C) A histogram of the bit score distributions of non-target promoter sequences (blue) and target sequences (red).

cially in the case of highly expanded TF families – do not have the discriminatory power to determine the target gene relationships of a TF (see Section 1.1 for a discussion on the various additional gene regulation mechanisms additional to TF binding).

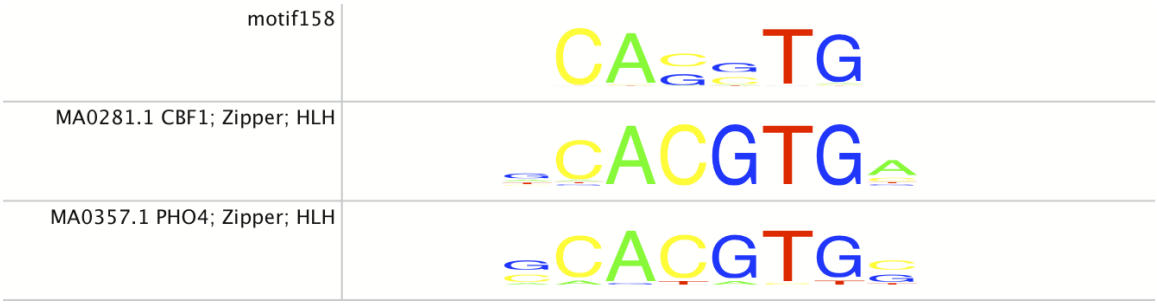


Figure 5.14: Motif158 is closely similar to both the CBF1 and PHO4 motifs.

Different stringency of calling genes either TF targets or non-targets can affect this analysis: if large number of TF targets are found in the non-target set, or vice versa, the separation between the target and non-target scores diminishes. This can be also caused by limitations in our knowledge of targets of some less studied TFs when compared to others, when dealing with hand-curated datasets. I considered three different TF target gene datasets in this study: a manually curated YEASTRACT dataset (Teixeira et al., 2006), the gene expression study based target set by Reimand et al. (2010), and the ChIP-chip data by Harbison et al. (2004). As a fourth set, I also attempted to retrieve the TF target predictions by Beyer et al. (2006), which are a result of integrating diverse lines of evidence into a probabilistic TF target prediction. Unfortunately however the dataset originally made available by the authors at `http://www.fli-leibniz.de/tsb/tfb` was not found anymore (authors were contacted). Several datasets were considered here because the coverage and confidence of TF–target associations included in each of them is not necessarily uniform across the TFs that each covers. The environmental states (e.g. growth conditions) covered by the datasets for instance are a factor: some factors bind their targets in an environment specific manner. According to Harbison et al. (2004), TFs fall into four groups with regards their target gene sets:

- Condition-invariant housekeeper TFs that bind target genes regardless of conditions. For instance Leu3, which regulates amino-acid biosynthesis (Kirkpatrick and Schimmel, 1995))
- Condition-enabled, for instance MSN2 which only enters nucleus to regulate target genes when the cell is under stress (Beck and Hall, 1999; Chi et al., 2001).
- Condition-expanded, which bind an expanded set of target genes under specific conditions. These include for instance Gcn4, which binds an expanded set of target genes under limited nutrients (Albrecht et al., 1998).
- Condition-altered, for instance Ste12 whose targets vary depending on condition-specific interaction partners (Zeitlinger et al., 2003).

Given the above categorisation of TFs by their ranges of target genes, one can imagine that there is variation between TFs in the power to detect a difference between promoters of target genes and non-target genes with high-scoring TFBS motif matches.

The largest number of TFs with a significant difference between the maximum bit score distributions of the target and non-target genes is seen consistently for all the algorithms with TF calls from the YEASTRACT dataset. This could be attributable for the manually curated YEASTRACT dataset being the most extensive and accurate resource of TF target calls, as it considers evidence from several sources. The ranking of motif inference algorithms relative to each other varies considerably depending on the source of TF target calls, with NestedMICA performing the best with the YEASTRACT and ChIP-chip based TF target calls, both in the case of the Kolmogov-Smirnov and the Mann-Whitney tests. AlignACE, with its mere 16 predicted motifs, also performs also remarkably well with this metric, outperforming all of MEME, SOMBRERO and the 100 motif NestedMICA prediction with the YEASTRACT dataset (Figure 5.15A). With the Reimand et al. (2010) expression based target calls, AlignACE outperforms NestedMICA with eight TFs ($p < 0.05$), with NestedMICA identifying only six differences at the same significance level. AlignACE and NestedMICA share the top rank with the Mann-Whitney test at this same significance level. Interestingly

though the reference JASPAR motifs identify a significant difference for only two more TFs than AlignACE, with this same dataset and statistical test. One feasible interpretation for this general failure of a motif match based approach to identify differences between the two populations of promoters with the Reimand et al. (2010) TF target calls is that the target list contains indirect downstream targets of the actual TF (possible because the dataset is expression effect based).

As an alternative to studying the closest JASPAR matches, all motifs could have been tested 'blindly' against all TF target sets. This however would necessitate a considerably larger number of statistical tests and make correcting for multiple testing more difficult. Furthermore, combinatorial regulation by TFs could potentially lead to statistical associations being called between TFs and motifs that are unrelated in binding specificity, but which tend to co-occur in promoters with the real motif.

5.3.4 Clustering of motifs and their binding sites

Some closely related patterns are expected amongst *de novo* predicted TFBS motifs, due to the shared evolutionary history of TFs. However, when challenged to infer a collection of motifs from a large series of genomic sequence, a motif inference algorithm should ideally find a wide spectrum of motifs, instead of predicting large numbers of redundant copies of a small number of patterns. I therefore measured the relatedness of motifs, not only to the JASPAR reference motifs, but also to other predicted motifs. I did this in two different ways: firstly by computing distance matrices between motifs with the SSD motif distance metric (Down et al., 2007), and secondly with an genomic match overlap score (Section 5.3.5). To begin with, I studied motif relatedness in a visual, qualitative way by drawing dendrograms of all of the motif sets together with JASPAR motifs (Figure 5.16), and with each of the sets separately with JASPAR motifs (Figure 5.17).

The dendrogram of all predicted motif sets with JASPAR shows – similarly as the analysis presented in Section 5.3.2 – that overall there are few large clusters of experimentally validated motifs with no related predicted motifs from one of the inferred motif sets. Redundant clusters by some of the motif predictions

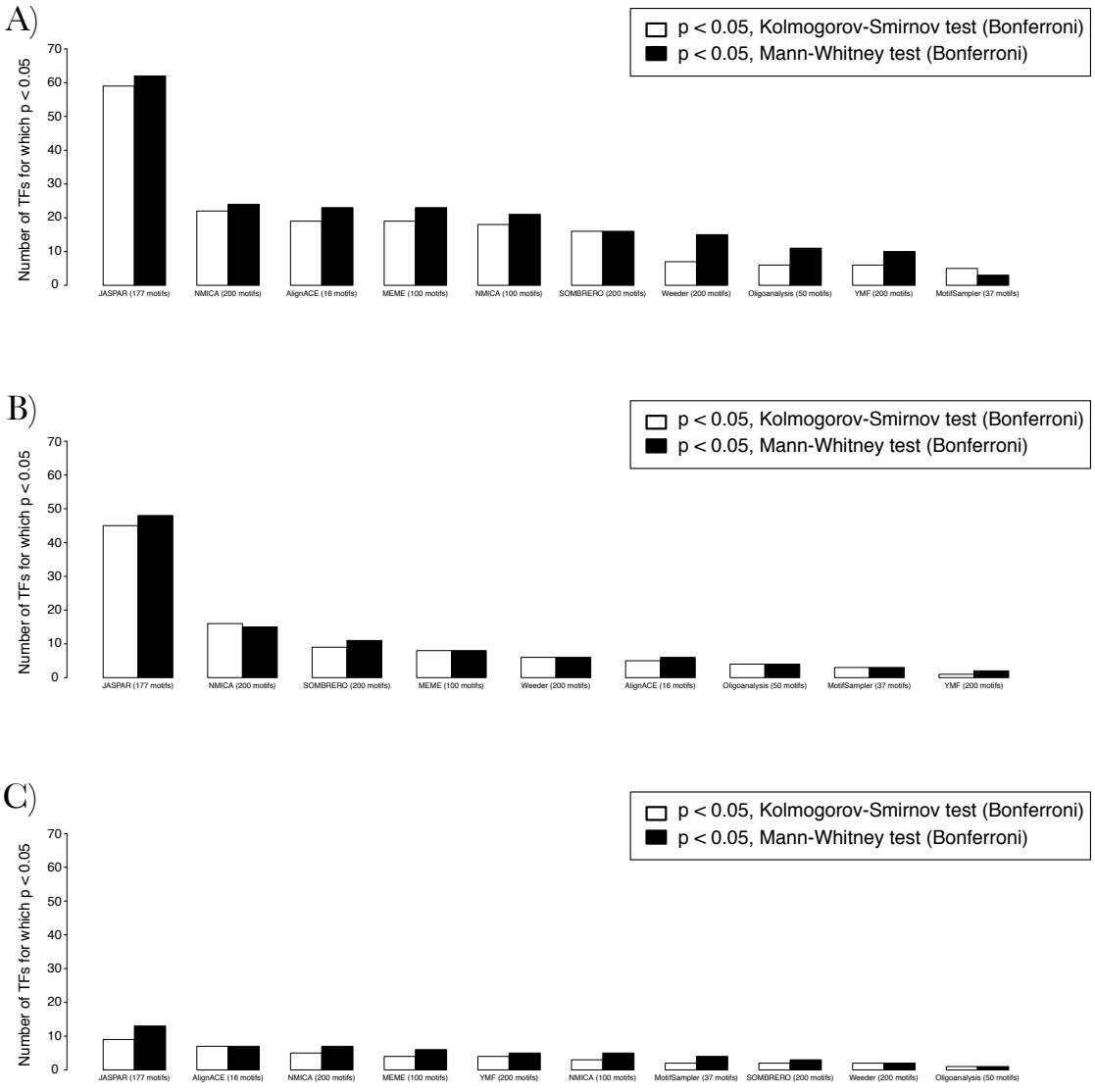


Figure 5.15: TF–target associations of the inferred motifs, when compared to JASPAR motifs (leftmost). The bars represent the number of TFs for which the computationally inferred motif shows a significantly different distribution of maximum bit scores, when target and non-target genes are compared. Motif sets are ordered by decreasing number of TFs with a significant effect. The p-values are Bonferroni corrected (divided by 176, which is the number of TFs tested).

are also apparent, especially in the case of YMF and Weeder. Conversely, the clustering pattern of NestedMICA and SOMBRERO motifs shows the predicted motifs much more 'intertwined' with the reference JASPAR motifs. Individual dendrograms are drawn in Figure 5.17 for each of the motif sets to make this pattern clearer to see.

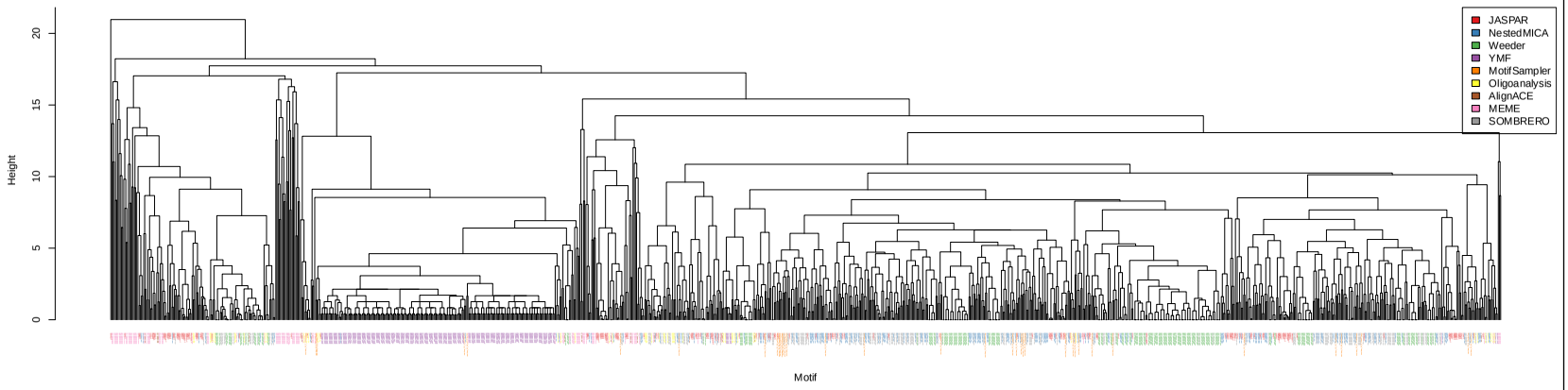


Figure 5.16: Dendrogram of a complete linkage clustering of all predicted motif sets with the JASPAR motifs, with the SSD metric from Down et al. (2007).

The motif clustering tree can be cut at different heights. I counted the numbers of cases where a JASPAR motif is clustered together with any of the other methods at varying heights. By this measure, SOMBRERO and especially NestedMICA perform favourably to the other algorithms (Figure 5.18).

Whereas Figure 5.18 measures inferred motif similarity to JASPAR motifs, the closest pairings of motifs within the predicted sets can also be studied using the distance matrix of the predicted motifs with each others (Figure 5.19). As one would already predict based on the motif dendrograms in Figures 5.16 and 5.17, YMF and Weeder predict considerably larger numbers of overlapping patterns than the other methods. At the 2.0 SSD distance cutoff for example, the average clique size of the motif distance matrix for YMF is above 40, compared to roughly 5 for Weeder, and between 2 and 1 for all of the other methods. Weeder and YMF appear essentially incapable of large scale motif inference as conducted in the present study, either due to my parameter choices for running the tool, or due to intrinsic problems with the algorithms.

The empirical significance values presented in Section 5.3.2 can be estimated for the closest pairs of motifs within each predicted sets, with the same protocol as used for comparing predicted motifs to reference motifs in Section 5.3.2. The 'uniqueness' of motifs varies considerably: almost all of Weeder motifs contain a statistically significant match, whereas MotifSampler and MEME have hardly any statistically significant matches regardless of the significance chosen. The JASPAR motif set also contains many motifs with close pairs; depending on the significance scores used, roughly 45% to 75% of JASPAR motifs have at least one match (Figure 5.20). This fraction is in fact higher for JASPAR than any of the other analysed methods but Weeder (the consensus based YMF and Oligoanalysis methods were omitted from the significance score analysis).

5.3.5 Comparing motifs by the overlap of their genomic matches

I measured the fraction of overlapping binding sites shared by motifs as a measure of motif similarity, complementary to the SSD distance matrices and motif clustering shown above in Section 5.3.4. I studied binding site overlap patterns

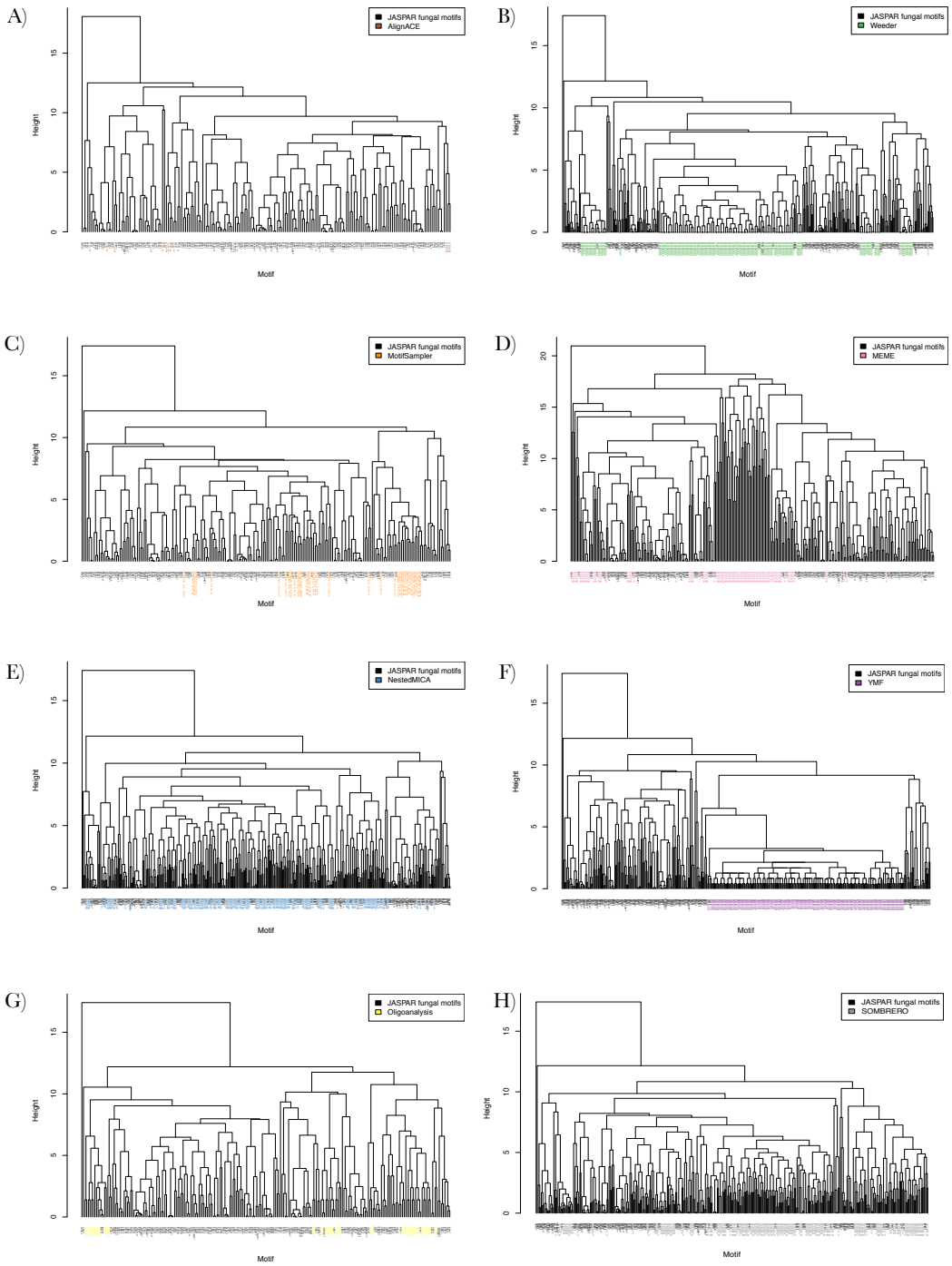


Figure 5.17: Clustering of JASPAR motifs with results of A) AlignACE, B) Weeder, C) MotifSampler, D) MEME, E) NestedMICA, F) YMF, G)Oligoanalysis H) SOMBRERO. The motif names are coloured according to the motif set where they originate from. They are shown as a quick visual summary of the clustering of the inferred motifs, rather than trying to present readable names.

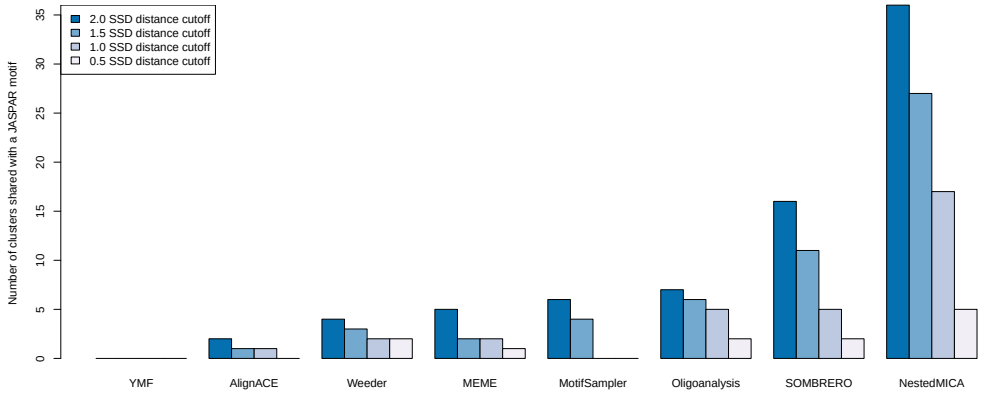


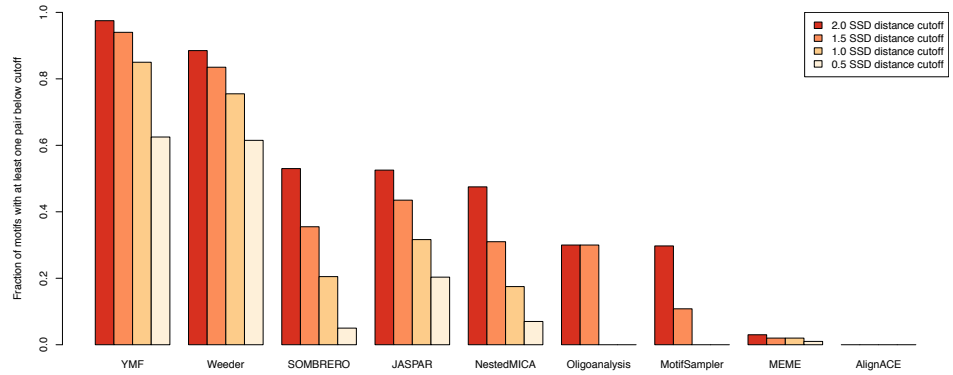
Figure 5.18: Numbers of clusters that contain at least one or more inferred, and one or more JASPAR motifs. Four different distance cutoffs are shown.

firstly visually, using dendrograms and heatmaps. The binding site overlap of two different inferred motif sets with the JASPAR reference motif set are shown in Figure 5.21.

Visual inspection of the heatmaps in Figure 5.21 suggests a higher overlap between NestedMICA and the JASPAR motifs, than between SOMBRERO and the JASPAR motifs. I quantified the binding site overlap by counting the numbers of motifs output by each of the eight methods, which overlap a JASPAR motif above a binding site score overlap. I repeated this analysis with five different overlap score cutoffs (Figure 5.22). The results are largely consistent with the clustering based motif similarity measures, suggesting NestedMICA is the method with the highest fraction of overlapping binding sites by this measure, followed by SOMBRERO, Weeder and MEME. Note that this similarity measure between the inferred and the reference motifs does not account for motif redundancy. This is the reason that Weeder for instance receive relatively high overlap scores with JASPAR motifs, when in fact its motifs map to a relatively small number of known TFBS motifs in the JASPAR set. The motif match significance score cutoff parameter, of both the reference and the inferred motifs, can also affect the results of this analysis.

Overlap of genomic matches between motifs can also be used as another means

A)



B)

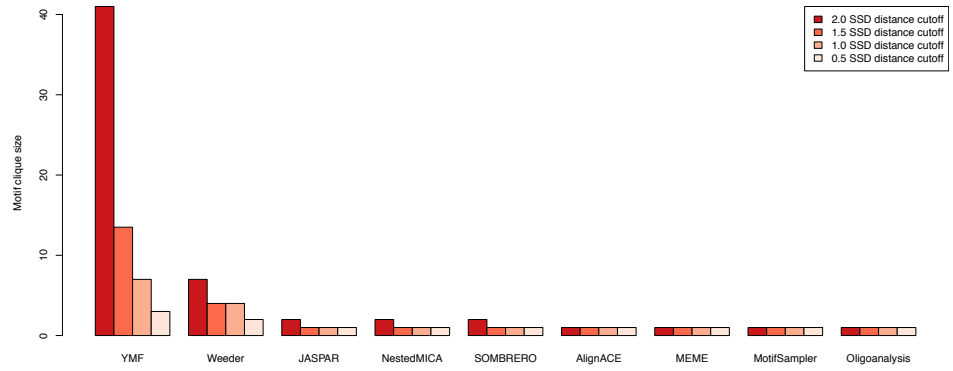


Figure 5.19: Motif redundancy as judged by the motif-to-motif SSD distance. A) Fraction of motifs which have at least one pair B) Average motif clique size.

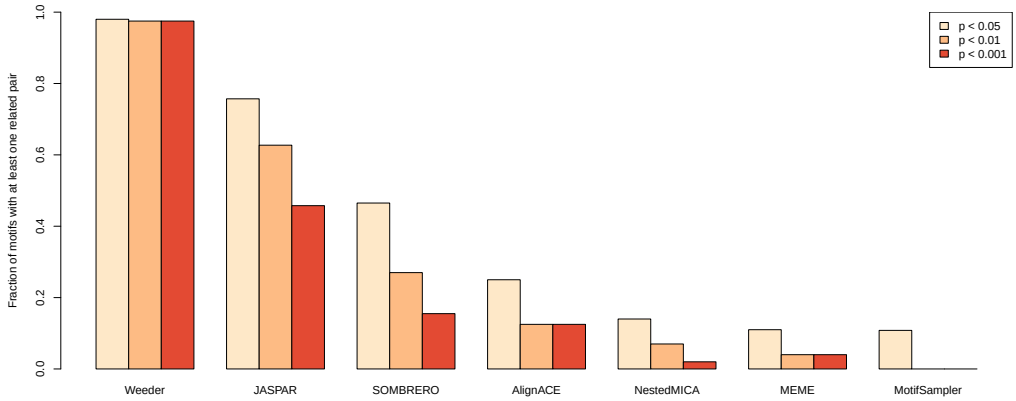


Figure 5.20: The fraction of motifs with at least one matching pair, at three different significance cutoffs. The consensus string based YMF and Oligoanalysis are omitted from this analysis, because the empirical significance score used here does not behave reliably for PWMs derived from IUPAC consensus strings.

of measuring motif similarity within sets. To illustrate this, Figure 5.23 shows the genomic match overlap of the SOMBRERO, Weeder and JASPAR motif sets. As discussed in Section 5.1.3, binding site level comparisons are not necessarily robust to the significance cutoffs used for genomic motif matches, and I do not advocate the use of these measures for ranking inference methods.

By this measure, Weeder receives the highest 'redundancy scores': for instance at the 10% overlap score cutoff, nearly all of the 200 weeder motif predictions have at least one motif pair which overlaps (Figure 5.24). The average number of motifs which all share a given fraction of their binding site matches (the motif clique size) however varies dramatically depending on the chosen binding site cutoff.

The present analysis of genomic match overlap between motifs is indeed a cautionary tale of assessing motifs based on their binding site overlaps: performance measures derived from genomic matches are not robust to the bit score significance cutoff chosen for a motif. This is an especially pressing concern for motif inference assessments such as Tompa et al. (2005), where experts applied many of these same algorithms, each with independently chosen motif match significance

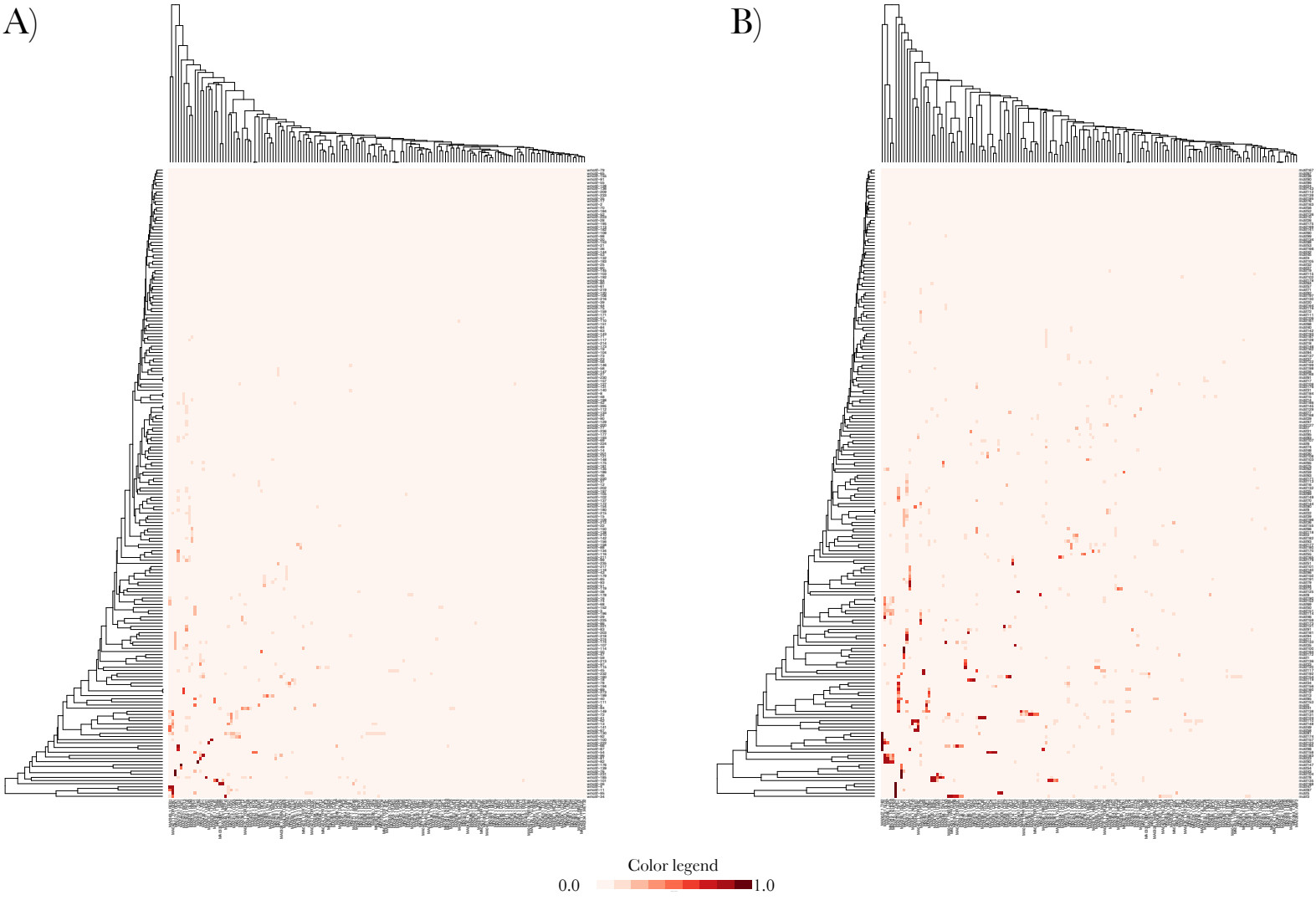


Figure 5.21: Motif binding site overlap of A) SOMBRERO and B) NestedMICA motifs. The rows represent inferred motifs, and the columns are JASPAR motifs. They are ordered based on an euclidian distance between the overlap patterns, with complete linkage clustering (Johnson, 1967).

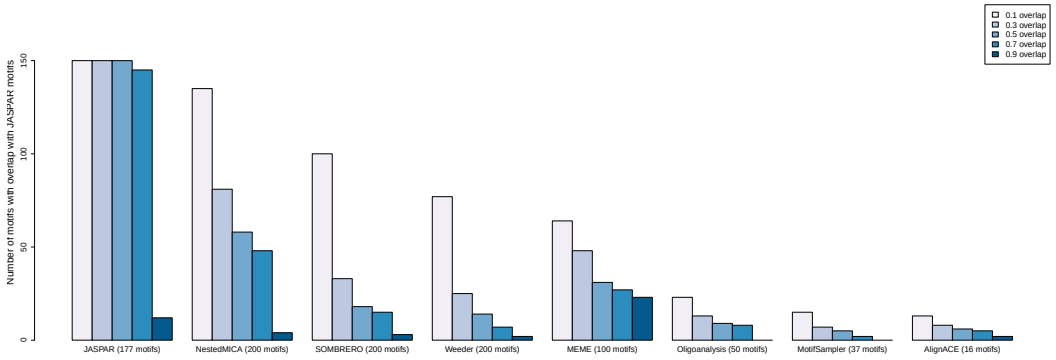


Figure 5.22: Predicted motif similarity to JASPAR motif set on the level of binding site overlap. The bars represent the numbers of motifs which show overlap above $0.10, 0.30, 0.70, 0.90$ to JASPAR motifs with the metric described in Section 5.2.5.

parameters.

5.3.6 Looking for evidence of function for the inferred motifs

On top of the 177 TFBS motifs included in JASPAR, the yeast genome contains others. The transcription factor database DBD (Wilson et al., 2008a) for instance contains 177 likely regulatory TFs for the genome, but its DNA binding domain model based predictions are estimated to cover only 2/3 of the genome Wilson et al. (2008a). The Harbison et al. (2004) ChIP-chip study on the other hand includes the binding profile of 203 putative regulatory TFs. It is therefore possible, even likely, that the promoters used in the study contain motifs for TFs which are not included in the 177 motifs of the JASPAR database. Therefore, I do not believe that all the apparent false positives (which do not match reference motifs) are false positives, and I wanted to identify a subset of particularly likely functional motifs from these unknown motifs.

I studied three different aspects of the computationally predicted motifs as signs of potential function: interspecies conservation (Section 5.3.6.1), SNP rate in yeast strains (Section 5.3.6.2), and positional bias of the motifs with respect

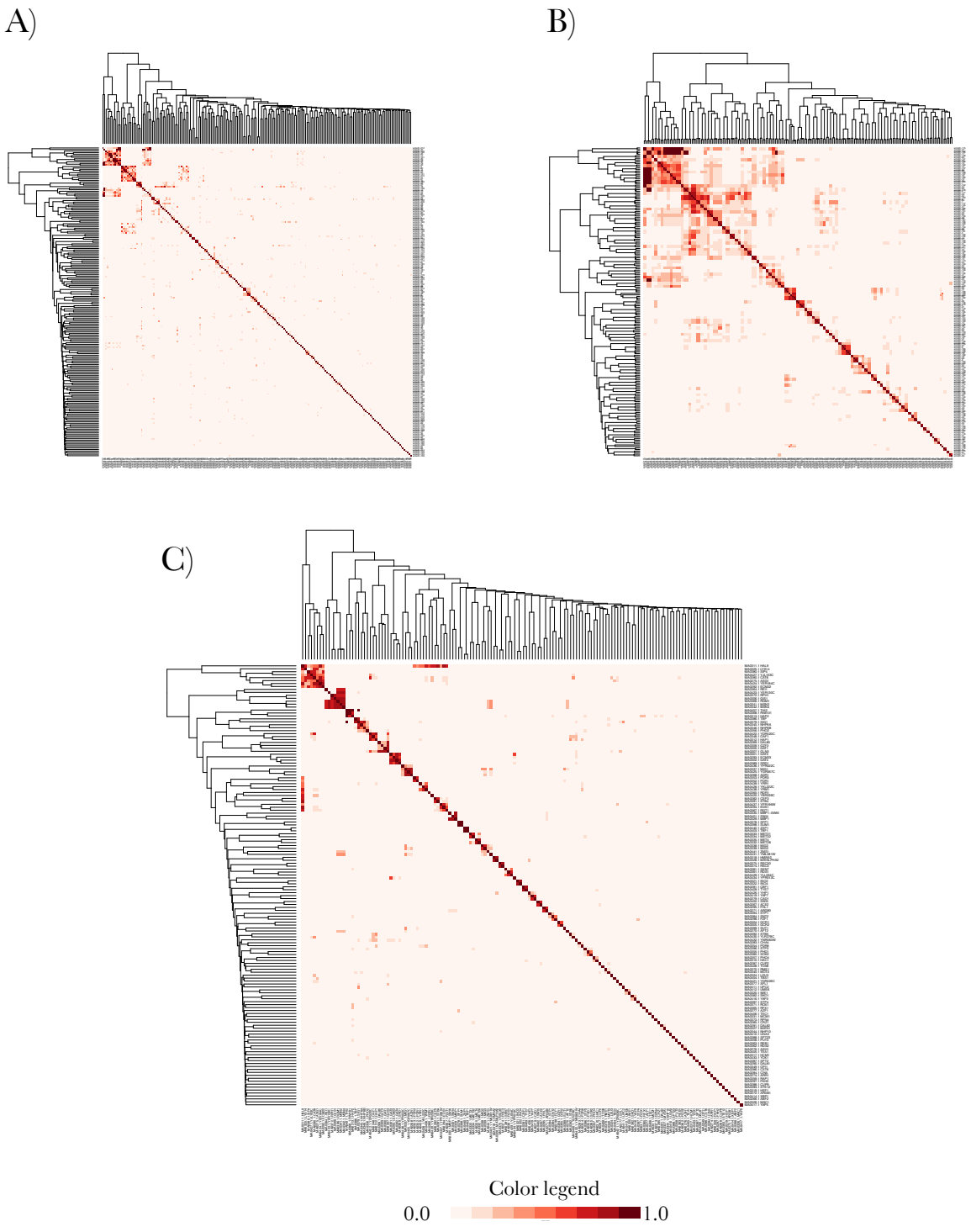


Figure 5.23: The overlap of genomic matches within motif sets. A) SOMBRERO and B) Weeder motifs are shown as examples of the predicted motif sets, and binding site overlap of JASPAR motifs are in panel C. SOMBRERO and Weeder differ in the degree of redundancy amongst the motif set. 500bp upstream sequences were analysed.

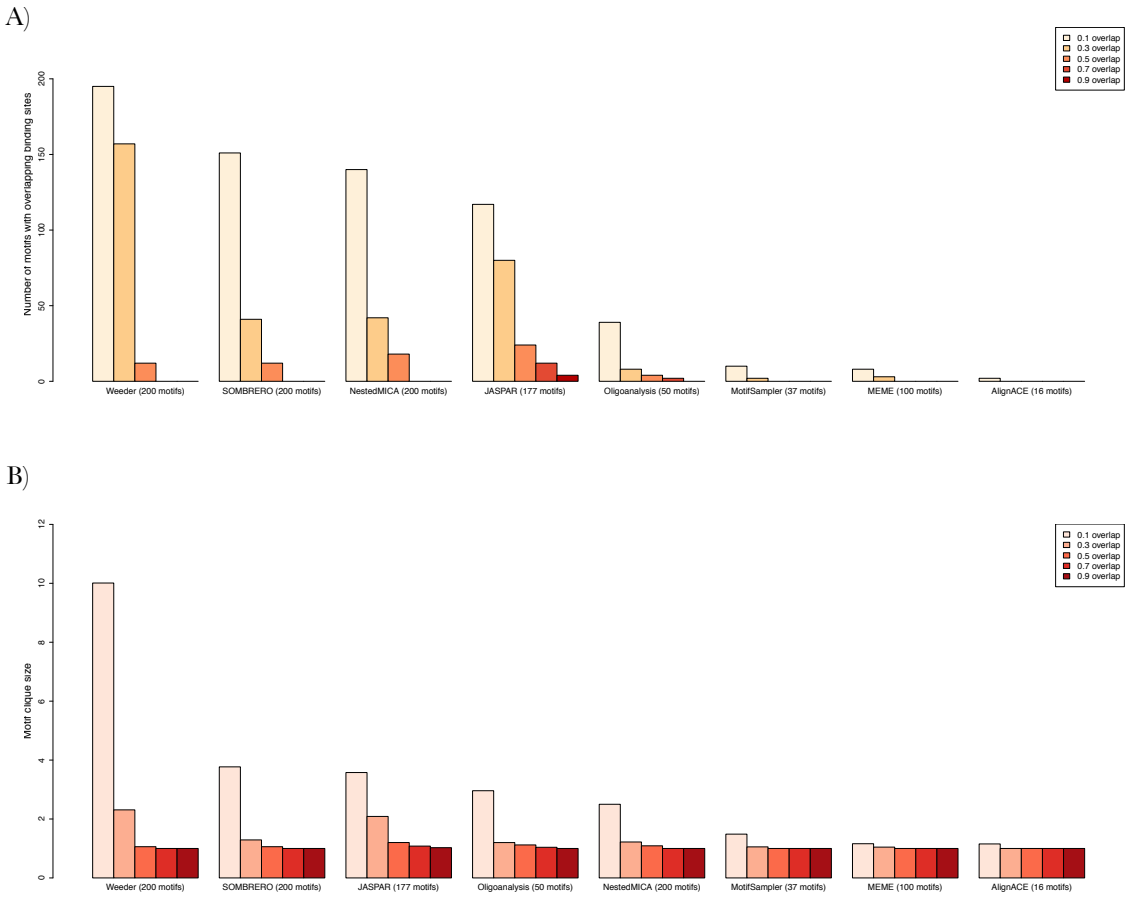


Figure 5.24: Predicted motif redundancy on the level of binding site overlap. The bars represent the numbers of motifs which show binding site overlap with the metric described in Section 5.2.5.

to the closest transcription start sites (Section 5.3.6.3). The motifs which fit all the criteria of high conservation, low SNP rate, and positional bias were then analysed in Section 5.3.6.4. Furthermore, I attempted to use the `metamatti` motif classification framework presented in Chapter 4 to predict the domain family of the motifs as a further sign of function (Section 5.3.6.5).

5.3.6.1 Inter-species conservation of the inferred motifs

The conservation scores for all of the 200 NestedMICA motifs at a 0.05 significance level are shown in Figure 5.25, as an example. A similar analysis was conducted also for all of the other methods (summarised in Figure 5.26). Figure 5.26 shows the fraction of motifs predicted by each method with a significantly higher conservation rate than random intergenic sequences of the same length. Note that for some of the methods, the fraction which matches known TFBS motifs in the JASPAR database (Section 5.3.2) is much smaller than the fraction which shows excess conservation. This could be explained by some of the predicted motifs being weak, undetected matches to real TFBS motifs, or artifacts of the multiple alignment based conservation PhastCons scores. Alternatively it could be that there are other potentially functional motifs within the motif predictions.

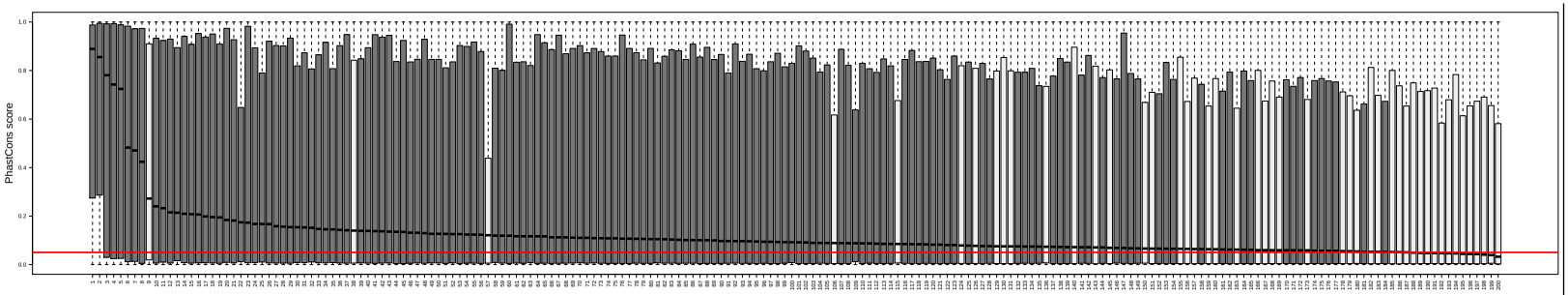


Figure 5.25: Conservation of motifs predicted by NestedMICA. The red horizontal line represents the mean PhastCons score for random intergenic sequence fragments of length 14 (the maximum length of motifs predicted by NestedMICA). The black markers are the median scores. The darker bars are motifs which are significantly more conserved than the intergenic sequence of the same length. The lighter bars are motifs for which this is not the case. Conservation of each motif was tested with a single-tailed two-sample Kolmogorov-Smirnov test between matches of each of the predicted motif sets, and random intergenic sequence positions of the matching length.

NestedMICA and Weeder show a roughly comparable fraction of significantly conserved motifs, between 60% and 80%, depending on the significance threshold which is varied between $p < 0.01$ and $p < 0.0001$. Overall the fraction of conserved motifs fits between 40% to 80% for all but two methods, which are overliers in the opposite ends of the scale; all of the YMF motifs show excess conservation, whereas only 8 of motifs inferred by MEME are significantly conserved. The results seen for YMF are in part explained by its highly redundant motif set, which shows variants of essentially one evidently highly conserved motif. The remarkably low figure of 8 motifs in the case of MEME is most likely due to its long motifs with high information content. This in combination with the stringent bit score cutoff determination method I used (Section 5.2.4) causes only a small number of hits to be reported and compared with the intergenic sequence regions, decreasing the sensitivity to detect differences between the distributions. An inspection of the median motif hit counts indeed shows alarmingly low figures for MEME's motifs at the 0.01 confidence threshold used: median motif hit count with the 200 base long upstream sequences is 2. This means that the significance score determination method used in the present study has largely failed with the motifs output by MEME. This, yet again, is an indication of problems associated with genomic hit based assessment of computationally inferred motifs.

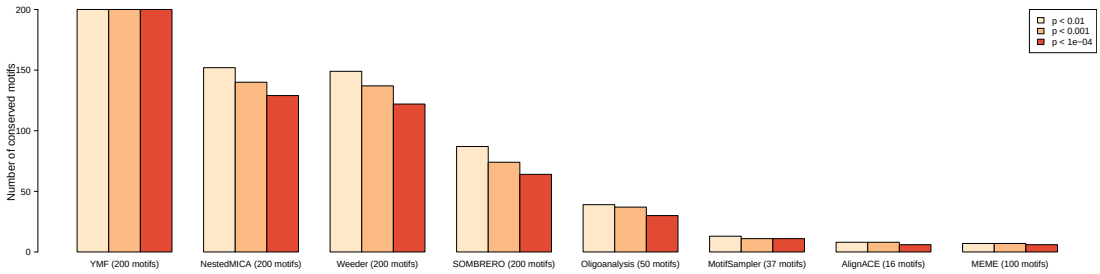


Figure 5.26: The number of motifs from each of the predicted motif sets that are found more conserved than intergenic sequence of the same length. Three different significance thresholds are shown. See Section 5.2.6.4 for details regarding the statistical testing.

5.3.6.2 SNP rates of the inferred motifs

A summary of the SNP rate analysis is shown in Figure 5.27. YMF and MEME are at the opposites of this scale, similarly as in the case of conservation patterns in Section 5.3.6.1. When compared with the inter-species conservation patterns, smaller fraction of motifs inferred by any of the methods show a significant difference to intergenic sequence. NestedMICA, SOMBRERO and Weeder identify the largest numbers of motifs with a significant difference. Similarly as in the case of interspecies conservation, the redundancy of motifs is not taken into account in the numbers reported, and they are not to be interpreted as a measure of the relative performance of the tools.

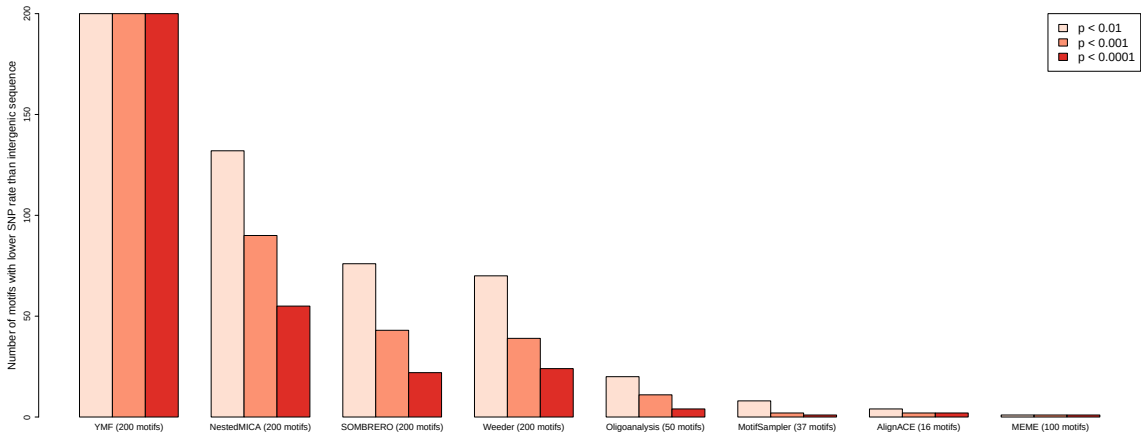


Figure 5.27: The number of motifs predicted by each of the methods with lower SNP rates than randomly selected intergenic sequence of the matching length. See Section for a description of the bootstrapping based significance scores.

5.3.6.3 Positional bias of motif matches close to the TSS

Many of the computationally inferred motifs were found to match preferentially upstream of the TSS. As examples of the typical positional bias trends which were seen, Figure 5.28 show the positional bias patterns in the case of SOMBRERO and Weeder. A summary of the positional bias trends of all of the methods are shown in 5.29, as the fraction of motifs with a statistically significant preference for positions -500 to 0. It is perhaps not surprising that a positional bias is seen

for many of the motifs, given that the motif search was made in the space of promoter sequences that span -200 to 0 from TSSs.

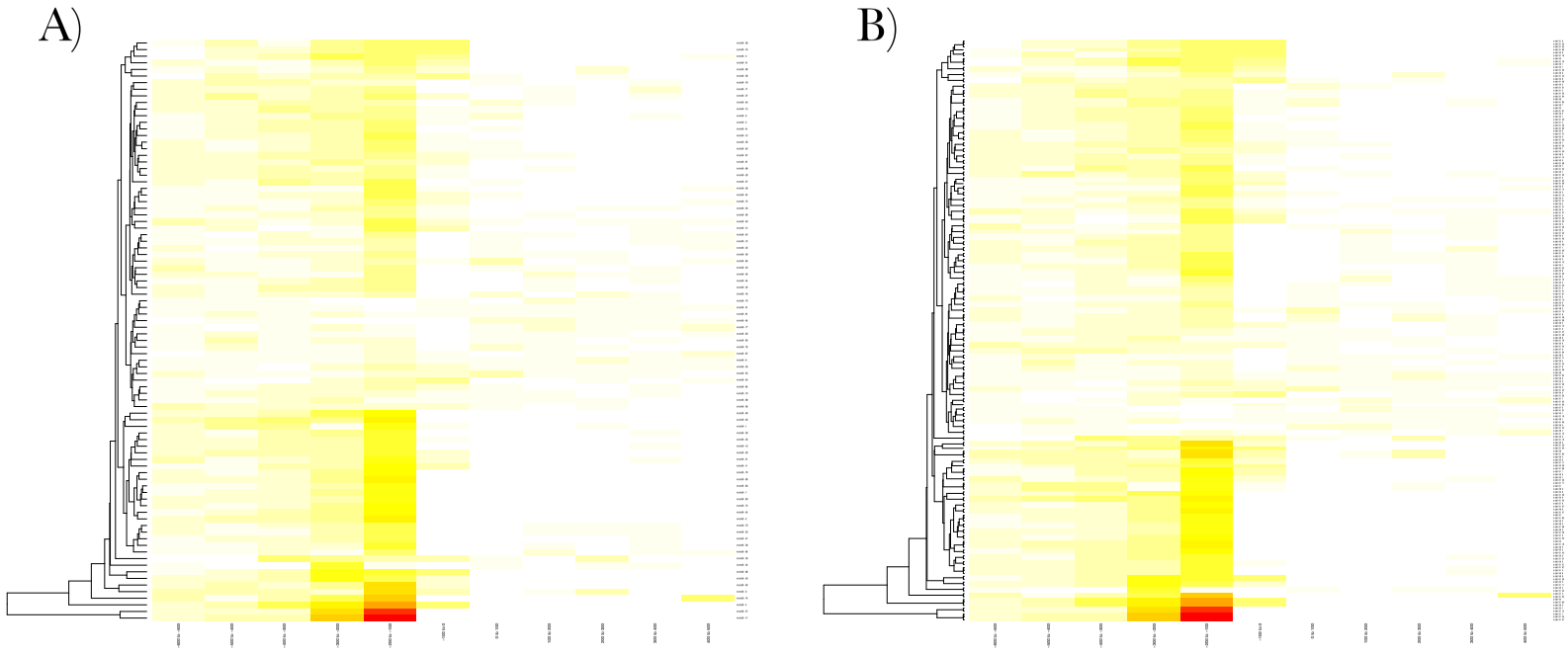


Figure 5.28: A heat map depiction of the positional bias trends of the motifs inferred with the A) SOMBRERO and B) Weeder algorithms. The columns in the heat map are 100 nucleotides long bins from -1000 to 1000, with respect to the TSS. Rows are individual motifs. Rows are ordered by a complete linkage clustering with an Euclidian distance of the relative frequencies at each bin.

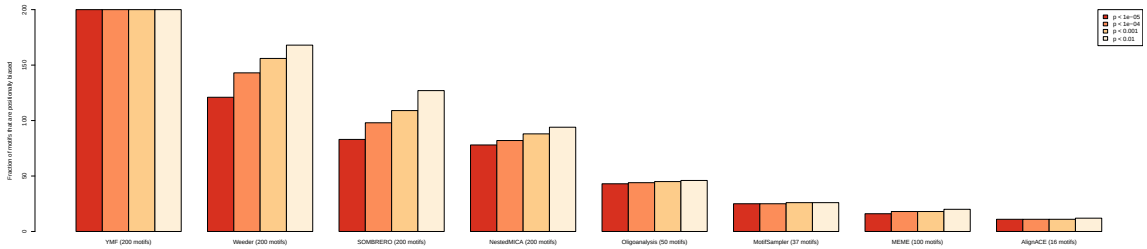


Figure 5.29: The fraction of motifs output by each of the eight methods, which show a preference for positions -500 to 0. See Section 5.2.7.1 for details regarding the method.

5.3.6.4 Combining the conservation, SNP rate and positional bias to highlight potentially functional motifs

I measured three aspects of the computationally predicted motifs as a sign of potential function: interspecies conservation, SNP rate, and positional bias of the motifs with respect to the closest transcription start sites. These properties do not appear to be randomly distributed amongst the motifs, with many motifs showing combinations of these features (Figure 5.30 shows SOMBRERO and NestedMICA motifs as an example). As also found by Down et al. (2007) in the *de novo* inference study of *D. melanogaster* regulatory motifs, a large fraction of motifs exhibit excess inter-species conservation, when compared to other intergenic sequence. The SNP rate and inter-species conservation are also closely associated, as expected.

I selected and counted motifs predicted by each of the methods which are not matches to JASPAR motifs, but show a combination of higher inter-species conservation than intergenic sequence ($p < 0.0001$), lower SNP rate than intergenic sequence in *S. cerevisiae* strains ($p < 0.0001$), and preferentially match close to the TSS ($p < 0.001$). Motifs which fit all of these criteria are shown in Figure 5.31. MEME, most likely because of its low total number of hits above the stringent bit score cutoff, did not find any such motifs.

NestedMICA found the largest number of unknown motifs of potential function (20), apart from YMF with its 182 highly redundant motifs with no sig-

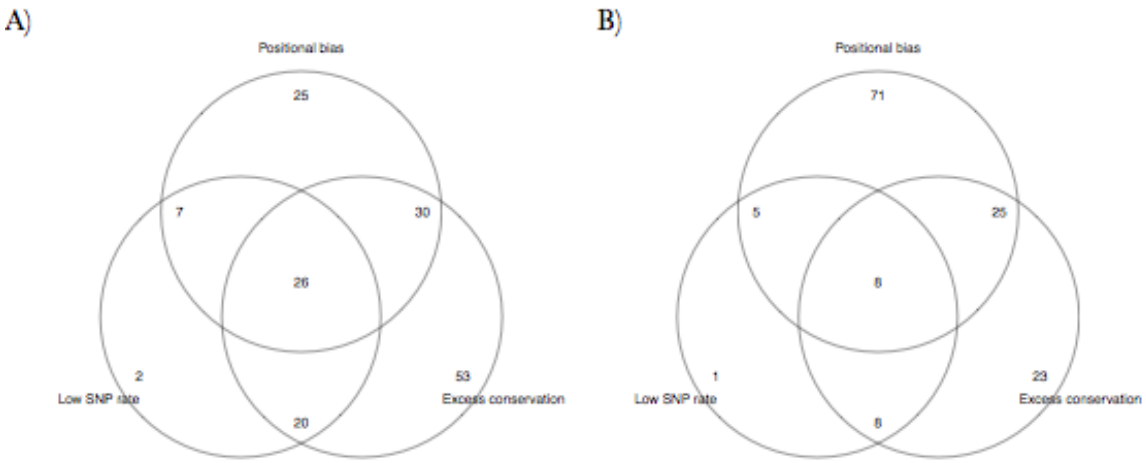


Figure 5.30: Overlap of motifs predicted by A) NestedMICA and B) SOMBRERO, that have lower SNP rate than intergenic sequence ($p < 0.0001$), higher conservation than intergenic sequence ($p < 0.0001$), and are preferential placed within -500 to 0 of TSS ($p < 0.001$).

nificant matches to known TFBS motifs (Figure 5.31G). I conducted literature searches to look for potential supporting information about the function of each of these motifs.

The TGAAAAATT motif (motif12 in the NestedMICA set, motif24 in the OligoAnalysis set) is perhaps the most interesting of the patterns. It is found by two previous *S. cerevisiae* motif inference studies (Li et al., 2005; Sudarsanam et al., 2002) to be associated with the TF ABF1. The ABF1 motif in the JASPAR database, derived from the high-throughput study by Badis et al. (2008), is however markedly different (Figure 5.32).

Other potentially functional motifs are also amongst the set. NestedMICA motifs motif152 and motif190 have the consensus TATAAAA and TATAAAG. Both of these sequences have been found to bind the TATA-binding protein (Kim and Burley, 1994; Starr and Hawley, 1991). The motifs both also show show a highly significant orientational bias. 60% of the 1864 hits of both motif152 and motif190 in 200bp upstream sequence regions appear as TATAAAA and TATAAAG – as opposed to TTTTATA and CTTTATA – on the same strand as the closest ORF ($p = 2.94 \times 10^{-17}$, binomial test).

Figure 5.31: Motifs predicted by different methods which have lower SNP rate than intergenic sequence ($p < 0.0001$), higher conservation than intergenic sequence ($p < 0.0001$), and preferential placement close to the TSS ($p < 0.001$). Motifs have been aligned with iMotifs (Piipari et al., 2010b).

Figure 5.32: The ABF1 motif in the JASPAR database. Data originates from the CSI, PBM and Dip-CHIP based study by Badis et al. (2008).

The NestedMICA motif33 (consensus NNTAAAC) matches the motif TAAAC, which has been suggested as the motif for the yeast TF 'Swi five factor', or SFF (Pic et al., 2000; Tamada et al., 2003). The 1252 instances of this motif in 200bp upstream sequence regions of the yeast genome show a highly significant bias in their orientation with respect to the closest ORF (56% of its instances are NNTAAAC, $p = 8.55 \times 10^{-5}$). SOMBRERO also finds a motif with a related, weaker consensus of ATAAAC.

Motif173 from the NestedMICA set has the consensus TAATAA. It has been described as a motif for the BAS2 homeobox TF (Rolfes et al., 1997; Tice-Baldwin et al., 1989). Interestingly, matches of this motif are also associated with the orientation of the closest gene (54% of its instances are TAATAA, $p = 4.00 \times 10^{-5}$).

The AAAGAAA motif (motif9 in NestedMICA's set, motifs 8 and 32 in the OligoAnalysis set) has been previously described in a phylogenetic foot printing study of the *S. cerevisiae* genome as a motif associated with genes involved in amino acid transport (Cliften et al., 2003). The reverse complement of motif motif138 (TTTGTT) corresponds to the consensus string of an HMG like TF domain (Grosschedl et al., 1994).

Several of the methods also find A- or T-rich motifs, such as AAAAAA, AAAAAAAAA, AAATAAA or AAATAA. Although I did find publications linking some of these sequence signals, or their reverse complements, to transcriptional control, it could also be that the high conservation and low SNP rate observed for these are artefacts caused by for example the genomic multiple sequence alignment procedures which both of the conservation and SNP rate criteria depend on.

In summary, the motif inference methods studied here find several putatively functional motifs not covered by the JASPAR motif set. NestedMICA – which

is consistently the top performer in the JASPAR based performance measures shown in Sections 5.3.2, 5.3.3 and 5.3.4 – finds a varied selection of 20 motifs with high conservation, low SNP rate and a preference for matching close upstream to the TSS, but with no known regulatory motif matches in the JASPAR database. Several of the other algorithms found different subsets of these 20 motifs identified by NestedMICA. SOMBRERO finds the second largest set of motifs which fit the criteria (6 motifs).

5.3.6.5 Classification of the inferred motifs with metamatti

I used the **metamatti** motif classification framework presented in Chapter 4 to predict the domain family of the motifs as another way of assigning function to them (see Section 5.2.8 for a description of the method), and comparing the motifs inferred by different methods to what is known about the yeast regulatory motifs.

The random forest based **metamatti** classifier outputs a probability for each classification decision, based on votes that each of the classes received in its ensemble of classification trees. This allows for the classification to be made at a chosen level of confidence. To aid the choice of the classification probability cutoff, I plotted a number of diagnostic curves, shown in Figure 5.33. Based on the analysis, I chose the lowest classification probability cutoffs for classifying the motifs predicted by each of the eight *de novo* motif prediction methods. I set the lowest probability at 0.60. I did this because the classification accuracy drops dramatically below this probability, and effectively plateaus after it, whereas the recall stays rather stable around this classification probability, but drops rapidly from around 70%. Results were also reported at 80% classification probability.

I profiled the importances of predictor variables in a separate JASPAR motif family classification exercise, to show that several different metamotifs per class contribute strongly to the classification (see Section 1.3.4 for a discussion of the variable importance measure used). The results of this analysis are shown in Figure 5.34. For instance, all of the top ranked six features are from different fungal Zinc cluster derived metamotifs.

The classification results at the 0.6 probability cutoff are shown in Figure 5.35.

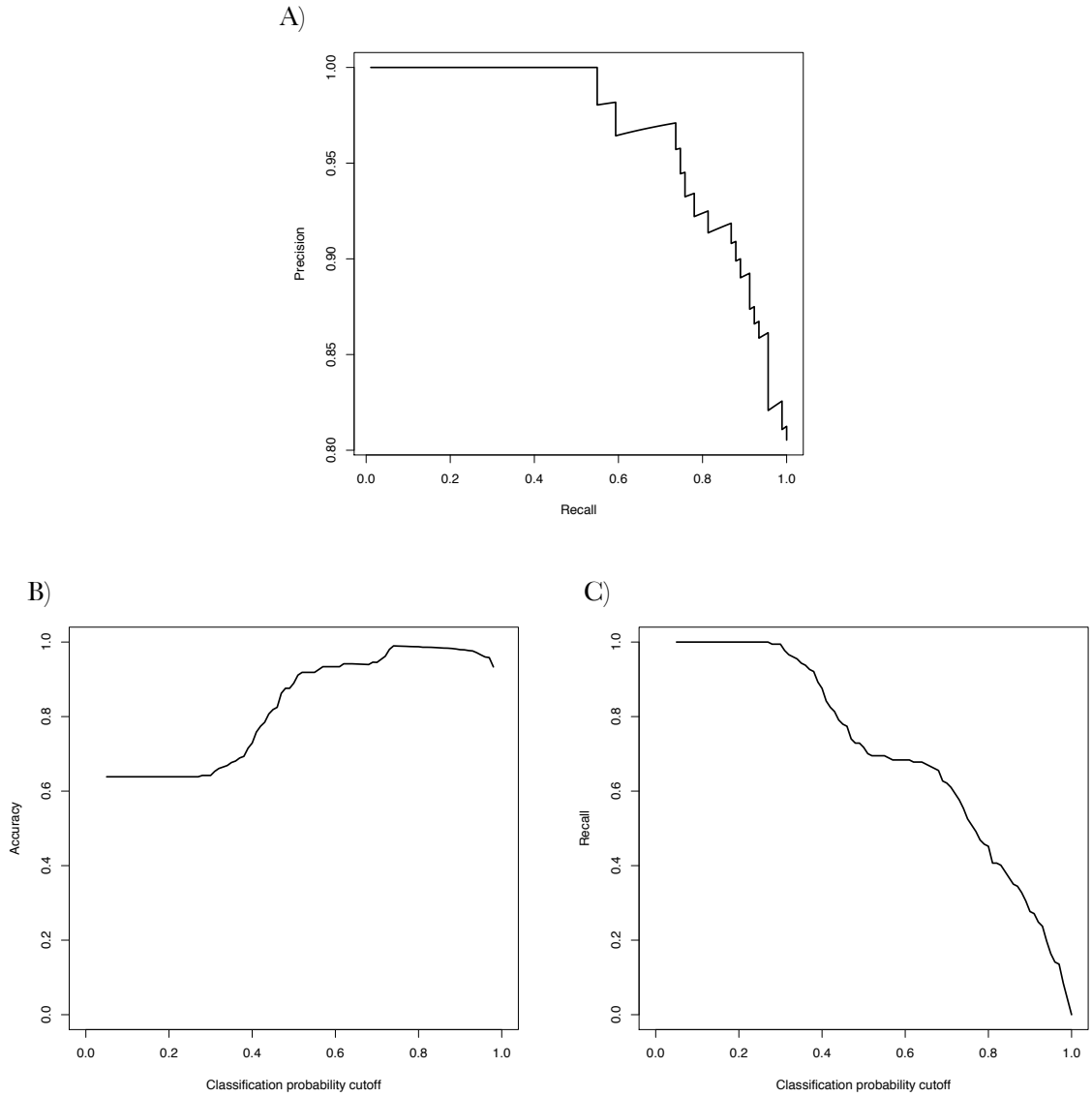


Figure 5.33: Performance measures of **metamatti** classification of JASPAR motifs. A) Precision-recall curve of 5-way JASPAR family classification training with fungal motifs in the JASPAR database. B) Accuracy as a function of the random forest classification probability cutoff. C) Recall rate as a function of the classification probability cutoff.

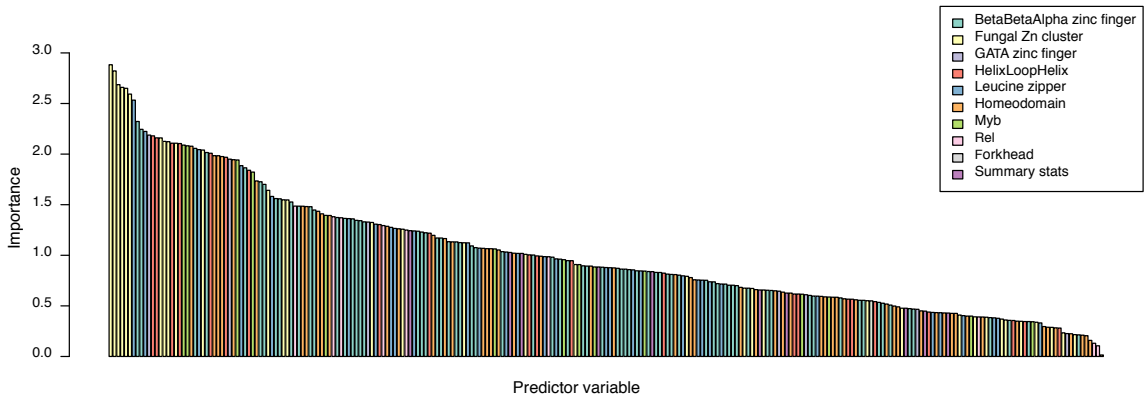


Figure 5.34: Variable importances of a JASPAR family classifier. The importance measure is described in Section 1.3.4. Metamotifs derived from ten major families of motifs in the JASPAR database were included in this exercise. One bar in the classification represents one metamotif density feature.

Instances of only two of the motif families in the 5-way classifier were found to be predicted above the cutoff, by any of the motif inference algorithms (Figure 5.35). It is disappointing that only fungal Zinc cluster motifs and $\beta\beta\alpha$ zinc finger motifs – which dominate the DNA binding domain of JASPAR motifs (Section 5.3.2) – can be detected from the *de novo* predictions at this probability cutoff. These two DNA binding domain families dominate the distribution of DBD families in the JASPAR motif set. It is however reassuring to see that in cases where there is a statistically significant close match to a JASPAR motif, the predictions are largely consistent between the `metamatti` TF family prediction (6 / 8 in the case of NestedMICA, 4 / 6 in the case of SOMBRERO, 3 / 3 in the case of Weeder), and the family of the closest JASPAR motif match. Furthermore, NestedMICA and SOMBRERO, which both show remarkably low distances to their closest JASPAR matches (Section 5.3.1), output the largest numbers of motifs which can be classified by `metamatti` at this confidence cutoff, followed by Weeder (18, 14 and 9, respectively).

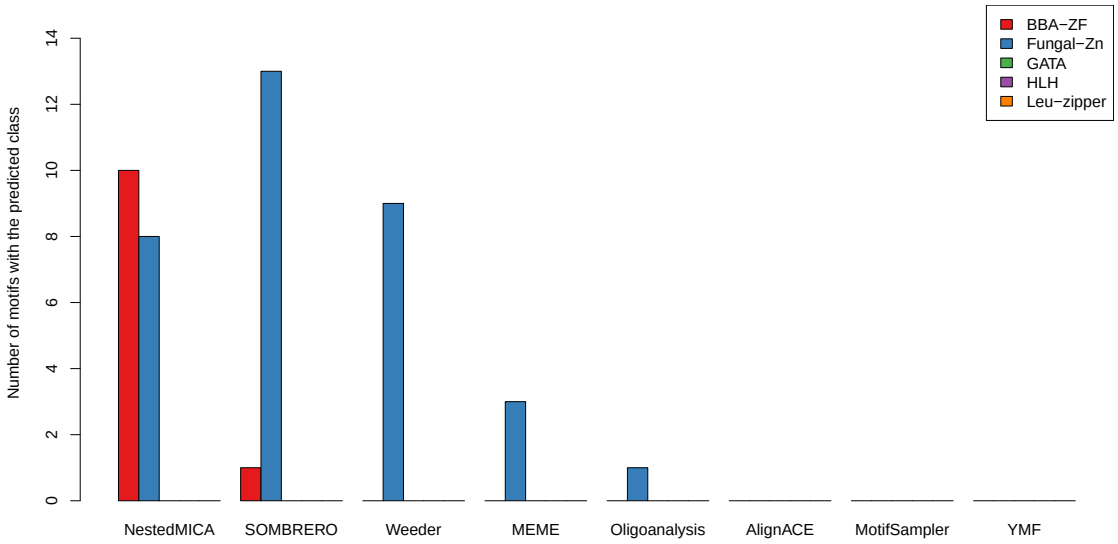


Figure 5.35: Metamatti classification of the predicted motifs at the 0.6 classification probability cutoff.

5.4 Summary

The work described in this chapter deals with large scale prediction of regulatory motifs in the *S. cerevisiae* genome, with the primary focus being a motif level performance assessment of several previously published *de novo* algorithms. The large scale motif comparison based performance assessment shown in Section 5.3.2 is in notable contrast to the binding site or nucleotide level assessments that is commonplace in motif inference literature (see Section 5.1.2). The association of a large collection of *de novo* predicted motifs with putative target genes (Section 5.3.3) has also not been previously tested in a comprehensive manner between a number of algorithms. The results of the performance assessment are rather consistent: especially NestedMICA but also SOMBRERO and MEME appear to perform adequately in finding motifs matching known regulatory motifs. None of the tested algorithms shows strong performance with the yeast genome to suggest wide applicability of *de novo* motif inference algorithms for large scale study of higher eukaryote regulatory genomes. NestedMICA's 54 statistically significant

matches to the 177 TFBS motifs in the JASPAR database is still however a surprisingly positive result for a *de novo* method when it is compared to previous work. For instance the ChIP-chip study by Harbison et al. (2004) reports a confident motif for 31% of 203 TFs, based on the output of six motif inference algorithms in the much easier case of finding motifs from sequence regions with ChIP based evidence of TF binding. It is also interesting to see that the arguable top performer of the (Tompa et al., 2005) assessment, Weeder, performs rather weakly using the metrics presented here. Indeed, NestedMICA, SOMBRERO and MEME are consistently the top performers in my assessment.

In addition to the performance assessment, I also profiled the conservation, SNP rate and positional bias trends of the motifs, to find motifs unknown to the JASPAR motif database but which are particularly likely to be functional (Section 5.3.6.4). This analysis also showed NestedMICA with the largest collection of conserved motifs with low SNP rates and evidence for preference to genomic positions close to TSSs. This analysis however depends on the criteria used for determining a significance cutoff for genomic matches of motifs, a parameter which the especially motifs predicted by MEME did not show robustness to.