Bioinformatical and Experimental Analysis of Gene Expression Regulation through RNAi and Alternative Polyadenylation

Margarita Schlackow

DPhil Thesis

Keble College; DTC SysBio; OCCAM, Mathematical Institute

University of Oxford, Oxford, UK

04.06.2014

Abstract

Polyadenylation signals in yeast are not very well defined and are believed to be largely degenerate. Here, we present a computational and experimental genome-wide analysis of polyadenylation signals in *Schizosaccharomyces pombe (S. pombe)*, identifying the canonical AATAAA motif as the most frequent and functional signal. RNA-Seq data from cells grown under various physiological conditions were used to map 3'UTRs, which classify as commonly heterogenic. We have shown that many genes have alternative 3'UTRs. Our results are summarised and can be accessed in a user-friendly online database *Pomb(A)*.

It has been shown that convergent genes require *trans* elements, like Cohesin, for efficient transcription termination. We demonstrate that convergent genes lacking Cohesin are generally associated with longer overlapping transcripts. Furthermore, we analysed ChIP-chip data of Rad21 and Mis4 as well as other Cohesin and loading complex subunits and show that regions of Rad21/Mis4 co-localisation are generally associated with highly transcribed genes. They are also cohesive, while sites with Rad21 only are less cohesive. Rad21/Mis4 co-localisation sites are in close proximity to annotated origins of replication, suggesting that cohesive sites may facilitate replication.

microRNAs (miRNAs) are well studies in higher eukaryotes and participate on post-transcriptional gene silencing by degrading target mRNA or blocking translation. It is believed that miRNAs do not exist in yeast. We reanalyzed miRNA presence in yeast using recently available small RNA data sets. Potential miRNA genes and targets in *S. pombe* were computationally predicted based on the described alternative 3'UTR data and further experimentally tested.

Dicer is an enzyme, which recognizes long dsRNA substrates and cleaves them into siRNA effector molecules, essential for gene silencing. Dicer has been thought to be a purely cytoplasmic protein. However, we employed ChIP-Seq and dsRNA RNA-Seq data to show that Dicer localises in the nucleus of mammalian cells and associates with the chromatin on numerous loci. Furthermore, we present evidence that Dicer processes long dsRNA into siRNA in the nucleus and the lack of Dicer causes the accumulation of long dsRNA. This consequently induces the interferon response pathway, which ultimately leads to apoptosis and cell death.

Acknowledgements

I would like to thank those who supported me throughout the completion of my D.Phil. My gratitude primarily goes to my supervisor Monika Gullerova and her great project ideas and guidance as well as my supervisor Radek Erban. I would like to thank Nick Proudfoot for his scientific and financial support.

I would also like to thank Konrad Krawczyk for his advice and reliability and my brother Waldemar Schlackow for his patience.

This thesis is dedicated to my parents without whom this work would not have been possible.

This work gave rise to the following publications:

Schlackow, M., Marguerat, S., Proudfoot, N. J., Bähler, J., Erban, R. and Gullerova, M. (2013). Genome-wide analysis of poly(A) site selection in *Schizosaccharomyces pombe*. *RNA*, **19**(12), 1617-1631.

Schlackow, M. and Gullerova, M. (2013). Understanding non-coding DNA regions in yeast. *Biochemical Society transactions*, **41**(6), 1654-1659.

White, E., Schlackow, M., Kamienarz-Gdula, K., Proudfoot N.J. and Gullerova, M. (2014). Human nuclear Dicer restricts the deleterious accumulation of endogenous double stranded RNA. *Nature Structural & Molecular Biology*. (Advance online publication 11.05.2014).

Further publications in preparation:

Bhardwaj, S., Schlackow, M., Yanagida, M. and Gullerova, M. (2014). Role of transcription in the establishment of cohesion on chromosomal arms.

Schlackow, M., Tzika, K. and Gullerova, M. (2014). Identification and structural characterisation of miRNA in *S. pombe*.

Credit

RNA-Seq data in Chapter 2.1 was generated by the Bähler Lab, UCL. GO analysis in the same chapter was performed by Samuel Marguerat at the Bähler Lab, UCL.

Experiments relating to Cohesin in Chapter 2.2 were performed by Shweta Bhardwaj, post-doctoral research fellow at the Gullerova Lab, Oxford.

Experiments relating to miRNAs in Chapter 3 were performed by Kelly Tzika, visiting student at the Gullerova Lab, Oxford.

Experiments relating to nuclear Dicer in Chapter 4 were performed by Eleanor White, post-doctoral research fellow at the Proudfoot Lab, Oxford. Dicer ChIP-Seq data were mapped back and peaks were called by Kinga Kamienarz-Gdula, post-doctoral research fellow at the Proudfoot Lab, Oxford.

All experiments in Chapter 2.1, as well as all bioinformatical analysis throughout the thesis, with the minor exception mentioned above, were performed by myself.

Table of abbreviations

3'RACE	3' Rapid Amplification of cDNA Ends
3PC	3' Poly(A) site mapping using cDNA Circularisation
A	Adenosine
A. thaliana	*Arabidopsis thaliana*
Ago1	Argonaute
APA	Alternative Polyadenylation
ATP	Adenosine Triphosphate
bp	Basepair
C	Cytosine
C. elegans	*Caenorhabditis elegans*
CAGE	Cap Analysis Gene Expression
cDNA	Complementary DNA
CTCF	CCCTC-binding factor, transcriptional repressor
C-Peaks	Peaks of Rad21 only occurrence
CPSF	Cleavage and Polyadenylation Specificity Factor
cRNA	Complementary RNA
CS	Cleavage Site
CstF	Cleavage Stimulatory Factor
CTD	Carboxyl Terminal Domain
CUT	Cryptic unstable transcript
DAPI	4',6-diamidino-2-phenylindole
Dcr1	Dicer
DNA	Deoxyribonucleic Acid
DRS	Direct RNA Sequencing
dscDNA	Double Stranded Complementary DNA
DSE	Downstream Element
dsRNA	double stranded RNA
EE	Efficiency Element
EMM	Edinburgh Minimal Medium
EST	Expressed Sequence Tag
EVD	Extreme Value Distribution
FISH	Fluorescence *in situ* hybridization
FRAP	Fluorescence recovery after photobleaching
FUE	Far Upstream Element
G	Guanine
imr	Innermost repeat
lncRNA	Long non-coding RNA
mRNA	Messenger RNA
MFPT	Mean First Passage Time
miRNA	microRNA
M-Peaks	Peaks of Rad21/Mis4 co-occurrence
N	Undetermined nucleotide
NLS	Nuclear localisation signal
nt	Nucleotide
NUE	Near Upstream Element
ORF	Open Reading Frame

Table of abbreviations - continued

ORI	Origin of replication
otr	outermost repeat
PAB	Poly(A) Binding protein
PAP	Poly(A) Polymerase
PAS	Polyadenylation Signal
PCR	Polymerase Chain Reaction
pri-miRNA	primary miRNA
pre-miRNA	precursor miRNA
PTGS	Post-transcriptional gene silencing
PolII	RNA Polymerase II
P-ura4	*ura4* promoter and ORF sequence
Py	Pyrimidine (Adenosine, Thymine or Uracil)
qPCR	quantitative PCR
RDRC	RNA-dependent RNA polymerase complex
RITS	RNA-induced transcriptional silencing complex
RISC	RNA-induced silencing complex
RNA	Ribonucleic Acid
RNAi	RNA interference
RPKM	Reads per kilo base per Million
rpm	Revolutions per minute
RT	Reverse Transcriptase
rRNA	Ribosomal RNA
Ser2/Ser5	Serine no. 2 or 5 (on PolII CTD)
SAGE	Serial Analysis of Gene Expression
sRNA	small RNA
shRNA	small hairpin RNA
siRNA	small interfering RNA
snoRNA	small nucleolar RNA
TIF-Seq	Transcript Isoform Sequencing
T-ura4	*ura4* terminator sequence
UMI	unique molecular identifier
S. cerevisiae	*Saccharomyces cerevisiae*, budding yeast
S. pombe	*Schizosaccharomyces pombe*, fission yeast
SUT	Stable uncharacterised transcript
T	Thymine
TGS	Transcriptional gene silencing
T_m	Melting Temperature for oligonucleotides
tRNA	Transfer RNA
U	Uracil
uORF	Upstream open reading frame
USE	Upstream Element
UTR	Untranslated Region

Contents

Chapter 1

Introduction

Over the past century molecular genetics has grown in importance as a biological discipline. Focus is on the study of heredity and variation in living organisms composed of one or more cells. It is known, that the fate of every cell is determined by its contained protein-composition, which is the result of gene expression. Genes are stretches of deoxyribonucleic acid (DNA), which contain all the necessary information for the development and functioning of living organisms. The transfer of genetic information from DNA to protein is explained in the Central Dogma of Molecular Biology. In its simplest form it states that DNA is irreversibly transcribed into RNA (ribonucleic acid), which is translated into proteins. This process will be described in more detail later in this chapter.

One of the primary interests in science is to understand the human organism and to prevent and cure diseases. Many eukaryotes have a comparable genomic structure to human cells and its molecular machinery is similarly organised, but it is crucial to understand the differences, as well as similarities. One of the very important model organisms is the unicellular fission yeast *Schizosaccharomyces pombe* (*S. pombe*).

The first step leading to gene expression is the transcription of DNA into RNA. The DNA is read by processing enzymes called RNA polymerases (I, II and III), which transcribe it into RNA. During this process the bases Adenine (A), Cytosine (C) and Guanine (G) are copied into the same base, but Thymine (T) is copied into Uracil (U).

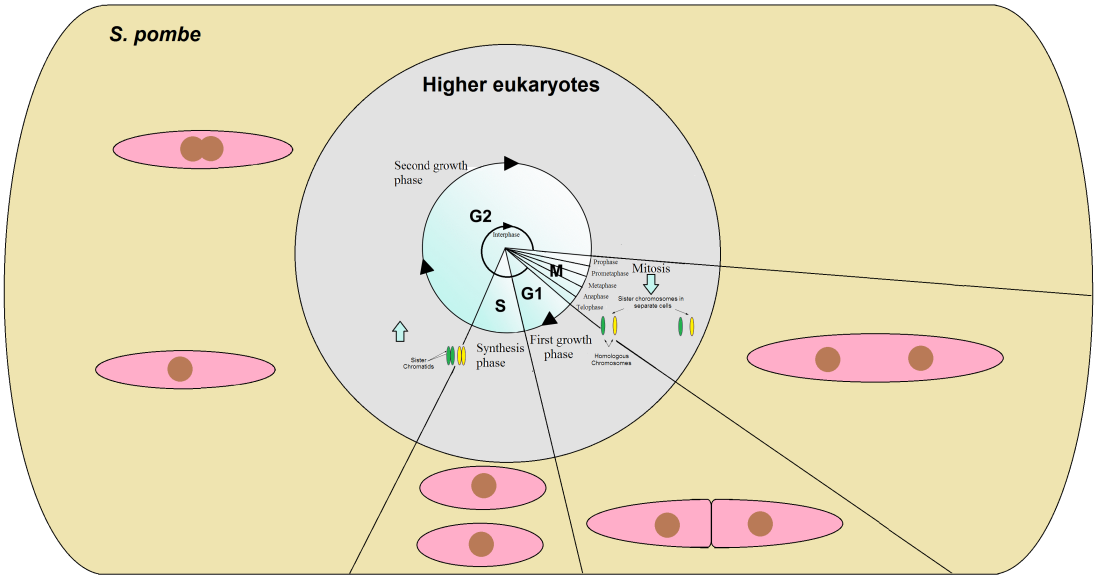

Figure 1.1: Illustration of the cell cycle. During G1 the cell grows and performs normal metabolic roles. S-phase marks the replication of the DNA. In G2 the cell grows and prepares for Mitosis. Finally Mitosis is divided into 5 sub-phases: During Prophase the chromatin condenses into the highly ordered chromosomes. In Prometaphase the nuclear envelope breaks and disappears and the microtubules reach the chromosome. In Metaphase the chromosomes align in the cell-middle to be separated and pulled to the spindle poles in Anaphase. Finally in Telophase the two daughter nuclei form and Pro- and Prometaphase are reversed. *S. pombe* (outer layer) has a closed mitosis cycle and the nuclear envelope does not break down during mitosis. The two daughter cells do not separate after M-phase, but at the end of G1/during replication in S-phase (Forsburg and Nurse, 1991).

RNA polymerase II (PolII) produces the carrier of genetic information called messenger RNA (mRNA), which is later translated into proteins. While RNA and DNA are similar molecules, an important difference is apparent in the backbone of the molecules: the RNA backbone consists of the ribose sugars and DNA of deoxyribose sugars. The ribose has an extra -OH group on the 2nd carbon of the sugar ring, which is a -H group in the deoxyribose (Figure 1.2). The sugars in the backbone of both molecules are joined by phosphodiester bonds. The -OH group of the RNA sugar makes the molecule more reactive than DNA and hence less stable and more dynamic in its production and degradation. Another important difference is that RNA is generally single stranded and forms shorter molecules, allowing easy transfer between the nucleus and cytoplasm within

the cell.

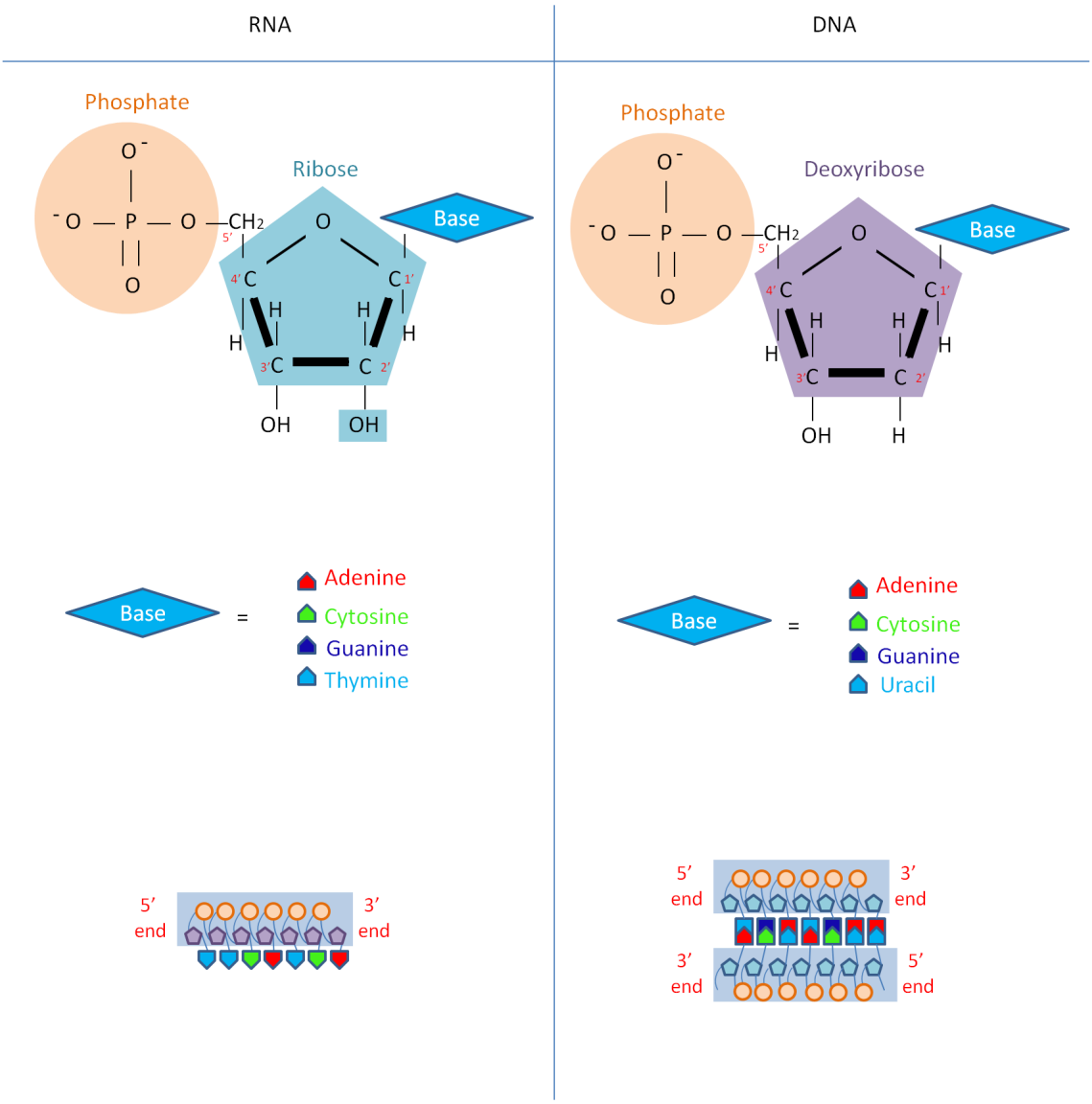

Figure 1.2: Illustration of the difference between RNA (left) and DNA (right). The RNA backbone consists of ribose sugars and the DNA backbone of deoxyribose sugars. The DNA base Thymine is replaced by Uracil in the RNA molecule. DNA is usually double stranded, while RNA is single stranded.

The process following transcription is translation which is the reading and translating of the mRNA sequence into an amino acid sequence. Translation comes about due to the activity of a ribosome in synchrony with several RNA molecules. Each amino acid is encoded by three bases called codon (several different codons can define the same amino

acid). Once the amino acid sequence has been formed it undergoes a folding process resulting in the mature three-dimensional protein.

The above described procedure is summarised in the Central Dogma of Molecular Biology: For higher eukaryotes DNA undergoes replication into DNA and transcription into RNA. RNA translates into proteins, which is the final stage of the gene expression path.

1.1 Transcription by RNA polymerase II and gene expression regulation

PolII is an enzyme responsible for transcribing the genes encoding mRNA and certain small nuclear non-coding RNAs. In *S. pombe* PolII has twelve subunits, which are functionally and structurally conserved from yeast to human (Mitsuzawa and Ishihama, 2004). PolII in *S. pombe* possesses a carboxyl-terminal domain (CTD), essentially a tail of a repeated sequence of seven peptides. The phosphorylation/dephosphorylation of the amino acid serine in position 2 and 5 of the repeat (Ser2 and Ser5 respectively) corresponds to the transcriptional stage it is in (Mitsuzawa and Ishihama, 2004). There are known to be three such stages: transcription initiation (Ser2 and Ser5 dephosphorylated initially, Ser5 is known to phosphorylate at the beginning of the gene), elongation (Ser2 phosphorylates, Ser5 dephosphorylates), and finally termination featuring pausing (Ser5 and Ser2 dephosphorylate, Bai et al., 2006).

Transcription initiation is the most studied process in gene expression, where certain transcription factors are recruited to the promoter region by recognition of a DNA-motif called the TATA-box. These factors in turn recruit PolII. The collective assembly of transcription factors and PolII is termed the transcription initiation complex. After several abortive initiation processes (truncated transcripts), Ser5 of the CTD is phosphorylated by

the transcription factor TFIIH and subsequently capping enzymes are recruited to form the 5' cap of the pre-mRNA (Goodrich and Tjian, 1994; Mandal et al., 2004).

At the elongation stage PolII moves away from the promoter region and produces a growing transcript, the pre-mRNA. It has been shown that PolII moves one base pair at a time. The exact process of how PolII moves along the DNA is not resolved, but a few models have been proposed regarding the transcriptional stages. It is known that transcription does not proceed at a uniform rate, but it is most likely DNA sequence dependent (Neuman et al., 2003). For example transcriptional pausing can happen due to a certain nucleotide composition in regions of the DNA. It is likely to have regulatory purposes by slowing down the transcription velocity (Darzacq et al., 2007). Hence it allows transcription factors to bind and modify gene expression, especially when expression is high. The presence of transcription inhibitors, which do not necessarily affect the active transcription velocity, can affect the frequency of pausing.

Finally, termination is to date the least studied stage of the cycle. Once entering the termination stage PolII releases the new RNA transcript (this process is termed cleavage) and disengages with the DNA (transcription termination). Both events are associated with pausing of PolII (Gromak et al., 2006). The cause of cleavage and termination can either be encoded in a particular DNA sequence and/or mediated by protein factors. It is thought that after cleavage the exonuclease Rat1 (in yeast) degrades the PolII associated transcript. Once Rat1 catches up with the still elongating PolII, transcription termination occurs (reviewed in Proudfoot, 2011).

1.2 Cohesin

In order for a cell to proliferate, it needs to replicate its entire genome in S-phase, which results in formation of sister chromatids, tightly held together. In anaphase these sister chromatids are pulled apart by the mitotic spindle into two identical daughter cells. For

this process to occur correctly the efficient cohesion of the chromatids before anaphase is crucial. Cohesion allows the correct alignment of chromosomes and correct attachment of the mitotic spindle, therefore ensuring that each daughter cell secures the correct amount of chromosome copies. Too many chromosomes within a cell, referred to as aneuploidy, is known to lead to birth defects in mammals, such as Down's syndrome.

Several proteins have been shown to be essential for sister chromatid cohesion, jointly forming the ring-like complex Cohesin (Losada et al., 1998; Sumara et al., 2000; Tóth et al., 1999). Several models have been investigated of how Cohesin might hold the sister chromatids together. The most prevalent one is the topological encircling of the two sister chromatids (Gartenberg, 2009). Other functions of Cohesin include demarcation of silent chromatin domains and stabilization of the genome by preventing undesirable recombination within the repetitive DNA of the heterochromatin (Gartenberg, 2009).

The structure of Cohesin is widely conserved across eukaryotes and comprises multiple subunits (Figure 1.3). Cohesin forms a tripartite ring responsible for the cohesion of sister chromatids, its structure in different organisms is reviewed in Peters et al. (2008). In *S. pombe*, the ring is formed by the proteins Psm1 and Psm3 subunits and these are bound tightly by a hinge domain. At the ATPase domain (the "head") they are bound by Rad21. Psm1 and Psm3 close the ring upon ATP binding. The ring opens again with ATP hydrolysis (reviewed in Nasmyth and Haering, 2009). There can be multiple types of Cohesin: in *S. pombe* Rad21 further associates with Psc3 or Pds5, which interact with the Cohesin loading complex Mis4/Ssl3 (also known as Kollerin). The loading complex is responsible for recruiting Cohesin to the chromosome. It has been shown that Pds5 is not essential for sister chromatid cohesion, indicating a redundant function (Losada et al., 2005). Generally its association with Cohesin remains to be elucidated. There is evidence of its association with Wpl1 (Kueng et al., 2006), a protein responsible for removing Cohesin from the chromatin.

It has been shown that Cohesin promotes PolII termination between convergent genes.

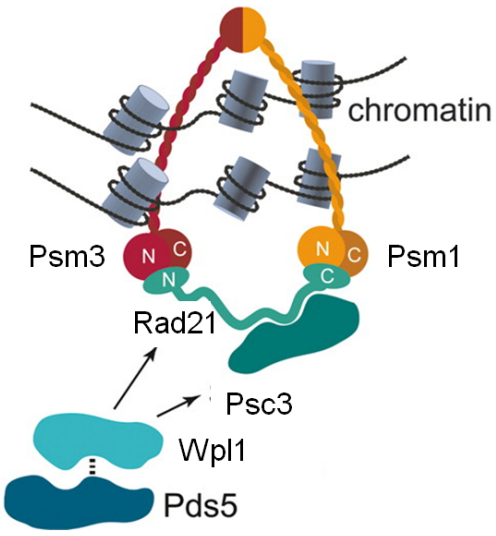

Figure 1.3: Illustration of the Cohesin protein complex. The Psm1 and Psm3 arms are tighly bound at the hinge domain. Rad21 binds them at the head. Psc3 associates with Rad21 and is responsible for the interaction with the chromatin loading complex Kollerin, as well as the separase Wpl1 (Figure modified from Peters et al., 2008 to fit the *S. pombe* notation). Reproduced under the Creative Commons license available at http://creativecommons.org/licenses/by-nc/4.0/legalcode.

The topological encircling of the DNA enables Cohesin to slide along the less stably packaged chromatin. It appears that while the replication fork is small enough to pass through the Cohesin ring, PolII is too large and pushes the Cohesin ring along the DNA (Gullerova and Proudfoot, 2008; Lengronne et al., 2004; Losada, 2007). Cohesin is not loaded onto the chromatin in G1 of the cell cycle, hence convergent genes may result in dsRNA, which induces RNA interference (see Section 1.3). However in G2, Cohesin is loaded and PolII pushes the ring-like structure to the intergenic region between the convergent genes. Here it acts as a block and causes PolII to dissociate (Gullerova and Proudfoot, 2008, Figure 1.4).

In higher eukaryotes transcription termination does not rely on *trans* elements like Cohesin. The polyadenylation signals (PAS) are signals consisting of one or more motifs, which are encoded in the DNA, transcribed into the pre-mRNA molecule and act there to direct cleavage and polyadenylation (the addition of an adenosine-tail, the so-called

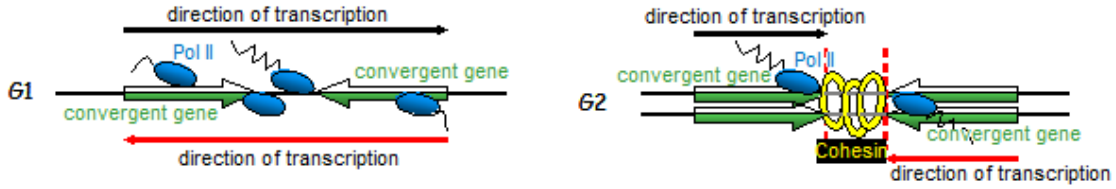

Figure 1.4: Illustration of Cohesin acting as a block to avoid overlapping complementary transcripts when it is loaded in G2. It acts as a block for PolII in intergenic regions of convergent genes.

poly(A)-tail, to the mRNA). In higher eukaryotes PAS are responsible for pre-mRNA cleavage and transcription termination (the physical dissociation of PolII from the DNA). In yeast transcription termination may not happen due to PAS sequence recognition, therefore a *trans* element is required, such as Cohesin.

A non-canonical or weak PAS leads to a frequently observed phenomenon called alternative polyadenylation, resulting in alternative cleavage events: transcripts of different lengths are derived from the same gene. The longer transcript can then acquire additional regions serving as binding platforms for regulatory protein complexes and microRNA (miRNA) target sites. A miRNA is an important gene expression regulator in higher eukaryotes, triggering RNA interference (RNAi). Even though yeast is believed to lack miRNAs, *S. pombe* does feature an RNAi pathway.

1.3 RNAi, miRNAs and gene expression regulation

RNAi is a widespread mechanism in eukaryotes and serves to silence genes either on a transcriptional (TGS) or post-transcriptional level (PTGS, Mello and Conte, 2004). It is driven by small interfering RNA (siRNA) produced from double stranded RNA (dsRNA). Endogenous dsRNA can result from natural antisense transcription (Faghihi and Wahlestedt, 2009), but can also be caused by overlapping transcripts from convergent genes (Gullerova and Proudfoot, 2008). Moreover in mammals imperfectly paired dsRNA is known to originate from the stems of hairpin structures, created by transcription

of an inverted repeat sequence. The dsRNA is cut into short fragments, which may complement regions of mRNA molecules. The genomic region where the RNA-copy originates from is then silenced either by heterochromatin formation (TGS) or inhibition of translation/mRNA cleavage (PTGS).

RNAi is known in higher eukaryotes as well as in fission yeast, yet it is absent in *Saccharomyces cerevisiae* (*S. cerevisiae*). The current belief is that gene silencing in mammals happens on a PTGS level, referring to the inhibition of translation or cleavage of the already mature mRNA. Key players in this process are microRNAs (miRNAs), which are identified in many higher eukaryotes. miRNAs are short RNA molecules originating from the mentioned hairpin structures (also called stem-loop or foldback structures). These hairpin structures are encoded in the genome, often in introns or non-coding genes (Rodriguez et al., 2004). An illustration of the miRNA and dsRNA into small RNA (sRNA) processing pathway is shown in Figure 1.5 (He, 2004). Once a miRNA-gene is transcribed it is known as primary miRNA (pri-miRNA), which is processed in the nucleus by the Drosha-enzyme into the precursor miRNA (pre-miRNA, see Figure 1.5, He, 2004). Drosha is assisted by DGCR8, also known as Pasha, which contains an RNA-binding domain and stabilises the pri-miRNA for Drosha-processing (Kim et al., 2008). Pre-miRNAs have the characteristic stem-loop structure. They are exported into the cytoplasm by Exportin-5, and further cleaved by Dicer into small fragments. One of the strands of these fragments, the mature miRNA, is incorporated into the RISC complex. The other strand is referred to as miRNA*. RISC driven complementarity to the miRNA targets mRNA or pre-mRNA. This complementarity, however, is far from perfect in organisms other than plants and makes miRNA target and gene detection a problematic process.

In fission yeast and some other eukaryotes, a different RNAi pathway exists, known as transcriptional gene silencing (TGS). TGS refers to the inhibition of transcription within the nucleus, so that an RNA molecule does not mature. Formation of tightly packed

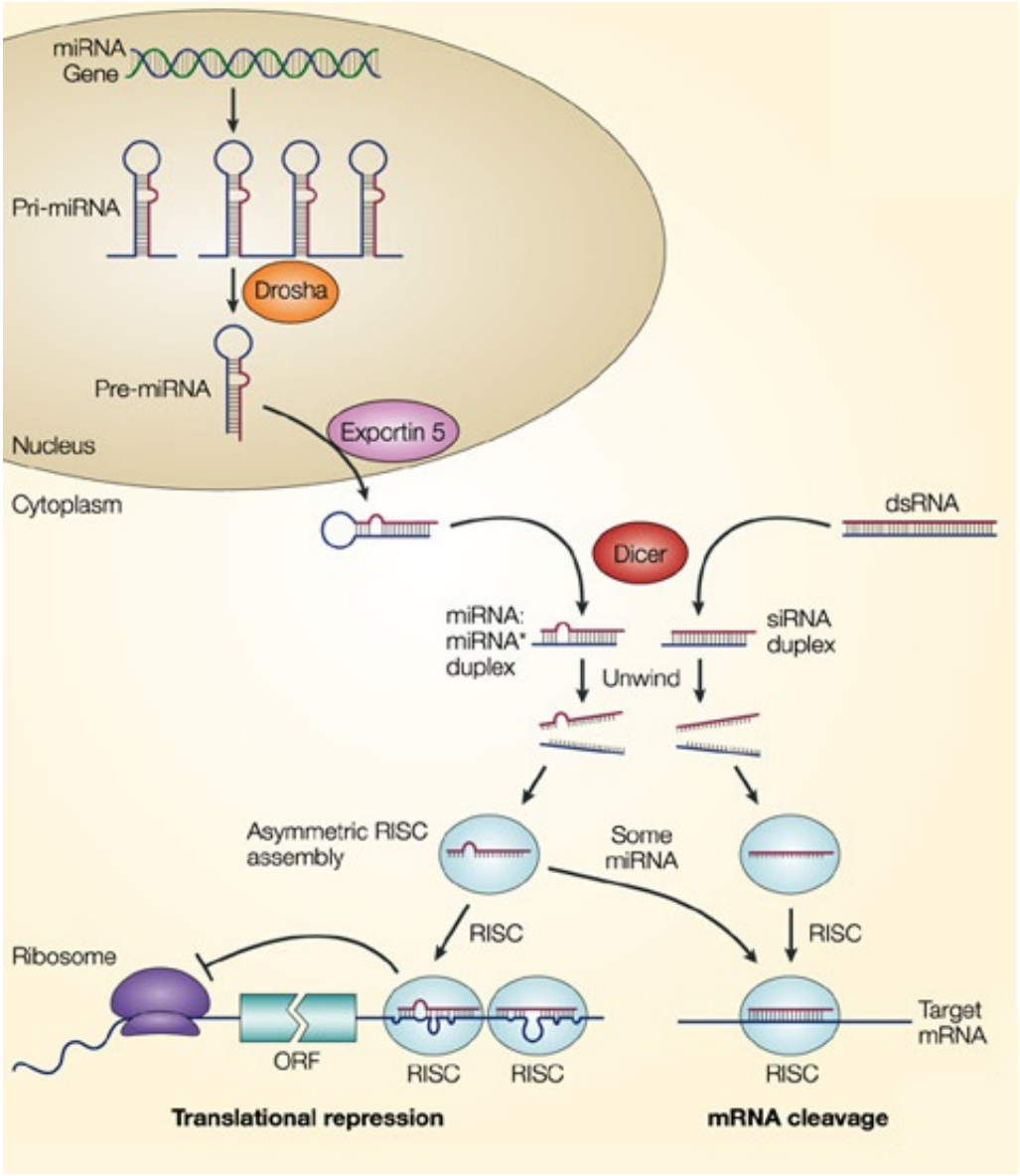

Figure 1.5: Pri-miRNA transcripts are obtained from introns or non-coding genes, and processed inside the nucleus by Drosha into a ~70 nt hairpins termed pre-miRNAs. Exportin 5 transports the hairpin into the cytoplasm, where Dicer cuts it into ~22bp miRNA:miRNA* duplexes, as well as processes long dsRNA into siRNA. RISC incorporates one of the strands of the miRNA:miRNA* duplex or the siRNA duplex. RISC then acts by partial (miRNA) or full (siRNA) complementarity to the incorporated strand on its target and causes translational repression or mRNA cleavage. ORF refers to Open Reading Frame. Adapted by permission from Macmillan Publishers Ltd: *Nature* (He, 2004), copyright (2004).

chromatin, called heterochromatin, prevents PolII access to the DNA and hence prevents transcription and leads to downregulation of gene expression.

1.4 RNAi and heterochromatin formation in *S. pombe*

Heterochromatin refers to tightly packed chromatin, which is considered to be transcriptionally inactive. In fission yeast it is mostly found in telomeres, silent mating-type loci and the centromeres. The central kinetochore binding region of the centromere is flanked by repetitive DNA, the innermost and the outermost repeats (imr and otr respectively). The otrs are composed of sequences which are coated by histone H3, methylated at lysine 9 (H3K9) serving as docking sites for other proteins such as Swi6 to maintain the silent state (Chang et al., 2012). Swi6 acts to recruit Cohesin by interacting with Kollerin and the Cohesin subunit Psc3 (Nonaka et al., 2002).

A self-enforcing loop driven by the RNAi components RNA-dependent RNA polymerase (Rdp1), Dicer (Dcr1) and Argonaute (Ago1) as well as further chromatin modifying components forms the heterochromatin at the centromeres. An illustration of the loop is depicted in Figure 1.6. dsRNA can also be formed by convergent genes (genes facing each other on opposite strands) with overlapping 3'UTRs or by antisense transcription (Horn and Peterson, 2006). Dcr1 recognises the dsRNA and cleaves it into ~22bp long fragments. One strand of these short fragments is loaded onto Ago1 and forms the RNA-induced transcriptional silencing complex (RITS in *S. pombe*, equivalent to RISC). RITS binds to the complementary region of a nascent transcript and associates with RNA-dependent RNA polymerase complex (RDRC) and generates a second strand from the pre-mRNA. This produces more dsRNA and the loop feeds back into itself.

Furthermore, the Clr4-Rik1-Cul4 complex (CLRC) also associates with RITS. When RITS binds to the nascent transcript, CLRC comes close to the chromatin. Clr4 methylates H3K9, which then recruits a protein Swi6. Swi6 recruits Cohesin and ensures the formation and stabilisation of heterochromatin. Clr4 and Chp1 stabilise CLRC and RITS at the heterochromatin, re-enforcing the spreading of heterochromatin.

For a long time there was no evidence of what the first step driving the self-enforcing heterochromatin loop might be. In 2010 Halic and Moazed attempted to answer this

question by identifying primal small RNAs (priRNAs). These originate in a Dicer independent manner from transcription degradation products. Possibly they are the result of endo- or exoribonuclease activity. priRNAs randomly associate with Ago1 in order to enforce the initial uprise in siRNAs.

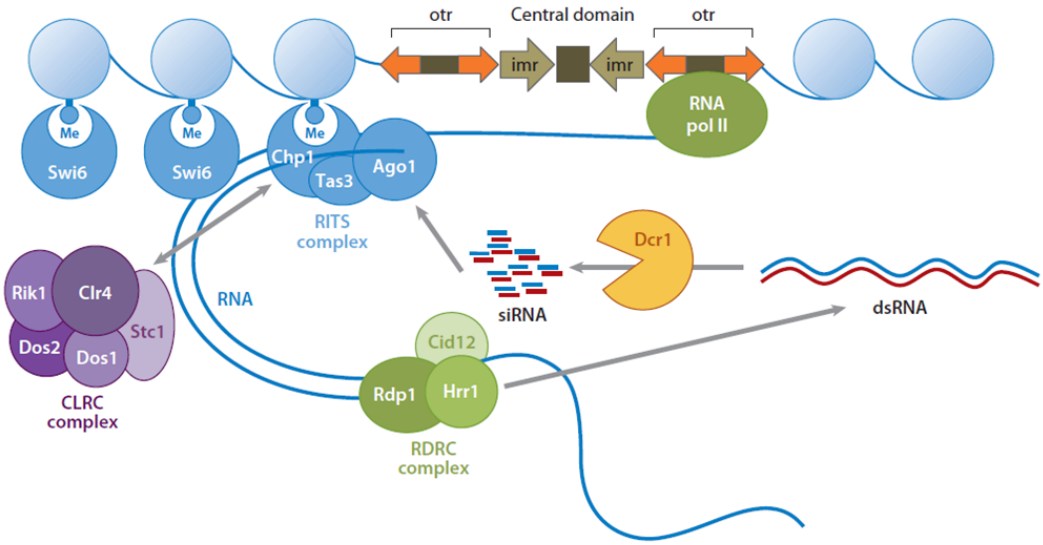

Figure 1.6: A schematic representation of the RNAi pathway for heterochromatin formation is *S. pombe*. dsRNA arises from overlapping transcripts, which are recognised by Dicer (Dcr1) and cleaved into short fragments. These are incorporated and sliced by Argonaute (Ago1) to form the RITS complex. The Ago1-incorporated short RNA molecule targets the nascent mRNA by complementarity and the RNA dependent RNA polymerase complex (RDRC) generates a complementary strand for the pre-mRNA, thus enforcing a feedback loop. A full description of the pathway is given in the main text. Figure adapted from Chang et al. (2012).

1.5 From single-gene to the genome-wide perspective

Gene expression is determined by transcription into mRNA and its subsequent translation into proteins. However, many RNAs, which are also transcribed by PolII are not translated into proteins and serve to regulate gene expression. While miRNA and dsRNA are obvious such candidates, parts of the coding mRNA is also not processed into proteins. These parts are fittingly termed 5' and 3' untranslated regions (UTRs), respectively corresponding to its location before or after the ORF or coding DNA sequence (CDS)

in transcriptional direction. Other regulatory sequence elements are introns lying within the CDS, which are spliced out before translation.

Starting from the analysis of individual genes, followed by DNA microarrays (David et al., 2011), ESTs (Graber et al., 1999), RNA-Seq (RNA sequencing, Marguerat et al., 2012; Wilhelm et al., 2008), DRS (direct RNA sequencing, Ozsolak et al., 2009) and TIF-Seq (transcript isoform sequencing, Pelechano et al., 2013), genomic studies are increasingly advancing technical aspects of experiments and consequently resulting in larger and more detailed datasets. At the commencement of this work RNA-Seq was the most advanced technique suitable for the analysis. Below I provide an explanation why DNA microarrays and ESTs were overpowered by RNA-Seq and hence are less desirable for data generation to study various aspects of the regulation of gene expression. DRS and TIF-Seq were recently implemented and exceeded the data quality of RNA-Seq in several ways, but as sequencing depth grows, RNA-Seq remains a powerful tool.

1.5.1 Expressed sequence tags

ESTs are partial reads of cDNA sequences. They provide the possibility of evaluating gene expression dependent on the cell cycle stage and cell type (Parkinson and Blaxter, 2009). Isolated eukaryotic mRNA is reverse-transcribed into cDNA, which is inserted into a suitable vector (Figure 1.7A). Inserted fragments are sequenced by priming their ends in randomly selected clones. Bioinformatical analysis removes low sequence reads and contaminating vector sequence. Individual EST reads can vary greatly in length, up to 800 nt (Parkinson and Blaxter, 2009). Although ESTs have proven to be useful in studying gene expression, several disadvantages have pushed them to the sideline of currently used techniques. In particular, low numbers of full-length ESTs, chimaeric sequences due to cDNA template switching (Cocquet et al., 2006), internal cDNA priming events and low quality sequences at the EST ends (Aaronson et al., 1996, reviewed in Nagaraj et al., 2007) led to the development of further approaches to study gene expression.

1.5.2 DNA microarrays

DNA microarrays became one of the most popular tools to measure transcriptional activity and genotype in multiple genomic regions. Fluorescently labelled cDNA or cRNA (complementary RNA) can hybridise to microscopic spots on a surface, each containing one specific fluorescent DNA probe (Figure 1.7B). A laser-scanning microscope detects the fluorescent signal corresponding to sampleprobe hybridisation (Schena et al., 1998; Shalon et al., 1996). Microarrays enable parallelisation of data acquisition, allowing gene comparison. A major challenge of these microarrays is the elimination of crosshybridisation. Additionally, microarrays depend on specific probes to allow the detection of particular gene expression. ESTs, in this case, are helpful for probe design. The more recent tiling arrays are independent of gene annotation and can probe for any genomic sequence. This allows the representation of non-repetitive DNA at various sequence resolutions, thus enabling the discovery of novel transcripts and regulatory elements (Bertone and Snyder, 2005).

1.5.3 RNA sequencing

DNA tiling arrays could fail to identify short exons and precise UTR boundaries. They may give high false positive rates of transcribed regions and cannot detect post-transcriptionally modified sequences (Nagalakshmi et al., 2008). RNA-Seq is a more recent quantitative method allowing estimation of gene expression levels, where cDNA reads are mapped back to genomic regions (Marguerat and Bähler, 2010). Unique DNA adapters are attached to both ends of sheared DNA fragments (<500 nt, Figure 1.7C),

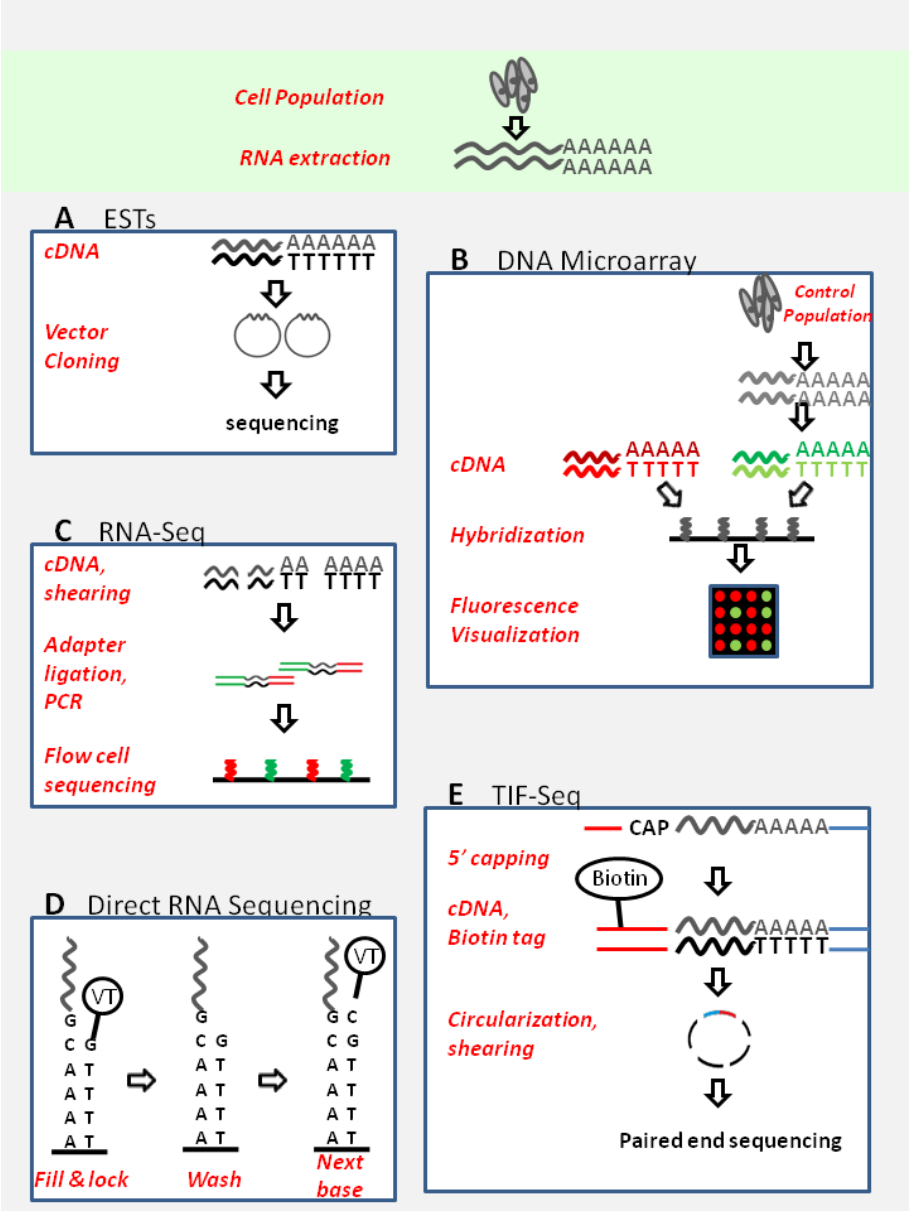

Figure 1.7: Schematic illustration of experimental techniques. RNA is extracted from a population of cells. **A ESTs.** The extracted RNA is converted into cDNA and inserted into a plasmid. Random clones are sequenced by 5' and 3' end priming. **B DNA Microarray.** The RNA from a population of interest and a control population is extracted and reverse transcribed into cDNA. Each population is differently fluorescently labeled. The samples are spotted onto a microarray slide, whose spots contain different probes to which the samples can hybridise by complementarity. Different levels and colors of fluorescence imply where and how strong the probed region is expressed. **C RNA-Seq.** After reverse transcription and shearing of the cDNA, the fragments undergo adapter ligation. Size selected fragments are added to a flow cell, which is covered with oligonucleotides binding to the adapters. Bound sequences are locally amplified and sequenced via addition of reversibly chemically blocked nucleotides. Strand-specific and non strand-specific RNA-Sequencing is possible nowadays. **D DRS.** poly(dT) oligonucleotides cover a surface which binds the polyadenylated RNA. A Fill and lock step fills in the overhanging poly(A) tail and locks the first non-A nucleotide with a complementary fluorescently labeled VT-nucleotide, which is imaged. The dye-nucleotide is then chemically cleaved and the next VT nucleotide can be incorporated and imaged. This process is then repeated. **E TIF-Seq.** The 5' end is capped. Full length cDNA is generated and the 5' end is Biotin tagged. The molecules are then circularised and sheared. The sequence is obtained by paired end sequencing.

allowing their selection on beads or on a slide in an adapter-dependent manner. Early RNA-Seq protocols used adapter ligation to double-stranded cDNA, which caused loss of information about transcriptional directionality. Strandspecific RNA-Seq and paired-end protocols were therefore developed later (reviewed in Marguerat and Bähler, 2010). RNA-Seq enables the investigation of unique sequences, even if there is only a difference of one nucleotide, and can detect and quantify DNA transcription at much lower levels than DNA microarrays. In addition, sequenced junctions between poly(A) tail and transcript allow a precise mapping of the 3' ends.

1.5.4 Direct RNA sequencing

The major advantage of the DRS technology (Ozsolak et al., 2009) is transcriptome sequencing using small amounts of RNA without cDNA conversion, eliminating several experimental artefacts associated with it. DRS provides the opportunity to map precise poly(A) sites (Ozsolak et al., 2010), giving insight into alternative polyadenylation and heterogeneity. Moreover, any polyadenylated RNA is observable, such as a small fraction of rRNAs and snoRNAs (small nucleolar RNAs, Ozsolak et al., 2009). DRS employ a single-molecule sequencing approach. Polyadenylated mRNAs are captured on the surface of a poly(dT)-coated flow cell. The poly(dT) primers initiate sequencing steps, using the exposed chain as a template to construct a complementary strand, reading each new base identity at its addition (Figure 1.7D). The protruding poly(A) tail is filled with overabundant dTTP. The complementary strand is constructed from fluorescent nucleotides containing a VT (virtual terminator) group, preventing any further nucleotide addition. If a nucleotide presented to the RNA strand is complementary to the first nucleotide in the strand, it will be captured and produce a fluorescent signal. The solution is then washed away, and the VT is cleaved from the incorporated nucleotide. The process is repeated multiple times, with a fluorescent signal indicating the growing chain (Ozsolak et al., 2009). DRS was developed using the Helicos Biosciences platform. Unfortunately

Helicos Biosciences declared bankruptcy at the end of 2012, so while the technique is still used and might continue to be used for a while, it is likely to disappear shortly.

1.5.5 Transcript isoform sequencing

The above techniques allow sequencing of transcript fractions. The problem remains to map 3' ends to their corresponding 5' ends. A novel technique called TIF-Seq allows simultaneous sequencing of both transcript ends (Figure 1.7E; Pelechano et al., 2013). It enables the observation of full-length transcripts and hence transcript and protein heterogeneity via different combinations of 5' and 3' ends. mRNA molecules with capped 5' ends and polyadenylated 3' ends are transcribed into full-length cDNA. The 5' end is tagged with biotin. The cDNA molecules are circularised and fragmented, but the biotin allows the capture of the 5' end 3' end junction on beads. The sequencing of the captured molecules is performed with a standard DNA-Seq library generation and paired-end sequencing (Pelechano et al., 2013).

1.5.6 3' Poly(A) site mapping using cDNA Circularisation

A mature mRNA is polyadenylated at its 3' end. A CS can therefore be mapped back to the DNA considering the junction of the RNA-Sequence and the poly(A) tail (up to precision of A-bases around the CS). While RNA-Seq data is a reasonable approach to map cleavage sites on the genome as some reads contain the junction, it is not the most efficient way, as only a fraction of the produced reads can be used. A recently developed technique called 3PC (3' Poly(A) site mapping using cDNA Circularisation) employs Illumina sequencing to obtain a more exhaustive CS set in *S. pombe* (Mata, 2013).

Total RNA is fragmented and magnetic oligo(dT) beads are employed to pull down polyadenylated RNAs. These RNAs are reverse transcribed using a specifically engineered primer: it contains an oligo(dT) sequence and two more primer binding sites.

These two binding sites are separated by a spacer, that blocks the DNA polymerase. The resulting molecules are circularised. Due to the spacer, DNA polymerases produce already linear molecules, with the help of PCR primers containing Ilumina-specific adaptors. Moreover, each cDNA molecule is individually tagged with a unique molecular identifier (UMI, Kivioja et al., 2012), which was contained in the RT primers. This UMI allows inference of each original individual cDNA, independently of PCR amplification. By sequencing single-endedly on the Illumina Platform from the oligo(dT) primer side, each mRNA-poly(A) tail junction can be detected. An illustration is provided in Figure 1.8.

1.6 Outline of the thesis

This project aims to elucidate the transcriptional and post-transcriptional regulation of gene expression in fission yeast and mammals. In particular it is demonstrated that fission yeast PAS are not well-conserved on a genomic level and therefore this organism seems to rely on other *trans* acting factors, such as Cohesin (Chapter 1). Moreover, the potential existence of a PTGS pathway in fission yeast is presented in Chapter 3, while Chapter 4 describes the existence of TGS in mammals.

This work concentrates on non-coding DNA sequences. To analyse PAS, I inspected 3'UTRs, where cleavage happens. In order to examine fission yeast for miRNA, I explored small RNA (sRNA) molecules. For our study of Dicer action in human cells I inspected dsRNA. The presented findings have been possible by recent technological advances of experimental strategies to obtain a genome-wide perspective of gene expression. Three decades ago, studies of gene expression regulation were based mostly on biological and biochemical analyses of a specific gene. Recent techniques involve quantitative transcriptome sequencing and subsequent bioinformatical and statistical analyses. These offer a much deeper and more general understanding of transcriptional

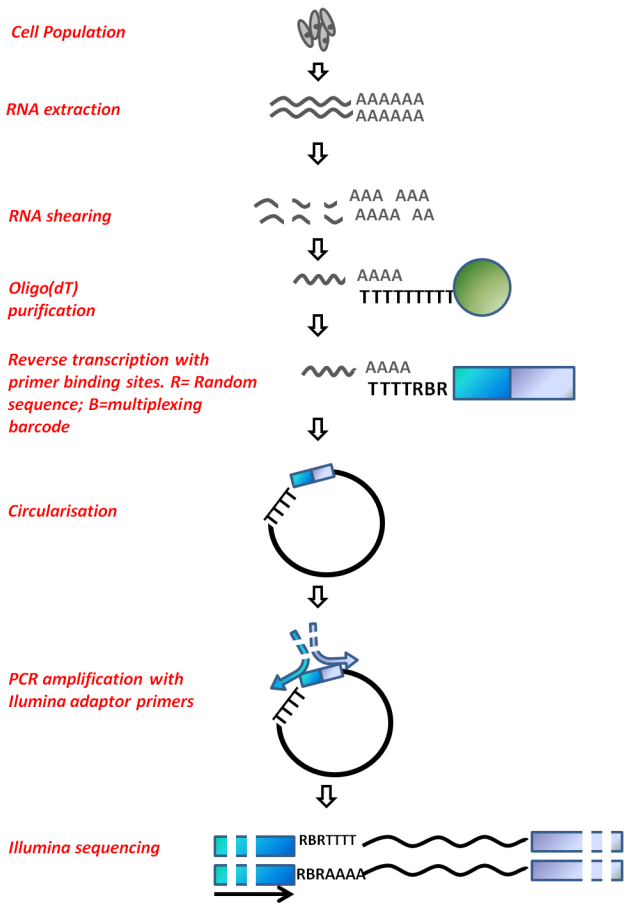

Figure 1.8: Outline of the 3PC protocol, based on Mata, 2013. Isolated RNA is pulled down on magnetic beads with an with an oligo(dT) primer, which contains two more spacer separated primer binding sites. Molecules amplified with UMI-containing primers and circularised and further amplified with Illumina adaptor primers for Illumina sequencing.

regulation, which is a major focus of this research.

I concentrated on the analysis of RNA-Seq data. While EST databases were existent at the commencement of this work, they were highly incomplete for yeast. Multiple studies focused on high-density oligonucleotide tiling arrays to re-annotate gene boundaries and the estimation of levels of coding and non-coding transcripts in *S. cerevisiae* (David et al., 2006) and *S. pombe* (Dutrow et al., 2008). Our interest in the precise locations of cleavage sites in *S. pombe* and short RNA sequences for the miRNA study required us to use RNA-Seq data. This technique had successfully been implemented to quantify gene expression in yeast species by multiple studies. Nagalakshmi et al. (2008) applied

RNA-Seq to *S. cerevisiae*, and Wilhelm et al. (2008), Rhind et al. (2011) and Marguerat et al. (2012) used RNA-Seq (with different protocols) in *S. pombe* to define transcribed genomic regions. Though 3PC would be favourable to RNA-Seq, the latter provides a sufficiently large dataset to statistically analyse signals before the CS. Hence I used RNA-Seq data to investigate PAS in *S. pombe*. A conserved hexamer AATAAA, or a close variant, promote mRNA cleavage in about 80% of all human genes (Beaudoing et al., 2000). Interestingly, in fission yeast such a highly conserved element has not been observed and AATAAA was believed to be non-functional (Humphrey et al., 1994). However, my experiments suggest differently. In Chapter 2 I describe in detail the meaning of polyadenylation and revise what is known about PAS, in order to classify PAS in fission yeast on a genome-wide level. I also come to the conclusion that genes in fission yeast are widely alternatively polydenylated and the CS experience severe heterogeneity. Moreover, Cohesin can have a role as a *trans* element in transcription. We investigated further roles of cohesin in the chromatin and its connection to transcription termination, cohesion and gene expression.

So far in unicellular organisms little is known of the purpose of multiple transcript isoforms. In mammals longer transcript isoforms often provide a miRNA target site, which is absent in the shorter isoform. This allows the gene expression to be post-transcriptionally regulated. I wish to bridge a gap of PTGS in mammalian cells and TGS in fission yeast. On the one hand, miRNA have yet not been identified in *S. pombe* and the cells lack the Drosha enzyme. Nevertheless, *S. pombe* does have one copy of a Dicer enzyme and Argonaute, which are part of the RNAi pathway. It has also been shown that it can process hairpin structures into siRNA (Simmer et al., 2010). Additionally plants, which lack a Drosha homologue, still have a miRNA pathway. I believe this may point at the existence of miRNA in the unicellular yeast and investigate this in Chapter 3. I identified possible miRNA candidates computationally, which are supported by experimental analysis. The possibility remains, though, that these miRNA

would act at transcriptional level, however their presence would provide evidence for transition from *S. pombe* TGS to mammalian PTGS.

Another indicator of the greater similarity of RNAi pathways between mammalian and yeast cells are past studies, which demonstrated that siRNAs targeting various mammalian gene promoter regions have the capacity to induce repressed chromatin structures, implying TGS effects (Janowski et al., 2005; Morris et al., 2005). siRNA, which target certain gene exons also have the potential to induce heterochromatic marks, hence transcriptional elongation decelerates and influences alternative splicing (Alló et al., 2009; Ameyar-Zazoua et al., 2012; Saint-André et al., 2011). It was shown that TGS effects are possible in the mammalian genome via the transfection of plasmids containing specific gene fragments placed between convergent PolII promoters (Gullerova and Proudfoot, 2012).

A further indicator of mammalian TGS would be the nuclear localisation of Dicer (Billy et al., 2001; Jakymiw et al., 2010; Kotaja et al., 2006; Provost et al., 2002). Recent investigations have implied the possible presence of Dicer in the nucleus (Haussecker and Proudfoot, 2005). We have therefore reconsidered the issue of human Dicer localisation in Chapter 4 via microscopic and molecular experiments, as well as genomic analysis. We demonstrate that Dicer indeed localises and binds to the chromatin in the nucleus and I show that Dicer binding loci are associated with both dsRNA and small RNAs. The absence of Dicer does correlate with the accumulation of dsRNA. Nuclear Dicer function is another step into the direction of assimilation of RNAi pathways of mammalian and yeast cells and brings us closer to the possible existence of mammalian TGS, strengthening the case of *S. pombe* to be chosen as a model organism.

Chapter 2

Polyadenylation and Transcription Termination *cis* and *trans* Elements

This chapter is divided in two separate sections: Firstly, I present the Polyadenylation Signal analysis in *S. pombe*, which has been published under the title "Genome-wide analysis of poly(A) site selection in *Schizosaccharomyces pombe*" in the *RNA Journal* (Schlackow et al., 2013). I showed, that polyadenylation signals in fission yeast are not well conserved. It is therefore likely, that polyadenylation and transcription termination in fission yeast relies on additional *trans* factors. Previously it has been shown that Cohesin can promote transcription termination and our analysis supports previous findings. In the second section of this chapter, which contributes to a manuscript currently in preparation, I go into more detail of understanding the roles of Cohesin on the chromosome. These are primarily involved in cohesion of the sister chromatids and go far beyond the control of transcription termination

2.1 Polyadenylation Signals

Transcription of coding genes is an essential process for every cell. The nascent RNA is co-transcriptionally cleaved at its 3'end and then further modified by poly(A)

addition. This so-called poly(A) tail was discovered in 1973 (Winters and Edmonds, 1973a,b). Around that era it was also shown that transcription proceeded beyond the coding region and the existence of the 3'UTR was established (Proudfoot, 2011). Transcription continues after cleavage until a termination site is reached (Proudfoot, 1976), the transcript originating from the region between CS and termination site is then degraded. There is evidence that transcription termination and cleavage/mRNA 3' end processing are indeed interconnected and that Pol II CTD recruits cleavage/poly(A) factors, which aid transcription termination (Birse et al., 1998). It was also proposed that the CTD interacts with a 5' exonuclease, called Rat1 in yeast. Rat1 degrades the RNA, which is still attached to the elongating PolII after cleavage. Rat1 binds to PolII causing a conformational change at the PolII active site and hence promotes termination (reviewed in Proudfoot, 2011). This process is usually further enhanced by transcription pause sites (Gromak et al., 2006).

Polyadenylation of eukaryotic mRNA plays a critical role in its nuclear to cytoplasmic export, mRNA stability and translation (Zhao et al., 1999). The importance of this process is further emphasised by the fact that the 3' end processing factors are encoded by essential genes, whose mutation is lethal to the organism (Proudfoot, 2011). Apart from a few non-polyadenylated examples, such as the histone genes (Gick et al., 1986), failed polyadenylation usually leads to mRNA destruction (Zhao et al., 1999). Correct mRNA cleavage, polyadenylation and transcription termination depend on the multipartite polyadenylation signal (PAS, Gick et al., 1986; Proudfoot, 1989; Wilusz and Spector, 2010; Winters and Edmonds, 1973a,b), encoded in the DNA and recognised by a group of polyadenylation factors (Zhao et al., 1999). In mammals, these include the cleavage and polyadenylation specificity factor (CPSF) and cleavage stimulatory factor (CstF), which attach to the transcribed PAS on the RNA-copy (Figure 2.1) and promote pre-mRNA cleavage. The poly(A) polymerase (PAP) and the polyadenine binding protein (PAB) are responsible for synthesizing a poly(A) tail (a sequence of multiple A-abases)

at the end of the pre-mRNA to produce a mature mRNA molecule.

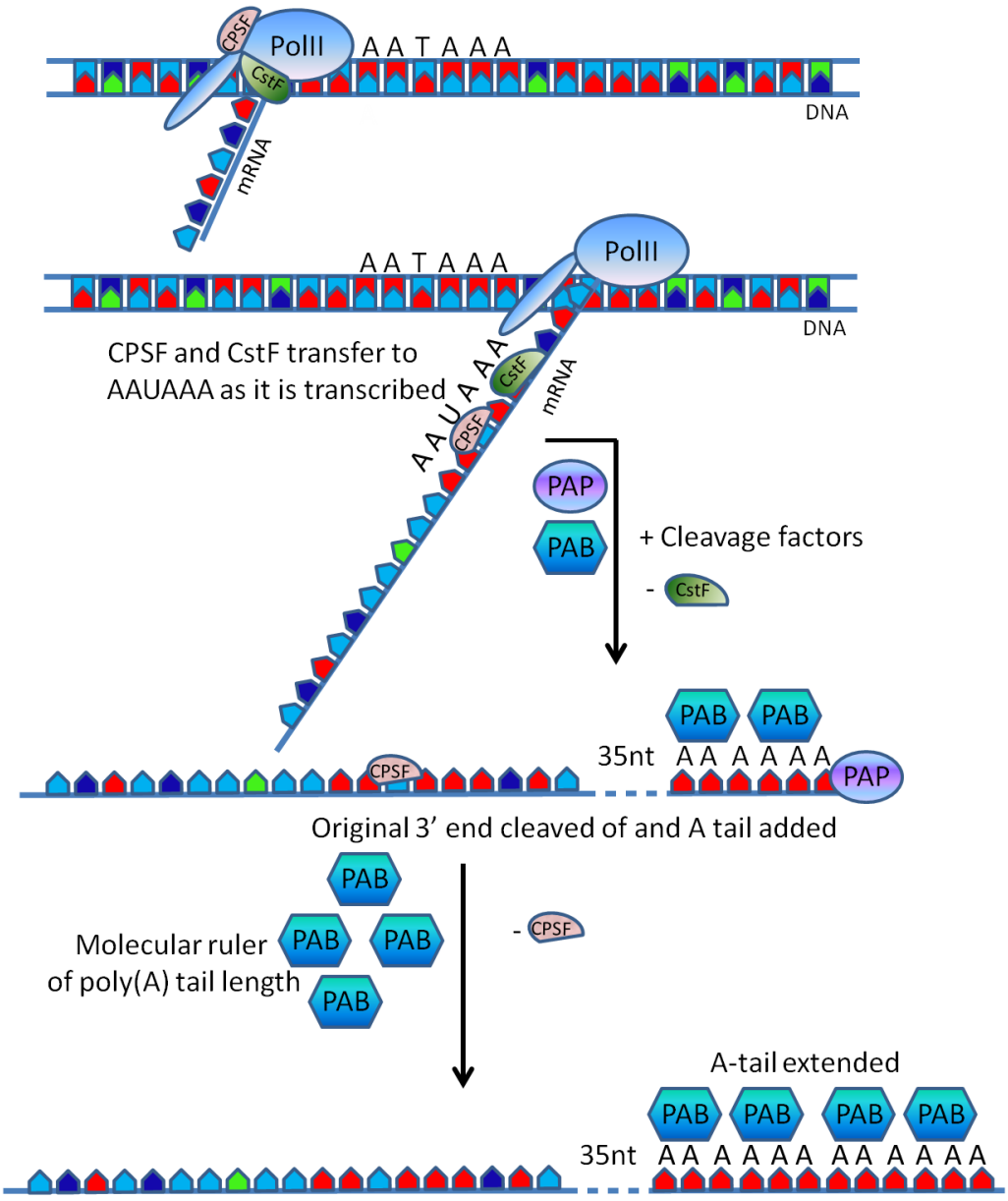

Figure 2.1: Illustration of the polyadenylation process. CPSF and CstF attach to the transcribed PAS on the RNA copy and promote cleavage. PAP and PAB synthesize the poly(A) tail.

PAS comprises an obligatory Near Upstream Element (NUE) usually A(A/T)TAAA (Proudfoot, 2011) and an auxiliary G/T-rich Downstream Element (DSE) in higher eukaryotes (Gil and Proudfoot, 1984; McLauchlan et al., 1985). By using EST data Beaudoing et al. (2000) established that nearly 80% of human genes have the canonical

NUE or a close variant. This illustrates how highly conserved human PAS are and hence how important functional polyadenylation is. It was later demonstrated that a strong T/GT-rich DSE does not require a canonical NUE, in fact an A-rich NUE suffices for efficient 3' end processing in human cells (Nunes et al., 2010). Similar predictions were made in *Caenorhabditis elegans* (*C. elegans*, Jan et al., 2011). PolII accumulation at the DSE causes PolII pausing and deceleration and hence promotes transcription termination (Gromak et al., 2006). However, to cause termination and cleavage, a functional PAS is necessary. These phenomena have been verified in *S. pombe* (Birse et al., 1998; Humphrey et al., 1991). Moreover, it has been shown that the distance between the NUE and the DSE is important for the strength of the terminator sequence (Aranda and Proudfoot, 1999).

In plants and budding yeast, the DSE can be replaced by a T-rich Efficiency Element (EE; Graber et al., 2002; Sherstnev et al., 2012; Zhao et al., 1999). Here the EE refers to a *cis* element near the CS. Note that various publication use this term differently. The cooperative function of all PAS-elements is crucial, as defective 3' end formation and termination due to mutation of the PAS may result in transcriptional read-through (Connelly and Manley, 1988; Dye and Proudfoot, 1999), with consequent gene silencing (Gullerova and Proudfoot, 2008; Gullerova et al., 2011). Furthermore, an absent or short poly(A) tail can mark the transcript for degradation, as shown in bacteria and plants (Chen et al., 2011; Dreyfus and Régnier, 2002; Lange et al., 2009; Steege, 2000). In higher eukaryotes, alternative PAS can lead to production of longer or shorter transcript isoforms (Tian et al., 2005). It was verified by EST analysis in human, mouse and rat that alternative PAS can act in conjunction with alternative splicing and exon truncation to yield alternative protein isoforms (Yan and Marr, 2005). PAS located within the coding region can result in pre-mature cleavage and transcription termination. Genome-wide studies reveal widespread alternative polyadenylation (APA) in yeast and human genomes (Ozsolak et al., 2010) that generate mRNA of different lengths, which may contain different regulatory regions. These alternative transcripts can encode different proteins

(di Giammartino et al., 2011) or selectively provide a binding platform for regulatory elements, such as miRNAs. Thus miRNAs may recognise a target seed sequence in a longer transcript, which leads to its selective degradation (Bartel, 2009; Mayr and Bartel, 2009). APA may occur at a specific cell cycle stage, or may relate to the developmental stage of the organism. This was established in *C. elegans* (di Giammartino et al., 2011; Mangone et al., 2010), as well the plant *Arabidopsis thaliana* (A. thaliana, Sherstnev et al., 2012). A related phenomenon to APA is CS heterogeneity: in mammals and plants (Sherstnev et al., 2012; Tian et al., 2005) multiple CS often exist in close proximity to each other, yet share the same PAS.

3' UTR sequences are different according to species. In higher eukaryotes a bipartite or tripartite signal encoded in the DNA is required for cleavage and termination. The NUE is located around 10-30 nt before the actual cleavage site. In humans the NUE is well defined by A(A/T)TAAA encoded in the DNA and in reasonably few cases variations of this consensus motif (~20%, Beaudoing et al., 2000). In many other organisms the A/T rich nucleotide composition of the 3'UTR is similar to the one in humans, as established in worm (Loke et al., 2005), fly (Retelska et al., 2006), plants (*A. thaliana* and rice, Loke et al., 2005; Shen et al., 2008a) and fungi (*Aspergillus oryzae* and *S. cerevisiae*). Surprisingly in the alga *Chlamydomonas reinhardtii*, whose genome is very G-C rich, the 3'-UTR is generally dominated by G. This implies a G- rich FUE, but the NUE region still shows an A/T richness (Shen et al., 2008b), with an apparent 5 nt NUE. Considering the G-C rich 3'UTR and the NUE-length, the alga appears to be an exception to the general picture portrayed by PAS in different organisms.

It appears that with lower complexity of the organism, also the level of conservation in NUE decreases. With AATAAA present in ~50% of human and fly sequences before the CS (Retelska et al., 2006), it decreases to ~40% in nematode, and to less than 15% in fungi and plants (Loke et al., 2005; Tanaka et al., 2011). Interestingly in the fungus *Aspergillus oryzae* the most dominant NUE motif is AATGAA, though present at less

than 10% of CS.

Even though fission and budding yeast are evolutionarily divergent and fission yeast is considered closer to higher eukaryotes it has been suggested that the RNA 3' end processing machinery of fission yeast has higher similarity to budding yeast than to humans (Humphrey et al., 1991). While *S. pombe* have human PAS present in their 3'-UTRs, they appear to be non-functional: two different human PAS were cloned into two *S. pombe* genes but appeared to be not recognised. On the other hand fission yeast PAS are recognised in budding yeast, as demonstrated by integrating an *S. pombe* pre-mRNA sequence into *S. cerevisiae* (Humphrey et al., 1991).

Nowadays, the fission yeast *S. pombe* has a fully sequenced genome. Its genomic organisation resembles that of higher eukaryotes and its high number of homologous genes related to human diseases promote investigation in *S. pombe*. A significant step towards understanding genomic function is the description of gene transcript 3' end formation and associated factors. In 2013 two parallel conducted studies have described CS and PAS in *S. pombe* at a genome-wide level: Mata (2013) developed an experimental method for effective extraction of CS (3PC) and our study (Schlackow et al., 2013) has mapped and analysed CS, which were obtainable from various sets of RNA-Seq data. In *S. cerevisiae* and *S. pombe* PAS are degenerate and other *trans* acting factors may be required for correct transcription termination. Thus it has been shown that convergent genes in *S. pombe* fail to terminate after proximal PAS, resulting in transcriptional read-through, producing overlapping, long mRNAs and consequently long double stranded RNA (dsRNA, Gullerova and Proudfoot, 2012). The Cohesin complex is recruited to chromatin in the S-phase of the cell cycle, followed by G2, where it is concentrated by PolII to intergenic regions between convergent genes (Schmidt et al., 2009). Here it blocks transcriptional read-through and promotes correct transcription termination (Gullerova and Proudfoot, 2008).

Only a few gene specific PAS, such as *ura4* (Humphrey et al., 1994), have been studied

experimentally in *S. pombe*. In contrast in this study, I have employed RNA-Seq data sets isolated from cells grown under different physiological conditions. I bioinformatically extracted polyadenylated reads and mapped them back to the *S. pombe* genomic sequence. My results re-annotate 3' UTRs of 4535 genes including extensive examples of APA and heterogeneity. Strand-specific RNA-Seq reads were used to analyse cleavage sites, polyadenylation signals, alternative polyadenylation and 3' end heterogeneity, showing condition-specific cleavage sites. Furthermore, I observed general preference for the canonical AATAAA PAS in fission yeast genome. I validated my genomic analysis experimentally using quantitative PCR (qPCR) analysis of the *ura4* gene expressed from a plasmid. Different PAS were positioned downstream of *ura4* coding region to measure relative efficiency. Changes in transcript levels correlated to PAS efficiency (Gehring et al., 2001). I have also examined, whether *S. pombe* genes containing AATAAA PAS, possess any significant functional similarity. Finally, I investigated the extent of overlapping transcripts derived from convergent genes on chromosome II and their coincidence with Cohesin peaks.

For the identified CS I have created a database *Pomb(A)* (www.pomba.co.uk) which allows review of all the mapped CS and the computed NUE. It also allows comparison of CS usage between conditions. Provided with the more extended scope of the 3PC data, I also included the CS identified by Mata (2013) into our genome browser. Finally, the user can also browse overexpressed motifs and user-defined motifs in their region of interest around the computed cleavage sites.

2.1.1 Materials and Methods

2.1.1.1 Data generation

Polyadenylated RNA molecules of various physiological conditions were subjected to Illumina sequencing by our collaborators at the Jürg Bähler lab at UCL, London. Data for non strand-specific cycling cells was obtained from Wilhelm et al. (2008), for strand-specific cycling cells and for quiescent (24 h) cells from Marguerat et al. (2012). A summary of their generation and Arrayexpress accession numbers can be found in Table 2.1.

dataset	**strand-specific**	**platform**	**read length (nt)**	**Arrayexpress Accession Number**	**reference**
Proliferating (Cycling)	yes	Illumina	51	EMTAB-1154	Marguerat et al. 2012
Proliferating (Cycling)	no	Illumina	trimmed to 30	EMTAB-5	Wilhelm et al. 2008
Quiescent (24 h)	yes	Illumina	51	EMTAB-1154	Marguerat et al. 2012
Quiescent (7 d)	yes	Illumina	76	E-MTAB-1824	Schlackow et al. 2013
meiotic	yes	Illumina	trimmed to 50	E-MTAB-1824	Schlackow et al. 2013

Table 2.1: Summary of all used data sets (including Arrayexpress accession number and references)

The meiotic pool and quiescent cells (7 d) were prepared in this study and sequenced with the following procedures (sequencing and sample preparation was carried by the Bähler lab, UCL, as described in Marguerat et al. 2012).

Meiotic Fission yeast *pat1-114* cells (*ade6-M210/ade-M216 pat1-114/pat1-114 h+/h+*) were grown to mid-log phase at 25°C in Edinburgh Minimal Medium (EMM). Cells were washed twice in EMM without nitrogen source (EMM-N) and cultured for 12 h in EMM-N at 25°C . The culture was then supplemented with 0.5 g/L NH4Cl and shifted to 34°C . Samples were collected for RNA preparation just before, and every hour for 8

hours after shifting the culture to 34°C . Equal RNA amounts from each time- point were pooled and used for sequencing library preparation.

Quiescent (7 d) 972 h- cells were grown to mid-log phase at 32°C in EMM. Cells were washed twice in EMM without nitrogen source EMM-N and grown for a further 7 days in EMM-N at 32°C .

2.1.1.2 Extraction of CS

Mapping RNA-Seq data Strand-specific RNA-Seq data from 2 combined data sets of cycling cells, meiosis arrested cells, quiescent cells after nitrogen depletion for 24 h and 7 d were filtered for polyadenylated reads (**Step 1**, Figure 2.2) and mapped back to the *S. pombe* genome. I have chosen my own mapping technique due to the high variability of mappable sequence before the poly(A)-tract. Current mapping software allows a fixed number of mismatches, while I have used a fractional mismatch allowance based on sequence length. The minimum of 5 consecutive adenine (A) residues (poly(A) tail) at the end of each read was taken as indicative of a polyadenylated read (compare Mangone et al., 2010). The same was done for non strand-specific data from cycling cells, but the original set of reads was also complemented and reversed. I permitted one non A in the poly(A) tail (before the final 5 nucleotides (nt)), as this could be a misread in the RNA-Seq procedure.

In **Step 2** each strand of each *S. pombe* chromosome was scanned for each sequence before the tail. The expected overlap of the sequence g with the genome ("correctness" for the sequence g) is given by $c(g) = \frac{l(g)-\sum_{i=1}^{l(g)} \epsilon_i - n}{l(g)}$, where $l(g)$ is the length of g before the start of the poly(A) tail, n is the number of unrecognised bases N, ϵ_i is be the error-probability derived from the quality score for base i (which is not N, nor in the poly(A) tail). In order to map a read g before the poly(A) tail to the chromosomes the minimal agreement between the two sequences must be at least $c(g)$. An agreement of 85% was chosen for computational efficiency, which corresponds to the expected smallest

"correctness" for at least 92% of the polyadenylated sequence reads (Step 2a). Due to the short length of reads in non strand-specific data this agreement was raised to 95% in that set. I refer to the first nucleotide after cleavage as the CS. If an agreement of 85% (or 95% in non strand-specific data) was reached, a preliminary CS was proposed under the following conditions:

- The CS was mapped to the chromosomal region, which follows any A-s in the chromosomal sequence overlapping the first nt of the poly(A) tail (Step 2b).
- If the nt of a proposed misread in the poly(A) tail is also found after RNA-Seq read mapping in the corresponding position of the chromosome (Step 2c) and the chromosomal sequence has 65% of A-s overlapping the poly(A) tail before the supposed misread, then CS is moved accordingly to after this falsely proposed misread. This step also occurs in combination with the previous condition. For both conditions the poly(A) tail length of the RNA-Seq reads is recalculated.

In **Step 3** I have eliminated internal priming by excluding sequences, whose recalculated poly(A) tail length was less than 5 nt. The chromosomal sequence must have less than 65% of A residues overlapping with the poly(A) tail, to eliminate the possibility that the read is not polyadenylated, but ends in a stretch of A-s, which are also found within the genome (in Steps 2 and 3 65% threshold was chosen because one sequence of correctness 35% has been observed).

In **Step 4** If a sequence maps multiple times to the genome, the map with the closest correspondence or closest downstream of an ORF was selected. See Figure 2.2 for an illustration of this data filtering procedure. CS were attributed to a gene, if they mapped inside its ORF or 1000 nt downstream of the stop-codon.

Elimination of heterogeneity Heterogeneity of CS was eliminated by grouping CS with an individual separation between two consecutive CS of at most 6 nt into either

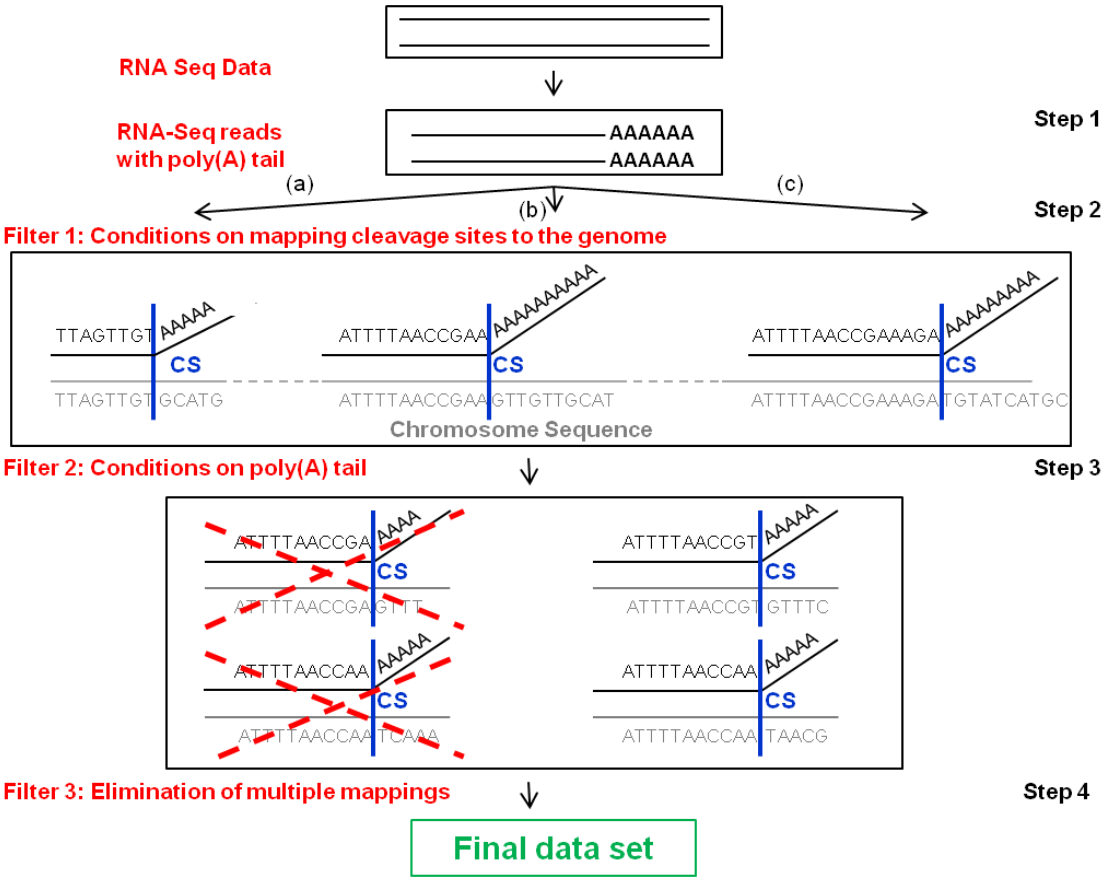

Figure 2.2: Illustration of the procedure to map polyadenylated RNA-Seq reads back to the genome. The complete description can be found in Materials and Methods.

the one with most RNA-Seq reads mapping to it, or otherwise the one with the shortest distance to the stop codon. The CS grouping based on an individual separation of 6 nt means that two different CS are at least 7 nt apart. I chose the individual separation for grouping to be 6 nt due to following biological and computational reasons. Firstly, my investigations of miRNAs in *S. pombe* (Chapter 3) suggest that possible miRNA existence in yeast should not be disregarded, as RNAi machinery is present and hairpin structures were shown to induce gene silencing. In higher eukaryotes, mature miRNAs bind to transcripts via imperfect base-pairing. The most important binding region for the ~22 nt miRNAs is a ~7 nt long seed region (Lewis et al., 2003). A 7 nt separation of CS would therefore be enough to provide a binding platform for a potential miRNA. Furthermore, in Figure 2.3A the total number of detected CS versus the individual

separation is plotted. It is apparent by visual inspection that the decrease in number of CS, becomes smaller after the individual separation of 6 nt is passed. Finally, the maximal distance, between first and last CS of one group, is more than 30 nt, with the individual separation of 6 nt (Figure 2.3B). It increases further, with increasing individual separation. CS of 30 nt apart are unlikely to use the same NUE, as the location of NUE is expected approximately ~30 nt before the CS.

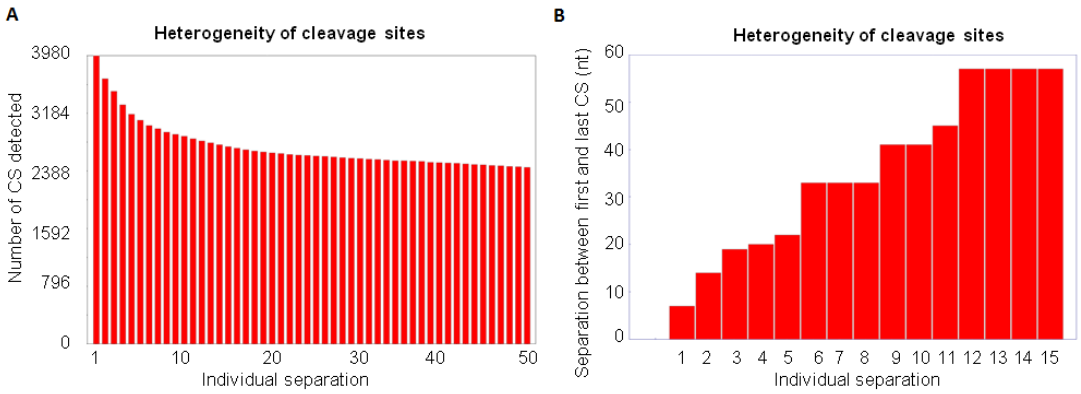

Figure 2.3: **A** Total number of identified CS after elimination of heterogeneity based on the individual separation. The difference in CS numbers decreases after an individual separation of 6 nt. **B** The maximal distance between the first and the last CS throughout the whole data set, which are grouped into one representative, is given on the vertical axis, while the individual separation is shown on the horizontal axis.

When considering CS positioning profiles with respect to their number per gene and order (Figure 2.12), unidentified sites following the ORF could affect the number of CS per gene, the order as well as the mean and standard deviation.

2.1.1.3 CS-usage and APA across physiological conditions

The number of polyadenylated RNA-Seq reads mapping to the CS, was counted for each strand-specific data set separately and normalised to the total number of RNA-Seq reads within the data set. Numbers were rescaled by a factor 10^8 to give the computed hit-scores in the CS-usage comparison across physiological conditions. A bar plot

across the whole genome of the hit-scores against CS coordinate was created and can be accessed in the developed web-interface *Pomb(A)* (presented in the Results Section 2.1.2 in Figure 2.17A and B). To compare CS-usage in meiotic and quiescent cells (24 h and 7 d) to cycling cells the absolute value of the difference in hit-scores for each CS was calculated and plotted as bars against CS coordinates. If the hit-score was larger in cycling cells, the bars are pointing upwards. For any other data set they are pointing downwards (see database *Pomb(A)*, www.pomba.co.uk).

The same scoring was performed for determination of cross-conditional APA. Genes with CS detected in both cycling cells and meiotic/quiescent cells (24 h or 7 d) were candidates for APA (if one of the conditions did not display a CS, this either speaks of condition specific genes or renders the cross-conditional comparison impossible).

Genes were considered to have alternative CS, if there were at least two distinct CS, one of which displayed greater usage in one condition and the other in the other condition. Greater usage means:

- If both conditions had the same CS identified, one of them must show at least a 3-fold higher hit-score compared to the other.
- If the CS was identified only in one condition, no CS was mapped within a distance of 20 nt around it to avoid it falling within the heterogeneity window of the other.

2.1.1.4 Motif-search

Within each region around the CS I consider the interval I_1 of interest of length l_1. A possible signal of length W was searched within I_1. I_1 was scanned for all possible motifs of length W consisting of 4 nt (A, T, C and G implying 4^W possible motifs). A profile of motif-frequency in each nt position close to the CS was created for each identified motif. The motifs were then ranked according to the difference between the maximal occurrence

within I_1 and the median in the 3000 nt region. The top 10 ranked motifs were re-ranked: The one with the smallest p-value (see below) was considered to be the most significant. All CS with the most significant motif in the interval of interest were excluded and the process was repeated on the remaining examples until 10 motifs were identified. These were then plotted for their distribution profile around the CS.

2.1.1.5 Statistical analyses

Alternative Polyadenylation To estimate the significance of alternative CS usage (APA), I have performed a permutation test and a Fisher Exact. The permutation test assumed the cleavage sites to be independent. Comparing CS-usage between cycling and (e.g.) meiotic cells, there are 6612 and 2000 polyadenylated reads respectively (as well as 9582 in quiescent 24 h and 8236 in quiescent 7 d). Looking at one particular CS, which encounters n_1 hits in cycling cells and n_2 hits in meiotic cells. The statistic was taken to be $\frac{n_2}{2000} - \frac{n_1}{6612}$. If the CS was not detected in cycling or meiotic cells, then $n_1 = 0$ or $n_2 = 0$, respectively. 1000 permutations were performed by splitting all polyadenylated reads of meiotic and cycling cells combined randomly into groups of 2000 and 6612 elements. This yielded a distribution for the statistic. If the observed value falls in the left tail of the permutation test distribution, then the CS is significantly more used in cycling cells. If it is in the right tail, then the CS is significantly more used in meiotic cells. The equivalent tests were performed for the quiescent data sets. The Permutation Test provides an approximation to the p-values. To compute them exactly, I also performed a Fisher Exact test based on the contingency table as illustrated in Table 2.2. Note that if the cycling cells use the particular CS more, I computed the p-value according to the left-tailed test, otherwise according to the right-tailed test.

All Fisher Exact Test p-values less than 0.1 can be found in Appendix Tables App.C.2-App.C.4.

	Meiotic	**Cycling**	**Total**
CS hits from poly(A) reads	n_2	n_1	$n_1 + n_2$
Other poly(A) reads	$2000 - n_2$	$6612 - n_1$	$8612 - (n_1 + n_2)$
Total	2000	6612	8612

Table 2.2: Illustration of the contingency table used for the Fisher Exact Tests

Motif Search Cumulative Binomial Distribution: p-values were calculated based on a first order Markov Model and a cumulative binomial distribution. Taking all I_1 next to each CS the probability P of a motif abcdef (each letter representing one nt) occurring in a fixed position was approximated by the first order Markov model

$$P = f(ab) \times \frac{f(bc)}{f(b)} \times \frac{f(cd)}{f(c)} \times \frac{f(de)}{f(d)} \times \frac{f(ef)}{f(e)},$$

where $f(x)$ is the frequency of the nt(s) x in all regions between stop codons and mapped CS. Then the probability p of the whole interval containing the motif was determined to be $p = 1 - (1 - P)^{l_1}$. From this the p-value for the observed number of intervals containing the motif was determined using the cumulative binomial distribution.

Improbizer (an Expectation Maximization algorithm by Jim Kent, UCSC) results are based on 512 3' UTR sequences in the region of interest around the CS for each dataset. The background model is based on a zero order Markov Model.

Fisher Exact Test: p-values were calculated based on the Text-NSP- 1.25 CPAN Perl Module.

2.1.1.6 RNA analysis, 3' RACE and PCR

RACE stands for Rapid Amplification of cDNA ends and can be either 5' or 3' end based. RACE is generally used to sequence full-length transcripts within a cell, but I used it to measure the distance of the cleavage site to a known region before. Initially the transcript

is reverse transcribed into cDNA by RT-PCR: this is done by priming the poly(A) tail by phased d(N)T, which is an oligonucleotide consisting of one C, G or A, followed by a stretch of Ts.

RNA from the *S. pombe* wild type 972 strain, was purified by acid phenol and treated by DNase I (Roche). The extracted RNA (2 µg for cleavage site verification and 500 ng for motif analysis) was reverse-transcribed by using the phased d(N)T oligonucleotide, 10 mM dNTPs, 1 µl Superscript III reverse transcriptase (Invitrogen), 4 µl 5×first strand buffer, 0.1 M DTT, in final volume 20 µl. Real time PCR reaction was set up with Sensimix (Qiagen) and a 1:2 dilution of the RT-PCR product. qPCR was performed with 250 ng of DNA. The amount of plasmid in each strain was quantified by using a primer specific to plasmid sequences in the PCR reaction. *ura4* expression was determined by the amount of detected *ura4* polyadenylated transcripts with respect to the amount of plasmid.

The cDNA was PCR amplified by gene-specific forward oligonucleotides (1 µl) and linker (1 µl) added to 5 µl of 10×Thermo Buffer, 75 mM MgCl, 10 mM dNTPs, 2 µl Taq Polymerase (Invitrogen) in final volume 50 µl. The PCR mix was cycled 40 times to: Heat activation at 95°C for 5 min, denaturation step at 95°C for 1 min, annealing step at 50°C -54°C (depending on primer Tm) for 1 min, elongation step at 72°C for 1 min, final elongation at 72°C for 5 min. 20 µl of the product was visualised on a 1.5% agarose gel. The exact protocol is described in Appendix A.

2.1.1.7 Cloning and *S. pombe* transformation

S. pombe cells were transformed with plasmids, which contained different PAS to analyse their effect on cleavage efficiency. A cloned *ura4* promoter and ORF served as a reporter. The terminator region on the plasmid was replaced by various PAS obtained from *ura4, pyp1, pep1, sid4 and pyp3* 3' UTRs. The exact transformation protocol is given in Appendix B. *ura4* promoter and ORF (*P-ura4*) were cloned into plasmid pJR1-3XH

(Figure 2.4) into PstI and XhoI restriction sites. *nmt1* was first cut out of the plasmid by inoculating it with the PstI and XhoI restriction enzymes. The outcome was run on a gel, the restricted plasmid was cut out of the gel (Figure 2.5) and recovered with the Gel DNA Recovery Kit.

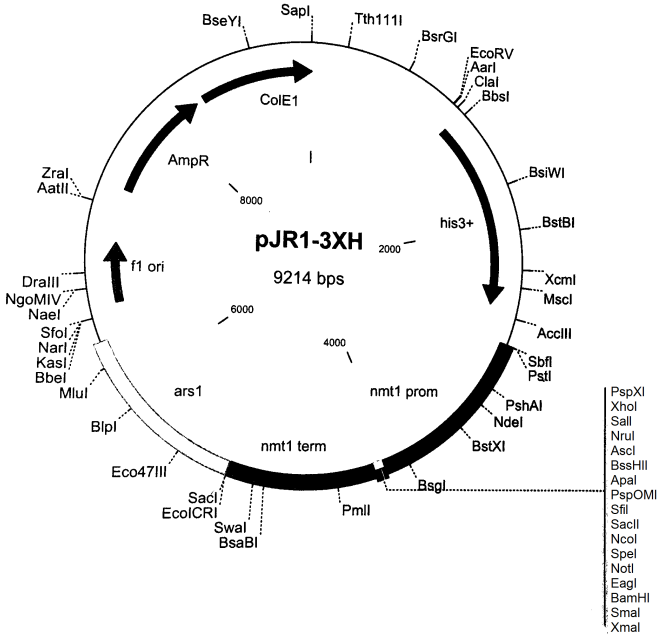

Figure 2.4: The plasmid which was used for transformations is shown. AmpR denotes the gene responsible for ampicillin resistance of bacteria. "*nmt1* prom" was replaced with *ura4*-Promoter and ORF (*P-ura4*), while "*nmt1* term" was replaced with different terminator regions. The restriction sites between "*nmt1* prom" and "*nmt1* term" ensure that these two fragments can be removed separately.

All inserts (*P-ura4* for "*nmt1* prom" region and various terminator regions for "*nmt1* term") were PCR amplified in purified *S. pombe* RNA (Appendix A.1). The agarose gel containing the PCR products after purification is depicted in Figure 2.6A. The PCR primers contained restriction sequences at the 5'end. These will ligate to the restriction sites on the linear plasmid sequence. The PCR products were purified with the PCR Purification Kit (Quiagen). T4 DNA ligase was used for ligation. The proportion of vector to insert (2 µl to 7 µl) was determined by the concentration of each of them (Figure 2.6B).

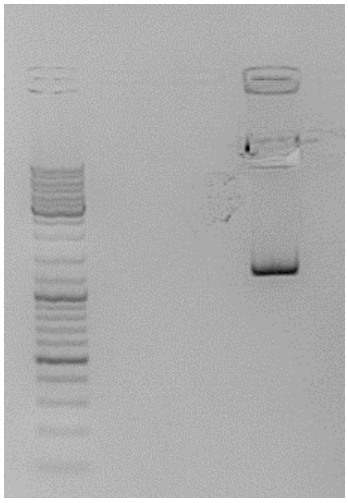

Figure 2.5: Depiction of restricted plasmid run on a gel. One can see the band for "*nmt1* prom" and the top band corresponds to the restricted plasmid and has already been cut out for gel DNA recovery

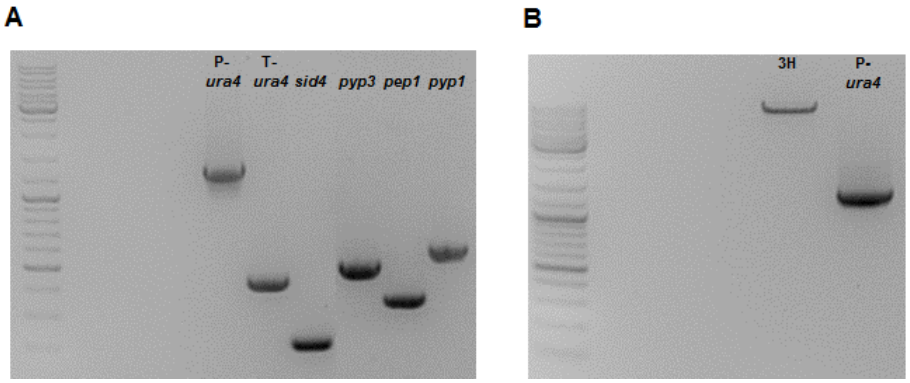

Figure 2.6: **A** PCR amplification of all products to be inserted in "*nmt* term". **B** PCR amplification of all inserts to be used, after PCR purification.

Bacterial XL1 competent cells were transformed (Appendix B) and grown on ampicillin plates. The bacteria are not ampicillin resistant, but the plasmid contains an ampicillin resistance gene. Therefore, growing colonies are the ones containing the plasmid. Colonies formed on the plates were diluted in LB medium, grown overnight and processed with the Qiagen Miniprep Kit to extract the DNA. This DNA was restricted with the previously used restriction enzymes and run on a gel. If two bands are observed, bacteria were successfully transformed with the plasmid that contains the required insert. If only one band is observed, the plasmid re-ligated before the insert incorporation. We

obtained three positive colonies (Figure 2.7).

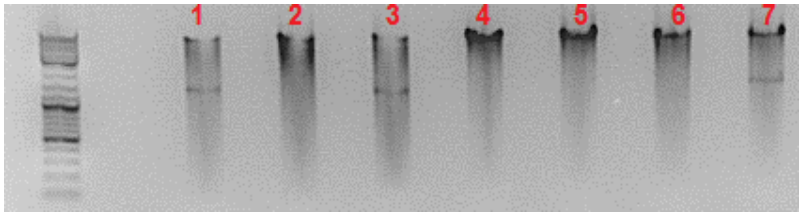

Figure 2.7: Selection of primer with a *P-ura4* insert. Two bands indicate successful replacement of "*nmt1* prom" with *P-ura4*. Samples used for the lanes 1, 3 and 7 were used for further experiments.

Different PAS (*T-ura4* denoting *ura4* PAS, *pyp1, pep1, sid4*) and a truncated *pyp3* terminator (described in more detail in the following Section), were PCR amplified from *S. pombe* genomic DNA and purified by Qiagen PCR Purification Kit, then they replaced the "*nmt1* term" by repeating the previous cloning process using restriction sites SacI and XmaI. Before isolating the DNA from the bacteria, colony PCRs were performed to detect the terminator inserts - if a band is present, the bacteria were further processed with the Miniprep Kit (Quiagen). An illustration is given in Figure 2.8A, only one out of four colonies was positive.

An *S. pombe* strain lacking endogenous *ura4* gene (ura4-D18) was grown in rich medium and transformed with plasmid constructs by LiOAc transformation (Bähler et al., 1998, Appendix B). Samples were plated on -His plates, to select positive colonies containing the plasmid (as the plasmid contains the *his3* gene). After transformation I performed a colony PCR to identify positive transforms. An example of the outcome is illustrated on the colony containing the *pyp1*-plasmid (Figure 2.8B).

Plasmids were submitted for sequencing and the outcome was checked for the correct *P-ura4* and terminator sequences.

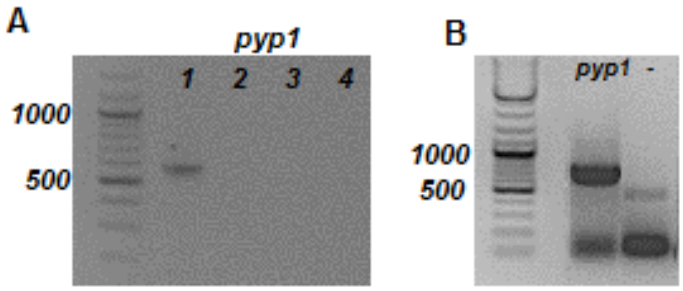

Figure 2.8: **A** One positive colony in lane 1 out of four overnight cultures showed a PCR signal for the *pyp1* terminator. This colony was further processed with the Miniprep Kit to isolate DNA. As before, the plasmid was checked with restriction enzymes and gel electrophoresis. **B** A successfully transformed yeast strain with a plasmid containing *P-ura4* and a *pyp1* terminator.

2.1.2 Results

Poly(A) reads from RNA-Seq data sets were extracted from RNA and mapped back to the *S. pombe* genome sequence to identify the CS (see Figure 2.2 in Section 2.1.1). The PAS analysis presented here uses strand-specific data for cycling cells in the main figures of the manuscript, while data from different growth conditions are summarised in the Tables App.C.2- App.C.4 in Appendix C. Results from all data sets are combined and accessible in our *Pomb(A)* database (www.pomba.co.uk), which also allows customisable motif search around the CS. I frequently observed multiple CS only a few nucleotides apart, a phenomenon I refer to as heterogeneity. Such heterogeneity is considered as a single CS to avoid motif overrepresentation (see section 2.1.1). Following these filtering criteria I identified CS in 4741 genes, out of which 4535 have CS in their 3' UTR (Figure 2.9). This corresponds to approximately 90% of all *S. pombe* genes. In the strand-specific data set of cycling cells 3093 CS were identified for 1964 genes, of which 843 were in tandem and 1121 in convergent orientation relative to their closest gene (Appendix Table App.C.1). CS for cycling cells (strand-specific and non strand-specific data), were verified by 3' Rapid Amplification of cDNA Ends (3'RACE, Figure 2.10). Three previously defined CS for *act1* (1190 nt, 1550 nt and 1800 nt downstream from the start codon, Rissland and Norbury 2009) were also detected in our data sets. Overall these data provide an independent validation of our computational approach.

2.1.2.1 Cleavage sites

I present the analysis from strand-specific data, as this allows direct comparison with other RNA-Seq data, generated using the same protocol, for cells grown under different conditions (meiotic and quiescent cells, see Section 2.1.1). While most *S. pombe* genes in cycling cells possess a CS following their ORFs (1964 out of total 2360 genes), CS within an ORF can also occur (925, Figure 2.9 and all other data sets in Table 2.3). However, such intragenic CS are underrepresented in ORFs and are very rare in introns (85 out of

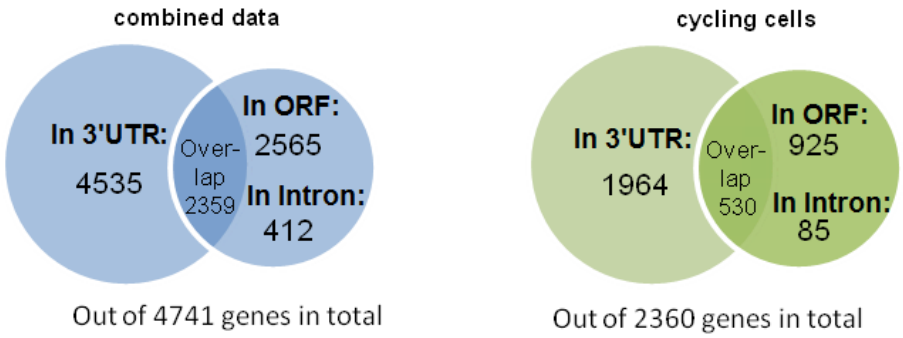

Figure 2.9: RNA-Seq data for all analysed data sets mapped to 4741 genes, out of which 4535 had the CS mapped after the stop codon and 2565 within the ORF. The strand-specific RNA-Seq data for cycling cells, which was analysed for significant PAS mapped to 2360 genes, of which 1964 had the CS mapped after the stop codon and 925 within the ORF. Out of these 85 mapped to annotated introns within the ORF.

925).

I show this underrepresentation by comparing the total number of detected internal cleavage events, to the total number of RNA-Seq reads, mapping to the same gene (Figure 2.11). Internal cleavage events increase with the number of RNA-Seq hits per gene. The number of these hits depends on gene expression level and length. More highly transcribed genes appear to result in more prevalent occurrence of internal cleavage events. Possibly degraded transcripts are detected as RNA-Seq hits, since these can be oligo-adenylated as part of the RNA turnover process (Schmid and Jensen, 2008). APA in 3' UTRs was observed in 36.6% of identified genes (Figure 2.12 and Table 2.4). As in mammals, *S. pombe* APA is common in 3' UTRs but occurs less frequently in ORFs, where it causes different exon lengths and consequent protein variability (Yan and Marr, 2005). 44.4% of genes are alternatively polyadenylated and I have plotted the average position of CS following ORFs, based on the number of alternative sites (up to six CS). The high standard deviations indicate that CS positions along the genes are highly variable (Figure 2.12). I also mapped the distance between multiple CS in one gene and show that they still occur in close proximity to each other (Figure 2.13).

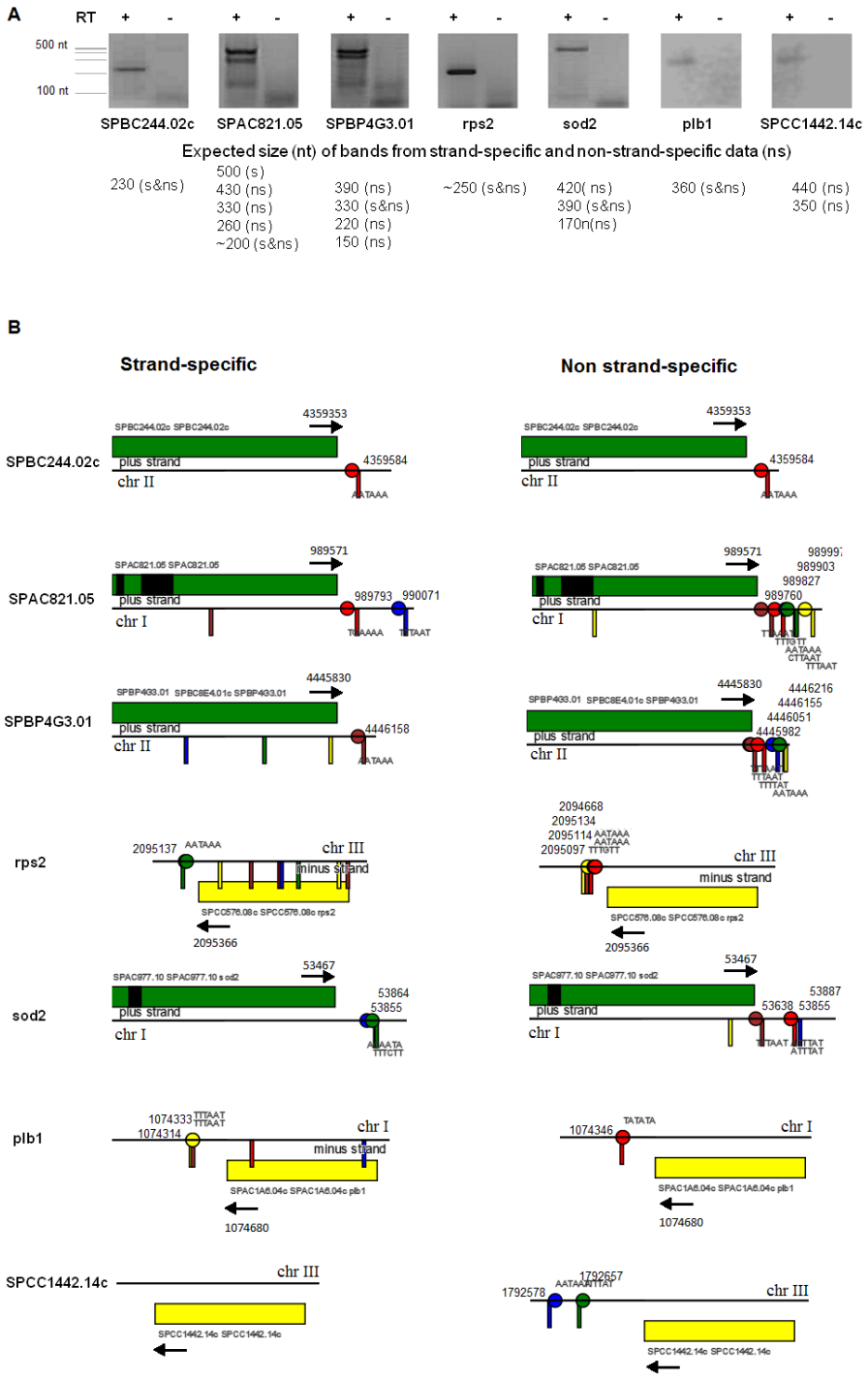

Figure 2.10: **A** 3'RACE analysis of mapped CS. Observed bands and predicted CS from strand-specific (s) and non strand-specific (ns) RNA-Seq data from cycling cells of size less than 500 nt are shown for seven randomly chosen genes (their names given below each image). + and − indicate the presence or absence of reverse transcriptase (RT) respectively. **A** Database snapshots of genes used in A of all mapped CS for strand-specific data (left) and non strand-specific data (right). All CS less than 500 nt away from the forward oligo (arrow) are indicated and presented with their NUE.

Genes / Data Set	Cleavage in 3'UTR	Cleavage in ORF	Cleavage in intron	Total identified	UTR-length Median, overall	UTR-length Median, tandem	UTR-length Median, convergent
Cycling, strand-specific	1964	925	85	2360	187.5	199	179
Cycling, non strand-specific	4198	821	204	4321	209	231	192
Meiotic	812	286	58	1041	196	210.5	186
Quiescent (24h)	2627	732	90	2903	204	227	191
Quiescent (7d)	2023	1354	97	2676	192	207	178

Table 2.3: Statistics on how many genes have an identified CS and how many had CS mapped outside and inside of the ORFs. UTR-length in the final three columns refers to the medians of the distances of CS from the stop codons.

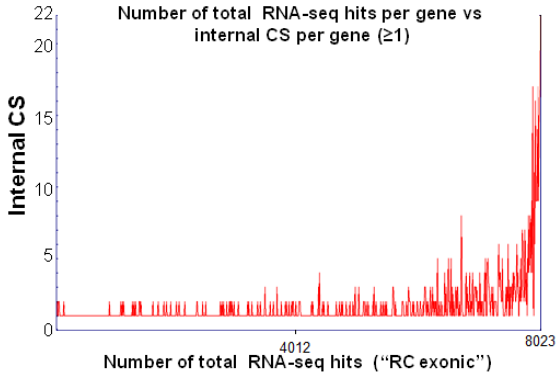

Figure 2.11: The number of RNA-Seq hits plotted against the internal cleavage number for genes which display internal cleavage (i.e. internal cleavage number presented on the vertical axis >0).

Multiplicity of CS	% from total genes identified									
	in 3'UTR					Overall				
	Cycling, strand-specific (1964)	Cycling, non strand-specific (4196)	Meiotic (811)	Quiescent (24h) (2609)	Quiescent (7 days) (2024)	Cycling, strand-specific (2346)	Cycling, non strand-specific (4319)	Meiotic (1041)	Quiescent (24h) (2903)	Quiescent (7 days) (2677)
1	63.4%	27.01%	83.62%	55.29%	63.42%	55.7%	24.97%	81.56%	51.41%	52.88%
2	24.0%	24.68%	12.93%	26.47%	23.28%	24.5%	23.79%	14.22%	26.14%	24.25%
3	7.3%	18.43%	2.34%	10.56%	8.26%	8.6%	18.23%	2.31%	11.90%	11.06%
4	3.3%	12.34%	0.62%	4.65%	2.57%	5.1%	12.80%	0.67%	5.55%	4.63%
5	1.2%	6.77%	0.37%	1.71%	1.53%	2.5%	7.98%	0.38%	2.41%	2.54%
6	0.7%	4.84%	0.00%	0.57%	0.54%	1.1%	4.95%	0.29%	1.17%	1.72%

Table 2.4: Alternative polyadenylation: the fraction of number of CS for the genes, which had CS mapped outside the ORF as well as mapped anywhere, is shown for all analysed data sets.

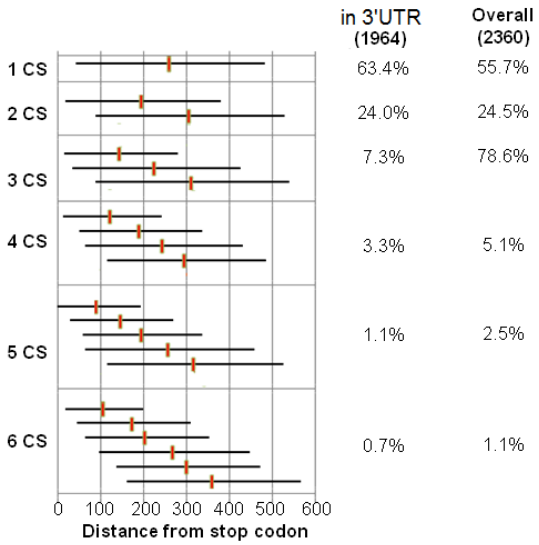

Figure 2.12: Average position of mapped CS relative to the stop codon dependent on the number and order of CS. Horizontal error bars correspond to one standard deviation. Percentage of genes with the displayed number of CS in the 3' UTR, as well as all identified genes are indicated.

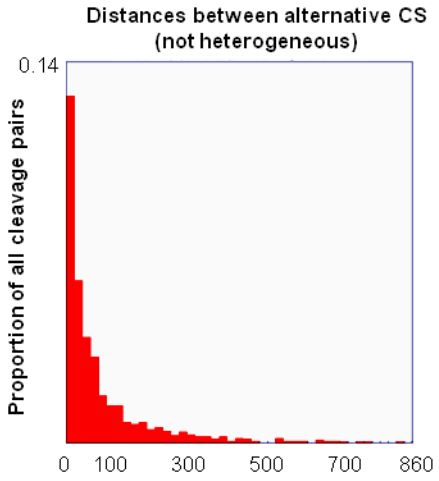

Figure 2.13: Distance between CS in genes with multiple CS after elimination of heterogeneity.

Next, I analysed the distribution of CS downstream of ORFs, using RNA-Seq CS-map coordinates, relative to gene orientation (Figure 2.14). In all tested data sets, median length for 3' UTRs is significantly longer between tandem genes compared to convergent genes (p <0.05, two-tailed student t-test to compare medians of 3'UTR lengths across different growth conditions). In detail, the overall median distance is 187.5 nt. Tandem

genes showed a slight preference for longer 3' UTRs (median 199 nt) compared to all genes (median 187.5 nt), while the convergent genes had shorter 3' UTRs (median 179 nt). By considering close CS (separated by 6 nt, see Materials and Methods in Section 2.1.1 for justification) as one, the large majority of CS (81.3%) displays no heterogeneity (Figure 2.15 and Appendix Figure C.1). Two CS are grouped in 11.2% of cases and less than 8% are groups of more than two CS. I also mapped the sequence distribution around all identified CS (Figure 2.16 and Appendix Figure C.2). Three interesting regions can be observed from this distribution, where the most frequent nucleotides A and T swap: region around -40 to -20 nt, containing potential NUE, and regions around -15 to -1 nt and 0 to 20 nt, possibly containing additional regulatory cis elements.

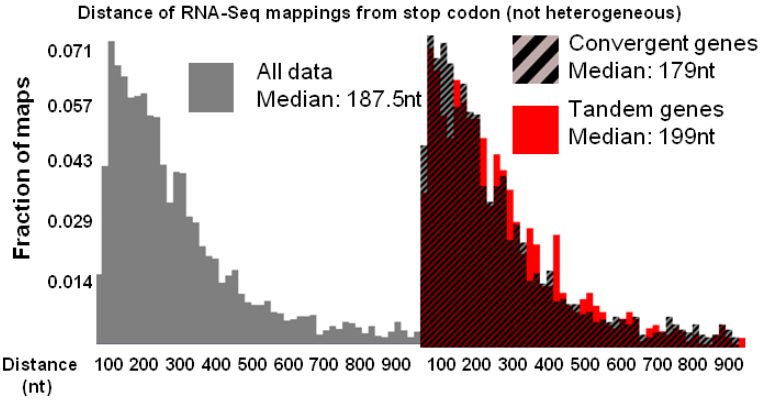

Figure 2.14: Positions of the RNA-Seq derived CS with respect to the upstream stop codon (<1000 nt). Left: outcome for all CS. Right: the outcome for tandem and convergent genes; data are normalised to respective total number of sequences.

2.1.2.2 Variation of cleavage site usage under different growth conditions

All mapped CS can be found in the *Pomb(A)* database, with the corresponding usage profiles under different growth conditions for any gene of interest. Here, I present a few illustrative examples. It is intrinsic to RNA-Seq data acquisition that certain sequence compositions will show a bias towards or against them. However, this bias is the same in every RNA-Seq experiment. One therefore cannot conclude if one CS is used more

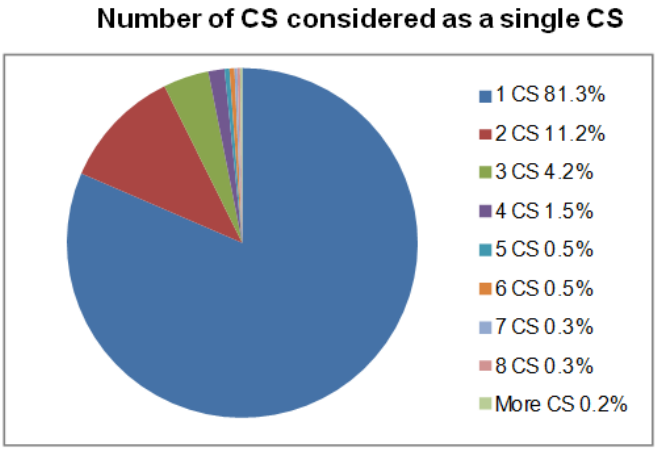

Figure 2.15: Heterogeneity of all CS: proportion of all CS numbers grouped into one CS for an individual separation of 6 nt.

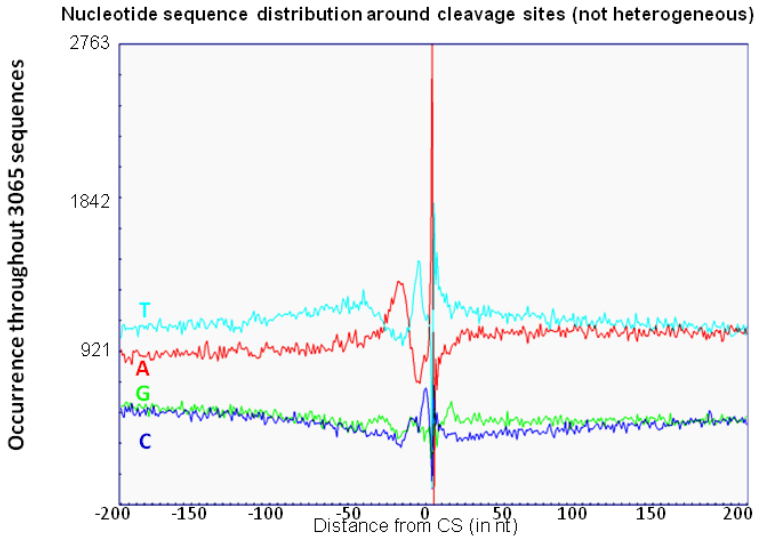

Figure 2.16: Nucleotide composition around mapped CS, denoted as 0 (1st nt after cleavage).

with respect to another CS. However, one can compare the usage of the same CS between growth conditions. I chose two examples from our database to illustrate the difference in CS usage under different growth conditions. The gene *SPBC16A3.02c* has multiple CS, detected in each physiological condition (Figure 2.17A). It is apparent that the major 3' end CS is predominantly used in quiescent cells (24 h), possibly because this gene is more expressed in quiescent cells than in cycling cells (Bähler lab Transcriptome Viewer,

Wilhelm et al., 2008). I also detected multiple CS for *meu4*, which has previously been described as a meiosis up-regulated gene (Watanabe et al., 2001) and its CS has been mapped 419 nt past the stop codon (Cremona et al., 2011). Our analysis confirms (within heterogeneity margin) and extends these findings by defining further alternative CS at low levels within the ORF (Figure 2.17B). However, we have shown in Figure 2.11 that the occurrence of internal cleavage events correlates with gene expression. *meu4* is highly expressed in meiosis, therefore the large amount of mapped internal cleavage events may be due to oligo-adenylated mRNA degradation products.

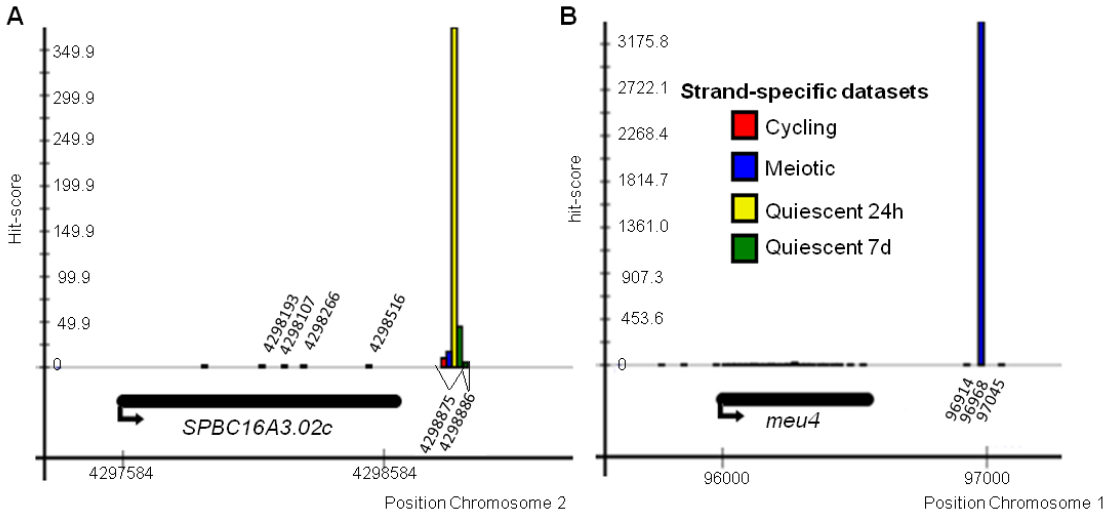

Figure 2.17: Snapshot from the *Pomb(A)* database. Comparison of CS-usage between data sets. The vertical axis presents a hit-score, which directly correlates with RNA-Seq hits for the depicted CS. The hit-score is obtained from the number of poly(A) hits normalised to sequencing depth and scaled by a constant (see Materials and Methods). The hit-scores are derived for all data sets and all identified CS in the strand-specific data. All CS positions can be found in the database, where the precise hit-scores and data sets can be viewed more clearly by hovering over the bars. The profiles correspond to the genes **A** *SPBC16A3.02c* and **B** *meu4*.

I show three examples of 3 APA associated genes under different conditions (Figure 2.18). The differences in PAS usage between cycling and meiotic cells (top), cycling and quiescent cells (24 h of nitrogen starvation, middle) and cycling and quiescent cells (7 d of nitrogen starvation, bottom) are depicted. The vertical axis marks the CS hit-score,

while the horizontal axis represents the distance from the stop codon of the corresponding gene. Several examples of APA are detected: the same CS is used in one condition, but not in the other (e.g. *SPAC27F1.07* in meiotic cells, top); the same CS is used in both conditions, but with different intensity (e.g. *SPAC343.09* in meiotic cells, middle) or a combination of both (e.g. *SPBC660.11* in meiotic cells, bottom). The full table with candidate alternatively polyadenylated genes, including CS position and distance from stop codon, hit number and hit-score can be found in the Appendix Tables App.C.2, App.C.3 and App.C.4. I have computed *p*-values to estimate the significance of alternative PAS usage. I have performed a Permutation Test and a Fisher Exact Test and the results are in agreement. Full details of how either was performed can be found in the section 2.1.1. All Fisher Exact Test *p*- values less than 0.1 are reported in the Appendix Tables App.C.2, App.C.3, App.C.4, in the columns corresponding to the data set where the CS appears to be used more. One should note that many CS are determined by few reads ($\leq$2), many alternative CS are therefore reported without a significant *p*-value. Nevertheless, few reads do speak of a cleavage event and are reported as potential alternative CS, which require further experimental validation.

2.1.2.3 *Cis* elements close to CS

It is well established that *cis* elements play an important role in cleavage and polyadenylation of mRNA (Proudfoot, 2011). I therefore searched for the most significant motifs of lengths 4-10 nt by scanning the 25 nt after the CS (therefore avoiding regions containing the potential NUE before the CS). The motif is ranked based on an initial filtering of overrepresented motifs using a method employed by Loke et al. (2005). This method ranks the motifs according to the difference of the maximum occurrence of the motif at a position in the scanned 25 nt region, compared to a large interval occurrence median (see section 2.1.1). The top motif candidates are re-ranked to produce one top candidate according to the *p*-value of a cumulative binomial distribution throughout all

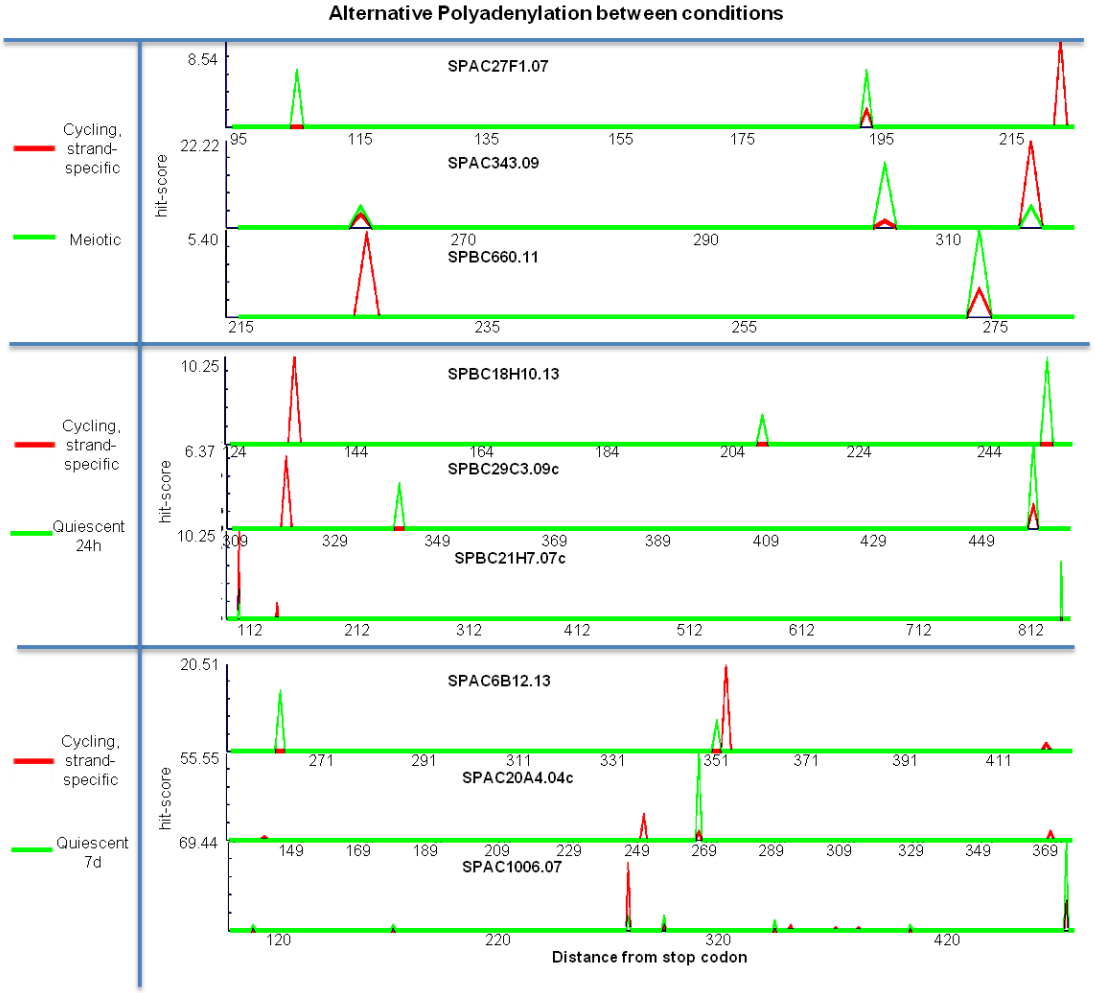

Figure 2.18: Examples of alternative polyadenylation between cycling and meiotic cells (top), cycling and quiescent cells after 24 h (middle) and 7 d (bottom) of nitrogen starvation (see Results and Materials and Methods section in the text for details). The hit-score on the vertical axis is obtained from the number of poly(A) hits normalised to sequencing depth and scaled by a constant (see Materials and Methods). The full table of APA associated genes can be viewed in Appendix Tables App.C.2, App.C.3 and App.C.4.

analysed sequences. This mimics the method used in the human PAS analysis performed by Beaudoing et al. (2000), which was also used in PAS analyses of other organisms (Sherstnev et al., 2012). The top 3 motifs of 4-7 nt length are summarised in Table 2.5. The most frequent motifs are T/G rich, and T-stretches. To verify our computational approach I used the Improbizer Expectation Maximisation Algorithm, which converged to TGTA, as did our own method (Table 2.6). An Expectation Maximisation Algorithm

(MEME) has also previously been used for PAS analyses (Ozsolak et al., 2010; Retelska et al., 2006). In the initial motif filtering, T-stretches are always present and they also display an interesting positional bias, before and after the CS (Figure 2.19). All other presented motifs in Figure 2.19 are over-represented hexamers in the 25 nt past the CS. Since the EE is also expected to occur before the CS (Loke et al., 2005), I have plotted a 40 nt window centred around the CS. TGTA containing motifs rank high among the overrepresented hexamers. Combining this with the Improbizer result, I also plotted 4 nt motifs with small *p*-values (Figure 2.20). TGTA motif displays a large peak just before the CS. G and T seem to be of particular importance for the significant motifs around the CS. This is also apparent from the rise of T and G in the nucleotide distribution after the CS (Figure 2.16).

Mot. size	Motif	P-value	% of sequences	Number of genes
4	TGTA	2e-221	24.0	619
	TTTA	2e-87	28.4	737
	TTTG	1e-58	14.1	380
5	TTGTA	3e-89	10.4	288
	TTTAT	2e-33	12.6	345
	AGTAA	4e-52	5.3	153
6	TTTGTA	4e-43	4.8	136
	TTGTTT	2e-21	4.8	134
	ATGTAA	3e-20	2.3	66
7	TTTTTGT	2e-15	2.5	71
	TTTGTAT	2e-11	1.4	38
	TGTAATT	5e-9	1.1	29

Table 2.5: Analysis of potential motif sequence mediating cleavage and polyadenylation. A 25 nt region after the CS was scanned for significant motifs. The top 3 ranked motifs according to lowest *p*-value of lengths 4-7 nt.

2.1.2.4 Upstream Polyadenylation signals (NUE)

I scanned the region -50 to -6 nt (from the CS) for the most significant motifs of length 4-10 nt using the same method as described previously. I summarise the top ranked signals

TGTA				
A	0.271	0.003	0.266	0.657
C	0.153	0.003	0.216	0.006
G	0.097	0.990	0.003	0.003
T	0.279	0.003	0.515	0.334

Table 2.6: Analysis of potential motif sequence mediating cleavage and polyadenylation. A 25 nt region after the CS was scanned for significant motifs. Improbizer Maximisation Expectation Algorithm result when scanning the above indicated region for overrepresented motifs.

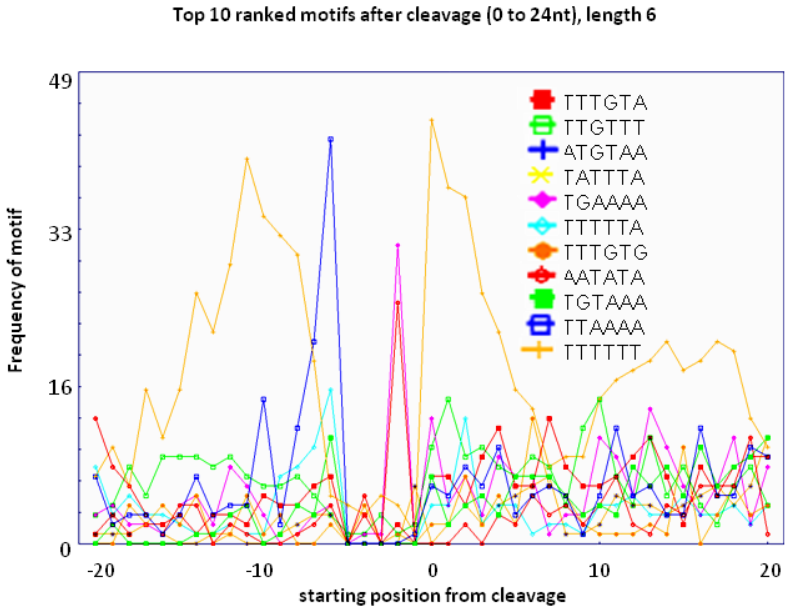

Figure 2.19: The distributions of the top 10 ranked motifs of length 6 nt around the CS (located at 0) are shown. The location of the motifs relates to the start of the motif. The A-rich peak before cleavage is partially due to the analytical procedure of considering the Adenosines as transcribed and not part of the added poly(A) tail. The T-stretches always rank highest in the preliminary ranking. This ranking is based on **max**(T-stretch)-**med**(T-stretch), where max denotes the maximal occurrence frequency at a nt-position (0 to 25 nt), while med denotes the median occurrence frequency over 3000 nt. The T-stretches were hence included in the analysis and plotted.

of lengths 4-7 nt in Table 2.7. The canonical human AATAAA NUE has the smallest *p*-value among all other hexamers, while the most dominant motifs of any other length are either sub-sequences of or contain AATAAA. This is supported by the Improbizer

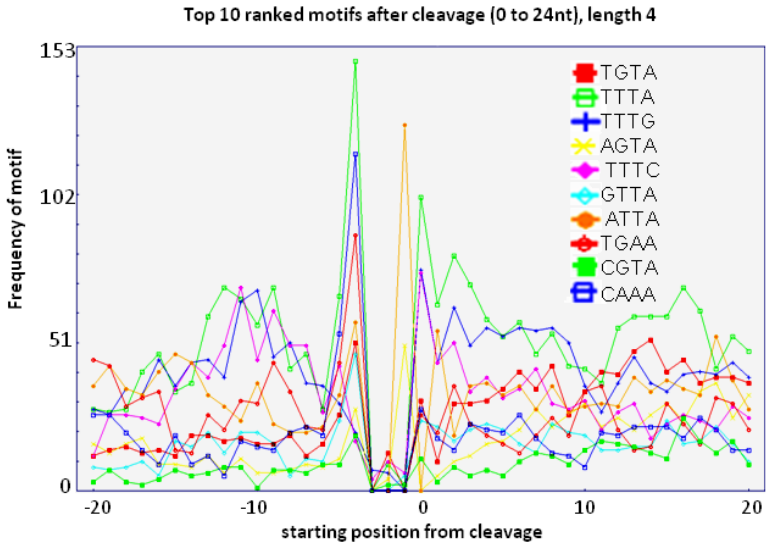

Figure 2.20: The distribution of the top 10 ranked motifs of length 4 around the CS (located at 0) is shown, as TGTA is the motif with the smallest *p*- value of all and the Improbizer outcome. T-rich EE have been shown to occur before and after the CS, so the distribution before the CS is also plotted.

outcome (Table 2.8, Appendix Table C.6 for all other datasets). These data represent a notable similarity between fission yeast and human. Due to high frequency of T-stretches around the CS, I tested how many PAS contain an AATAAA motif before the CS and how many sequences have a T-stretch around the CS (Figure 2.21). The low p-value (<0.05) shows that the simultaneous occurrence of both motifs is rare implying mutual exclusivity.

I show the distribution of the top 10 NUE motifs of 6 nt length in Figure 2.22 (Appendix Figure C.4). The NUE in *S. pombe* are located around 29 nt upstream of the CS, and the most frequent hexamer motif is AATAAA. I calculated the proportion of genes containing each of the top 10 motifs (exclusively) in their 3' UTR upstream to the CS (Figure 2.22, Appendix Figure C.5). While AATAAA is the most frequent NUE in *S. pombe* (as in humans) the next 9 most common NUE variants are different to human NUE variants (Beaudoing et al., 2000, Figure 2.23 A and B). A striking difference is the drop in rank and positional specificity of the AATGAA motif (also apparent in the Improbizer outcome in Table 2.2.6) in fission yeast (standard deviation of 7.9 nt) compared to humans

Mot. size	Motif	P-value	% of sequences	Number of genes
4	AATA	0	60.8	1405
	AATG	2e-102	14.9	382
	AATT	8e-63	14.8	395
5	AATAA	1e-316	30.6	772
	AATAT	2e-127	16.6	428
	ATGAA	1e-91	9.4	247
6	AATAAA	5e-240	16.9	436
	AATGAA	1e-110	8.1	214
	TTTAAT	5e-25	7.5	202
7	AATAAAA	2e-153	8.6	224
	AATAATA	8e-69	4.4	120
	TAATGAA	1e-62	3.6	96

Table 2.7: Analysis of potential PAS. The region from -50 nt to -6 nt was scanned and motifs were ranked as before. The top 3 ranked motifs according to lowest *p*-value of lengths 4-7 nt.

AATAAA						
A	0.927	0.928	0.003	0.534	0.465	0.551
C	0.003	0.042	0.017	0.089	0.141	0.110
G	0.048	0.027	0.016	0.227	0.147	0.168
T	0.022	0.003	0.965	0.150	0.247	0.171

Table 2.8: Improbizer Maximisation Expectation Algorithm results when scanning the region 50 to 6 nt before the CS for overrepresented motifs.

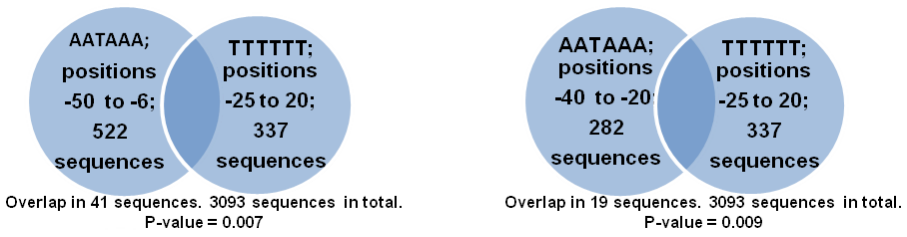

Figure 2.21: Illustration of how many sequences have the proposed most dominant NUE and the supplementary signal. The *p*-values of such an overlap were computed by a left-sided Fisher exact test.

(standard deviation of 10 nt). I also analysed, whether in alternatively polyadenylated genes, proximal and distal sites show differences in PAS preference, but none was observed (Appendix Figure C.3). This is in contrast to mammalian APA, where the

upstream PAS tends to be a non-canonical signal (Beaudoing et al., 2000). This analysis can be repeated for any of the described data sets and any region of interest in the provided database *Pomb(A)*.

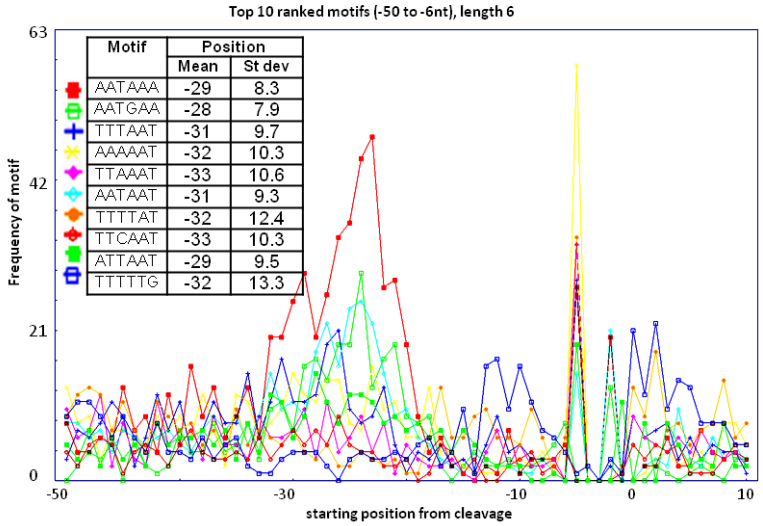

Figure 2.22: The distributions for the top 10 NUE motifs of 6 nt length ranked in the region -50 nt to -6 nt before the CS.

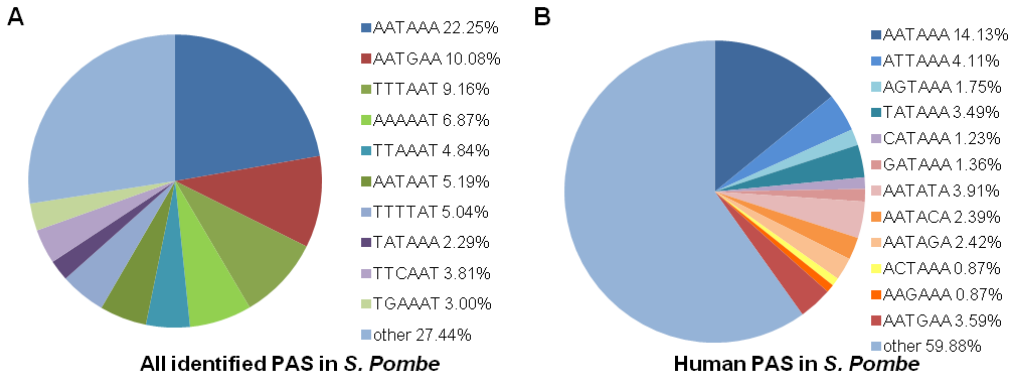

Figure 2.23: **A** The distributions for the top 10 NUE motifs of 6 nt length ranked in the region -50 nt to -6 nt before the CS.The proportions of genes with the detected potential NUE. Genes containing a higher ranked NUE are not considered for the lower ranked series. **B** For comparison, the proportion of genes with human NUE is shown. Again, genes containing more significant NUE in humans are not considered for the less significant ones.

2.1.2.5 Functional analysis of identified polyadenylation signals

I wished to verify the functionality of the computationally identified *S. pombe* PAS. Cells lacking the endogenous *ura4* gene were transformed with pJR1-3XH plasmid constructs, comprising a *ura4* promoter, *ura4* ORF and one of 5 different PAS as follows: wild type (WT) *ura4* PAS (positive control), no PAS (negative control), *sid4* PAS containing AATAAA, *pep1* PAS containing AATGAA, and an artificial PAS derived from a *pyp3* 3' flanking region fragment (Figure 2.24 and 2.25). This sequence consists of the 3' UTR following the mapped CS by RNA-Seq, which possesses several different potential PAS motifs (Figure 2.26 top). This allows us to perform a competition experiment to test which PAS motif is preferentially used (this is provided the proximal NUE candidate is not the strong signal implying that the distal weak signal of another NUE candidate would not be used). I used RNA isolated from cells transformed with each *ura4* plasmid construct. For the artificial *pyp3* PAS plasmid I detect bands corresponding to the two canonical NUE AATAAA sequences (Figure 2.26, bottom) with the shorter being more dominant. No products corresponding to AATGAA or the T-stretches were detected. To measure PAS efficiency, I performed a qRT-PCR experiment to quantify the amount of *ura4* transcription derived from each PAS containing plasmid, since PAS efficiency determines mRNA levels (Gehring et al., 2001). The cloned *ura4* PAS contains all sequence elements necessary for efficient 3' end formation (Humphrey et al., 1994) and produces the highest level of functional *ura4* transcripts (Figure 2.27). The most dominant CS of *ura4*, when located on the plasmid is preceded by an AATAAA motif (3'RACE, Figure 2.28). Similar levels of *ura4* mRNA were detected in strains with AATAAA PAS (*sid4* and *pyp3*, Figure 2.27). All *ura4* expression levels are based on three biological repeats with a *p*-value $p < 0.05$ versus the negative control, by a two-tailed student t-test. Finally, the strain containing the second most significant motif AATGAA (Table 2.7) shows *ura4* transcription levels very similar to the negative control.

One should note that these experiments indicate the efficiency of the whole PAS in the

terminator region (including EE and DSE) and not only of the proposed NUE. Moreover the qPCR is only indicative of PAS efficiency, but not conclusive, as this experiment does not provide a quantification of cleaved vs un-cleaved message at the CS and is influenced by isoform stability .

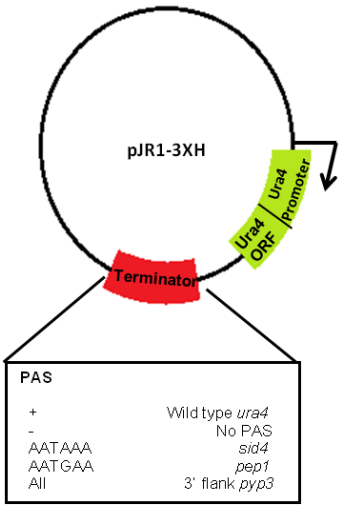

Figure 2.24: Functional analysis of identified PAS. Plasmid pJR1-3HX was modified to contain a *ura4* ORF and promoter. Indicated PAS candidates were cloned into the 3' region located 83 nt downstream of the *ura4* stop codon.

ura4
AAAAAGACTAATGTAAAATTTTTTTGGTTGGTTATTGAAAAAGTCGATGCCTTGTTTGCGTTTGTTTTCCTAGGCGTTTTATGTCAGAAGGCATTTAGAATTAGTATACAAG
TACTCTTTGGTAAAATTTTATGTAGCGACTAAAATATTAACTATTATAGATAAACACCTTGGGAATAAAAAGTAATTTGCTATAGTAATTTATTAAACATGCTCCTACAAC
ATTACCACAATCTTTTCTCTTGGATTGACATTGAATAAGAAAAGAGTGAATTTTTTTAGACTTGTAATGATAACTATGTACAAAGCCAATGAAAGATGTATGTAGATGAAT
GTAAAATACCATGTAGACAAACAAGATAAAACTTGGTTATAAACATTGGTGTTGGAACAGAATAAATTAGATGTCAAAAAGTTTCGTCAATGTCACAAGCCTGCA

sid4
ATGGATGCCAATGACTTACAACTCGATCAATAAGATTTTGGGATGCAGCTTATATTATACAAATGTTGATAAATATGAAATATATTATTAACTACTGACTGAATGAAAGAG
GCGTTAAATAAATCGATAGTAATAATAAGCGTATGGCAAAATGTACTATATTATGGCATCCTTTGCATAAAGAATTTTATCCTGCTGCTACACATC

pyp3 (sequence following mapped CS)
TAATTTGCTGCATGGAATAAAATTGACAAAGAAGTAACCACGAATTTAAATGTTCATCGCAAAAAAGTTGAATGAACGCTTTGGCAATTCTTTTAATTTTCTTGACCACGT
ACCATTTGACGTGCAAACTATTAGTACTTTGGCATATAGTGAAGCTGGGTATCTGTGGTTATTGATATTAATACTGTTTAGTACATATAAAAACCTTGTCAAGTATGAAATT
CAACTTTCCCATTTATTCCATTACATATCATATAAATTGCTATAAAAATAATAAACAAACAATAAACTTCGTTTAAGGTACTGAATTAAAAGTGCTATTCGGTAAGTTTAGT
CGCTTGTTGGTCGATACTACGATCCGCTAAGCCCAACAAGGCTGGTTCAGAGGAAGTAGCTTCTTTAACCGTACTACCCACTACTTCTATTACAGACAGTCGTTTTCACTG
TTTTTTCCATAAAGTTGTTGATTGAATGCAATTATTAAATTCAAAAAA

pep1
TTTAACTTAACCTGGGCATGTTTGTCAATCATTCCACTCATCATGTAAAGCAGCTATGCCACGTTCTTATATTAATGATCAGTTGATATATATGGTAGTTTAATGTATAAATA
TATATATTTACATACATACACTCATGCACATACGTGTAGTATAATGAACGAAATTTTCTTTCATGCCAACGTTGCAACCTGTCATTTATTACATAAGATATCAAACAAAGCT
TTCTAATACGTTTCTAGTTCGAACAGCTAAAGTGACCAGGCAACAAAATTTTGAAGTTTTTTAATTCATTTGCGGGATAAACCTATGAAATCTCATAAAGAATGGCGTAAT
AAAGGAAATTAT

Figure 2.25: Inserted PAS sequences. Blue, larger, underlined: CS. Red, underlined: Proposed NUE. *ura4*: RNA-Seq yielded first CS, previously identified other two CS, PAS and EE (red, underlined).

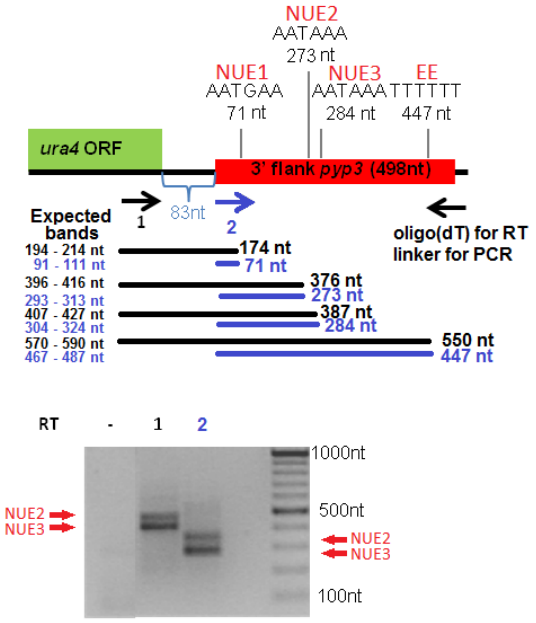

Figure 2.26: 3'RACE set up for the artificial PAS consisting of a *pyp3* 3' flanking region fragment. Primers: 1 20 nt before the end of the ura4 ORF; 2 at the start of the PAS. Distances for possible NUE from the start of the PAS are marked. 83 nt denote the distance from the *ura4* ORF to the inserted terminator, where the restriction sites were located. The distances from the primers to the potential NUEs are indicated. Expected bands are about 20 - 40 nt larger than the distances between primers and NUEs. The 3'RACE outcome is presented, where "-" is the control lane without reverse transcriptase (RT), lane 1 corresponds to Primer 1 and lane 2 to Primer 2 in the diagram above. Reverse primer was oligo(dT) for the RT-PCR and a linker for the PCR amplification (linker is a VT_n oligonucleotide, where V stands for a non T nucleotide and T_n for a stretch of Ts)

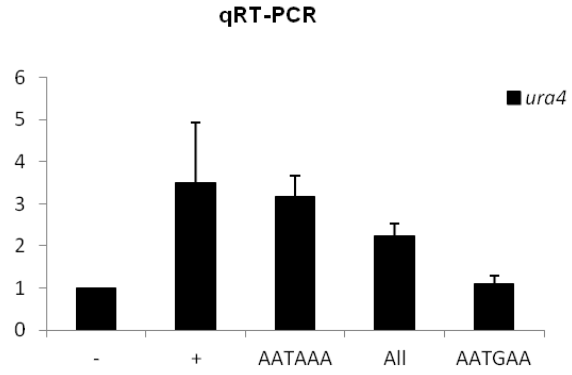

Figure 2.27: qRT-PCR outcome of levels of *ura4* transcription.

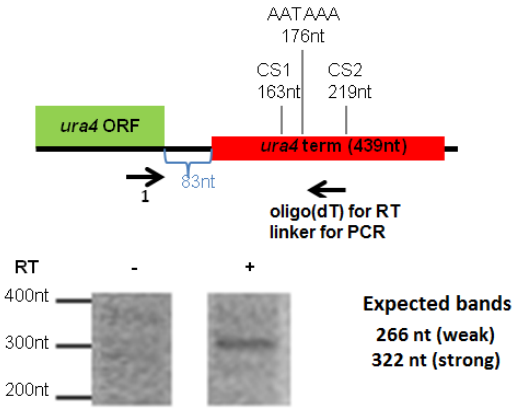

Figure 2.28: 3'RACE experiment set up for *ura4* transcription on the plasmid. Forward primer was 20 nt before the end of the *ura4* ORF. Band corresponds roughly to the second CS, which follows AATAAA. The 3' RACE outcome is presented below. Only the expected strong band is observed. Reverse primer was oligo(dT) for the RT-PCR and a linker for the PCR amplification (linker is a VT_n oligonucleotide, where V stands for a non T nucleotide and T_n for a stretch of Ts)

Next it was investigated whether genes possessing particular PAS can be grouped based on their biological function. Genes containing each of top 10 PAS were compared against ~3700 GO lists (Gene Ontology annotation lists, this analysis was performed in collaboration with Samuel Marguerat at the Bähler lab, UCL). Only genes containing AATAAA showed significant enrichment in GO category. This class of genes is associated with general cellular functions, mainly relating to translation (Table 2.9). Genes with important cellular functions are constantly expressed within the cell. This may point at AATAAA being an important regulator to ensure that essential genes have efficient 3' processing to create functional transcripts.

2.1.2.6 Analysis of overlapping transcripts derived from convergent genes

Transcription of some convergent genes in *S. pombe* fails to terminate after their proximal PAS in the G1 phase of the cell cycle. This results in transcriptional read-through, overlapping transcripts and consequent long dsRNA production. This in turn may activate the RNAi pathway and leads to gene silencing. It has previously been proposed that the

Motif in S. pombe	Function	P-value
AATAAA	Growth module (Chen et al.)	2e-19
	Cytoplasmic translation	2e-17
	cytosolic large ribosomal subunit	4e-11
	translation	6e-09
	cytosolic small ribosomal subunit	1e-08
	ribosome biogenesis	3e-05
	nucleolus	0.003
AAAAAT	Growth module (Chen et al.)	8e-04
All other motifs	P-value ≥0.01	

Table 2.9: Comparison of NUE containing genes to ~3700 GO lists. The *p*-value was computed by Fisher exact test and corrected for multiple testing by FDR. This analysis was performed in collaboration with Samuel Marguerat.

Cohesin protein complex promotes transcription termination between convergent genes in G2 (Gullerova and Proudfoot, 2008). Cohesin is loaded onto chromatin in S phase and is pushed by an active RNA polymerase II. Sites of Cohesin localization on the chromosome are nearly exclusively between convergent genes (Schmidt et al., 2009). Overlapping transcripts (overlap <50 nt) between convergent genes for 37 loci are identifiable on chromosome II in cycling cells. It is illustrated in Figure 2.29 that transcripts overlapping at Cohesin sites are smaller than at sites lacking Cohesin (student t-test, p <0.1). Most of cycling cells in *S. pombe* are in the G2 phase of the cell cycle with only 10% of cycling cells in G1. It is possible that the larger overlaps at Cohesin sites are derived from cells, which are in G1.

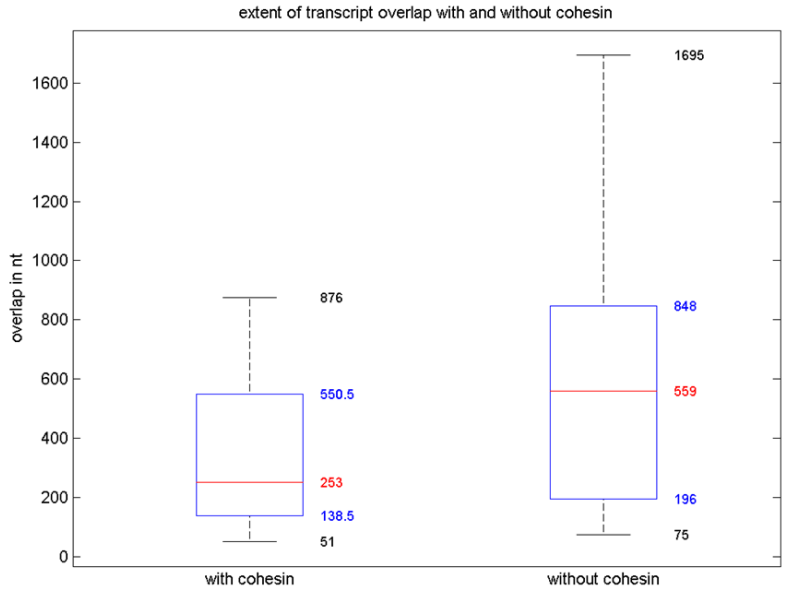

Figure 2.29: Boxplot for the sizes of the transcript overlaps (>50 nt) between convergent genes on chromosome II with or without Cohesin peaks. Red line represents the median, while the blue box marks the quartile. There are no outliers, so whiskers extend to the minimum and maximum. All values are marked in the same color as they are represented on the plot.

2.1.3 Discussion

I have analysed RNA-Seq data from five different data sets. When mapping CS one is always presented with the danger of not detecting CS, mapping of non-existent CS and mis-mapping CS. To analyse PAS I chose specificity over sensitivity. Therefore this study should be complemented in the future by further investigations, which are more sensitive to CS detection. An important step in this direction has already been completed by a parallel conducted study (Mata, 2013), which mapped over 8,000 CS in *S. pombe* WT cells. One should note that in the non strand-specific data the numbers of detected CS and identified genes are 4-fold higher in comparison to the strand-specific data sets, even though the initial number of reads is comparable. The experimental procedure to obtain the strand-specific data was different, as the non strand-specific reads come from libraries primed with a poly(dT) primer while the strand-specific ones use "random" RNA-RNA ligation. Data obtained by the first protocol is therefore bound to contain more poly(A) reads. As I am comparing several growth conditions only data derived from equivalent protocols is presented in the main results section. The analysis for the non strand-specific data can be viewed in the Appendix C and the database. A feature of RNA-Seq is the copying of RNA to cDNA, which comes with a set of caveats. As reviewed by Proudfoot (2011), these include template switching and internal priming. Moreover, RNA selection using the poly(A)-tail could cause the sequencing of transcripts, which were polyadenylated by alternative polymerases and marked for degradation (Schmid and Jensen, 2008). The latter two issues are addressed below, while template switching does not present a problem when mapping CS. Template switching refers to the process of a nascent cDNA changing RNA template by annealing to another template with a similar sequence. If the nascent cDNA were to re-anneal to a polyadenylated RNA, our mapping procedure would still discard it as it would not map to a genomic region perfectly. In another scenario it would still determine a real CS, regardless of the initial template because of sequence similarity. Eukaryotic PAS define the 3' ends of mRNA

and comprise independent *cis* elements such as NUE and EE. I have analysed NUE and motifs flanking the CS in *S. pombe* at a genome-wide level. Most *S. pombe* genes contain multiple alternative CS and/or heterogeneous CS both at the end and inside the ORF (Figure 2.9). This is consistent with a recent study performed in *S. pombe* (Mata, 2013) but contrasts with higher eukaryotes, where internal CS are infrequent. I identify 95 genes with apparent CS only inside the ORF under particular growth conditions (cycling cells). We have carefully excluded internal priming (Step 3 of the mapping procedure, see section 2.1.1 and Figure 2.2) so possibly these intragenic CS give rise to non-functional proteins and hence control expression of highly transcribed genes such as *act1* in cycling cells and *meu4* in meiotic cells (Figure 2.17 and 2.18). However, polyadenylation is well known to serve as a marker for degradation in bacteria (Dreyfus and Régnier, 2002; Steege, 2000) and plants (Lange et al., 2009). Chen et al. (2011) have further shown that the addition of an unusually long poly(A) tail ("hyperadenylation") to mRNA leads to degradation in *S. pombe*. Other cases of polyadenylation for RNA degradation are also known in *S. cerevisiae* (Kuai et al., 2004). The additional fact that internal cleavage amounts correlate with gene expression levels and/or gene length (Figure 2.11) suggests that these detected CS are likely to be derived from degradation products in most cases. This could therefore also extend to some CS identified in the 3' UTRs, though here they are more likely to be real. Identification of real truncated and functioning transcripts remains to be investigated further. These could be considered as non-coding, as they lack a stop codon, but they still have an important regulatory function. CS in introns of *S. pombe* are rare, similar to *A. thaliana* (Sherstnev et al., 2012), which could reflect a common mechanism that restricts intronic CS in these two eukaryotes. Such a process may involve U1 snRNA, which is known to restrict premature polyadenylation in mammalian introns (Berg et al., 2012; Kaida et al., 2010). The nucleotide distribution profile around the CS demonstrates the similarity of *S. pombe* 3' UTRs to other higher eukaryotes, fungi and plants (Ozsolak et al., 2010; Retelska et al., 2006). Another similarity includes a

median 3' UTR length of 187.5 nt, which is close to the 166 nt median 3' UTR length in *S. cerevisiae* (Ozsolak et al., 2010). Furthermore, the larger 3' UTRs for tandem genes compared to convergent genes, is consistent with their larger intergenic regions (Bähler and Wood, 2003). Certain CS are more common (more RNA-Seq hits) under certain physiological conditions (Figure 2.17). This confirms that CS selection also depends on the developmental and physiological stage of the cell (di Giammartino et al., 2011; Mangone et al., 2010; Sherstnev et al., 2012). I have computed *p*-values to determine the significance of CS usage under different growth conditions. Care needs to be taken as some genes might be considered alternatively polyadenylated by having three or more sites, which vary between conditions. This could show that a CS has a significantly higher usage in one condition over another. However, it might turn out that the other condition also has a highly used CS only few nt away, which might still be within the heterogeneity margin. Though I was careful not to consider this as alternative polyadenylation in its own right, it might arise as an artefact from considering the gene alternatively polyadenylated for other reasons. This kind of analysis, which assumes the independence of CS might fail to consider that one CS is discriminated in favour of another, which may then be reversed in another condition. However, by considering the CS to be independent, this occurrence will go unnoticed. Therefore the full list of possibly alternatively polyadenylated CS, regardless of the *p*-values, is presented in Appendix Tables App.C.2-App.C.4. Many CS have few reads ($\leq$2) mapping to them rendering the data insufficient to draw statistically significant conclusions of an alternative polyadenylation event. These CS are reported as candidate alternative CS, but require further experimental validation, possibly by improving sequencing depth. It has previously been noted that efficient 3' end formation of the *ura4* gene in *S. pombe* requires a site determining element (NUE) and an efficiency element (EE, Humphrey et al., 1994). Considering that the EE is following the CS, our motif search across 4-10 nt yielded mostly sequence motifs containing GT. I detect TGTA as the most significant EE in fission yeast. Interestingly this has previously been identified

as part of the far upstream element in humans and budding yeast (Graber et al., 1999; Venkataraman et al., 2005), as well as in plants (Rothnie et al., 1994), including rice (Shen et al., 2008a). In algae PAS are different from other eukaryotes, mainly due to their G-rich 3' UTRs. Notably TGTAA forms the NUE in this organism (Shen et al., 2008b). In fission yeast the TGTA containing motif occurs even further downstream. This is likely due to the higher frequency of Gs over Cs after the CS, which is different in plants (Shen et al., 2008a; Sherstnev et al., 2012) and budding yeast (Ozsolak et al., 2010). Furthermore the identified sequences (especially ranked 3 and 10) show high sequence similarity to the previously described RNA 3' end signals TTTTTT/TTTTAT/TTTTCT in budding yeast (Graber et al. 2002), suggesting an evolutionary conserved function in 3' end formation. OligoT-stretches always rank high in the first steps of the ranking procedure, but are disregarded in the later steps due to a high *p*-value. Since T and TT are frequent in the 3' UTR, this could be due to a bias in the Markov model, which approximates the probability of certain motifs based on the frequencies of single and di-nucleotides. In addition T-stretches and our determined most significant hexamer NUE AATAAA (Table 2.7) rarely occur together in one PAS. This fact, together with the high frequency of T around the CS, suggest an influence on CS efficiency or a potential compensatory function of T-stretches, if the NUE is not otherwise recognised by the polyadenylation machinery. AATAAA is suggested to be the most efficient functional motif to serve as a NUE element in *S. pombe* (Figures 2.26 and 2.27). However, the presented experiments are influenced by mRNA stability, which may depend on various *cis* elements in the cloned terminator regions. We also lack an estimate of the ratio between cleaved and un-cleaved message. Nevertheless, to support the efficiency of AATAAA, it is found 30 nt before the CS with most RNA-Seq hits in *meu4*. This CS is detected in the non strand-specific data from cycling cells (within the heterogeneity window), which include a minority of meiotic cells. However, even though AATAAA is the most conserved PAS throughout the genome, it only occurs in slightly above 20% of all 3' UTRs, which is a

lower frequency than found in flies, worms and mammals (Mangone et al., 2010; Retelska et al., 2006; Yan and Marr, 2005). Nevertheless, this still displays a higher conservation in fission yeast than in plants (Loke et al., 2005). It has been shown in human APA that the 3' distal PAS mainly uses the canonical AATAAA NUE, whilst the proximal PAS uses non-canonical signals (Beaudoing et al., 2000). In *S. pombe* I do not detect this positional preference for NUE, implying that PAS selection in genes that display APA, is an individual process specific to that gene. It has also been noted that in the case of mammalian APA the proximal PAS is preferentially used over the distal (Denome and Cole, 1988) unless distal PAS has a stronger NUE (Legendre and Gautheret, 2003). The 3' RACE data obtained for the *ura4-pyp3* gene construct confirms this. Two observed CS correspond to AATAAA, and the proximal CS is stronger than the distal one (Figure 2.26). Interestingly, AATGAA always ranks second for all strand-specific data sets in *S. pombe*. This motif ranks as the highest NUE motif in *Aspergillus oryzae* (Tanaka et al., 2011), emphasizing the evolutionary conservation of the PAS sequence motifs across some organisms. However, the mutation of AATAAA to AATGAA can lead to loss of PAS function in humans (Bennett et al., 2001). Similarly I see a loss in PAS function in our transformation experiment. Our experimental analysis shows that the wild type *ura4* AATAAA NUE functions more efficiently than the same AATAAA NUE derived from *sid4* (Figure 2.27). It should be noted that the *ura4* PAS contains a defined EE (Humphrey et al., 1994), which acts with the NUE to promote efficient RNA 3' end formation. There is no detectable EE in the *sid4* PAS, even though it does contain TG-rich elements. The importance of AATAAA is underlined by the observation that only genes containing the canonical AATAAA NUE are enriched for a functional GO category. These GO categories of enrichment are essential for the cell, indicating that AATAAA may mediate efficient 3' end formation and ensure correctly formed mature mRNA.

It has been suggested that PAS efficiency depends on the secondary structure formed by the mRNA (Loke et al., 2005). This could provide an explanation for the loss of

polyadenylation function of AATGAA. I suggest that yeast does not possess efficient PAS elements throughout the whole genome and may therefore rely on additional *trans* acting. For example, it has been shown that the Cohesin protein complex prevents transcriptional read-through between convergent genes and consequently leads to transcription termination after the proximal PAS (Gullerova and Proudfoot, 2008).

Overall our data redefine previously annotated 3' UTRs in *S. pombe*. Our analysis can be generally accessed by operating our user-friendly database *Pomb(A)*. This is a tool to visualise all CS and identified PAS in *S. pombe* and so allows comparison of CS usage between physiological conditions. Our database permits the user to search for most significant sequence motifs as well as defined motifs of interest. I believe the *Pomb(A)* database will be a new tool for the *S. pombe* community, especially for studies requiring identification of PAS and indication of conditional isoforms.

2.2 Cohesin and its roles in Transcription, Cohesion and Replication

The primary role of Cohesin is the establishment of sister chromatid cohesion. However, previous roles have also been implicated. In particular, it has previously been shown that Cohesin may affect transcription. In the previous section we have shown that indeed, overlapping 3'UTRs are longer at sites lacking Cohesin. I have demonstrated that *S. pombe* polyadenylation is heterogenic and often multiple CS are present in the 3'UTR. I have also shown that *S. pombe* lacks a well conserved NUE. This indicates that fission yeast may rely on other factors for efficient 3' end formation which act in *trans*. Cohesin has been demonstrated to be one of these factors.

Cohesin is known to play a role in gene expression (Gullerova and Proudfoot, 2008), recombination (Gerlich et al., 2006) and DNA damage repair (Birkenbihl and Subramani, 1992). In higher eukaryotes Cohesin co-localises with Kollerin, strongly present at transcription start sites (Dorsett and Merkenschlager, 2013). Cohesin also co-localises with the transcription factor CTCF in mammals (Wendt et al., 2008) and with other tissue specific transpiration factors (Schmidt et al., 2010). In fission yeast, the centromeric, pericentromeric and telomeric regions are strongly associated with Cohesin (Schmidt et al., 2009). It mostly localises at Kollerin associated sites, coinciding with strongly transcribed genes (Schmidt et al., 2009) and in transcription termination regions between convergent genes (Gullerova and Proudfoot, 2008).

Cohesin is loaded onto the chromatin in S-phase of the cell cycle and is removed during mitosis (Losada, 2007). However, a substantial amount can be detected also during mitosis. This suggests the existence of multiple Cohesin subfunctions involved in processes such as cohesion-maintenance, transcriptional regulation, homologous

recombination and DNA damage repair. It has been suggested that Cohesin is in a dynamic binding state at Cohesin loading sites, while away from these sites it is stable (Schmidt et al., 2009). However, we challenge this perception.

2.2.1 Materials and Methods

2.2.1.1 Data Analysis

It has previously been described, that Cohesin localises nearly exclusively between convergent genes (Schmidt et al., 2009). I have extracted the herein generated ChIP-chip data and analysed it. Firstly, it should be noted, that the data presented in the publication is inconsistent with any *S. pombe* annotation I was able to obtain. In Figure 2.30 I present the most striking inconsistencies: Firstly, the "pi"-genes are mostly not annotated in the *S. pombe* genome (e.g. *pi066*). Moreover, the ones that are, appear to be annotated on the opposite strand than presented (e.g. *pi053*). Finally, the ones, that are annotated occur actually downstream of the gene *h3.2*. I therefore went back to analyse the original data. The tiling array, which was used, was designed in 2004 (Affymetrix S_pombea520106F). In the past years the *S. pombe* genome has changed its annotation, in particular 1,000 "N"s, which represented unknown sequence sections were replaced by 100 "N"s, as well as the first 80,000 nt were unknown then. Therefore towards the beginning, the annotation was shifted by 80,301 nt, and towards the end by 79,401 nt. Taking all of this into consideration almost made the data consistent with the current genome annotation. I then examined the region described in Figure 2.30. It turns out, that ignoring the "pi "-rich region and patching up the presented gap, which introduces a negative shift by 38,000 solves the problem. The re-plotted data (in $\log_2$ scale) is consistent with the current *S. pombe* genome annotation and can be viewed in the Appendix F.4. I also analysed the previously neglected chromosome III data, presented in the Appendix F.5.

Schmidt et al. (2009) have called the peaks from their data. However, I believe that certain peaks might have been missed as illustrated in Figure 2.31. Strangely, a peak has not been called, while an adjacent, much smaller one has been. This may be due to calling peak by the employed data smoothing. Moreover, I wanted to analyse all data for Rad21, Mis4, Pds5, Psc3 and Ssl3 (Schmidt et al., 2009) consistently. I therefore used my own peak calling algorithm, which does not use data smoothing.

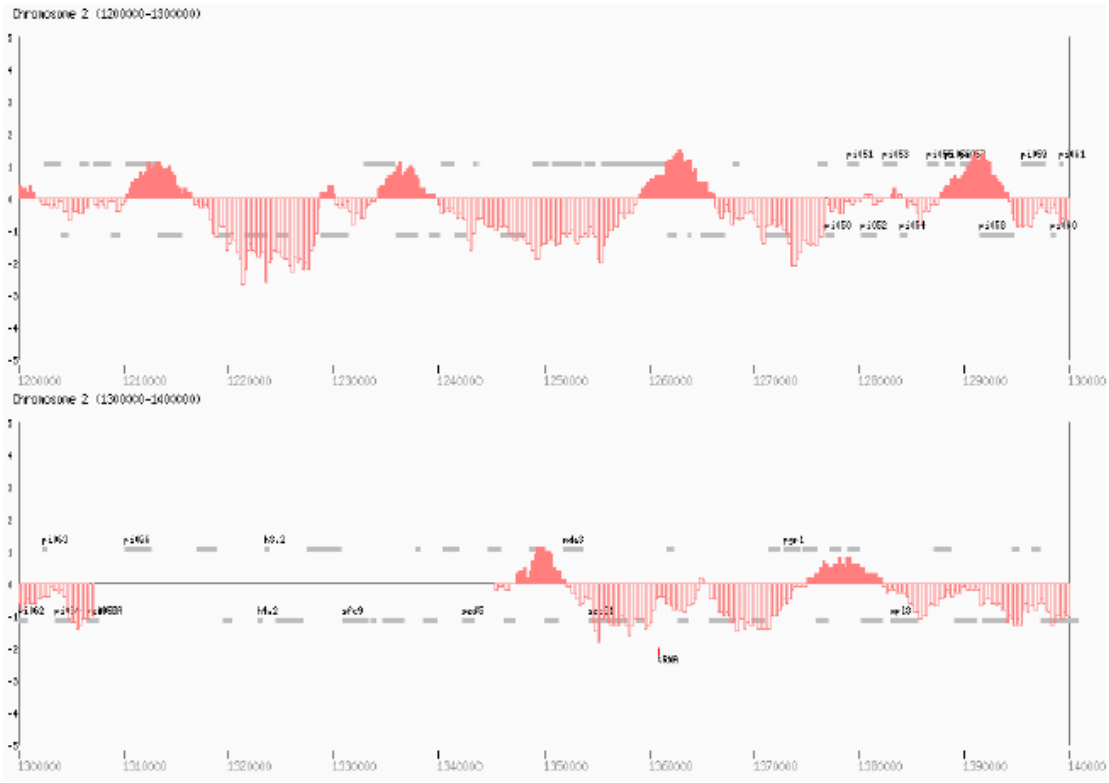

Figure 2.30: Illustration of and inconsistent data annotation region in Schmidt et al. (2009)

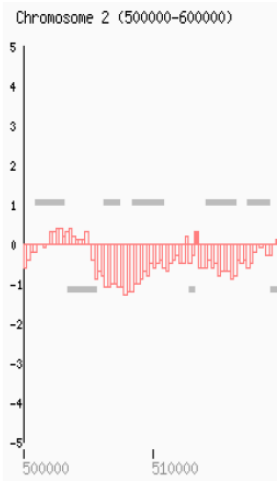

Figure 2.31: Illustration of a called peak (filled bar) and a non-called peak (hollow bars) in the data presented by Schmidt et al. (2009). This is an expected occurrence in data which is smoothed by a moving average prior to analysis. I decided to employ my own peak calling analysis

2.2.1.2 Peak-caller

Our developed peak-caller involves two steps. Firstly, the MATLAB peakfinder algorithm is employed to $\log_2$ normalised signal data

(http://www.mathworks.co.uk/matlabcentral/fileexchange/25500-peakfinder). $\log_2$ scaling is primarily used because we normalise the data. Therefore, if normalisation data is larger than the signal, everything is pooled between 0 and 1, in the opposite case the data can spread between 1 and any large value. This implies a vast discrepancy between the intervals $[0, 1]$ and $[1, \infty)$ where data is pooled based on whether normalisation or signal data is larger. $\log_2$ scaling makes a difference out of division and therefore signal and normalisation value are displayed without bias. In addition, 0 is the value to distinguish between the two scenarios. Moreover, large peaks may display large discrepancies between adjacent signals. $\log_2$ would smooth these and avoid calling the same peak multiple times. In contrast, smaller peaks become more pronounced compared to the surrounding. Background and noise, if small enough would become even negative or remain small.

The algorithm uses two criteria to call peaks: firstly, a threshold is specified, above which the peak values must lie. Secondly, a cut-off value is given, by a factor of which the peak must be higher than the average data of a surrounding island. This algorithm, depending on the criteria, is very sensitive, but not necessarily very specific. I preferred sensitivity at the first instance and therefore set the cut-off to be 0.2. Compared to other cut-offs it was still computationally efficient, but provided a considerable amount of candidate peaks. To restrict the number of peaks further, I subjected them to a custom written perl code. Peaks were extended to both sides until $\log_2$ reached 0. The criteria employed there were a minimal peak-width of 1,000 nt and an average signal throughout the peak (with the summit value counting double). Rad21/Mis4/Ssl3/Pds5/Psc3 were labelled with a Pk9 epitope for the ChIP experiment. Upon correspondence with Frank Uhlmann (Schmidt et al., 2009) I was informed that the Mis4/Ssl3 ChIP due to technical reasons contains very low concentrations of specific DNA. This is primarily due to Mis4 and Ssl3 being enzymes, hence they do not directly interact with the DNA, while Cohesin encircles the DNA. Untagged Pk9 data provides information on preferentially amplified

DNA regions, when only very small amounts of specific DNA are recovered in the ChIP. The Mis4/Ssl3 data is therefore more noisy and I imposed a higher threshold than for the Rad21 data. I additionally compared it to untagged Pk9 binding on the chromatin. I expected a real protein binding signal to be twice higher than the untagged data and not to overlap with a called Pk9 peak. For the remaining datasets (Rad21/Pds5/Psc3) the ChIP is more efficient and contains high amounts of specific DNA, hence the untagged Pk9 profile could be disregarded and was not taken into account. The threshold used for each of the datasets was picked based on a plotted data distribution: the noisy background data would represent the bulk of the data, while the real peaks would be in the tail (illustrated on Mis4 on chromosomes II and III in Figure 2.32 and 2.32). I picked threshold values 0.3 for Rad21, 0.5 for Mis4, 0.7 for all other datasets. Based on the same observation the average signal value throughout a peak had therefore been chosen as 0.2 for Rad21, 0.3 for Mis and 0.4 for all other datasets.

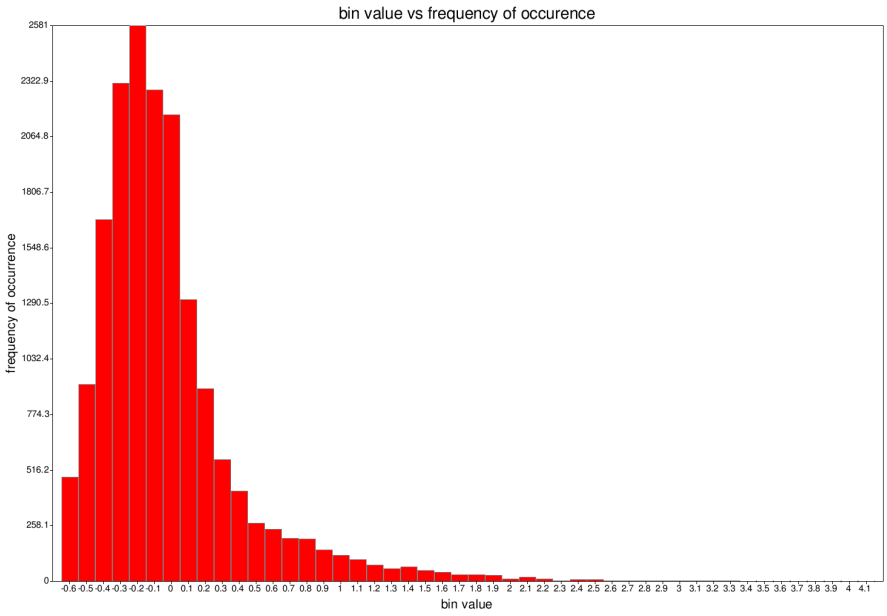

Figure 2.32: Mis4 data value distribution on chromosome II, normalised to whole genome DNA sample (x-axis in $\log_2$ scale). I estimate the tail of the distribution starts at around 0.5.

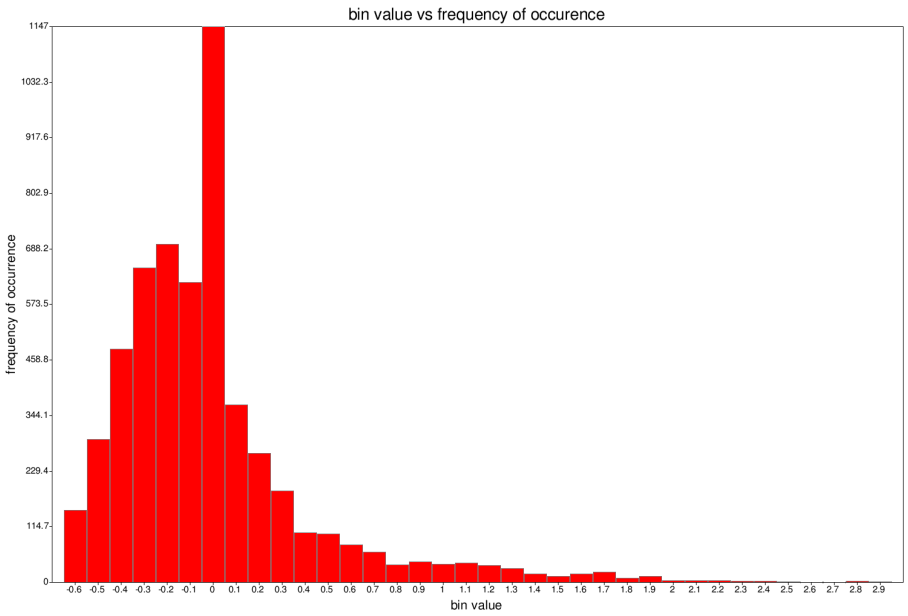

Figure 2.33: Mis4 data value distribution on chromosome III, normalised to whole genome DNA sample (x-axis in $\log_2$ scale). I estimate the tail of the distribution starts at around 0.5. Note that the large value of the bin at x-value 0 is due to large gaps in the data, which bias it towards the 0 score.

2.2.1.3 Gene Expression Analysis

To identify the Gene Expression distribution in a set of called peaks, I identified all the genes each peak was overlapping. The Gene Expression assigned to each peak was taken to be the maximally expressed gene. Gene Expression values are based on RNA-Seq derived reads per kilo base per million (RPKM), which were obtained from Marguerat et al. (2012).

2.2.1.4 Origins of Replication

Origins of Replication coordinates were obtained from OriDB (Siow et al., 2012) and classified into "Dubious ", "Likely" and "Confirmed". I disregarded the "Dubious" ORIs and proceeded with the "Likely" and "Confirmed" ones for the analysis.

2.2.1.5 Statistical Analysis

To compare various features of two different sets of peaks to each other I employed different approaches to test for statistical significance. Let M-Peaks designate those Rad21 peaks, which overlap Mis4 and let C-Peaks denote Rad21 peaks, where Rad21 is detected on its own ("M" for Mis4 and "C" for Cohesin only). Together M-Peaks and C-Peaks are all detected Rad21 peaks. The difference in nature of different Cohesin peaks is analysed, hence loci where Mis4 localises independently of Cohesin are not of interest for this project.

All analysed peaks are located on chromosome II. I performed the same analysis for chromosome III and both chromosomes together (chromosome I was not present on the array). However, it is apparent that the chromosome III data only covers about half of the chromosome. Moreover, large regions are skipped in the on the array, in particular the centromeric regions (Appendix F.5). This skews the data. For example, gene expression is lower in centromeric regions. Not taking it into account biases the data towards higher transcribed regions. I therefore present all conclusions based on the chromosome II data.

Background distribution based on sampling DNA regions To identify if M-Peaks or C-Peaks are in significantly higher expressed regions than to be expected, I sampled DNA regions based on a uniform distribution. To generate a background distribution for M-Peaks, I sampled an M-Peaks-set size amount of DNA-regions, which have the same length as the peaks in the set. The same was done for C-Peaks. Sampling was performed more than 1,000 times. p-value p_{right} is based on the number of times the average from the trials is $\geq$ than the observed average (right-handed test), hence giving an estimate of whether the regions contain significantly higher expressed genes. To estimate whether the regions contain significantly lower expressed genes, the left-handed p-value can be approximated by $p_{\text{left}} \sim 1 - p_{\text{right}}$.

Background distribution and dataset comparison based on permutation test M- and C-Peaks were combined into one set and resampled into random groups of C-Peak-set size and M-Peak-set size. Sampling was performed ≥10,000 times. On the one hand this yielded a background distribution for each peak type, on the other hand it provided a comparison statistic (defined as **average**(M-Peak expression)-**average**(C-Peak-expression)) for the two sets. The *p*-value is based on the number of times the average from the trials is ≥ than the observed average (right-handed test).

Fisher Exact Test Right-handed Fisher Exact Test was performed as described previously (Section 2.1.1) and is described in more detail in the main text.

2.2.2 Results

To analyse how Cohesin and Kollerin are coupled to transcription I investigated if there is an enrichment of highly transcribed genes at Cohesin and/or Kollerin sites. To identify where Cohesin and Kollerin localise, previously published ChIP-chip data (Schmidt et al., 2009) of Cohesin and Kollerin subunits were used to locate genomic sites of their enrichment. The previously published binding profiles (Schmidt et al., 2009) and called peaks were re-computed to make them consistent with the current *S. pombe* genome annotation.

2.2.2.1 Peak detection in ChIP-chip data

My peak-caller yielded 294 Rad21 peaks and 161 Mis4 peaks (as opposed to the previously defined 228 Rad21/77 Mis4 peaks by Schmidt et al. (2009), >90% of which are also contained in our identified peaks). Moreover, I identified 174 Ssl3, 174 Pds5 and 274 Psc3 peaks. All called peaks for Rad21 and Mis4 are listed in the Appendix Tables App.F.3.1 and App.F.3.2. The distribution profiles are attached in the Appendix F.4 for Rad21 and Mis4 versus untagged. The profiles for Ssl3, Psc3 and Pds5 data were also generated (data not shown).

2.2.2.2 Rad21 and Mis4 co-localise at highly transcribed regions

Interestingly, Mis4 peaks do not necessarily co-localise on the chromatin with Rad21 peaks, even though it is responsible for loading Cohesin onto the chromatin. This confirms the mobile nature of Cohesin on the chromatin. The Rad21 peaks overlapping Mis4 peaks I will refer to as M-Peaks from now on, otherwise (Rad21 only) I will call them C-peaks. I analysed whether Rad21 peaks, which overlap Mis4 (M-Peaks) are more associated with strongly transcribed genes. I classified the extent of transcriptional activity by the most strongly expressed gene overlapping the peak, based on RPKM values of the RNA-Seq data employed in the previous section (Marguerat et al., 2012). Indeed M-Peaks are more

strongly transcribed (Figure 2.34).

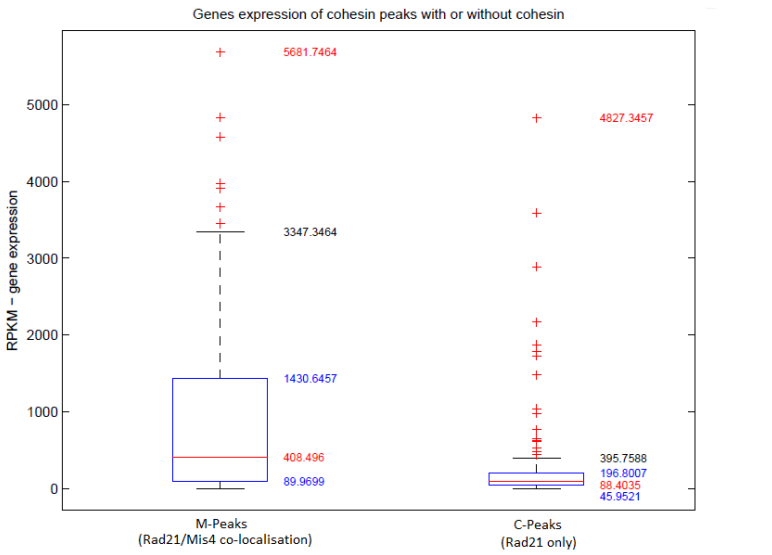

Figure 2.34: Gene expression of Rad21/Mis4 overlapping peaks versus Rad21 peaks only. Gene expression was determined as the most highly expressed gene overlapping the peak. The difference between the two sets is significant based on a permutation test

I wished to classify the significance of this outcome and employed three different approaches: To compare gene expression to a background distribution I generated it by two different methods. On the one hand I used a background distribution based on randomly picked genic regions (see Materials and Methods). Alternatively I generated a background distribution due to a permutation test approach, which also yielded a comparison statistic between M-Peaks and C-Peaks (see Materials and Methods).

Firstly, note that based on the DNA-region sampling, I observed that Mis4 generally localises in regions with significantly higher gene expression than would be expected by chance. I therefore wished to analyse whether gene-expression in M-Peaks is significantly higher than in C-Peaks. In Table 2.10 the outcome of all tests is presented. p-values are based on the number of sampling outcome averages, which are higher than our observed average value (hence I present a right-handed p-value, although a value close to 1 would mean, that expression is significantly low). To contrast this, I have performed the same

analysis between Rad21 peaks and Psc3 peaks, which are both components of the same complex. The difference in gene expression is not significant according to the random DNA region sampling, nor the permutation test background distributions. However, the comparison of the two averages of the overlapping peaks and non-overlapping peaks does seem to be mildly significant (Table 2.11, last row). Similarly the two loading complexes Mis4 and Ssl3 show insignificant differences based on the permutation approaches, but significant results based on DNA region sampling. Intriguingly, cross-comparing Rad21 and Mis4 with all the other data-sets often gives significantly highly expressed genes in overlapping peaks based on the random DNA region sampling (data not shown). This phenomenon will be discussed later in the chapter. The extent of overlap between Rad21/Mis4 peaks to all other protein peaks in the analysis is presented in Table 2.12.

	Mis4 overlap Rad21 expression			**Rad21 only expression**		
	Observed average gene expression (RPKM)	Expected	P-value (right-handed)	Observed average gene expression (RPKM)	Expected	P-value (right handed)
DNA Region sampling (N>1000)	959	431	0	246	373	0.99
Background distribution (permutation) (N>10000)	959	493	0	246	492	1
	Mis4-Rad21		**Expected**		**P-value**	
Permutation test between two samples (N>10000)	713		1		0	

Table 2.10: Statistical analysis and outcome of gene expression difference between C-Peaks and M-Peaks. Test are based on a background distribution generated from random DNA regions of the same sizes as the peaks, or permutation tests (see Materials and Methods). Permutation Test between two samples was performed with the statistic **average**(Rad21/Mis4 gene expression)-**average**(Rad21 only gene expression). All reported *p*-values are right-handed.

2.2.2.3 Experimental Analysis of Cohesin function

In this section I present experiments, which were performed by Shweta Bhardwaj from the Gullerova Lab, Oxford. The experiments illustrate the predictive power of the genomic analysis in the previous section. In particular, they indicate that M-Peaks are strongly

	Rad21 overlap Psc3 expression			Rad21 only expression		
	Observed average gene expression (RPKM)	Expected	P-value (right-handed)	Observed average gene expression (RPKM)	Expected	P-value (right handed)
DNA Region sampling (N>1000)	569	452	0.08	386	312	0.18
Background distribution (permutation) (N>10000)	569	492	0.14	386	492	0.9
	Psc3-Rad21		**Expected**		**P-value**	
Permutation test between two samples (N>10000)	182		1		0.04	

Table 2.11 As Table 2.10, but Rad21 overlapping Psc3 was analysed.

Number of overlapping peaks between datasets	**Rad21 (294)**	**Mis4 (161)**
Mis4 (161)	102 (35%)	/
Ssl3 (174)	107 (36%)	97 (60%)
Pds5 (134)	122 (41%)	68 (42%)
Psc3 (274)	174 (59%)	95 (59%)

Table 2.12: Extent of overlapping peaks between two samples.

cohesive, while C-Peaks are less cohesive. This implies a function for C-Peaks different from sister chromatid cohesion (e.g. regulation of transcription). Moreover, experiments presented in the appendix confirm that Mis4/Rad21 co-localisation indeed depends on the high gene expression of their localisation loci.

The function of M-Peaks and C-Peaks remains to be illuminated. Therefore, DNA-Fluorescent *in situ* hybridisation (FISH) experiments were performed to elucidate the function of Rad21/Mis4 co-localisation. A summary of the experimental procedures is given in the Appendix F.1. Six sites of each, C-Peaks and M-Peaks, were selected for experimental analysis. The C-Peak sites were taken to be the genes *med8*, *psh3*, *cut3*, *pla1*, *mug37* and *swi3*. The M-Peak sites were *dis2*, *tif11*, *ubp9*, *srp1*, *rpl3801* and as a positive control the strongly cohesive centromeric degenerate (cendg) repeats. Each site was labelled with a fluorescent probe. We hypothesise, that if sister chromatids are indeed stably held together within a 40 nm diameter, we should see only one dot. On the

other hand, non-cohesive sites should make two dots visible. Indeed, in the Mis4/Rad21 co-localisation loci we see only one dot in 100% of all cells (≥200 cells, Figure 2.35A), while in the Rad21 only we see 50%/50% one or two dots (≥100 cells, Figure 2.35B). These results were further confirmed by the use of temperature sensitive cells, which after shifting to 37°C experience loss of Cohesin or Mis4. Loss of either shows similar results as the presented analysis of Rad21 peaks only (data not shown).

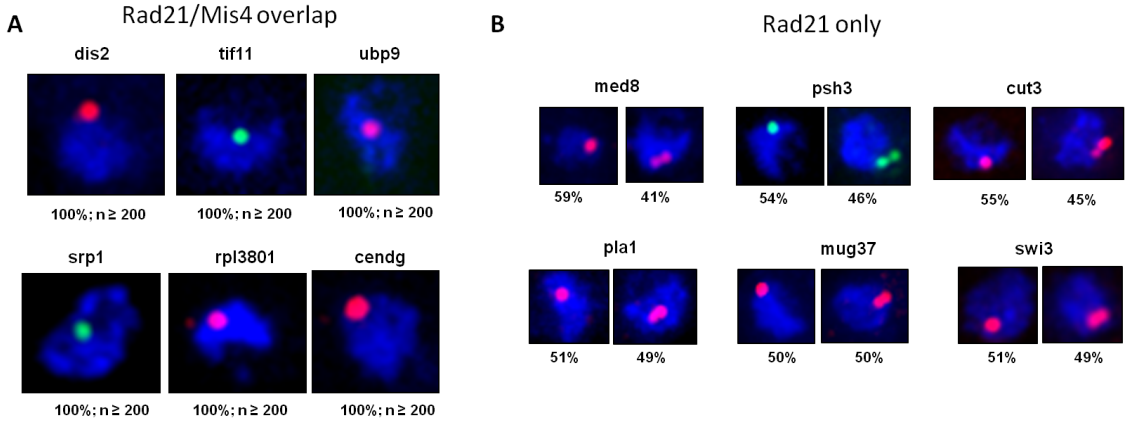

Figure 2.35: Fish analysis of individual loci on regions selected from chromosome II. PCR products were roughly 10kb long to span the loci and were pooled in equimolar quantities. The FISH-Tag kit (Invitrogen) was used for labelling. Blue regions show DAPI staining of the nucleus, each dot within the nucleus is a DNA- FISH signal. Green probes were labelled with Alexa-488, red probes with Alexa-555. Images represent maximum intensity projection of the z-section taken at 200 nm resolution. **A** Analysis of Mis4/Rad21 overlapping loci. More that 200 cells were counted for each experiment. We see 100% of cohesion, similar to the centromere (cendg) positive control. **B** Analysis of Rad21 regions without Mis4 binding. Each locus exhibits only 50% of proper cohesion, indicating cohesion instability. These experiments were performed by Shweta Bhardwaj from the Gullerova Lab, Oxford

We therefore conclude that the Rad21/Mis4 sites are the true strongly cohesive sites, while all other sites have travelling Cohesin, which can bind and unbind again and therefore is likely to have a different role that sister-chromatid cohesion. I therefore challenge the previous hypothesis of unstable cohesion on Cohesin/Kollerin sites, but stable cohesion at Cohesin only sites (Schmidt et al., 2009).

In *S. cerevisiae*, genomic analysis has revealed the presence of DNA-polymerase

in the ORF of highly transcribed genes (Azvolinsky et al., 2009; Tuduri et al., 2009). Moreover, the process of replication would imply the sister chromatids are in close proximity. Given the overlap of M-Peaks with highly transcribed genes and their establishment as "strongly cohesive" , I decided to check if I see a significant proximity of M-Peaks to Origins of Replication (ORI), compared to C-Peaks. Multiple datasets are available, which provide genomic coordinates of ORIs, these are curated into a file available from OriDB (Siow et al., 2012, see Materials and Methods). The first attempt to identify significant differences in proximity was by ordering ORIs and peaks by coordinates, and assigning each ORI to its closest peak. I checked whether the distances of ORIs to their assigned M-Peaks and ORIs to their assigned C-Peaks were significantly different via a permutation test as described before (see Materials and Methods). I did not get a significant result, although small differences are observed (Figure 2.36). This test, however, has a considerable flaw: peaks which have no ORI assigned to them, are not taken into consideration. I therefore performed a Fisher Exact Test on M- and C-Peaks with the categories "assigned ORI <2000 nt away" and "assigned ORI >2000 nt away or no assigned ORI" on all M-Peaks and C-Peaks together. The used contingency table is presented in Table 2.13. This indeed shows a mildly significant preference of ORIs to localise close to M-Peaks ($p = 0.029$, right handed Fisher Exact Test).

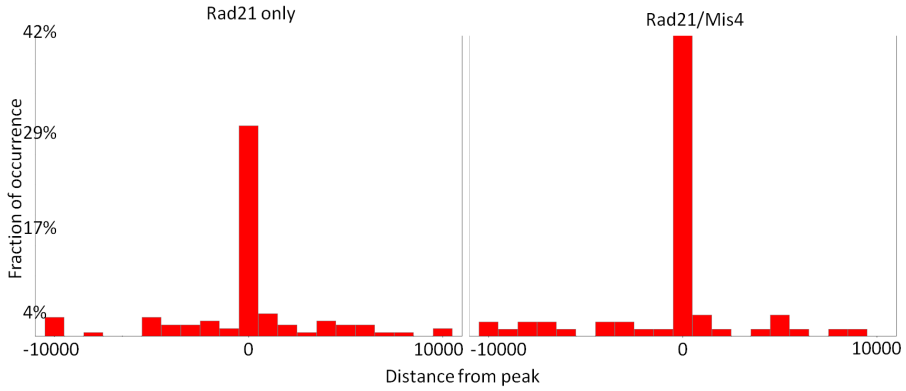

Figure 2.36: Distance of ORIs to their assigned peaks. Any distance $\geq$10,000 was binned into the $\pm 10,000$ bins

	M-Peaks	**C-Peaks**	Sum
<2000 distance	47	64	111
>2000 distance or no ORI	55	118	173
	102	182	294

Table 2.13 Contingency table used for classification of ORI proximal to M-Peaks or C-Peaks

2.2.3 Discussion

In the past it has been established that Cohesin has several subpopulations. FRAP Analysis in higher eukaryotes revealed that in G1 Cohesin associates and dissociates from the chromatin, with an average binding time of less than 30 mins. However, as cells enter G2, Cohesin can be stably bound (binding time >6h) or weakly bound as before (Gause et al., 2010; Gerlich et al., 2006; McNairn and Gerton, 2009; Onn and Koshland, 2011). It is possible, that the stably bound Cohesin is indeed responsible for cohesion maintenance while weakly bound Cohesin may perform other functions associated with transcription termination (Gullerova and Proudfoot, 2008) or DNA damage repair (Birkenbihl and Subramani, 1992).

We have identified strongly cohesive (Mis4 and Rad21) and less cohesive sites (Rad21 only). These sites show a difference in gene expression, where cohesive sites are associated with a significantly higher gene expression. This observation is based on different statistical tests (described in the Materials and Methods). Note that I detected a significantly higher gene expression for the of all pair-wise overlaps of all analysed datasets (except Rad21 and Psc3), based on the random DNA sampling, which was not necessarily observed in the permutation test based approaches. It is expected that Ssl3, being part of the same complex as Mis4, follows the same pattern in preferentially binding to highly transcribed regions. Therefore, any Rad21 peaks overlapping either one of these components are expected to also follow the same pattern. The significantly higher expressed genes in peaks of Rad21 and Pds5 is intriguing, but as mentioned before the function of Pds5 remains to be elucidated and we cannot make a meaningful deduction. A Wpl1 and Pds5 ChIP-Seq profile would in future studies help to draw a conclusion.

Given that Cohesin loading sites prefer highly transcribed regions, the sampling of random DNA-regions should be handled with care when looking at Rad21 peaks. Firstly, the sites available for Cohesin binding are not random. I therefore complemented this test with permutation test, to compare subpopulations of Rad21 peaks (either co-localising

with one of Mis4/Pds5/Psc3/Ssl3 or biding on its own). Interestingly I observe that gene expression in the Rad21 peaks overlapping Psc3 peaks is higher with a mild significance. I suspect this might be due to the higher gene expression of Rad21 peaks overlapping Mis4 (Random DNA Region Sampling): possibly some Psc3 still interacts with Kollerin.

We identified Mis4/Rad21 loci as strongly cohesive, while Rad21 only showed a less stable cohesion. We hence hypothesised this might be due to Replication Fork activity. Moreover in budding yeast an enrichment of DNA-polymerase in highly transcribed regions was observed (Azvolinsky et al., 2009; Tuduri et al., 2009). Our comparison to available ORI coordinates shows that indeed, there is a possibility that Kollerin/Cohesin peaks are correlated with ORI coordinates. However, these are only preliminary results. Moreover, many predicted ORIs have not actually been confirmed by 2D Gel-Electrophoresis and are merely predictions, which makes any conclusions based on the data questionable. In addition, many ORIs probably remain to be identified. Predictions rely on AT-rich islands (Segurado et al., 2003) and ChIP-chip of pre-RC components (Hayashi et al., 2007), increase of DNA-content in early S-phase via microarrays (Heichinger et al., 2006) and ssDNA mapping (Feng et al., 2006), which are subjected to challenges of data analysis (e.g. peak-calling). In recent years more data has become available through deep sequencing and shall be incorporated in our future analysis (Xu et al., 2012). Nevertheless, the predicted origins remain to be confirmed: only 14 confirmed loci are reported on chromosome II.

Schmidt et al. (2009) also provide non-analysed data for chromosome III. I have carefully checked the re-annotation of the *S. pombe* genome and I have included the data into our analysis. The dataset is unsuitable for statistical analysis though, as it contains large gaps. However, one can use it to identify strongly cohesive and less cohesive sites on chromosome III.

Eso1 is a protein which has been shown to interact with DNA replication components

(Mayer et al., 2004; Petronczki et al., 2004; Skibbens, 2004; Skibbens et al., 1999). During replication, Eso1 acts on the Smc3 subunit via acetylation to stabilize the Cohesin binding to DNA (Rudra and Skibbens, 2013) and enhance its resistance against Wpl1, which is responsible for Cohesin removal. Eso1 is therefore of high importance during S-phase. I showed that ORIs are closer to Mis4/Rad21 peaks than just Rad21 peaks. It was moreover shown, that highly transcribed regions are replicated in early S-phase (Schübeler et al., 2002; White et al., 2004; Woodfine et al., 2004). The predicted and observed connection of Mis4/Rad21 compared to Rad21 only peaks with highly transcribed regions and ORIs raise the possibility, that in G1, Mis4 associates with highly transcribed genomic regions, where the euchromatin structure allows replication fork assembly. The replication fork recruits Eso1, which then acetylates Smc3. This process stabilizes Cohesin, until it is removed in anaphase. To confirm this, an Eso1 chromatin binding pattern would be of high interest. We are currently working on Eso1 ChIP-Seq, which has not been performed before. The raw data will be processed by mapping it to the genome using maq or Bowtie. The obtained coverage data would be processed similarly to the data analysed here.

The association of Mis4 with highly transcribed genes raises the question of a Kollerin interaction with PolII. As preliminary experiments, PolII transcription was chemically inhibited with thiolutin. Indeed targeting Rad21-GFP and Mis4-GFP with an anti-GFP-Antibody, ChIP analysis reveals a drop in Rad21 and Mis4 in centromeres, telomeres, and especially chromosomal arms (Appendix Figure F.1). This serves as evidence that Cohesin affects transcription and *vice versa*. The exact interplay between Cohesin and transcription remains to be elucidated.

Chapter 3

Detection of potential miRNA genes and targets in *S. pombe*

In the previous chapter I showed that a large proportion of fission yeast genes are alternatively polyadenylated or at least heterogenic. Genes with overlapping 3'UTRs may be silenced in the absence of Cohesin by producing long overlapping transcripts and consequently dsRNA. In the presence of Cohesin, transcriptional readthrough is blocked and the genes are expressed (Gullerova and Proudfoot, 2008). However, on the analysed chromosome II only few genes had a sufficiently large overlap of 3'UTRs. This suggests a different form of gene expression regulation. Alternative polyadenylation in higher eukaryotes is known to provide binding platforms for regulatory elements, such as miRNA. The purpose of longer and shorter isoforms in *S. pombe* remains to be elucidated. In this chapter I investigate the possible existence of miRNA in *S. pombe*, which are very important regulators of gene expression in other eukaryotic cells. Their disfunction has been found to lead to many human diseases, including a variety of cancer types (Akao et al., 2007; Iorio et al., 2005; Porkka et al., 2007; Yanaihara et al., 2006; Yang et al., 2008; Zhang et al., 2009) as well as Alzheimer's (Hébert et al., 2009), multiple sclerosis (Cox et al., 2010) and schizophrenia (Beveridge et al., 2009).

The first miRNA identified in *C. elegans* was *lin-4*. It turned out to be non-coding RNA, regulating the expression of *lin-14* by antisense complementarity (Lee et al., 1993; Wightman et al., 1993). It has taken seven years to reject the hypothesis that this type of regulation was organism specific to *C. elegans*. The second discovered miRNA *let-7*, again in *C. elegans* (Reinhart et al., 2000), proved to be evolutionarily conserved across *bilateria* (Pasquinelli et al., 2000). Moreover, many 3'UTRs contain evolutionarily conserved sequences complementing miRNA sequences (Lewis et al., 2005), which mediate not only gene repression (Lee et al., 1993), but also upregulation (Vasudevan et al., 2007). These observations, along with other regulatory roles of miRNA, initiated the search for miRNA prediction methods and the development of computational tools.

Nowadays there is a fairly well defined consensus of when a gene is recognised as a miRNA (Li et al., 2010):

1. Northern blotting should demonstrate the existence of a roughly 22 nt mature miRNA.

2. The mature miRNA originates from one arm of a precursor hairpin structure.

3. With the disruption of Dicer function, the precursor miRNA should be observed, while the mature molecule should not be detected.

4. Phylogenetic conservation is a desired, though not a required, criterion.

The database for miRNA mature forms, miRBase (Kozomara and Griffiths-Jones, 2014) release 20, currently contains 24521 miRNA loci from 206 species. However, with the consensus of a recognised miRNA in mind, miRBase entries merely require a 22 nt sequencing product and its predicted, not necessarily verified, pre-miRNA.

Plant and animal miRNAs are believed to have evolved independently, but it appears that some miRNAs are shared after all (Arteaga-Vázquez et al., 2006) and that the human miRNA processing machinery recognises rice miRNAs upon ingestion (Zhang et al., 2011). To date, miRNAs have not been found in fungi (Drinnenberg et al., 2009).

However, I believe that there are several reasons why this should be investigated in more detail. Firstly, although *S. cerevisiae* lacks the necessary RNAi components, *S. pombe* does have a functional RNAi pathway. Secondly, although Drosha is absent in *S. pombe*, it has been shown that Dicer can process exogenous hairpin structures originating from a plasmid into siRNA, which can also induce heterochromatin formation on the plasmid (Simmer et al., 2010). Moreover, the siRNAs act in *trans* (on a region with sufficient complementarity within the genome), although it is suggested that a patch of heterochromatin nearby is necessary to provide a silencing foothold. The silencing can spread to adjacent regions of the targeted area. I anticipate that Dicer might act similarly to DCL1 (Dicer like protein 1) in *A. thaliana*, which also lacks a Drosha-homologue. DCL1 performs all miRNA maturation steps in the nucleus (Kurihara and Watanabe, 2004). Also in animals miRNA can be formed in a Drosha independent fashion, e.g. if an intron happens to be a pre-miRNA (miRtron, Okamura et al., 2007; Ruby and Bartel, 2007).

Target and miRNA gene discovery can be approached from various angles. Generally confirming computationally predicted target sites is tricky, as this is often done by attaching a target site sequence to a reporter gene and checking its expression with or without the miRNA. However, the high-throughput or deep sequencing approaches together with ultraviolet crosslinking and immunoprecipitation (CLIP-seq, HITS- CLIP, Chi et al., 2009; Hafner et al., 2010; Leung et al., 2011; Zisoulis et al., 2010) allows identification of endogenous target sites that co-immunoprecipitate with the RISC complex (Pasquinelli, 2012) at nucleotide-level resolution. Yet, mRNA bound by RISC is not enough to guarantee nor classify the regulation.

Next to the outlined biochemical target site identification and computational target site predictions, which is presented in detail in the next Section 3.1, genetic methods can be employed. Genetic methods make use of the phenomenon that a miRNA mutant can be

recovered by a target mutation, as illustrated with the first discovered miRNA in *C. elegans* (Lee et al., 1993). Though advantageous in identifying physiologically relevant target genes, a major disadvantage lies in the incapability of differentiating between direct and indirect miRNA targets. Moreover, the identification of individual regulators if several targets cause the observed phenotype becomes problematic (Pasquinelli, 2012).

3.1 Computational miRNA target discovery

Computational target site prediction is difficult due to the low numbers of verified miRNA-target interaction on an endogenous level (Pasquinelli, 2012). However, after the first few miRNA and their targets were identified, a pattern of complementarity between miRNA and targeted mRNA could be identified, which led to many *in silico* methods of target prediction (reviewed in Maziere and Enright, 2007).

Many miRNA target prediction algorithms are available nowadays. The initial algorithms such as miRanda (Enright et al., 2003), DIANA-microT (Kiriakidou et al., 2004), RNAhybrid (Rehmsmeier et al., 2004), MicroInspector (Rusinov et al., 2005), TargetScan/TargetScanS (Lewis et al., 2003) and Stark's *et al.* method (2003) have three main constituents (Li et al., 2010):

1. 5' positions 2-7/8 of the miRNA, termed the 'seed', has high complementarity to the target 3'UTR (several different seed types have been identified, Figure 3.1).
2. The negative folding free energy is higher in the resulting RNA-RNA duplex.
3. There is a high inter-species conservation between miRNAs, targets and miRNA:mRNA duplex.

The rules of miRNA-target matches have been verified experimentally, summarised in John et al. (2004): Asymmetry in 5' and 3' end of binding, with the 5' seed region showing higher complementarity, is a key feature. Additionally G:U wobble pairs are

less frequent in the 5' end. A correlation in the binding energy of the first 8 miRNA nucleotides and translational repression has been implied (Doench and Sharp, 2004). Finally, cooperativity contributes to miRNA efficacy, where targets sites may overlap (Doench and Sharp, 2004).

More recent methods such as PicTar (Krek et al., 2005), MicroTar and PITA(Thadani and Tammi, 2006) as well as RNA22 use machine learning approaches to predict miRNA targets, some of which disregard evolutionary conservation.

A few general problems in target site prediction became obvious. In many organisms exact 3'UTR boundaries are not available (Hubbard et al., 2002). Considering conservation of target sites can increase specificity, but unfortunately reduces sensitivity for species-specific targets. The advantage of specificity lies in facilitation of large-scale miRNA target prediction, however at the cost of hindering single gene analyses (Maziere and Enright, 2007).

The important seed region can be distinguished into 3 types: 5' dominant-canonical, 5'-dominant seed only and 3'-compensatory (Maziere and Enright, 2007). These are illustrated in Figure 3.1.

With so many options for target prediction available, it is difficult to choose the right one for novel, potential miRNA target site prediction in *S. pombe*. Many are unsuitable as they are only applicable to pre-defined species and only available as web servers. As I am looking for a yet unknown phenomenon, I prefer sensitivity over specificity. miRanda was found to be among the best computational methods with sensitivity of about 65% (Sethupathy et al., 2006). However, due to issues with specificity is has not scored highly compared to other programs in the past (Baek et al., 2008; Selbach et al., 2008). This issue has been addressed in the most recent release, by more stringent criteria on seed region complementarity and an incorporated statistical model. Given the availability of miRanda as an open resource and its high scoring performance amongst other software as well as the non-restriction to certain organisms, I chose miRanda to look for possible

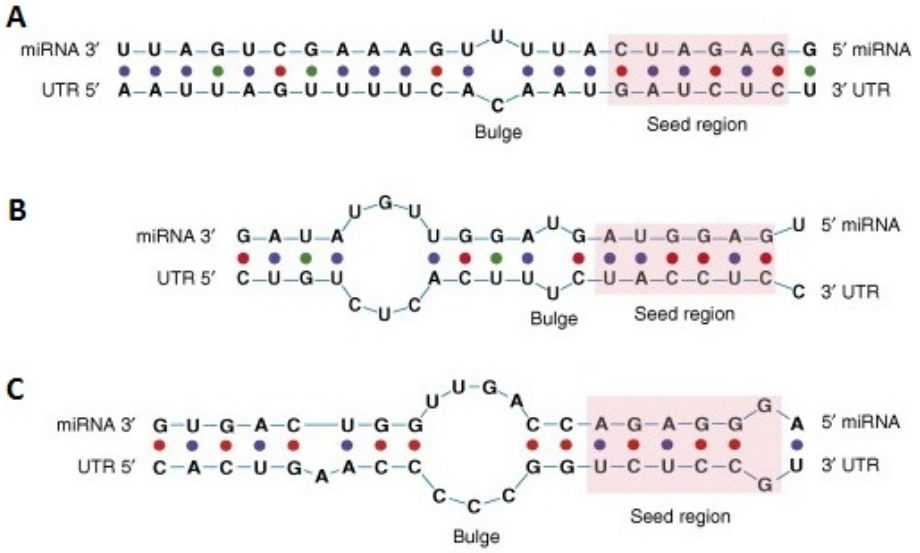

Figure 3.1: Illustration of possible miRNA-target binding. **A** the canonical site has perfect pairing in the seed region, as well a the 3' end of the miRNA. **B** Though miRNA 3' complementarity is poor, the seed region has perfect complementarity. **C** The seed region contains a mismatch or a G:U wobble pair, but the 3' miRNA end shows excellent complementarity to the mRNA target. Figure adapted from Maziere and Enright (2007) with permission from Elsevier [1].

miRNA targets. I will therefore describe it in more detail.

3.1.1 miRanda

miRanda was amongst the first published algorithms for miRNA target prediction (Enright et al., 2003). It looks for potential target sites by searching for regions of high complementarity in a set of submitted 3'UTRs. The scoring matrix for complementarity rewards matches in the 5' end more than matches in the 3' end. The obtained binding sites are evaluated due to thermodynamic binding stability computed by the package *RNAfold* (Wuchty et al., 1999).

In more detail, as default values the scoring matrix assigns scores of +5 for perfect Watson-Crick complementarity, +2 for G:U wobble pairs and -3 for mismatches. Complementarity scores and mismatch penalties in the first 11 base pairs of the miRNA

[1]Reprinted from Drug discovery today, 12(11), P Maziere and AJ Enright, Prediction of microRNA targets, 452458, Copyright (2007), with permission from Elsevier.

5' end are doubled (default parameter). The algorithm also assigns score-penalties to gaps: an initial -8 for gap opening and additional -2 for any gap extension. Moreover, the complementarity in the seed-region requires: no mismatches in positions 2-4, no more than five mismatches in positions 3-12, at least one mismatch after position 9 and before the last 5 nt, finally only 2 mismatches in the last 5 nt. The complementarity score is optimised by summing over all aligned positions and ranking all hybridisation alignments, which do not overlap, in decreasing order to a cut-off value of 80. A minimal binding energy according to *RNAfold* can be set (Enright et al., 2003).

A more recent miRanda version includes a statistical background model, which improves its specificity and is equivalent to the one in PicTar (Rehmsmeier et al., 2004). So far this new feature has not been described in detail anywhere, hence I will outline it here. After the initial alignment against a control-set of submitted sequences (often the sequences of interest, but shuffled), each sequence gets its maximal miRanda score assigned. Let N denote the random variable of miRanda scores. According to probability theory the maximal values follow an Extreme Value Distribution (EVD, Gumbel, 2012) given by:

$$P(N \leq t) = \exp\left(- \exp\left(- \frac{t - \xi}{\theta} \right) \right), \tag{3.1}$$

where ξ is the location parameter and θ is the scale parameter. Equation 3.1 transforms into a line by applying $\log(-\log(\cdot))$. Fitting a line with slope a and intercept b to the transformed maximal miRanda scores by a least squares fit, yields estimators for the parameters θ and ξ by

$$\hat{\theta} = -\frac{1}{a} \qquad \text{and} \qquad \hat{\xi} = b\hat{\theta}, \tag{3.2}$$

respectively. To censor the set of miRanda scores, the top 10% and bottom 10% are removed before fitting. Once miRanda is run against the actual sequences of interest

(usually 3'UTRs), a p-value p is deduced from the EVD distribution. The p-value is given by

$$p = P(N \geq S) = 1 - P(N \leq S) = 1 - \exp\left(-\exp\left(-\frac{t-\hat{\xi}}{\hat{\theta}}\right)\right), \tag{3.3}$$

where S is the observed maximal score. p essentially represents the area under the curve of the EVD, beyond the point S.

If sequences with potential targets are available from multiple related species, it has been the practice that after identification of potential targets, these binding sites are checked for evolutionary conservation to improve specificity (Enright et al., 2003; John et al., 2004). I did not include conservation analysis as there are no known miRNA in the fungi-kingdom to compare to. Comparing to plants or animals appears difficult as the miRNA are considered evolutionarily divergent, therefore criteria to predict an RNA molecule as a miRNA differ and are more stringent in plants. Furthermore, conservation analysis would discard of possible species-specific miRNA, whose existence would be an explanation why they have yet not been discovered. Therefore, I avoided conservation analysis on miRNA targets and employed the more relaxed animal binding criteria. More details on our methods on discovery of possible miRNA targets can be found in Section 3.3.1.

3.2 miRNA gene prediction

Assessment of miRNA expression in specific conditions or cell types is not straight forward. The low-throughput qPCR and fairly high-throughput microarrays have generally been used for assessment of miRNA expression, but they are constricted to known miRNAs. High-throughput sequencing of small RNAs, on the other hand, allows for novel miRNA detection and offers a wider and more dynamic range than microarrays. This technique, however, goes hand-in-hand with computational and bioinformatical

analysis.

Several approaches of computation gene prediction have been developed, which can be categorised into based on sequence/structure conservation, machine learning or experimental data (Li et al., 2010). The key descriptors are the foldback structure, a high minimal folding free energy and evolutionary conservation. It is clear that all computational approaches rely on known miRNA and miRNA conservation (Liu et al., 2012).

Conservation based approaches focus on predicting new miRNA genes based on similarity of sequence and secondary structure to verified miRNAs (Lim et al., 2003; Wang et al., 2005). However, the typical filtering of evolutionary non-conserved hairpins prevents the detection of species-specific miRNA (Bentwich et al., 2005).

Machine-learning based approaches are usually algorithms such as support vector machine, neural network, Hidden Markov Model and Naïve Bayes. These are considered quite unsatisfactory (Li et al., 2010) due to the fact that a positive sample of real miRNAs is easy to compile, a negative on the other hand fairly difficult. Among the best ones is considered MiRFinder (Huang et al., 2007).

Experimental data based approaches evolved with the high throughput sequencing techniques. Small RNA (sRNA) molecules can be sequenced and mapped back to the genome. Comparing these loci with known annotation categories, could identify miRNAs. MiRDeep (Friedländer et al., 2008) probabilistically scores (and estimates false positives) the suitability of frequency and position of the sequenced RNA molecule with the possible structure of the pre-miRNA. MiRanalyzer (Hackenberg et al., 2009) combines sequencing analysis with machine learning algorithms to detect new miRNA from known ones.

An interesting "backwards" approach was developed by Chang et al. (2008). Screening for differentially expressed genes between tissues (via microarray or EST data) and scanning the 3'UTRs for motifs, they identified possible miRNA seeds. Motifs, which

did not map known miRNA, were successfully utilised to search for novel miRNA genes. Similarly "Sylamer" (van Dongen et al., 2008) searches for motif enrichment/depletion across an expression-experiment driven gene list.

A recently developed tool MapMi was initially intended to allow querying pre-defined genomes for a supplied set of (known) miRNA sequences, possibly from other organisms (Guerra-Assunção and Enright, 2010). *S. pombe* is not among those pre-defined genomes. However, the web-server independent stand-alone version allows for providing the desired genome sequence. Given that I am looking for miRNA in an organism, where they are believed to be absent, this is the major advantage of using MapMi.

3.2.1 MapMi

Guerra-Assunção and Enright (2010) developed the program MapMi, which maps miRNAs within and across species. The primary goal is to map known miRNA to their most likely orthologues in various species. The user supplies the program with mature miRNA sequences and a genome. The genome is masked for repetitive elements to avoid them being classified as miRNAs in the cases where they have similar sequences to known miRNAs. The supplied mature miRNA sequences are mapped to the genome using Bowtie (Langmead et al., 2009). The program allows no gaps, but few mismatches. Each match is extended to 70 nt at 5' end and 40 at 3' end in case the miRNA lies on the right arm. Similarly, in case the miRNA lies on the left arm, and extension of 40 nt at 5' end and 70 nt at 3' end is applied. Each extended sequence is folded with *RNAfold* (Hofacker et al., 1994). To determine potential candidates, a scoring function and threshold are used to determine the best candidate for the pre-miRNA.

The scoring function depends on $Mismatches$, $Matches$, $PerfectMatches$ and $MatureMismatches$ which describe the structurally unpaired bases, structurally paired bases, perfect Watson-Crick basepairing and mismatches of the mature miRNA structure to the genome, respectively. Moreover a $MismatchPenalty$ (Penalty for mismatches

ensures distinction between sequences which can map to the same loci) is employed as well as the negative free folding energy $-\Delta G$ from the hairpin structure is taken into consideration in the score. The Score S is then given by:

$$S = \frac{Perfectmatch * Match}{Match + Mismatch} + (1 - Maturemismatches) * MismatchPenalty - \frac{\Delta G}{2}$$

MapMi was validated by creating a negative dataset through dinucleotide shuffling the initial miRNA sequences and comparing the scores of positive (initial miRNA) and negative dataset. The positive dataset had a significantly better performance than the negative. Moreover many miRNA from one species could be reproduced with MapMi by providing miRNA from another species to the software, while *S. cerevisiae*, where there is no RNAi, did not yield any miRNA gene candidates. This high level performance as well as the nature in the program of providing customised mature miRNA sequences makes it a valuable tool for our investigation.

3.3 Quest for miRNA in *S. pombe*

3.3.1 Materials and Methods

3.3.1.1 Computational methods

sRNA data acquisition Previously published small RNA (sRNA) sequencing data from two publications sequencing experiments (Halic and Moazed, 2010; Yamanaka et al., 2013, GEO accession numbers GSM492813, GSM1020595, respectively) was used and mapped back to the genome using the maq package (http://maq.sourceforge.net/). It is a package to map short sequences to the genome, allowing two mismatches (upon specification). Multiple sequences can map to the same area of the genome, hence the matched genomic sequence was taken as the representative sequence of those small RNA and considered a potential mature miRNA sequence, if its length was between 18 nt and 24 nt. Note that Halic and Moazed (2010) sequenced the RNA associated with Ago1, while Yamanaka et al. (2013) sequenced sRNA.

3' UTR extraction 3'UTR sequences were extracted based on the CS dataset described in Chapter 2. If there was an apparent alternative CS usage between cycling and cell cycle stage arrested cells (meiotic, quiescent- 24 h of nitrogen depletion and quiescent- 7 d of nitrogen depletion) the 3'UTR sequences were considered to contain a potential target site. The sequences between the most downstream CS in cycling cells and the most downstream CS in cell cycle stage arrested cells was queried for potential miRNA targets. This yielded 3 datasets of sequences for each of the data generated in Yamanaka et al. (2013) and Halic and Moazed (2010) corresponding to 6 datasets of sequences in total.

Control sequences To create a negative test dataset (control sequences) all such sequence-sets were shuffled using uShuffle (Jiang et al., 2008), a bioinformatical tool to shuffle sequences keeping the dinucleotide composition constant. Further the sequences were masked for low complexity regions by dustmasker

(http://nebc.nerc.ac.uk/bioinformatics/docs/dustmasker.html). This avoids using false positive possible miRNAs, which would come from genomic regions with a simple nucleotide composition.

miRanda The software miRanda (Enright et al., 2003) was run in profile mode to identify possible targets of the sRNAs. miRanda run in profile mode, provides a *p*-value for each sRNA complementarity to the 3'UTR based on comparison to the control sequences (this *p*-value is based on the EVD distribution of the miRanda scores as described previously in section 3.1.1). Possible miRNA with promising targets are identified, via the profile presented in Figure 3.2. Low *p*-values indicate a high likelihood of a non- random target match.

MapMi The sRNAs were also submitted to MapMi with the *S. pombe* genome sequence. MapMi computed potential hairpin structures with *RNAfold* (Hofacker et al., 1994) surrounding sRNA and provided the folding structure in "([." notation. Those sRNA that yielded a target with miRanda as well as mapped to a genomic region which could form a hairpin, were visualised with *RNAfold* software (Hofacker et al., 1994).

Comparison of gene expression between potential hairpin and target genes Microarray gene expression data was extracted from Pancaldi et al. (2010), which is available on the Bähler lab resources webpage. Data was collected from the same custom-spotted microarray platform (Lyne et al., 2003). Data describes a wide range of experimental conditions. Among these are environmental stress conditions, cell cycle stage arrest conditions and genetic perturbations. Isogenic control strains were used as a reference for all datasets and normalised to WT. In time series experiments data were normalised to time point 0, which itself was generally discarded to avoid biased data (Pancaldi et al., 2010). I used this data to analyse the correlation in gene expression between potential miRNA and target genes. To compute the Pearson's correlation

```
=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=
Performing Scan: 0_2793_128072 vs contig="SPBC365.06";
=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=

   Forward:        Score: 17.765960  Q:3 to 22  R:534 to 555 Align Len (20) (75.00%) (85.00%)

   Query:    3' taACCCACGTTGGATACAACCA 5
                  ||||||:     |:    ||||||||
   Ref:      5' gcTGGGTGTCCCTGATGTTGGT 3'

   P-Value: 9.344631e-03
   Energy:  -26.420000 kCal/Mol

Scores for this hit:
>0_2793_128072    contig="SPBC365.06";    17.77    -26.42    0.00    9.344631e-03 3    23    534 555    20    75.00%    85.00%

Score for this Scan:
   Seq1                Seq2                   Tot Score   Tot Energy   Max Score   Max Energy   Strand   Len1   Len2   Positions
>>0_2793_128072    contig="SPBC365.06";   17.77       -26.42       17.77       -26.42       10532    22     862    533
Complete
--
```

Figure 3.2: An example of an outcome in miRanda. The sRNA is labelled 0_-2793_128072 (Seq1), from which I can infer which sequence the sRNA had in the original dataset and where in the genome it mapped to. The sRNA could have a potential target in the 3'UTR corresponding to the gene *SPBC365.06* (Seq2), as the *p*-value of the match is comparatively low with respect to other siRNA matched to 3'UTRS. The total energy (Tot Energy) corresponds to the reduction in energy caused by the binding. Maximal energy (Max Energy) presents the reduction in energy caused by all bindings of the siRNA in the 3'UTR (different to total energy, if there are multiple targets present. Len1 and Len2 are the length of the sRNA and the 3'UTR respectively and Position indicates the position where the sRNA maps to the 3'UTR. The strand here is irrelevant, as I am not mapping to the whole genome but to separate 3'UTRs. The Tot Score and Max score are assigned based on the *p*-value and the Tot or Max Energy respectively. Note that the seed-region (6-7 nt starting from position 2 of the 5'end of the sRNA) in this case matches the 3'UTR perfectly.

coefficient, data points need to be independent from each other. The available data with 1272 experimental conditions were therefore further processed to keep only one time point from each time series experiment. One dataset was created where each initial time point was kept and another where each final time point was kept, yielding 341 experimental conditions. Correlation coefficients were computed for each potential miRNA-target pair.

Selection of miRNA candidates for experimental verification Correctly folded hairpins derived from chromosomal loci positive for the sRNA were selected for further

analysis. Non-essential genes containing predicted hairpins were selected for further experimental analysis.

3.3.1.2 Experimental methods

The candidate genes were used for deletion analyses. Experiments were performed by Kelly Tzika, visiting student at the Gullerova Lab.

Transformation The pre-miRNA hairpin structure was deleted and replaced with the *ura4* gene in the *S. pombe ura4* deficient *ura*4-D18 strain by Lithium Acetate transformation (Bähler et al., 1998). RNA purification and 3'RACE on the target gene was performed as described in 2.1.1.

3.3.2 Computational Results

I have identified potential targets from alternatively polyadenylated genes between cycling, meiotic or quiescence (24 h and 7 d nitrogen depletion) arrested cells, whose mature forms are either within the small RNA sequencing data by Yamanaka et al. (2013) or Halic and Moazed (2010).

For further analysis I chose several structures with a *RNAfold* predicted conventional hairpin structure and location of the potential mature miRNA on the stem.

The gene *SPAC1039.04* contains a potential hairpin structure (Figure 3.3) and is not essential to the organism for survival. The mature sRNA was detected in Yamanaka et al. (2013) and maps to the 5' end of the CDS on chromosome I. It has two mismatches to the genomic sequence. The sequence of the sRNA is TCCGTTTCAAGGCGTGA. It could also map to a region in chromosome II but with 3 mismatches. Moreover, the same genomic location within *SPAC1039.04* on chromosome I had another sRNA mapped to it of 20 nt length (with three mismatches). I searched for the occurrence of antisense transcription at the potential miRNA gene in the *S. pombe* Transcriptome Viewer (Wilhelm et al., 2008). I identified a small island of potential antisense transcription at position around 5456000 inside the gene *SPAC1039.04*. However, the sRNA maps to position 5454918-5454936, which excludes any overlap of the sRNA or its precursor with antisense transcription. The sRNA, unless it originated from a different genomic region, is unlikely to be derived from a dsRNA due to antisense transcription. This increases the chance of a real hairpin structure giving rise to the sRNA.

The target was predicted by miRanda within the *SPAC12G12.04* 3'UTR. In the Chapter 2 I mapped the growth condition-specific CS. Interestingly, *SPAC12G12.04* appears to have one longer transcript isoform, which is only present in cycling and quiescent (24 h nitrogen depletion) cells, but meiotic and quiescent (7 d nitrogen depletion) only display two (or three) highly used shorter isoforms. Hence I considered the sequence between the most downstream CS in meiotic/quiescent (7 d) cells and the

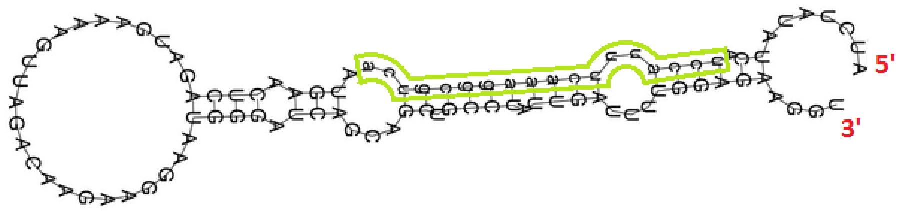

Figure 3.3: The predicted pre-miRNA identified in the gene *SPAC1039.04*. The potential mature miRNA form, which was sequenced by Yamanaka et al. (2013) is outlined in green. Figure was generated using *RNAfold* (Hofacker et al., 1994).

most downstream CS in cycling cells. The way that the sRNA maps to a part of the sequence is shown in Figure 3.4 as a miRanda output.

```
=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=
Performing Scan: 0_4023_835390 vs contig="SPAC12G12.04";
=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=

   Forward:     Score: 19.518090  Q:1 to 17  R:69 to 87 Align Len (18) (77.78%) (88.89%)

   Query:    3' ACTGC-GGAAACTTTACCt 5'
                ||::|  |||||||||||
   Ref:      5' TGGTGAGCTTTGAAATGGt 3'

   P-Value: 4.508284e-03
   Energy:  -21.180000 kCal/Mol

Scores for this hit:
>0_4023_835390  contig="SPAC12G12.04";  19.52   -21.18  0.00    4.508284e-03    1 18    69 87   18      77.78%  88.89%

Score for this Scan:
Seq1,Seq2,Tot Score,Tot Energy,Max Score,Max Energy,Strand,Len1,Len2,Positions
>>0_4023_835390 contig="SPAC12G12.04";  19.52   -21.18  19.52   -21.18  169797  18      304     68
Complete
--
```

Figure 3.4: miRanda prediction for the potential miRNA which maps within *SPAC1039.04* binding to a potential target in the 3'UTR of *SPAC12G12.04*. Results of the miRanda output are to be understood as described before (Section 3.3.1.1). The seed region 2-8 nt from the sRNA 5'end has perfect complementarity to the 3'UTR. This result stems from the differences in CS of cycling cells compared to quiescent (7 d) set of 3'UTR sequences. The result from the set of differences in CS between cycling and meiotic cells is the same, with a slightly different, but same order, *p*-value.

I further analysed whether the potential miRNA gene may affect the predicted target gene by altering its expression. To do so, I looked at the correlation between predicted miRNA and target expression across 341 experimental conditions. Time series experiments were used either at their initial time point or their final time point to ensure independent data points. *SPAC1039.04* and its predicted target *SPAC12G12.04* show a mild, but significant, negative correlation (Figure 3.5). This supports the view that the

predicted miRNA act on their target to degrade the mRNA.

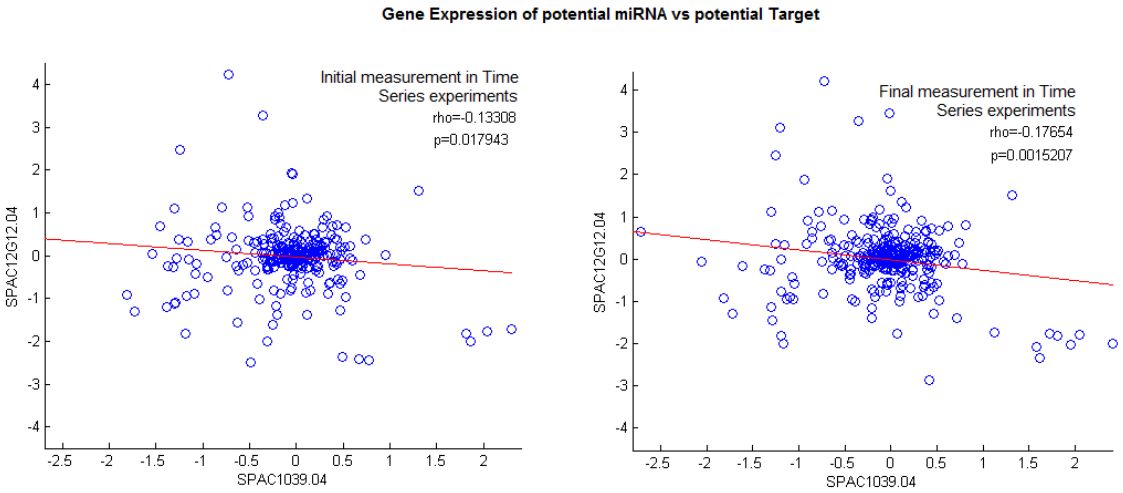

Figure 3.5: Correlation and fitted line between predicted miRNA *SPAC1039.04* and its predicted target *SPAC12G12.04*.

The gene *SPAC1039.04* is the first choice to investigate for effects on the *SPAC12G12.04*-target. More candidates are chosen and prepared for experimental verification and described in Chapter 5.

3.3.3 Experimental Results

All experiments in this section were performed by Kelly Tzika. They illustrate the predictive power of my computational results, and are therefore presented as a confirmation.

S. pombe ura4 deficient cells were transformed via LiAc transformation (Bähler et al., 1998) to delete the hairpin structure within *SPAC1039.04* from its genome. A schematic of the deletion is depicted in Figure 3.6. The cells were selected on EMM-Ura(-) plates.

3'RACE on *SPAC12G12.04* was performed, with the forward primer 200 nt before the stop codon in the CDS. The reproducible result is presented in Figure 3.7A. There is a clear band in *SPAC1039.04* deletion mutant in the 800 nt region, which is much weaker in WT. The *act1* loading control confirms that this is not due to differences in the RNA content. Another band is observed in the 900 nt region, which is entirely

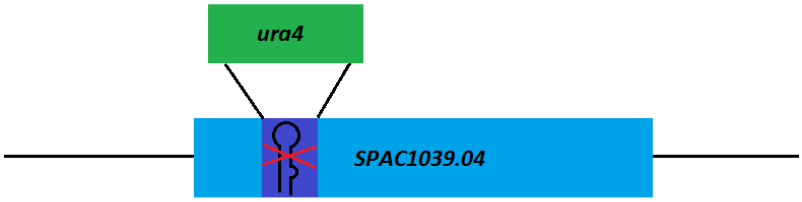

Figure 3.6: Illustration of the hairpin structure deletion from *S. pombe* cells.

absent in WT. Finally, a very faint band (indicated by the red arrow) is also observable in the 1200 nt region and not to be seen in WT. This suggests that *SPAC1039.04* acts on *SPAC12G12.04* and causes the longer transcripts to degrade, so that we do not observe it in WT. I cross-compared our experimental observation with the CS obtained in Chapter 2. *SPAC12G12.04* has coordinates 340107..341855 on chromosome I, minus-strand. A snapshot from the interconditional usage data from my database *Pomb(A)* is provided in Figure 3.7B. The forward primer for the 3'RACE was located 200 nt before its stop-coordinate 340107. The CS are marked at 339435, 339751, 339770 and 339912. Therefore, the bands to expect are ~400 nt, 530-550 nt (corresponding to the two close by CS) and ~900 nt. The thickness of the band in the 550 nt region is explained by multiple cleavage events happening at this distance from the forward primer. The stronger bands larger than 550 nt are explained by the genomic position of the target: 339677, which is about 100 nt after the two CS corresponding to the 550 nt band. As mentioned before, it is not possible to compare strength of CS between each other based on the hit-score (y-axis in Figure 3.7), however, the strength may be compared for the same CS between conditions. Note that all condition arrested cells display stronger usage of the CS than the cycling cells (WT). This may be due to miRNA inaction.

The same 3'RACE experiment was performed in Argonaute knock-out and Dicer knock-out cells (Figure 3.8). It appears, that the 900 nt band is present in elevated quantities in Dicer knock-out cells, but not in Ago1 knock-out cells, compared to WT. This suggests that the processing of the predicted miRNA is indeed Dicer dependent. Interestingly it appears to be Ago1 independent.

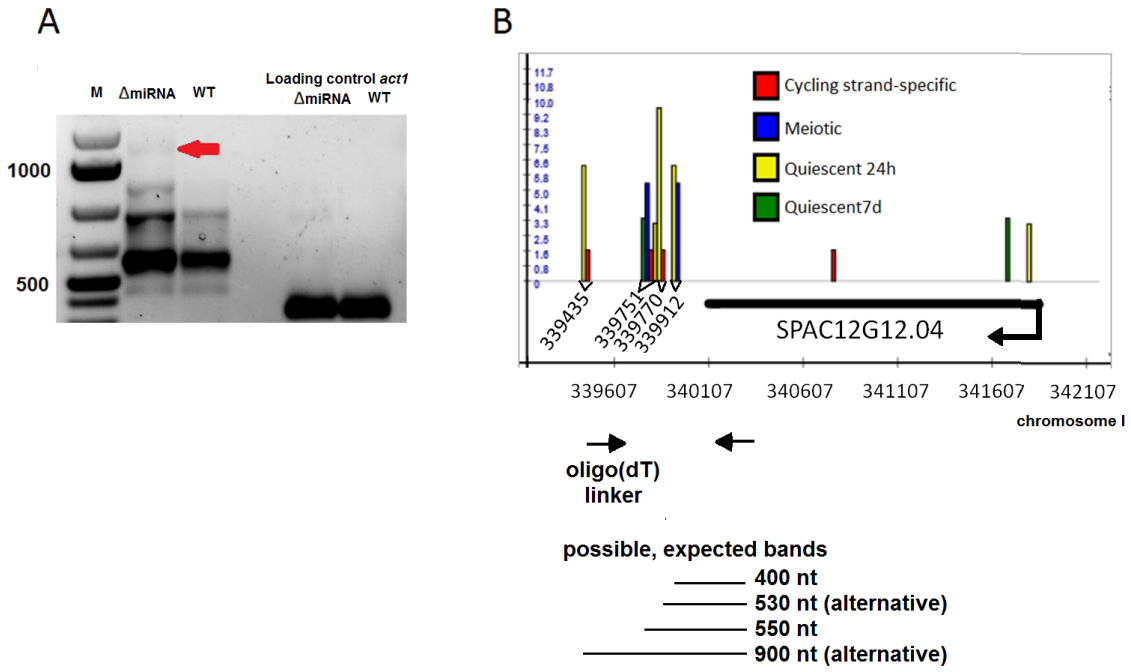

Figure 3.7: **A** 3'RACE outcome on *SPAC12G12.04* in mutant cells with deleted *SPAC1039.04* (ΔmiRNA) and WT cells. Experiment performed by Kelly Tzika. **B** Snapshot of the target gene in my *Pomb(A)* database. CS-positions are marked. All CS are used more in quiescent cells with respect to WT. Below is an illustration of the expected bands from the 3'RACE experiment in A, based on the mapped CS. oligo(dT) was used for RT-PCR, while a linker primer was used for PCR amplification, as described previously.

The target *SPAC12G12.04* (*hsp60*) is a gene involved in cellular response to heat shock, strongly upregulated at higher temperature. We therefore performed an experiment by shifting WT cells from 32°C to 42°C. Interestingly, temperature shifted WT cells show the same bands as the *SPAC1039.04* deletion strain, both showing longer bands than non-temperature shifted WT strain (Figure 3.9). This could be explained by two scenarios. Possibly the longer *hsp60* transcript codes for the functional heat shock protein and is attacked by the predicted miRNA under normal conditions and therefore degraded. With the temperature shift the predicted miRNA may be downregulated or is inhibited from targeting the *hsp60* mRNA otherwise. According to the Bähler lab Transcriptome Viewer, *SPAC1039.04* is not affected by heat shock. Therefore, the more likely scenario

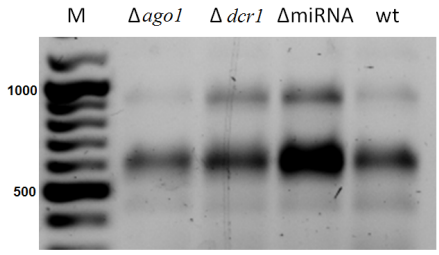

Figure 3.8: 3'RACE outcome on Ago1 and Dicer knock-out cells (Δ*ago1* and Δ*dcr1* respectively). It appears, that the larger band is also elevated in quantities in the Dicer mutant compared to WT. This is not the case for the Ago1 mutant. Experiment performed by Kelly Tzika.

is the possibility, that under heat shock *hsp60* is overexpressed to an extent, that the amount of miRNA targeting it is not sufficient to degrade a significant proportion of longer transcripts. Hence we observe the same bands in heat shock and *SPAC1039.04* deletion strain.

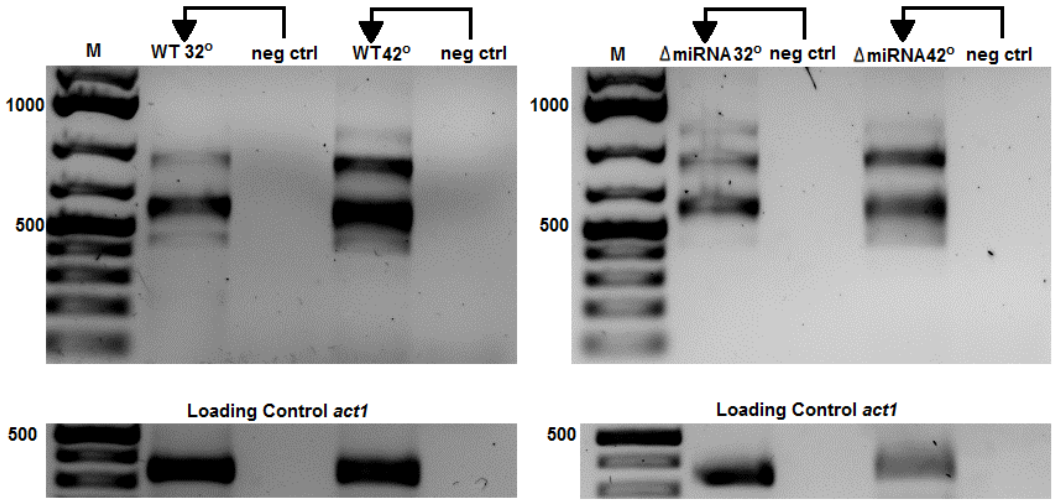

Figure 3.9: 3'RACE on WT (left) and deletion mutant (right) in 32°C and shifted to 42°C . Temperature shifted WT displays the same profile as the deletion mutant. Deletion mutant also shows a higher amount of longer *hsp60* transcripts in temperature shifted cells. neg ctrl denotes the negative control without reverse transcriptase. Bottom panel shows the loading control. Experiment performed by Kelly Tzika.

3.3.4 Discussion and Outlook

So far the appearance of miRNA in animals and plants is believed to have evolved independently (Shabalina and Koonin, 2008). Moreover, it is considered to be non-existent in fungi. While RNAi processing components Dicer and Ago1 are missing in budding yeast, they are present in *S. pombe*. In mammals the enzyme Drosha processes pri-miRNA into the hairpin like pre-miRNA but it is absent in *S. pombe*. However, it has been demonstrated that *S. pombe* Dicer can indeed process hairpin structures into small RNA molecules (Simmer et al., 2010). It is established, that intronic pre-miRNA can bypass Drosha processing (Ruby and Bartel, 2007), so even in animals Drosha is not a crucial element for miRNAs to exist. The absence of Drosha in *S. pombe* leaves space for a miRNA pathway to be similar to the one in plants. In contrast to animals, where Drosha cleaves the pri-miRNA into a pre-miRNA in the nucleus and Dicer acts in the cytoplasm to process the pre-miRNA into the miRNA:miRNA* duplex (Kim et al., 2009), plants lack a Drosha-homologue. Here, pri-miRNA is processed into the pre-miRNA by a Dicer-like protein *DCL1* (Kurihara and Watanabe, 2004; Ruby and Bartel, 2007), which also processes it in the nucleus into the miRNA:miRNA* duplex. Only then is the duplex exported to the cytoplasm by an Exportin-5 homologue for further processing by Ago1 (Park et al., 2005). Most plant studies have been performed on *A. thaliana*. Already in the previous chapter I have described a striking similarity between *S. pombe* and *A. thaliana*. The potential miRNA pathway may result in another one.

So far methods of miRNA detection have focused on evolutionary conservation either in miRNA gene detection or in miRNA target detection. This is illustrated by all the computational methods described before. These discard predicted hairpin structures or fitting target sequences, which are not orthologous to any known miRNA genes/targets in other species. In particular, MapMi (Guerra-Assunção and Enright, 2010), which I used here, expects known mature miRNA sequences from one organism to scan for possible hairpin precursors in other organisms. There is not even an opportunity to discover

miRNA in fungi: the online platform of MapMi offers 102 species to scan for miRNA, the only fungus is *S. cerevisiae* to serve as a negative control with its absent RNAi processing machinery.

I decided to approach the benefits of MapMi to scan for hairpin structures with a slightly different angle than it anticipates. Instead of submitting known miRNA sequences to serve as potential mature miRNA, I submitted small RNA sequences obtained by high throughput sequencing in *S. pombe* to check if they might originate from intragenic hairpin structures. To reduce the number of potential candidates and to facilitate experiments, I further submitted these sequences for miRanda processing, to provide us with potential targets in case these are real miRNA. From the sRNA, which potentially arise from a hairpin precursor and have a possible target according to miRanda and MapMi, I selected candidates for experimental verification.

Our predictions and experiments suggest the one candidate we have tested so far may be a miRNA targeting the heat shock gene under normal conditions. In WT the target gene is still transcribed but has shorter mRNA isoforms than in the miRNA mutant or heat shock affected cells. This explains, why the negative correlation in gene expression between the predicted miRNA and target is not larger: only a fraction of the transcripts would be degraded by the miRNA, but the short transcript isoforms are probably expressed across a wide range of experimental conditions. Interestingly, in mammals *hsp60* is regulated by two microRNAs (Shan et al., 2010). The miRNA and target prediction and indicative experimental data are completion of my work and computational analysis. Further experimental investigations, which is out of scope of this computational study, are currently being performed at the Gullerova lab to verify and classify the miRNA functionally. Northern Blots and the consequent detection of bands of about 70 nt and 22 nt corresponding to the pre-miRNA and the mature miRNA, respectively, are essential.

An interesting experiment to perform would be fractionation of cytoplasmic and

nuclear RNA and probing for the sRNA. Depending on whether fungi could have a miRNA pathway more similar to animals or to plants, the processing into the miRNA:miRNA* duplex would happen either in the nucleus or in the cytoplasm. Either way, we should see a difference between the two compartments. If it turned out to be more similar to animals, PCR of the sRNA would give a ~70 nt band in the nucleus and two bands of ~70 nt and ~22 nt in the cytoplasm. If it happens to be more similar to plants, the reverse scenario would turn out to be the case.

Until recently miRNA were believed to have evolved with multicellularity and independently in plants and animals (Allen et al., 2004). However, it has been demonstrated that a unicellular organism can indeed possess a miRNA pathway as it was demonstrated on the alga *Chlamydomonas reinhardtii* (Molnár et al., 2007). This was the first turn in the perception of miRNA evolution. If we were to be successful in demonstrating the existence of miRNA in fungi, we would overthrow a great volume of current understanding.

Chapter 4

Dicer localises in the nucleus in mammalian cells

In the previous chapter I have analysed the possibility that a lower eukaryote might be more similar to higher ones than previously thought. Here, I am presenting how it might be possible that higher eukaryotes might have kept some of the features as they are seen in lower eukaryotes, either damped in the course of evolution or having evolved independently in lower eukaryotes. The results of this investigation are published in the journal *Nature Structural and Molecular Biology* under the title "Human nuclear Dicer restricts the deleterious accumulation of endogenous double stranded RNA".

Dicer is an RNAse III enzyme and is known to play an important role in gene silencing. It processes dsRNA with perfect or imperfect basepairing. Perfectly paired dsRNA can come from antisense transcription of either overlapping coding genes or endogenous non coding transcription on the opposite strand. Imperfectly paired dsRNA is usually the result of the previously described hairpin structures in the genome. Moreover, dsRNA can enter the cell through a viral infection. In this case the interferon response pathway prevents the virus from spreading throughout the cell population by leading to cellular apoptosis of infected cells (Samuel, 2001).

Dicer is extremely important for the healthy development of an organism. It was

shown to be crucial for early development (Bernstein et al., 2003), oocyte maturation (Murchison et al., 2007; Tang et al., 2007), centromeric silencing in embryonic stem cells (Kanellopoulou et al., 2005), stem cell proliferation (Murchison et al., 2005) and tissue differentiation (Zehir et al., 2010).

Dicer processes dsRNA into siRNA and hairpin structures into miRNA. Both, siRNA and miRNA are required for mammalian gene silencing. These short RNA molecules bind by not necessarily perfect complementarity to target mRNA and induce RNA degradation or translational inhibition in the cytoplasm. Hence this process is referred to as PTGS (post-transcriptional gene silencing), as its occurrence in the cytoplasm excludes interaction with ongoing transcription. As outlined before in *S. pombe* TGS (transcriptional gene silencing) occurs in the nucleus and leads to transcriptionally repressed heterochromatin. In addition, Dicer in fission yeast associates with the nuclear pore complex at the nuclear periphery (Emmerth et al., 2010), the transport "gate" across the nuclear membrane.

TGS has so far been considered not to play a role in mammalian RNAi, but rather PTGS has been its only mechanism. This is due to several factors. Nuclear localisation signals (NLS) describe an amino-acid sequence, which interact with nuclear pore complexes, consisting of proteins called nucleoporins, on the peripheral nuclear membrane. NLS have not been detected in Dicer (Billy et al., 2001; Kotaja et al., 2006; Provost et al., 2002), suggesting that it does not travel across the nuclear membrane and remains in the cytoplasm. In addition, overexpressing GFP-Dicer only makes it detectable in the cytoplasm (Jakymiw et al., 2010).

However, more and more studies suggest that there might be a possibility of Dicer assisting TGS. The C-terminal dsRNA binding domain of Dicer has shown NLS in human cells (Doyle et al., 2013). Moreover, it has been demonstrated that Dicer might interact with nucleoporins independently of NLS (Ando et al., 2011b). In addition, a human Dicer knock-down strain shows an increase in intergenic ncRNA (Haussecker and

Proudfoot, 2005). Furthermore heterochromatic features were observed in mammalian genes. siRNA targeting specific promoter regions induce repressed chromatin (Janowski et al., 2005; Morris et al., 2004), siRNA targeting certain exons cause heterochromatic marks, which decelerate transcription and affect alternative splicing (Alló et al., 2009; Saint-André et al., 2011). In addition, transcriptional gene silencing has been observed by nuclear dsRNA formation through a plasmid with a particular gene fragment between convergent promoters, which can have TGS effects on endogenous homologous genes (Gullerova and Proudfoot, 2012).

4.1 Experimental results for nuclear localisation of Dicer

Our recent publication (mentioned above) investigated the Dicer distribution with human embryonic kidney cells (HEK293), which contained a chromosome-integrated cassette, capable of inducing Dicer small hairpin RNA (shRNA). Dicer could thus be knocked down via treatment with doxycycline. This cell-line has previously been described in Schmitter et al., 2006. Note that Dicer is needed to process the shRNA, which silence it, therefore induced cells still have minimal levels of Dicer. However, it was not detectable by Western Blot as opposed to non-induced cells (data not shown). Dicer localisation is presented in Figure 4.1. DAPI stains the DNA within a cell and therefore visualises the cell-nucleus. Dicer is visualised via a commercially available antibody (Abcam: 13D6). In Figure 4.1 one can clearly see that Dicer is not only on the nuclear periphery where it is necessary for miRNA processing, but also shows a clear signal within the nucleus, which is not detected in induced cells. This indicates that Dicer is indeed not restricted to the cytoplasm as it was previously believed.

In the next step localisation of Dicer was examined by transfecting GFP-tagged Dicer into HEK293 cells, non-induced cells lacked the Dicer signal inside the nucleus (Figure 4.2 upper panel), but GFP on its own was detected inside the nucleus (Figure 4.2 middle panel). Therefore the absence of signal in the nucleus is not a phenomenon specific to the

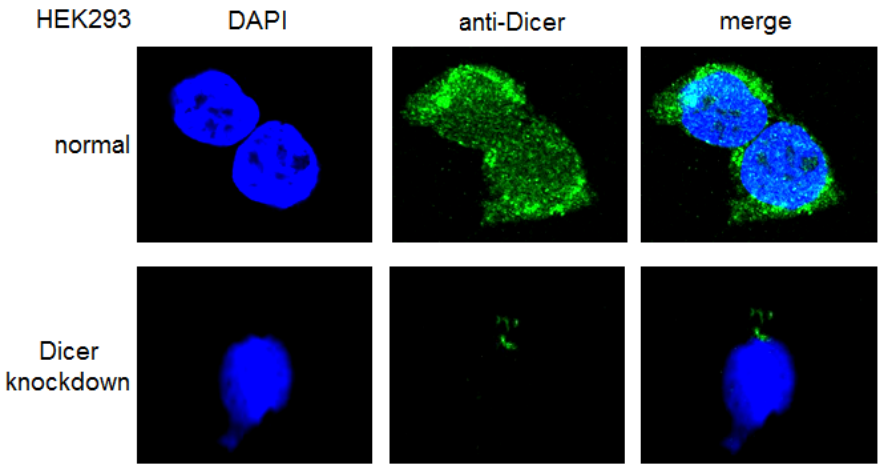

Figure 4.1: Immunofluorescence images of Dicer localisation in single sections. DAPI (blue) visualises the nucleus. **Upper panel:** HEK293 cells without Dicer shRNA induction. Green signal shows the anti-Dicer antibody, also present within the cell nucleus. **Lower panel:** HEK293 cells induced with Dicer shRNA. Significantly lower levels of Dicer are detectable in the cells. Figure adapted from White et al. (2014).

GFP. More crucially, once Dicer was knocked down with shRNA, GFP-Dicer is clearly detectable in the nucleus (Figure 4.2, lower panel). The absence of GFP tagged Dicer in uninduced cells and its presence in induced cells could be due to two possible reasons. On the one hand Dicer transport machinery could have a higher affinity to the untagged Dicer and therefore does not pick up the GFP-Dicer in uninduced cells. On the other hand and more likely, however, is the scenario of possibly controlled levels of Dicer inside the nucleus, preventing the entrance of transfected Dicer into the nucleus.

4.1.1 Dicer associates with PolII through dsRNA

We investigated whether Dicer associates with nascent transcription. Firstly, immunoprecipitation of Dicer co-immunoprecipitated PolII and *vice versa*. The exact percentages are depicted in Figure 4.3, lower panel. Note that two antibodies were employed for PolII precipitation. The 8WG16 differentiates between phosphorylated (PolIIO) and unphosphorylated (PolIIA) CTD and hence provides information whether PolII was transcriptionally active or inactive. Evidently, Dicer associates with

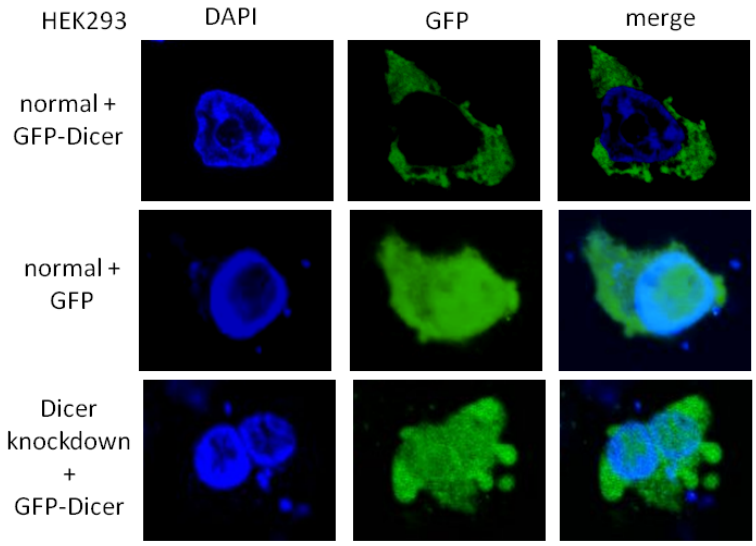

Figure 4.2: Immunofluorescence analysis of transfected GFP-tagged Dicer. **Upper panel:** GFP-Dicer does not localise in the nucleus of normal cells (top panel), while GFP does (**middle panel**). However, upon Dicer depletion GFP-Dicer is detectable in the nucleus. **Lower panel:** GFP-Dicer is also affected by Dicer shRNA expression, but is still detectable with the given expression levels. Figure adapted from White et al. (2014).

transcriptionally active PolII.

To investigate Dicer:PolII association with dsRNA, a dsRNA specific nuclease V1 was employed. After treatment with V1, a substantial reduction in Dicer:PolII association is observed (Figure 4.4). The unaffected interaction of PolII with the elongation factor Spt5 by V1 treatment serves as a control. We therefore conclude that dsRNA is required for Dicer association with transcriptionally active PolIIO.

It was analysed if transcriptional inhibition has an effect on Dicer levels at selected four, usually transcriptionally active, loci. Transcription was inhibited with a chemical α-amanitin. An equivalent reduction of Dicer and PolII at these loci was observed (Figure 4.5).

The same four loci were investigated for sense and antisense transcription with strand specific qRT-PCR. All loci show about tenfold more sense than antisense transcription in uninduced cells (Figure 4.6A). Upon induction, we observe a striking increase in both sense and anti-sense transcription (Figure 4.6B and C). Moreover, treatment of Dicer

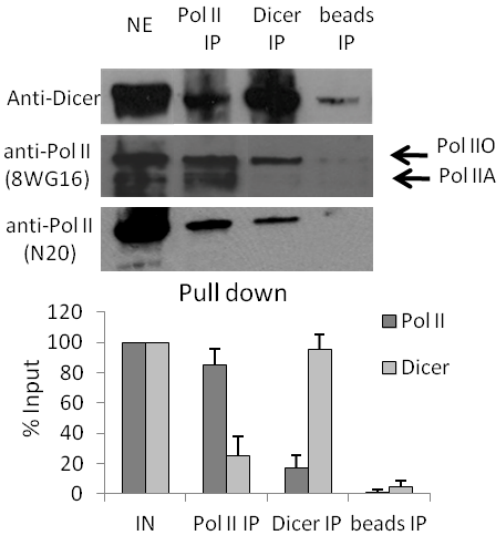

Figure 4.3: Western Blot analysis of PolII and Dicer. PolIIO signifies hyper-phosphorylated form of PolII, PolIIA signifies unphosphorylated, representing transcriptionally active and inactive PolII, respectively. Image-Quant software measure the levels of pull down and are expressed in percentages of input. They are depicted in the lower panel. The quantitation of PolII was measured on the upper band of the 8WG16 antibody experiment. NE signifies Nuclear Extract. Figure adapted from White et al. (2014).

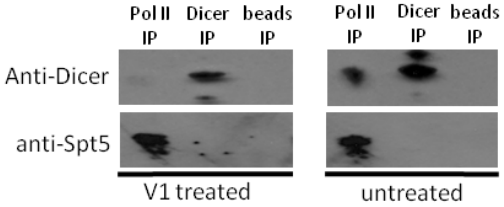

Figure 4.4: Co-IP experiment. Dicer:PolII association in is lost after dsRNA-specific nuclease (V1) treatment. Spt5 serves as a control. Figure adapted from White et al. (2014).

knock down cells with V1, which degrades dsRNA, showed a decrease in antisense transcription (Figure 4.6D). As a negative control the effects of Dicer knockdown on sense and antisense transcription were tested in two non-Dicer associated loci (data not shown). Knock-down of Dicer did not influence transcription in these loci, implying that knocking down Dicer does not influence transcription in general but only in Dicer associated chromatin loci. This suggests that transcription at Dicer loci forms dsRNA,

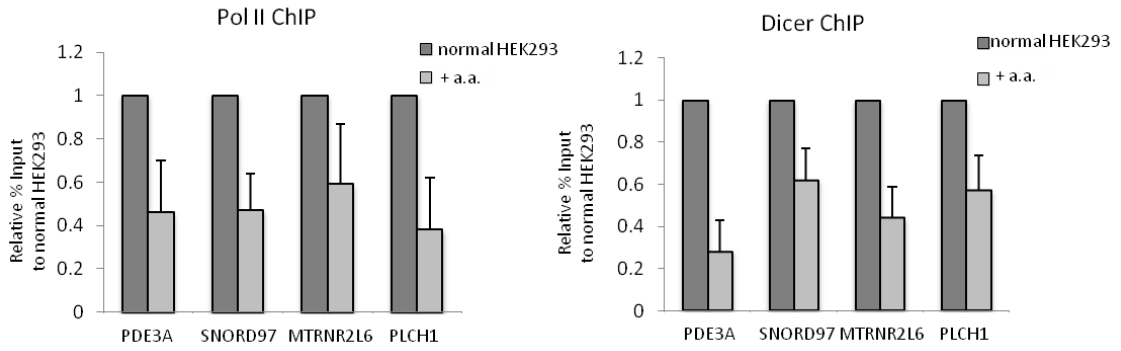

Figure 4.5: Left side panel show PolII ChIP experiment. The indicated genes with Dicer localisation were checked for PolII levels in HEK293 cells with and without α-amanitin treatment. The levels are computed as % of Input. Three independent biological experiments were performed and averaged, error bars indicate one standard deviation. Right-hand side panel shows a Dicer ChIP experiment. Figure adapted from White et al. (2014).

which accumulates upon Dicer knock-down.

To analyse possible TGS effects the four loci were subjected to ChIP experiments of Argonaute, which is necessary for siRNA-mediated TGS in human cells (Kim et al., 2006). Moreover, the heterochromatin mark H3K9me2 was investigated, normalised to H3 occupancy, to avoid imprecision due to nucleosome density. All signal was compared to the non-Dicer associated 28S rRNA locus. Substantial increase compared to the negative control in Ago1 (Figure 4.7A) and H3K9me2 (Figure 4.7B) was observed. Moreover PolII ChIP in Dicer knock down cells compared to uninduced cells provides evidence of increased PolII activity (Figure 4.8). This provides further evidence of potential loss of repressive chromatin. We therefore conclude that Dicer acts with Ago1 on specific mammalian gene loci to promote assembly of H3K9me2 heterochromatic marks.

A dsRNA specific antibody J2 provides the opportunity to observe dsRNA accumulation within the cell via immunofluorescence. The specificity of J2 to dsRNA was validated *in vitro*. In Figure 4.9 it is visible, that dsRNA does accumulate on the nuclear periphery and within the cell nucleus upon Dicer knockdown. Figure 4.9A presents this

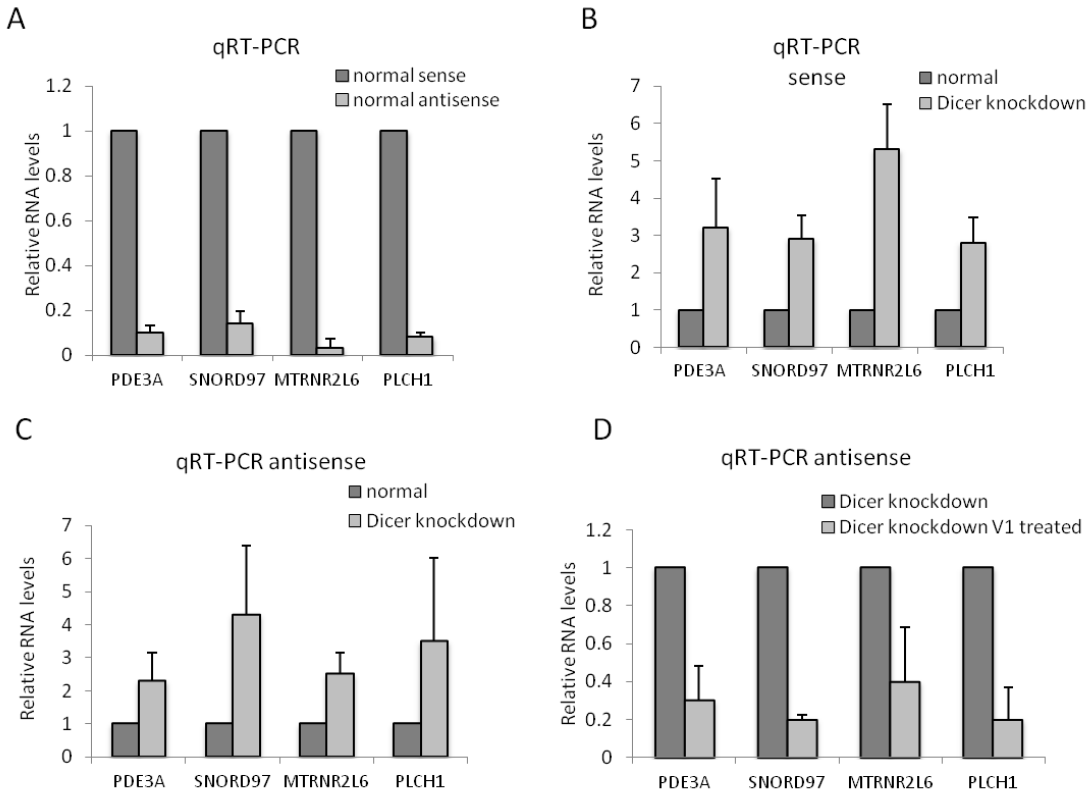

Figure 4.6: qRT-PCR analysis of transcripts in Dicer binding loci in HEK293 cells. All levels are based on 3 independent biological repeats, error bars indicate one standard deviation. **A** Levels of sense and antisense transcripts, normalised to the levels of sense transcripts. **B** Levels of sense transcripts in normal and Dicer knock-down cells, normalised to levels in normal cells. **C** As B but with antisense transcripts. **D** Antisense transcripts at Dicer binding loci are sensitive to V1 treatment, which digests dsRNA. V1 treatment occurred before RNA isolation. Levels are normalised to values in untreated cells. Figure adapted from White et al. (2014).

effect in human HEK293 cells, while B shows the same result for mouse embryonic stem cells.

It was investigated whether co-localisation of Dicer, PolII and dsRNA means that Dicer processes the dsRNA into siRNA. Three out of the four previously used Dicer localisation loci were checked. By using a Northern blot technique designed to enhance small RNA detection, it was identified that the three loci indeed have siRNA detectable, intriguingly their levels are notably reduced upon Dicer knock-down (Figure 4.10).

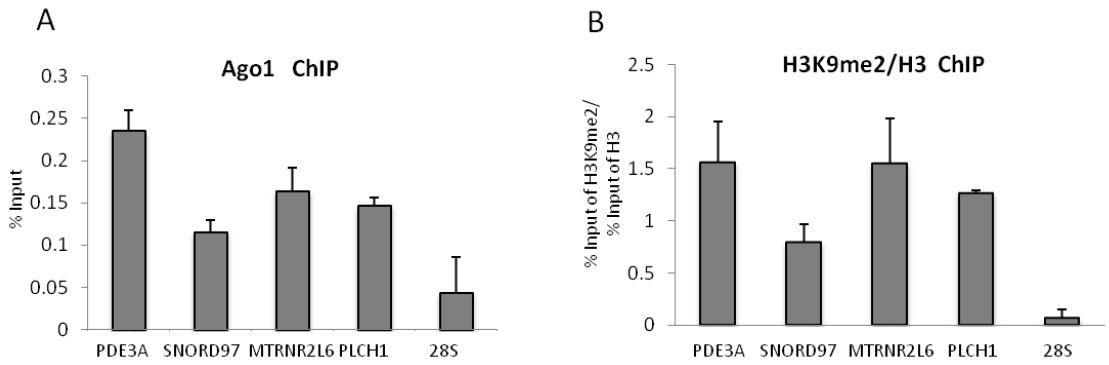

Figure 4.7: ChIP analysis in normal HEK293 cells at the four previously analysed Dicer binding loci. The levels of 28S rRNA, where no Dicer was detected, serve as a control. ChIP values are based on 3 biological repeats, error bars indicate one standard deviation. **A** Ago1 ChIP **B** H3K9me2 ChIP versus H3 levels. Figure adapted from White et al. (2014).

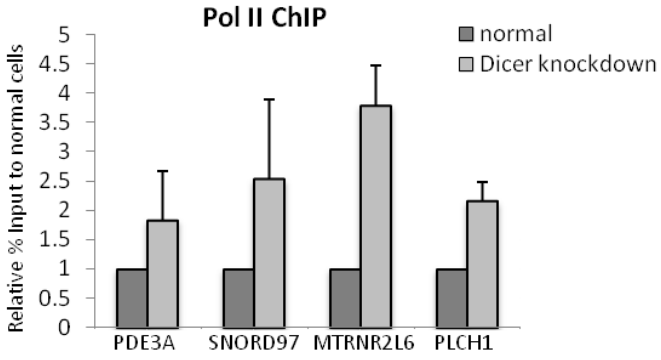

Figure 4.8: PolII ChIP, procedure as in Figure 4.7. Figure adapted from White et al. (2014).

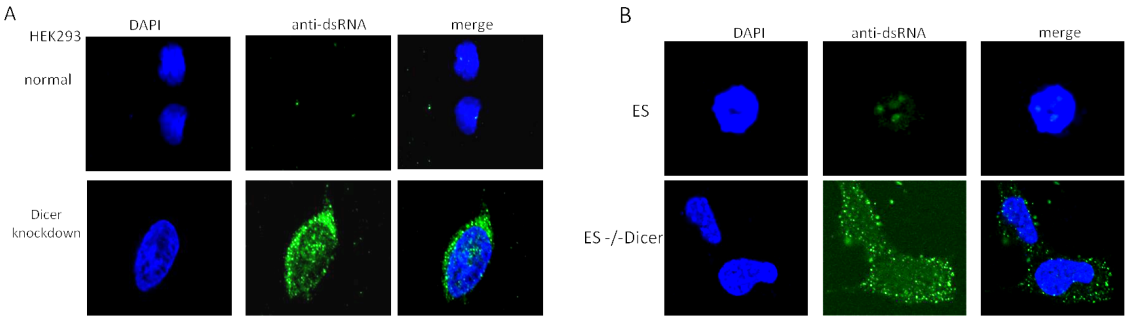

Figure 4.9: Immunofluorescence analysis of dsRNA, detected by a J2 antibody. Blue DAPI strains the nucleus, green signal depict dsRNA. Figure adapted from White et al. (2014).

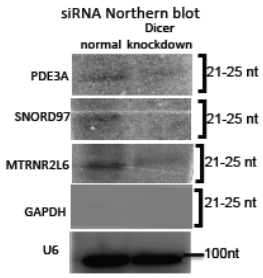

Figure 4.10: Small RNA isolated from HEK293 cells with and without Dicer knocked down. The sample was resolved on a 20% PAA gel and treated with EDC for Northern Blotting. A ^{32}P-radio-labelled oligonucleotide served as a probe for Dicer binding loci. GAPDH (not a Dicer binding locus) served as a negative control, while U6 snRNA served as a loading control. Visualisation occurred via PhosphoImager. Figure adapted from White et al. (2014).

4.1.2 The interferon response pathway is triggered by loss of Dicer

dsRNA may enter the cell in the form of a virus. Therefore mammalian somatic cells developed the interferon response pathway (Stetson and Medzhitov, 2006), which in order to prevent the virus from spreading through the cell population, leads to cellular apoptosis. We speculated that the observed accumulation of dsRNA upon loss of Dicer may activate the interferon response pathway. TLR3 and PKR1 are two key proteins in the interferon response pathway (Garcia et al., 2007; Kawai and Akira, 2011). After treatment of HEK293 cells with Dicer shRNA for one and two weeks, loss of Dicer was obvious and both proteins were elevated as observed by Western Blot (Figure 4.11A). Moreover, other interferon induced proteins (INFβ and OAS1) were also upregulated in Dicer knockdown cells (Figure 4.11B). The experiments were repeated with a different Dicer specific shRNA and confirmed the result (data not shown), leading us to believe that the interferon induction is Dicer specific.

Flow cytometry was implemented to confirm that Dicer knock down cells indeed undergo apoptosis. Cells were stained with an apoptotic marker (See Appendix E). Moreover, 7-AAD was used to bind dead cells, therefore cells undergoing apoptosis, but not dead yet, could be observed. Dicer knock- down cells showed a three-fold increase of cells undergoing apoptosis. To exclude the possibility that this is due to a disturbed

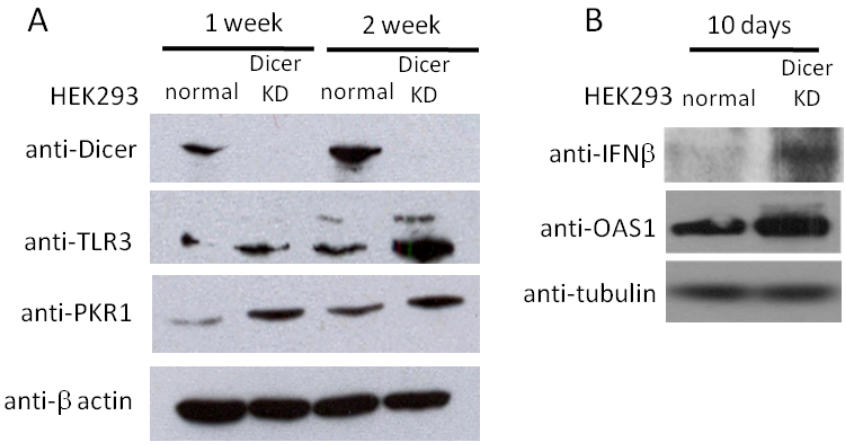

Figure 4.11: Western Blot of normal and Dicer knock-down cells. Antibodies are specific to each protein analysed. **A** knock-down cells were cultured for 1 and 2 weeks. We expect the upper band of the Anti-TLR3 antibody corresponds to a modified protein. **B** Dicer knock-down cells were cultured for 10 days. Figure adapted from White et al. (2014).

miRNA-processing pathway or shRNA expression, apoptosis was measured in Drosha and PKR knock-down cells. No difference in apoptotic cells was observed (Figure 4.12)

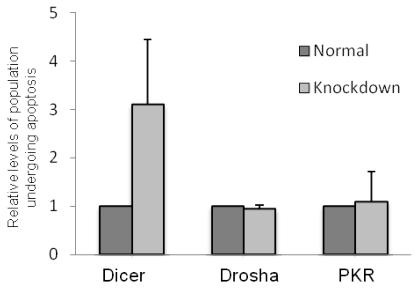

Figure 4.12: Flow cytometry to identify cell undergoing apoptosis. Vertical axis represents the fraction of cells undergoing apoptosis in Dicer, Drosha and PKR knock-down cells as marked on the horizontal axis. Data is based on three independent biological replicates, error bars represent one standard deviation. Figure adapted from White et al. (2014).

Evidence is presented that mammalian cells are indeed capable of TGS. It was shown, that Dicer also localises in the nucleus against previous perceptions, but its levels seem to be regulated. The above experimental data suggests that Dicer binds to the chromatin and interacts with PolII, as well as co-localises with dsRNA and small RNA. Moreover, in Dicer knock-down cells dsRNA accumulates at the co-localisation loci. I extend these

results to a genome-wide perspective by analysing ChIP-Seq and RNA-Seq data. Overall the joint experimental (described above) and genomic (presented below) demonstrate that Dicer acts as inhibitor of dsRNA accumulation in the cell to prevent programmed cell death by the interferon response pathway.

4.2 Materials and Methods

Materials and methods for all experiments described in the introduction, which were performed by Eleanor White, can be found in the Appendix E and in the publication White et al., 2014. These include preparation of tissue culture, microscopy, protein analysis and flow cytometry. Dicer ChIP-Sequencing (ChIP-Seq) was prepared by Eleanor White, the raw data was mapped back to the genome and the top 119 peaks (118, which are not mitochondrial) were called by Kinga Kamienarz-Gdula. The exact method is also described in Appendix E.

All analysis presented below was performed by me.

4.2.1 RNA-Seq

dsRNA samples from WT and Dicer-knockdown cells were submitted for sequencing and alignment at the Wellcome Trust Centre. The sample preparation is described in the Appendix E. The alignment was performed with BOWTIE to hg19 allowing maximally 2 mismatches. dsRNA mapping displays no continuous background signal, so peaks are considered to be non-0 hits extended to both sides until 0 was reached. To compare real dsRNA peaks to each other, I first combined multiple peaks into one, if they overlapped only one peak in the other dataset. This was done iteratively until no change in either dataset was observed. For further analysis I also imposed a cut off of 1500 reads in the peak summit. This guaranteed that I analysed only real dsRNA peaks, which were not due to background signal.

4.2.2 dsRNA peak length distribution

Boxplots were created using MATLAB. The median is presented as a red line, upper and lower quartiles (q_3 and q_1 respectively) are presented as a blue box. The whiskers are

given by $q_3+1.5(q_3-q_1)$ (upper) or $q_1-1.5(q_3-q_1)$ (lower). The default of 1.5 corresponds to approximately $+/2.7\sigma$ (where σ is the standard deviation) and 99.3% coverage if the data are normally distributed. Outliers are plotted as red crosses if their value is higher than the upper whisker or smaller than the lower. The plotted whiskers extend to the adjacent value, which is the most extreme data value that is not an outlier.

4.2.3 Metagene Analysis

To map sequences to a genomic region (distal promoter, promoter, gene body, terminator, intergenic), I used the human hg19 Refgene annotation, available for download at http://genome.ucsc.edu/. Promoter is defined as 1kb up and downstream of the annotated transcription start site (TSS). Distal promoter is between 5kb and 1kb before the TSS. Gene body is defined as the region between 1kb downstream of TSS and the annotated transcription end site (CS). Terminators are defined as the region 5kb downstream of the CS. Everything else is considered intergenic. Dicer peaks were mostly wide and flat, so I used their midpoints as a reference point, to assign them to a region. dsRNA peaks were often asymmetrical, so I used their peak summit as a reference point. The summit was computed as follows. If there is only one global maximum in the peak, that is its summit. If there are several points with the same maximal value, I take the point which is closest to the average of the positions of these maxima. All calculations were performed with custom perl scripts.

I plotted Dicer and dsRNA across a metagene. This refers to average data values and Dicer/dsRNA localisation with respect to the TSS and CS. To plot data values across a metagene, I used the average of all data points at a given distance to TSS and CS. To plot localisation, I only differentiated between “presence ” and “absence”, contributing a summand of 1 or 0 per position, respectively. The resulting sum was averaged over all considered peaks in dataset (Dicer and dsRNA).

4.2.4 Overlap of peaks with repetitive Elements

I extracted genomic repetitive elements from the RepeatMasker hg19 table. Repetitive elements overlapping a Dicer or dsRNA peak with 5 nt or more were included in the analysis. The separately analysed categories of repetitive elements were LTRs, LINEs, SINEs, as well as Satellite sequences, rRNA and tRNA encoding regions. I sampled a background distribution for each set of peaks (Dicer and dsRNA) based on random genomic sequences. Each trial had the same amount of sequences as peaks in the set. Moreover, they were taken of the same lengths as the peaks (depending on set under consideration). Each background distribution is based on sampling more than 1000 times. The p-value was computed based on how many trials have the same or greater amount of overlap with each of the categories of repetitive elements (right-handed p-value).

4.3 Results

4.3.1 Dicer localises in the cell nucleus

ChIP-Seq to identify Dicer localisation on the chromatin was performed. The sequencing provided candidate genomic loci for further experimental investigation, but also revealed the Dicer distribution on the chromatin. The top 118 non-mitochondrial Dicer peaks on the chromatin were identified (Appendix E). I have used the RefGene annotation to view the Dicer distribution with respect to gene regions. Roughly 50% of Dicer peaks fall into intergenic regions. The majority of the remaining 50% are in the gene body (Figure 4.13)

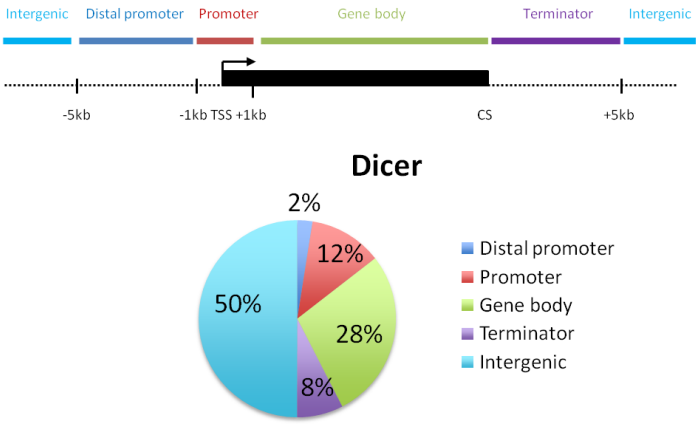

Figure 4.13: Distribution of Dicer peaks relative to genomic regions, as defined in Materials and Methods Section 4.2.

I then proceeded to analyse the distribution of Dicer across a metagene. Figure 4.14 shows the average distribution of Dicer ChIP seq hits with respect to the TSS and CS. The scores peak right after the TSS, and before the CS. However disregarding the number of hits and simply looking at the presence of Dicer or not, I present that Dicer preferentially localises just before the TSS and CS (Figure 4.15). The discrepancy between the two figures around the TSS could be due to the fact that one or few Dicer peak(s) have a particularly large summit after the TSS, hence skewing the average in that position. To gain insight into this issue, I have also plotted the median of Dicer scores in these positions (Figure 4.16). I conclude that indeed, while overall Dicer localises preferentially before

the TSS, some peaks reach their summit after the TSS.

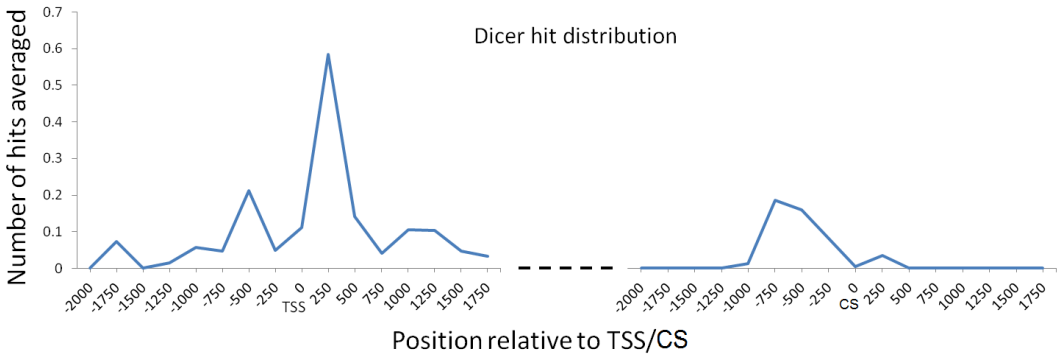

Figure 4.14: Metagene analysis of top 118 Dicer binding sites (ChIP-Seq peaks) showing the distribution of number or hits around TSS and CS of coding genes in the human genome. Data were grouped into 250 nt bins and averaged over 250 nt. Each bin-midpoint was plotted on the graph.

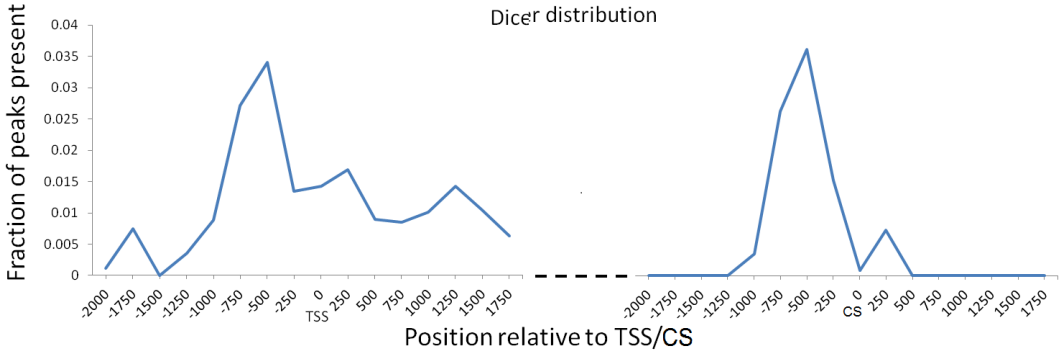

Figure 4.15: Metagene analysis of top 118 Dicer binding sites (ChIP-Seq peaks) showing the location distribution around TSS and CS of coding genes in the human genome. Data were grouped into 250 nt bins and averaged over 250 nt. Each bin-midpoint was plotted on the graph.

I then checked what kind of genomic regions are more prone to Dicer binding than others. Coordinates of repetitive Elements were obtained from the RepeatMasker table, any overlap of 5 nt with a Dicer peak was considered. It turns out that while LINEs, SINEs and LTRs are significantly under-represented among the Dicer peaks, rRNAs, tRNAs and Satellite sequences are significantly over-represented (Table 4.1). *p*-values and expected values are based on computed background distribution, as described in the Materials and Methods (Section 4.2). The significant overlap of Dicer with

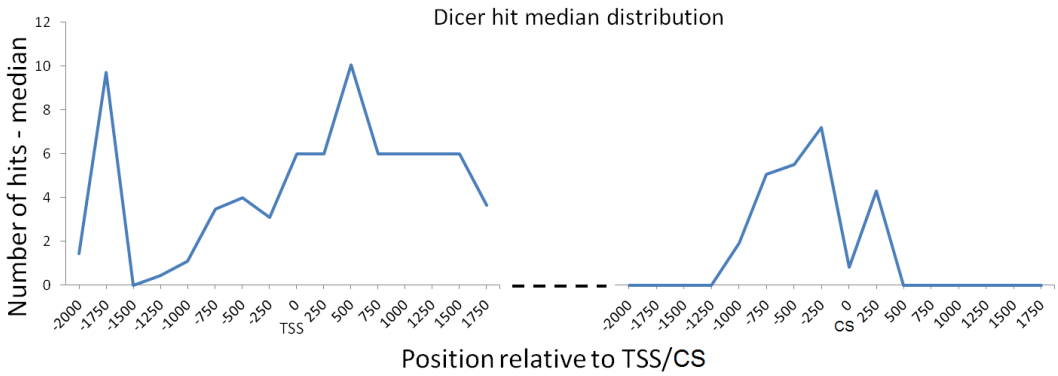

Figure 4.16: Metagene analysis of top 118 Dicer binding sites (ChIP-Seq peaks) showing the median hit distribution around TSS and CS of coding genes in the human genome. Data were grouped into 250 nt bins and averaged over 250 nt. Each bin-midpoint was plotted on the graph.

Satellite regions is interesting. Satellite regions are the main structural constituent of heterochromatin (Lohe et al., 1993). There is a possibility that Dicer might play a role in heterochromatin assembly or at least be recruited to heterochromatic regions. Moreover, these regions could be the reason why only a limited amount of PolIIO co-IPs with Dicer, as heterochromatic regions are transcriptionally inert.

Due to the previous observations of sense and antisense transcription elevation upon Dicer knockdown (see introduction to this chapter), dsRNA was subjected to RNA sequencing (RNA-Seq, dsRNA preparation described in Appendix E) and the sequences were mapped back to the genome (see Materials and Methods, Section 4.2). Indeed, as illustrated at a specific gene locus (5' flanking region of KPNA2), a striking co-localisation of Dicer and dsRNA on the chromosome is observed. Moreover, dsRNA levels increase upon Dicer knockdown (Figure 4.17).

I therefore checked whether this increase is common for Dicer loci. We should note that not all detected Dicer peaks co-occur with dsRNA as I have defined the dsRNA peaks (Section 4.2). Upon processing of our peaks as described in Materials and Methods, the peaks in normal and induced cells have comparable length distribution (Figure 4.18) and were subjected to further analysis.

I investigated the localisation relative to genomic regions of dsRNA, similar to the

type	number of dicer peaks overlapping the elements	expected	p-value (right handed)	significance
LINE	20	31	0.996	Significantly underrepresented
SINE	9	27	1	Significantly underrepresented
LTR	8	13	0.957	Significantly underrepresented
Satellite	16	1	0	Significantly overrepresented
rRNA	15	0	0	Significantly overrepresented
tRNA	38	0	0	Significantly overrepresented

Table 4.1: Table summarising the enrichment of repetitive DNA sequence elements in the Dicer ChIP-Seq data. *p*-values and expected values are based on a background distribution of sampling 118 random genomic sequences of the same length as Dicer peaks. More details can be found in Materials and Methods (Section 4.2). Sampling was performed more than 1000 times.

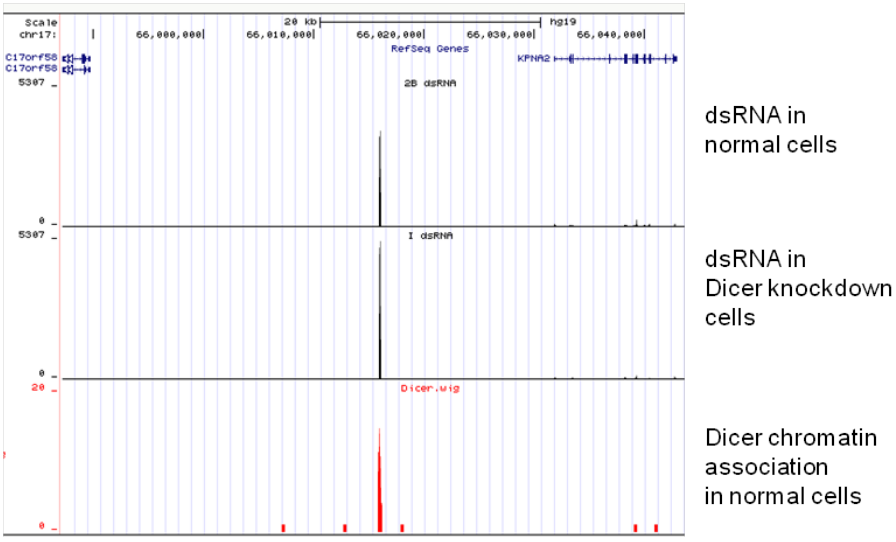

Figure 4.17: Snapshot from the uploaded data to the UCSC genome browser, depicting the intergenic region before KPNA2. From top to bottom the panels depict dsRNA sequencing hits in normal cells, dsRNA sequencing hits in Dicer knock-down cells and Dicer ChIP-Seq levels in normal cells. Note that levels are not normalised to sequencing depth, but dsRNA sequencing depth is comparable in both samples.

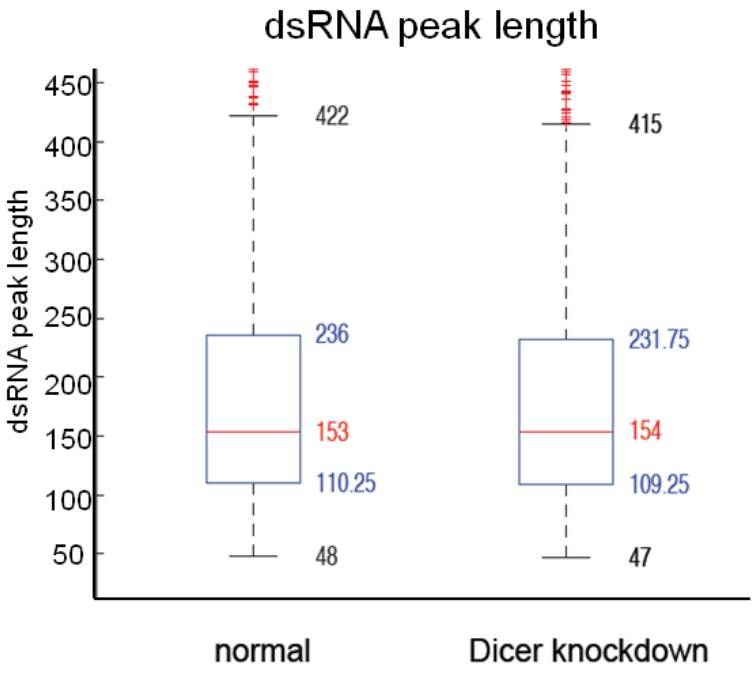

Figure 4.18: Boxplot of dsRNA peak length in normal and Dicer depleted cells. The red line indicates the median, lower and upper quartiles are denoted by the blue box. All outliers are plotted as red crosses, the whiskers extend to the adjacent value. For more details see Materials and Methods Section 4.2

Dicer peaks before (Figure 4.19). Roughly 60% of the dsRNA corresponds to genic regions. This number is slightly higher than for the Dicer peaks (Figure 4.13). I suspect this is due to background undigested mRNA in the dsRNA preparation.

I have checked the distribution of dsRNA across a metagene. Considering dsRNA hits, I see an increase in dsRNA before the TSS and the CS (Figure 4.20). However, as with Dicer, this could be due to one or few particularly dominant peak(s). I hence investigated general presence of dsRNA peaks (Figure 4.21). I still observe the preferential localisation before the TSS, but the CS displays a much broader distribution of dsRNA around it. Hence, in order to explain the enrichment in scores of the dsRNA before the CS and the suspicious extra peak about 2000 nt before the CS, I plotted the median of dsRNA hits (Figure 4.22). Indeed it appears the dsRNA generally has its summit before the TSS and CS, while the peak 2000 nt before the CS was an artefact of taking averages. Note that

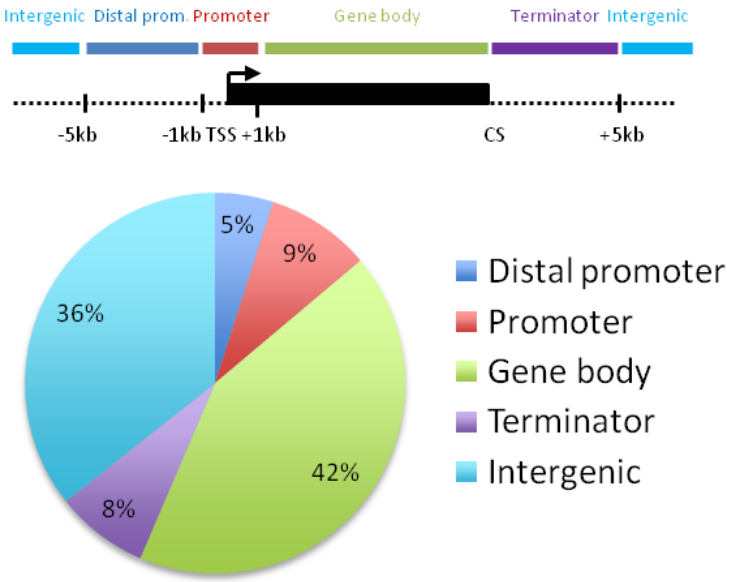

Figure 4.19: Analysis of dsRNA peak distribution relative to genomic regions as defined by the Refgene coordinates. Exact definitions of genomic regions can be found in the Materials and Methods Section 4.2.

this is in a slight contrast to Dicer peaks, where many peaks seem to reach their summit after the TSS (Figure 4.16).

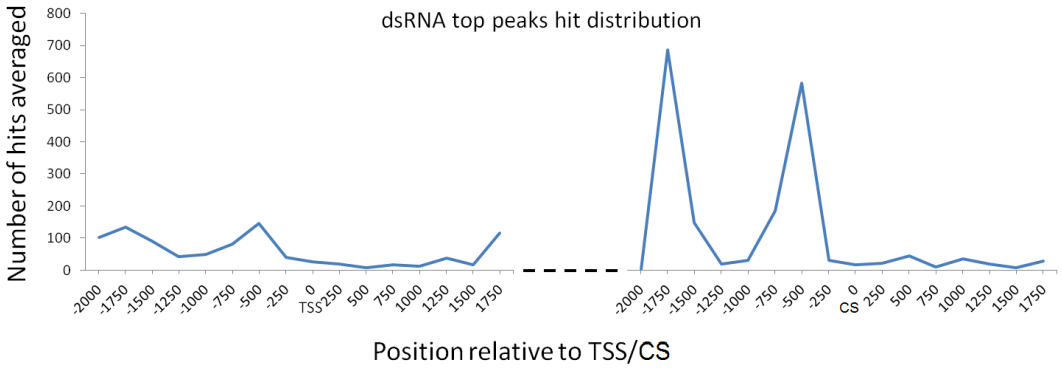

Figure 4.20: Metagene analysis of top dsRNA peaks (RNA-Seq peaks) showing the distribution of number or hits around TSS and CS of coding genes in the human genome. Data were grouped into 250 nt bins and averaged over 250 nt. Each bin-midpoint was plotted on the graph.

Similar to the Dicer peaks I have also investigated if dsRNA overlaps with repetitive elements in the genome (Table 4.2). dsRNA does not seem to be derived from repetitive elements such as LINES, SINES, LTR and Satellite sequences, which are

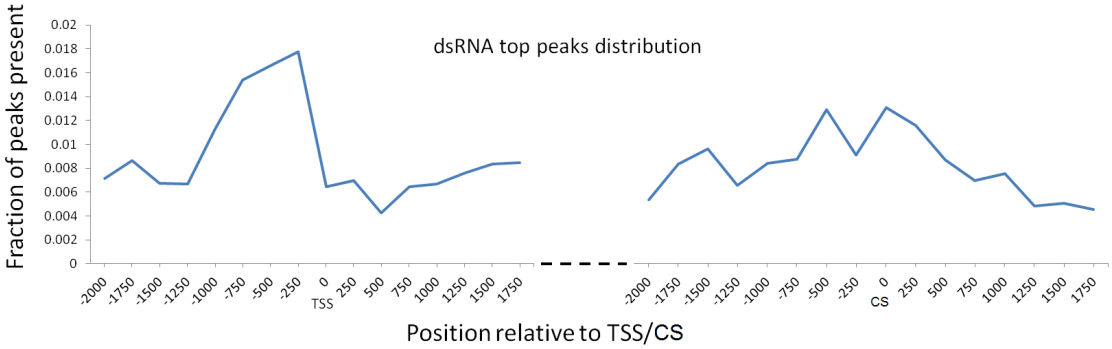

Figure 4.21: Metagene analysis of top dsRNA peaks (RNA-Seq peaks) showing the location distribution around TSS and CS of coding genes in the human genome. Data were grouped into 250 nt bins and averaged over 250 nt. Each bin-midpoint was plotted on the graph.

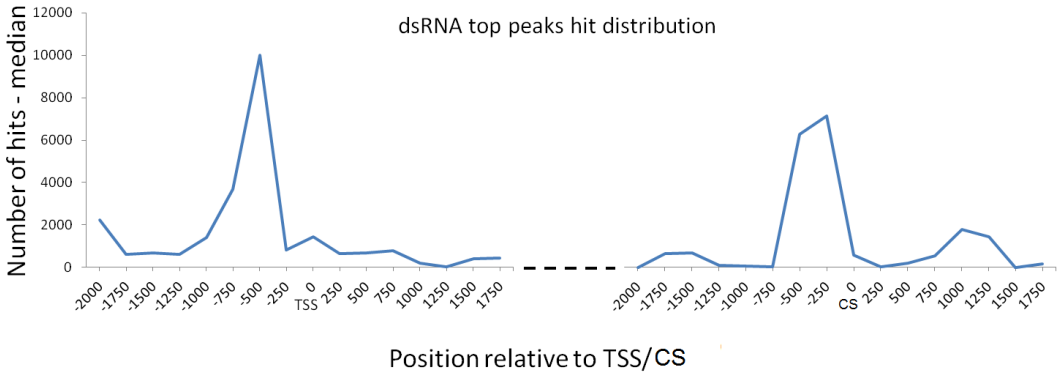

Figure 4.22: Metagene analysis of top dsRNA peaks (RNA-Seq peaks) showing the median hit distribution around TSS and CS of coding genes in the human genome. Data were grouped into 250 nt bins and averaged over 250 nt. Each bin-midpoint was plotted on the graph.

significantly underrepresented in the set. However, rRNAs and tRNAs are significantly overrepresented.

I have overlayed Dicer and dsRNA data with genomic PolII ChIP and Input data. These data were obtained from the ENCODE Transcription Factor Binding Scores by ChIP-Seq from Stanford/Yale/UCSC/Harvard. I have employed non stringent criteria to analyse a subset of Dicer peaks of more reliable size. I merely consider Dicer peaks where PolII signal > Input signal to ensure PolII presence. Out of the original 118 peak, I am left with 49 peaks to analyse. Within these 49 peaks, I actually observe a majority of them overlapping our defined dsRNA peaks (Figure 4.23A). Moreover, the majority of

type	number of dicer peaks overlapping the elements	expected	p-value (right handed)	significance
LINE	71	211	1	Significantly underrepresented
SINE	70	153	1	Significantly underrepresented
LTR	17	90	1	Significantly underrepresented
Satellite	0	5	1	Significantly underrepresented
rRNA	200	0	0	Significantly overrepresented
tRNA	188	0	0	Significantly overrepresented

Table 4.2

these Dicer loci display an increase of dsRNA upon Dicer knock-down. Strikingly none of them show a decrease. In correspondence to PolII presence, 39 peaks out of these 49 were genic Dicer peaks. The percentage of genic peaks, where dsRNA increases upon Dicer knock-down is remarkable (Figure 4.23B) and in stark contrast to intergenic Dicer peaks (Figure 4.23C).

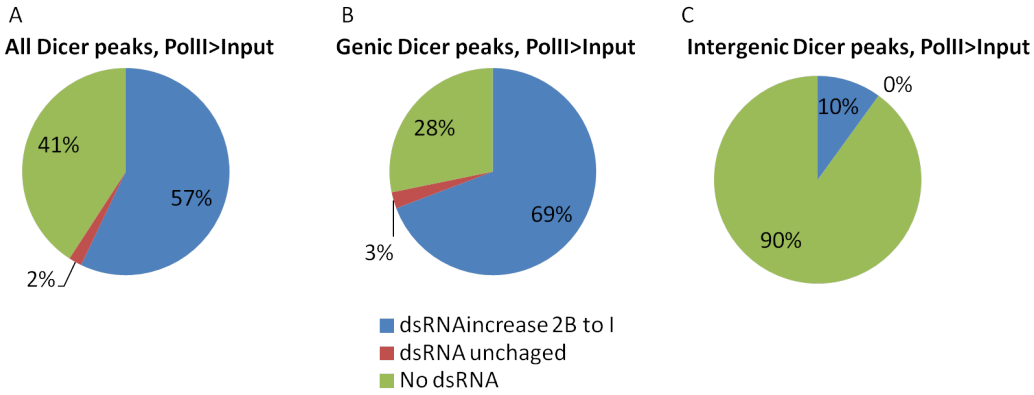

Figure 4.23: Co-occurrence of Dicer and PolII ChIP peaks with dsRNA. PolII presence was defined non-stringent: PolII must be greater than Input. This left us with **A** 49 peaks overall. A majority of these co-localise with dsRNA. Note that the dsRNA levels do not decrease but mostly increase at these loci. **B** As before, but Dicer peaks which localise inside genic regions: 39 peaks overall. An even larger fraction of these peaks display co-localisation with dsRNA. **C** Same as before, but intergenic Dicer peaks (10 in count)

I investigated the previously suggested processing of dsRNA to siRNA by Dicer

(4.10) on a genomic level. I overlayed the data of sRNA sequencing in IMR90 cells downloadable from the UCSC genome browser (Fejes-Toth et al., 2009). It does not contain data for chromosome Y, so I also took those Dicer peaks out of consideration. Two examples are shown in Figure 4.25. From the previously analysed loci of Dicer-PolII-dsRNA co-occurrence (29 peaks), I checked how many co-occur with sRNA. Note that I do not make a distinction between intergenic and genic, as there was only one intergenic. The intergenic peak happened to co-occur with sRNA. As one can see for all considered Dicer peaks (the figure does not change for genic Dicer peaks), almost all of them overlap with sRNA (Figure 4.24). Again, most of the Dicer-dsRNA-PolII-sRNA peaks are associated with dsRNA accumulation upon Dicer knockdown. This suggests that these sRNAs are in fact siRNAs. We deduct that loss of Dicer causes accumulation of dsRNA throughout the cell, which would usually be quickly processed by PolII associated Dicer. Moreover, we observe the accumulation at Dicer chromatin binding sites, which may correlate with overlapping sense and antisense transcription.

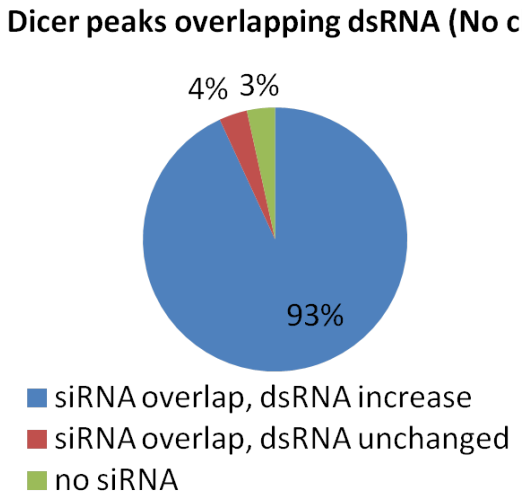

Figure 4.24: Overlap of loci with Dicer, PolII and dsRNA with sRNA. Nearly all such loci also give rise to sRNA. 29 peaks considered in total.

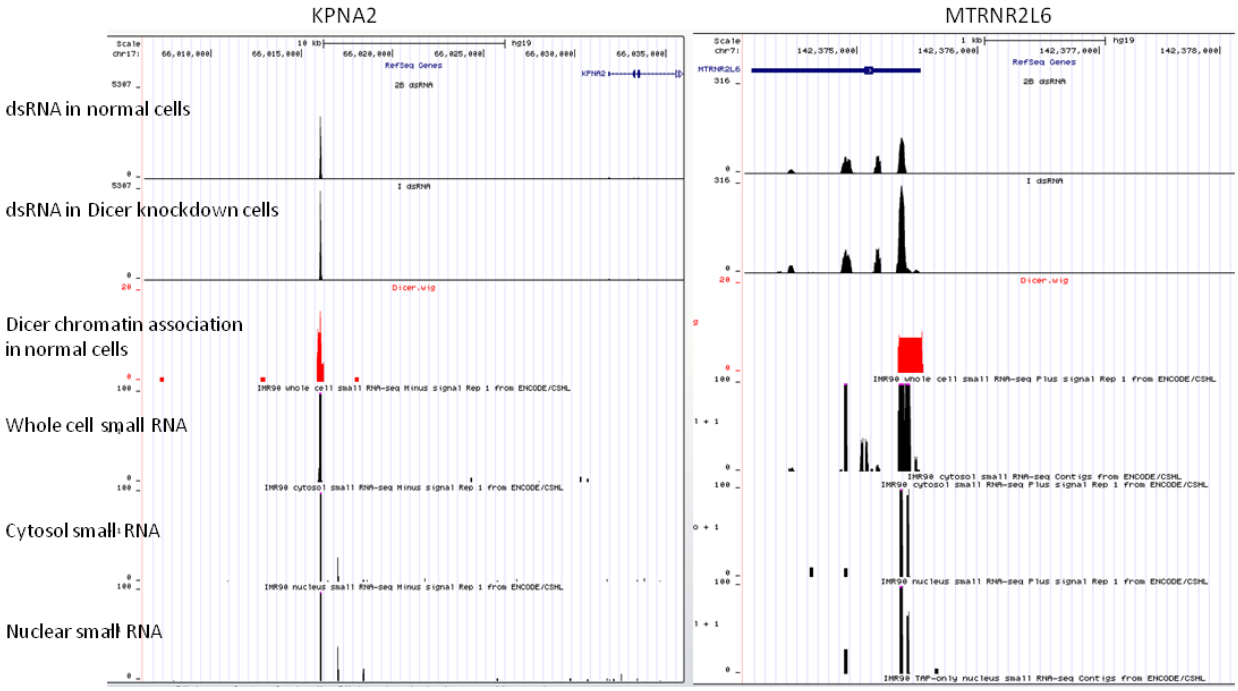

Figure 4.25: Data was uploaded to the UCSC genome browser and provides a visualisation of the data mapping to the genome. The intergenic region of KPNA2 (left) and the terminator region of MTRNR2L5(right) is depicted. In order from top to bottom: dsRNA in normal cells, dsRNA levels in knock-down cells, Dicer levels in normal cells. Note that levels are not normalised to sequencing depth, but dsRNA sequencing depth was equivalent in normal and Dicer knock-down cells. The last three panels, from top to bottom, display the small RNA isolated from whole cell, cytosolic and nuclear fraction in IMR90 cells.

4.4 Discussion

Dicer is an endoribonuclease type III enzyme, necessary for the processing of miRNAs and dsRNA into siRNAs, to promote gene silencing. Previously it has been believed, that Dicer merely processes hairpins and dsRNAs in the cytoplasm, and does not enter the nucleus. However, in fission yeast it migrates between cytoplasm and nucleus (Barraud et al., 2011), while in mammals no conserved NLS has been observed. We show that Dicer can be detected with a Dicer-specific antibody within the nucleus. More interestingly an exogenous GFP-Dicer is not detected in the nucleus in normal cells, but upon Dicer deletion the exogenous GFP-Dicer does indeed enter the nucleus. We suspect there is a chaperone element, which closely regulates Dicer levels within the nucleus. In the future the molecular basis of this chaperone shall be investigated. Moreover, Dicer localisation inside the nucleus, as well as other RNAi components (Gagnon et al., 2014) do point at the possibility of TGS within human cells.

dsRNA can arise within the cell from overlapping transcription units, natural antisense transcription and repetitive sequence elements. Interestingly Dicer did not show a significant overlap with LINEs, SINEs and LTRs. However, it should be noted that the number of Dicer peaks analysed is hardly enough to draw irrevocable statistical conclusions. Once a more exhaustive ChIP-Seq set of Dicer binding loci is available, this analysis should be repeated. Interestingly Dicer has a significant overlap with Satellite repeat sequences. This could point in the direction of a role for Dicer in heterochromatin assembly. Reduced Dicer levels were moreover observed in cancer cell lines (Passon et al., 2012) and aging animal cells (Anderson, 2012), which suggests that dsRNA turnover is defective in such cells.

In *S. pombe* Dicer is required for heterochromatin assembly and maintenance in the centromeres. PolII transcribes centromeric sequences, in cooperation with RNA-dependent-RNA polymerase it generates dsRNA. dsRNA is cleaved by Dicer and introduced back into the nucleus by the Argonaute complex. The heterochromatin

structure is set up by the H3K9 methylase Clr4 and Swi6. The process is described in more detail in Section 1.4. Mammalian heterochromatin also acts to repress gene expression in a large proportion of the genome (Castel and Martienssen, 2013). Two categories of mammalian heterochromatin exist: constitutive and facultative. The former is associated with Satellite sequences and defined by CG methylation and H3K9me3 histone marks, which recruit heterochromatic proteins such as HP1. Recently, constitutive heterochromatin has been associated with Transcription factor occupancy implying active transcription (Bulut-Karslioglu et al., 2012), although at low levels according to Martens et al., 2005. The same study describes dsRNA synthesis. This may explain why there is a significantly higher number of Dicer peaks overlapping Satellite regions. Facultative heterochromatin is of a more dynamic nature and affiliated with H3K9me2 histone marks. It has been demonstrated that this type of heterochromatin can be induced by exogenous siRNAs (Alló et al., 2009; Morris et al., 2004) and may also regulate transcription endogenously (Fagegaltier et al., 2009). Given our observation of H3K9me2 marks in the analysed Dicer loci, we expect Dicer, in cooperation with Ago, to be involved in mediating RNAi to loci of locally synthesised dsRNA, which sets up the heterochromatin structure.

We propose a model of Dicer function within the cell. We have strong evidence that Dicer knock down blocks the role of low-level dsRNA synthesis in TGS by causing an accumulation of cellular dsRNA. We predict the escape of overabundant dsRNA into the cytoplasm where it induces the interferon response and finally causes cellular apoptosis (Figure 4.26). Also mice cannot develop past the embryonic stage if no Dicer is present, however mouse embryonic stem cells lacking Dicer are viable (Murchison et al., 2005). The fact that embryonic stem cells do not have an interferon pathway supports our model. In combination with our studies we assume that Dicer is responsible for the correct balance between heterochromatin and euchromatin, by regulating dsRNA levels. If these increase past a sustainable level, the cells save themselves from cell pathologies

via cellular apoptosis of misregulated cells.

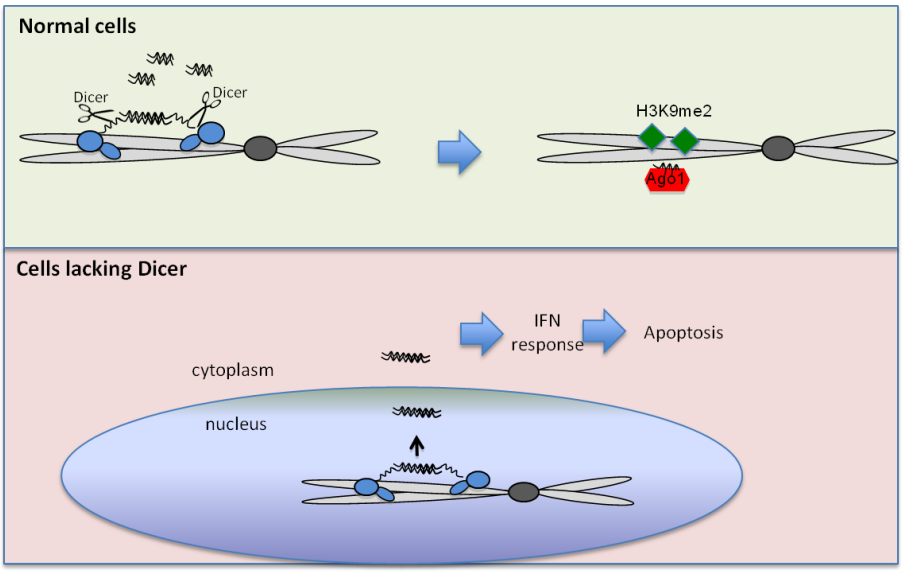

Figure 4.26: Model of Dicer action within the cell. In a healthy cell Dicer is recruited to regions of overlapping transcripts through association with PolII and dsRNA. This association cleaves dsRNA in siRNA, which with the help of Ago1 established H3K9me2 mark. In knock- down cells the accumulation of dsRNA results in the induction of the interferon pathway which ultimately leads to cell apoptosis.

Chapter 5

Conclusions

Over the years, increasingly sophisticated experimental and analytical tools have given scientists the opportunity to gain a deeper insight into gene regulation elements. These may range from the production of regulatory ncRNA, which cause a gradient of genetic repression, gene silencing and coding transcript variation. Transcript diversity and functionality may depend on the protein sequence (Nagalakshmi et al., 2008; Rhind et al., 2011), translational activity (Rojas-Duran and Gilbert, 2012) or deletion/retention of binding sites for RNA- binding proteins (Pelechano et al., 2013). Sequencing resources have not yet reached their limits. Their depth and precision grow with newly developing techniques. In the first few chapters I have taken advantage of these techniques and the knowledge they provided in order to analyse yeast 3' UTRs. However, these techniques have provided the opportunity for insight into other non-coding genomic regions in yeast.

5' UTRs It has been shown that 5' UTRs are significantly shorter than 3'UTRs (David et al., 2006; Nagalakshmi et al., 2008) and they were re-annotated and characterized using emerging technologies in the yeast genomes. Although RNA-Seq fails to determine the 5' end at nucleotide resolution and can only approximate it due to sharp signal transition (Nagalakshmi et al., 2008), TIF-Seq identified variations also in 5' UTRs. Moreover, previously considered short regulatory ORFs (uORFs or upstream ORFs) were

re-annotated as short coding regions (Pelechano et al., 2013). CAGE (cap analysis gene expression) is a popular method to map TSSs (Shiraki et al., 2003). However, it was not applied to yeast species. Alternatively, a newly developed method called 5'SAGE was used in *S. cerevisiae* (Zhang and Dietrich, 2005). In *S. pombe*, the TSS has been reported to lie 25 to 40 nt from the TATA element (Choi et al., 2002). In contrast, 5'SAGE maps the TSS for *S. cerevisiae* 50 to 125 nt away. The TSS consensus sequence was determined to be $(A_{rich})_5$NPy<u>A</u> (A/T)NN$(A_{rich})_6$, with the underlined letter representing the first transcribed nucleotide and Py representing a pyrimidine. Moreover, 5'SAGE identifies 24 genes with regulatory uORFs in their 5'UTRs. These findings were improved by a large-scale cDNA analysis (Miura et al., 2006), which revealed that the vast majority of genes have two or more TSS. These were classified as either having a single dominant TSS or multiple modestly used TSSs. The cDNA analysis also revealed at least one uORF in 2415 5' UTRs and introns in 32 5' UTRs. Recent small-scale experiments in *S. cerevisiae* (including 5' RACE) demonstrate that alternative TSS selection can have a significant impact on translational activity and possibly act as a switch between coding and regulatory RNA production (Rojas-Duran and Gilbert, 2012), as observed in *S. pombe* (Sehgal et al., 2008). This effect remains to be analysed on a genome-wide scale.

Intergenic ncRNA High-density tiling arrays were used to detect ncRNA in the entire budding yeast genome (Samanta et al., 2006). A large number of novel transcripts was detected in intergenic as well as in promoter regions, in both sense and antisense directions. It was suggested that the transcripts mapping to promoter regions have a regulatory role. Described below are CUTs (cryptic unstable transcripts) and SUTs (stable uncharacterised transcripts), which were mapped via RNA-Seq in combination with 3' long SAGE (serial analysis of gene expression, Neil et al., 2009; Xu et al., 2009). CUTs are short RNA molecules overlapping with the promoter regions, soon degraded after their synthesis. They are possibly a by-product of bidirectional transcription, initiated from nucleosome-free regions at promoters (Xu et al., 2009). The less abundant sense CUTs

are assumed to aid gene suppression. The proximal TSS is preferred under repressive conditions. If it belongs to the CUT, it can repress transcriptional initiation of the downstream gene. Alternatively, CUTs can be involved in transcriptional interference. Some CUTs share a TSS with the downstream gene and cause premature transcription termination, repressing the genes expression (Neil et al., 2009). Further investigation of CUTs via microarray data and Northern blots revealed that, although many CUTs are degraded in the nucleus, they can also be exported to the cytoplasm, where decapping and 5'-to-3' exonucleolytic digestion causes its degradation. Moreover, some of them enter translation, although they do not encode a functional protein (Thompson and Parker, 2007). SUTs usually arise from divergent transcription. The difference in stability between CUTs and SUTs (Xu et al., 2009) and the precise role of bidirectional transcription remain to be understood (Marguerat and Bähler, 2010).

Introns High-density tiling arrays can only detect sequences present in the genome, therefore RNA-Seq is better suited for annotation of splice sites, either by mapping the reads to predicted exon junctions or by de novo annotation via splitting the RNA-Seq reads into two parts (Marguerat and Bähler, 2010). RNA-Seq revealed high amounts of alternative splicing; however, not in the form of exon skipping or alternative exon incorporation (Rhind et al., 2011). Transcripts can be spliced on the basis of growth conditions in *S. pombe* (Wilhelm et al., 2008); interestingly, the unspliced variant may be the protein-coding isoform (Rhind et al., 2011). The splicing efficiency (i.e. ratio of spliced/unspliced transcripts) varies between genes and growth conditions. It has been determined that spliced or unspliced products affect predicted protein sequences (Nagalakshmi et al., 2008), and there is alternative splicing between vegetative growth and heat shock response in *S. cerevisiae* (Yassour et al., 2009). The mapping of a significant fraction of DRS reads to introns (or exons) indicates a dynamic interplay between polyadenylation and splicing (Ozsolak et al., 2010), hence diversifying the organisms transcriptome and probably proteome.

Antisense transcription DRS identified widespread antisense transcription in >60% of budding yeast genes (Ozsolak et al., 2010). Additionally, coordination between sense and antisense polyadenylated transcript levels plays a role in gene expression regulation. Genes that are expressed at low levels in budding yeast show a positive correlation between sense and antisense transcripts, whereas highly expressed genes demonstrate a negative correlation (Ozsolak et al., 2010). In contrast, in fission yeast, tiling array analysis suggests that most antisense transcripts are not polyadenylated (Dutrow et al., 2008). This method also detected a likely propensity of highly expressed genes and histone-depleted regions to show elevated antisense transcription (Dutrow et al., 2008). The antisense transcripts generally occur more frequently in the 3' UTR than in the 5' UTRs, which implies that they may also arise from overlapping 3' UTRs, as shown by RNA-Seq (Nagalakshmi et al., 2008). However, the identification of antisense transcription using any method that involves reverse transcription of RNA to cDNA requires caution, as this can cause secondary mispriming. Generally, it seems that antisense transcription is lower than sense transcription (David et al., 2006; Rhind et al., 2011). Also, some antisense transcripts are part of lncRNAs (long ncRNAs). These long antisense transcripts, as well as some long intergenic transcripts are present at less than one copy number per cell, which could reflect their tight repression at transcriptional, post-transcriptional or chromatin levels (Marguerat et al., 2012). An important difference between *S. cerevisiae* and *S. pombe* is the lack of an RNAi pathway in budding yeast. The RNAi pathway in fission yeast and in higher eukaryotes is induced by dsRNA formation. dsRNA is processed into siRNA, which target nascent mRNA via complementarity or induce heterochromatin formation and consequent gene silencing at the transcriptional level. Multiple protein complexes, including Dicer (Hannon, 2002), are involved in the RNAi pathway. Small RNA libraries are prepared by size-selecting molecules, followed by RNA-Seq, which determines sequence content (Yamanaka et al., 2013). It is apparent that RNAi in *S. pombe* acts jointly with the exosome to repress developmentally regulated

genes and retrotransposons. Furthermore, similar analysis in Dicer-knockout cells (Halic and Moazed, 2010) revealed Dicer-independent priRNAs (primary small RNAs). It appears that priRNAs originate from degradation of abundant transcripts, and they target antisense transcripts arising from bidirectional transcription of DNA repeats.

All of the above non-coding genomic regions may play a role in gene silencing and the RNAi pathway. It has been shown, that siRNAs targeted to promoter regions in human cells can give rise to TGS effects, provided they have access to the nucleus (Morris et al., 2004). Introns can give rise to Drosha independent miRNA, so called miRtrons, which will act on a PTGS level in mammalian cells (Ladewig et al., 2012 and references therein). Intergenic ncRNA, similarly to antisense transcripts, can give rise to overlapping transcripts (and hence dsRNA inducing RNAi), if it happens to fall in the 3'UTR or 5'UTR of a gene on the opposite strand. Moreover, many of these might be non-annotated miRNA. Genome-wide techniques are therefore crucial to understand the nature of non-protein-coding regions, which play an important role in gene expression regulation.

In eukaryotes siRNA and miRNA arising from dsRNA and hairpin structures are known to silence genes, often by complementarily binding to 3'UTRs of the mature mRNA. In particular, in animals alternative 3'UTRs may provide a miRNA binding domain in the longer transcript and therefore subject it to degradation. The shorter transcript may give rise to a modified or truncated protein, which may impair or enhance its function. I have demonstrated that alternative polyadenylation is very pervasive in *S. pombe*. This has been confirmed by other studies (Mata, 2013). It has to be noted that what is considered as true APA and what is considered as heterogeneity is very much a subjective matter. Mata (2013) has considered less than 25 nt as an individual separation between CS to still be considered as one CS. I have chosen a much smaller individual separation (6 nt), which may group CS into a cluster spanning more than 30 nt. Given the

NUE lies about 25 nt before the CS, two CS separated by 30 nt will probably require two independent NUE. Unfortunately for an exhaustive study of CS on a genome-wide scale RNA-Seq may be quite limiting. The problem is the fraction of reads, which contains a poly(A) tail, compared to all sequenced reads is quite small. Moreover, reads, which provide only the poly(A)-tail or a read, which only has a small non-A tract and therefore maps to the genome far too often (many such reads were also involved in our data), are useless for CS mapping. These problems may be remedied by increasing sequencing depth and they depend on the protocol (e.g. poly(A) selection) and the primers which are used (these may avoid the poly(A) tail overall). Even though RNA-Seq yielded a dataset of large enough size for statistical analysis of PAS an inter-conditional comparison as presented in the supplementary data in Appendix C may be unreliable. 3PC was used to obtain higher coverage set of CS in *S. pombe* (Mata, 2013). Ideally, the same technique, or a close variant, would be employed to map CS in other conditions. This would provide a truly meaningful and significant set for inter-conditional APA analysis, with more reliable statistical testing for significance.

I have identified many intra-genic CS, which were attributed to the gene they fall in. However, possibly they belong to the previous gene if it was in tandem. In addition, I did not consider 3'UTRs larger than 1000 nt, but possibly some are. A technique like TIF-Seq, so far only used in *S. cerevisiae*, applied to *S. pombe* would shed light on which gene the CS should be attributed to and help us to re-evaluate our data.

Transcriptional cleavage and termination in *S. pombe* may depend on additional *trans* factors such as Cohesin. I have analysed previous ChIP-chip data of Cohesin subunits and Kollerin subunits. I adjusted the data to the current genome annotation and avoided the previous inconsistencies. Our own peak-calling algorithm was more sensitive to peak-calling than the previous moving average approach. Indeed I confirmed, that the Cohesin loading complex localises in highly transcribed genes (Schmidt et al., 2009) and that the Cohesin associating with it is significantly higher transcribed than Kollerin

independent Cohesin enrichment. Moreover, the fact that co-localisation sites are more cohesive than just Cohesin sites, presents the possibility that the replication fork may start firing at these sites. In accordance, I observe a mildly significant proximity of ORIs to Mis4/Rad21 co-localisation sites. In the future I wish to confirm this data with more recent ORI predictions (Xu et al., 2012). Currently, we are investigating if the gene expression differences imply that Cohesin or Kollerin interact with the transcription machinery.

Cohesin is recruited to the transcriptionally silent heterochromatic regions by the protein Swi6. Between convergent genes it may act to stop readthrough transcription and hence inhibit gene silencing (Gullerova and Proudfoot, 2008), thereby unpacking the heterochromatin. The RNAi would have been induced by siRNA originating from overlapping convergent 3'UTRs. RNAi may also occur from natural antisense transcription in fission yeast. What is therefore the function of the detected alternative transcripts? I hypothesised that possibly miRNA have been overlooked in fission yeast until now. Again, a separation between two CS of more than 6 nt will already be enough for a miRNA seed region. The fact that miRNA remain undetected in *S. pombe* may be due to the absence of Drosha in yeast, which may also be the reason why the miRNA pathways in animals and plants are believed evolutionarily independent. Moreover, in plants miRNA do not just target 3'UTRs, but possibly ORFs and they experience much stronger complementarity than animal miRNA. Given these differences, possibly mature miRNA sequences between animals, plants and yeast may experience considerable variations. The evolutionary conservation based approaches of miRNA discovery, where one organism is scanned for possible targets and hairpin precursors of mature miRNAs derived from other organisms, may fail in fungi as appropriate mature miRNA candidates are not available. Recent development of high throughput and deep sequencing techniques allowed sequencing of sRNA in *S. pombe* (Halic and Moazed, 2010; Yamanaka et al., 2013) and provided us with a yeast specific set of possible mature miRNAs and I was able to circumvent the evolutionary conservation step. However, a certain amount of

evolutionary conservation needs to be taken into consideration for computational miRNA discovery, to allow conditioning on the form of hairpin structures and to stay true to the definition of a miRNA. In the future BLAST will be used to align the predicted miRNA genes (or just the pre-miRNA sequences) and targets to animal genes in the expectation of being a known miRNA. Moreover, real hairpin structures and targets are likely to be conserved across Schizosaccharamyces species, which can also be analysed via BLAST.

Here, I have presented computational results for possible miRNA genes and targets in *S. pombe*. They are currently being investigated in more depth. The 3'RACE of the target gene, which has been reproduced, provides some evidence that indeed, there may be a chance for miRNA mediated regulation of gene expression in fungi. Until recently miRNA were believed to be a feature of multicellular organisms, until it was disproved in a unicellular alga. Maybe we can disprove it on the unicellular fungus. Moreover, existence of miRNA in fungi will overthrow the belief of independent evolution of miRNA in humans and plants and produce a novel idea of how the differences in miRNA regulation may have come into existence. This is already corroborated by the fact that some miRNA are homologous between humans and plants. The question remains whether miRNA in *S. pombe* functions in a TGS level or PTGS level as in animals. I anticipate its function to be similar to the siRNA pathway as the same RNAi components are expected to be involved, hence affecting chromatin structure. A way to elucidate this would be to perform Western Blots and identify if protein structure or levels change between WT and mutant (normal and heat shocked) cells. One could also amplify the pre-miRNA and mature miRNA sequence of a predicted miRNA in cellular and nuclear fractionations of the cells. If a 70 nt and a 22 nt product are detected solely in the nucleus, it may indicate a TGS pathway. If a 22 nt product is also detected in the cytoplasm, the miRNA pathway may be similar to plants and be either PTGS or TGS by Dicer re-introducing it into the nucleus. Should 70 nt and 22 nt products be detected in the cytoplasm, but 22 nt product be absent in the nucleus, this may truly indicate a PTGS regulation pathway.

Fractionation of cells may indeed help us understand the miRNA pathway in *S. pombe*, should the miRNAs be confirmed.

Confirmation of miRNA needs to happen experimentally, as pure computational predictions are certainly not sufficient. I have considered miRNA genes and targets simultaneously. Moreover, to confirm the candidate being a real miRNA, we need to perform Northern Blots to detect pre-miRNA and mature miRNA bands.

I have picked more genes for testing. *SPBC530.02* has a predicted target in *SPAC2F3.11*, which has a longer UTR in quiescent 24 h and 7 d than in cycling cells, according to our APA data. *SPCC1682.11c* has two potential targets. One of them has a longer UTR in quiescent 24 h and 7 d, the other one only in cycling. This time, I picked them for the Ago1 pull-down experiment (Halic and Moazed, 2010). Another candidate from the general sRNA data (Yamanaka et al., 2013) was also chosen: *SPCC825.02* with two targets *SPBC29A10.01* and *SPCC663.02*. According to the usage profile of my developed *Pomb(A)* database, the former target shows several longer transcripts in meiotic and quiescent (24 h) and the latter one longer transcript in quiescent (24 h and 7 d) cells. These candidates are currently undergoing preparation for the same type of experiments as performed before. The described candidates show a mixed correlation in gene expression between potential miRNA gene and target (ranging in negative correlation, no correlation or positive correlation, Appendix Figures D.1-D.3). The choice was made to possibly provide more insight on discriminative criteria for miRNA prediction in *S. pombe*. One should note the strikingly high negative correlation in gene expression between the predicted miRNA *SPCC1682.11c* and its predicted target *SPBC56F2.12* (Appendix Figure D.2A). In contrast *SPCC1682.11c* shows a fair positive correlation in gene expression with its other potential target. Experimental analysis of these two targets will provide insight into whether miRNA may also upregulate genes in *S. pombe*, in particular if the same miRNA can act as and upregulator and downregulator at the same time on different targets. All target binding profiles, predicted hairpin structures

and gene expression correlation profiles between miRNA and target of chosen genes for experimental testing can be viewed in the Appendix Figures D.1-D.3.

Genes for experimental testing were selected based on total sRNA sequencing data and Ago1 pull-down sRNA data. While the total sRNA sequencing dataset (Yamanaka et al., 2013) provides more candidates and is probably more sensitive, the Ago1 pull-down would provide more specific candidates if miRNA processing in *S. pombe* requires Ago1. Requirement of Ago1 is still debatable based on our experiment presented in Figure 3.8. This raises the question of which protein is responsible for the slicing action and targeting. It has been shown in mouse cells, that siRNAs may target mRNAs independently of any Argonaute protein (Vickers and Crooke, 2012). However, which protein is then responsible for slicing has not been identified. Possibly a similar mechanism is present in *S. pombe* and will be further investigated. Choosing possible miRNA from Ago1-dependent and independent data will provide the first insight of whether two possible miRNA targeting mechanisms exist in yeast. I anticipate exciting and enlightening results.

With the absence of Drosha in yeast cells the question of how the hairpins are processed arises naturally. It has been shown that Dicer can indeed process hairpins (Simmer et al., 2010). Moreover, DCL1, the Dicer homologue in plants, takes over the Drosha function within the nucleus and processes the pre-miRNA into the mature form, which are then exported and loaded onto Ago. In addition, miRNA may arise from miRtrons. This renders the miRNA production independent of Drosha and Exportin-5. Interestingly in mouse cells, the mirtron pre-mmu-mir-1982 RNA with an 11 nt 5' overhang is not compatible with the export by Exportin-5 to the cytoplasm (Berezikov et al., 2007; Okada et al., 2009). Nevertheless, the mature miR-1982* is detectable in deep sequencing data, a product which *in vitro* is produced by Dicer (Ando et al., 2011a). Possibly a similar miRNA production pathway exists in *S. pombe*.

This also provides evidence that Dicer might not be a purely cytoplasmic protein

in animals as previously thought. In fact, TGS has been believed non-existent in mammals other than in germ lines (Reuter et al., 2011). But recent findings indicate that RNAi components are present and active in human cell nuclei (Gagnon et al., 2014). Moreover dsRNA and siRNA introduced into the nucleus of mammalian cells shows that mammalian cells are actually capable of heterochromatic gene silencing and hence TGS (Gullerova and Proudfoot, 2012; Morris et al., 2004).

In *S. pombe*, where TGS is a well-known phenomenon, until recently cross-linking Dicer to heterochromatin has failed (Volpe et al., 2002) and only recently succeeded (Woolcock et al., 2011). This provides evidence that Dicer may indeed mediate heterochromatin formation *in cis*. Possibly the same reasons, such as the absence of a reliable antibody, has prevented cross-linking Dicer to the chromatin in mammals. We have succeeded and have presented the effect, that loss of Dicer causes accumulation of dsRNA in the nucleus. However, Dicer entrance into the nucleus seems to be chaperoned and tightly regulated. Recent studies have shown that Dicer associates with nucleoporins detectable in the cytoplasm, but not the nucleus (Ando et al., 2011b). This points at a potential Dicer transport across the membrane and may account for its previously believed absence in the nucleus. siRNA introduced into the nucleus causes DNA-methylation at targeted regions, which are the markers of silent chromatin (Morris et al., 2004). Given the possibility of siRNA entering the nucleus, TGS is indeed a consequence in human cells. Possibly, Dicer transported into the nucleus with the aid of nucleoporins may be the actor upon endogenous dsRNA from antisense or convergent overlapping transcription, which are incompatible with Exportin-5. The resulting siRNA cause transcriptional gene silencing via heterochromatin formation.

As presented in Chapter 4 dsRNA in the cytoplasm will induce the interferon response pathway. Moreover, accumulation of dsRNA has been linked to the degradation of the human retina and consequent blindness (Kaneko et al., 2011). The regulation of dsRNA levels is therefore of utmost importance and may be best be achieved within the nucleus.

The presence of Dicer in the nucleus gives rise to a new method of gene silencing via introducing dsRNA into the nucleus through a plasmid, whose derived siRNA will impose TGS effects (Gullerova and Proudfoot, 2012).

We have demonstrated the presence of Dicer in the nucleus. In our ChIP-Seq data, one of our called peaks was detected on the mitochondrial chromatin. This raises an interesting new question, of whether Dicer has a possible role in the mitochondrial DNA and may be subject of investigation in the future.

The understanding of Dicer and miRNA is of great medical interest. Tumor cells have a significantly different miRNA profile to normal cells (Calin and Croce, 2006). Some miRNA may target tumor suppressors and are hence anti-apoptotic and pro-survival of the cells and are upregulated in tumor cells. These miRNA function as oncogenes. Other miRNA may function as tumor suppressors by targeting oncogenes and are downregulated in tumor cells. Clearly, miRNA expression abnormalities may be due to abnormal Dicer function. Understanding miRNAs and Dicer function may therefore provide important clinical tools. For example miRNA profiling is much easier than total mRNA profiling. It can function as a diagnostic tool (how far has the cancer developed and where is the primary site of metastatic cancer), a prognostic tool (how quickly is the cancer going to develop and how serious should be the applied treatment) and possibly as a predictive tool (it is likely that miRNA expression differences are inherited in the germ line, Calin and Croce, 2006). Finding new model organisms for miRNA research is therefore highly desirable.

This work demonstrates that evolutionary conservation between yeast and higher eukaryotes may be far stronger than previously thought. TGS and PTGS are not exclusive to yeast and mammals respectively but we provide evidence that they co-exist in both. It is possible that fission yeast is a better model organism for mammalian cells than currently

accepted. TGS in mammals is by now an irrevocable fact, confirmed by many independent studies. Our promising computational and experimental results for PTGS in yeast still require confirmation but are a step towards yeast similarity to higher eukaryotes.

Bibliography

JS Aaronson, B Eckman, RA Blevins, JA Borkowski, J Myerson, S Imran, and KO Elliston. Toward the development of a gene index to the human genome: an assessment of the nature of high-throughput EST sequence data. *Genome Research*, 6(9):829–845, 1996.

Y Akao, Y Nakagawa, and T Naoe. microRNA-143 and-145 in colon cancer. *DNA and cell biology*, 26(5):311–320, 2007.

E Allen, Z Xie, AM Gustafson, G Sung, JW Spatafora, and JC Carrington. Evolution of microRNA genes by inverted duplication of target gene sequences in *Arabidopsis thaliana*. *Nature genetics*, 36(12):1282–1290, 2004.

M Alló, V Buggiano, JP Fededa, E Petrillo, I Schor, M de la Mata, E Agirre, M Plass, E Eyras, SA Elela, et al. Control of alternative splicing through siRNA-mediated transcriptional gene silencing. *Nature structural & molecular biology*, 16(7):717–724, 2009.

M Ameyar-Zazoua, C Rachez, M Souidi, P Robin, L Fritsch, R Young, N Morozova, R Fenouil, N Descostes, JC Andrau, et al. Argonaute proteins couple chromatin silencing to alternative splicing. *Nature structural & molecular biology*, 19(10): 998–1004, 2012.

RM Anderson. A role for Dicer in aging and stress survival. *Cell metabolism*, 16(3): 285–286, 2012.

Y Ando, Y Maida, A Morinaga, AM Burroughs, R Kimura, J Chiba, H Suzuki, K Masutomi, and Y Hayashizaki. Two-step cleavage of hairpin RNA with 5'overhangs by human Dicer. *BMC molecular biology*, 12(1):6, 2011a.

Y Ando, Y Tomaru, A Morinaga, AM Burroughs, H Kawaji, A Kubosaki, R Kimura, M Tagata, Y Ino, H Hirano, et al. Nuclear pore complex protein mediated nuclear localization of Dicer protein in human cells. *PloS one*, 6(8):e23385, 2011b.

A Aranda and NJ Proudfoot. Definition of transcriptional pause elements in fission yeast. *Molecular and Cellular Biology*, 19(2):1251–1261, 1999.

M Arteaga-Vázquez, J Caballero-Pérez, and J Vielle-Calzada. A family of microRNAs present in plants and animals. *The Plant Cell Online*, 18(12):3355–3369, 2006.

A Azvolinsky, PG Giresi, JD Lieb, and VA Zakian. Highly transcribed RNA Polymerase II genes are impediments to replication fork progression in *Saccharomyces cerevisiae*. *Molecular cell*, 34(6):722–734, 2009.

D Baek, J Villén, C Shin, FD Camargo, SP Gygi, and DP Bartel. The impact of micro-RNAs on protein output. *Nature*, 455(7209):64–71, 2008.

J Bähler and V Wood. *The Genome and Beyond*, book section 2, pages 13–26. Springer, Heidelberg, Germany, 2003.

J Bähler, J Wu, MS Longtine, NG Shah, A McKenzie I, AB. Steever, A Wach, P Philippsen, and JR Pringle. Heterologous modules for efficient and versatile PCR-based gene targeting in *Schizosaccharomyces pombe*. *Yeast*, 14(10):943–951, 1998.

L Bai, TJ Santangelo, and MD Wang. Single-molecule analysis of RNA Polymerase transcription. *Annual Review of Biophysics and Biomolecular Structure*, 35(1): 343–360, 2006.

P Barraud, S Emmerth, Y Shimada, H Hotz, FH Allain, and M Bühler. An extended dsRBD with a novel zinc-binding motif mediates nuclear retention of fission yeast Dicer. *The EMBO journal*, 30(20):4223–4235, 2011.

DP Bartel. microRNAs: Target recognition and regulatory functions. *Cell*, 136(2):215 – 233, 2009.

E Beaudoing, S Freier, JR Wyatt, J Claverie, and D Gautheret. Patterns of variant polyadenylation signal usage in human genes. *Genome Research*, 10(7):1001–1010, 2000.

CL Bennett, ME Brunkow, F Ramsdell, KC O'Briant, Q Zhu, RL Fuleihan, AO Shigeoka, HD Ochs, and PF Chance. A rare polyadenylation signal mutation of the FOXP3 gene (AAUAAA–>AAUGAA) leads to the IPEX syndrome. *Immunogenetics*, 53(6):435–9, 2001.

I Bentwich, A Avniel, Y Karov, R Aharonov, S Gilad, O Barad, A Barzilai, P Einat, U Einav, E Meiri, et al. Identification of hundreds of conserved and nonconserved human microRNAs. *Nature genetics*, 37(7):766–770, 2005.

E Berezikov, W Chung, J Willis, E Cuppen, and EC Lai. Mammalian mirtron genes. *Molecular cell*, 28(2):328–336, 2007.

MG Berg, LN Singh, I Younis, Q Liu, AM Pinto, D Kaida, Z Zhang, S Cho, S Sherrill-Mix, L Wan, and G Dreyfuss. U1 snRNP determines mRNA length and regulates isoform expression. *Cell*, 150(1):53–64, 2012.

E Bernstein, SY Kim, MA Carmell, EP Murchison, H Alcorn, MZ Li, AA Mills, SJ Elledge, KV Anderson, and GJ Hannon. Dicer is essential for mouse development. *Nature genetics*, 35(3):215–217, 2003.

Pl Bertone and M Snyder. Advances in functional protein microarray technology. *FEBS Journal*, 272(21):5400–5411, 2005.

NJ Beveridge, E Gardiner, AP Carroll, PA Tooney, and MJ Cairns. Schizophrenia is associated with an increase in cortical microRNA biogenesis. *Molecular psychiatry*, 15(12):1176–1189, 2009.

E Billy, V Brondani, H Zhang, U Müller, and W Filipowicz. Specific interference with gene expression induced by long, double-stranded RNA in mouse embryonal teratocarcinoma cell lines. *Proceedings of the National Academy of Sciences*, 98(25): 14428–14433, 2001.

RP Birkenbihl and S Subramani. Cloning and characterization of Rad21 an essential gene of *Schizosaccharomyces pombe* involved in DNA double-strand-break repair. *Nucleic acids research*, 20(24):6605–6611, 1992.

CE Birse, L Minvielle-Sebastia, BA Lee, W Keller, and NJ Proudfoot. Coupling termination of transcription to messenger RNA maturation in yeast. *Science*, 280 (5361):298–301, 1998.

A Bulut-Karslioglu, V Perrera, M Scaranaro, IA de la Rosa-Velazquez, S van de Nobelen, N Shukeir, J Popow, B Gerle, S Opravil, M Pagani, et al. A transcription factor–based mechanism for mouse heterochromatin formation. *Nature structural & molecular biology*, 19(10):1023–1030, 2012.

GA Calin and CM Croce. microRNA signatures in human cancers. *Nature Reviews Cancer*, 6(11):857–866, 2006.

SE Castel and RA Martienssen. RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. *Nature Reviews Genetics*, 14(2):100–112, 2013.

S Chang, Z Zhang, and Y Liu. RNA interference pathways in fungi: mechanisms and functions. *Annual review of microbiology*, 66:305–323, 2012.

Y Chang, H Juan, T Lee, Y Chang, Y Yeh, W Li, and AC Shih. Prediction of

human miRNAs using tissue-selective motifs in 3' UTRs. *Proceedings of the National Academy of Sciences*, 105(44):17061–17066, 2008.

H Chen, B Futcher, and J Leatherwood. The fission yeast RNA binding protein Mmi1 regulates meiotic genes by controlling intron specific splicing and polyadenylation coupled RNA turnover. *PloS one*, 6(10):e26804, 2011.

SW Chi, JB Zang, A Mele, and RB Darnell. Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps. *Nature*, 460(7254):479–486, 2009.

WS Choi, M Yan, D Nusinow, and JD Gralla. *In Vitro* transcription and start site selection in *Schizosaccharomyces pombe*. *Journal of molecular biology*, 319(5):1005–1013, 2002.

J Cocquet, A Chong, G Zhang, and RA Veitia. Reverse transcriptase template switching and false alternative transcripts. *Genomics*, 88(1):127–131, 2006.

S Connelly and JL Manley. A functional mRNA polyadenylation signal is required for transcriptional termination by RNA Polymerase II. *Genes Dev*, 2:440–452, 1988.

MB Cox, MJ Cairns, KS Gandhi, AP Carroll, S Moscovis, GJ Stewart, S Broadley, RJ Scott, DR Booth, J Lechner-Scott, et al. microRNAs miR-17 and miR-20a inhibit T cell activation genes and are under-expressed in MS whole blood. *PLoS One*, 5(8): e12132, 2010.

N Cremona, K Potter, and JA Wise. A meiotic gene regulatory cascade driven by alternative fates for newly synthesized transcripts. *Molecular biology of the cell*, 22 (1):66–77, 2011.

Xr Darzacq, Y Shav-Tal, V de Turris, Y Brody, SM Shenoy, RD Phair, and RH Singer. *In vivo* dynamics of RNA Polymerase II transcription. *Nature structural & molecular biology*, 14(9):796–806, 2007.

L David, W Huber, M Granovskaia, J Toedling, CJ Palm, L Bofkin, T Jones, RW Davis, and LM Steinmetz. A high-resolution map of transcription in the yeast genome. *Proceedings of the National Academy of Sciences*, 103(14):5320–5325, 2006.

L David, S Clauder-Münster, and LM Steinmetz. *Genome-Wide Transcriptome Analysis in Yeast Using High-Density Tiling Arrays*, volume 759 of *Methods in Molecular Biology*, book section 7, pages 107–123. Humana Press, 2011.

RM Denome and CN Cole. Patterns of polyadenylation site selection in gene constructs containing multiple polyadenylation signals. *Molecular and Cellular Biology*, 8(11): 4829–4839, 1988.

DC di Giammartino, D Campigli, K Nishida, and JL Manley. Mechanisms and consequences of alternative polyadenylation. *Molecular cell*, 43(6):853–866, 2011.

JG Doench and PA Sharp. Specificity of microRNA target selection in translational repression. *Genes & development*, 18(5):504–511, 2004.

Da Dorsett and M Merkenschlager. Cohesin at active genes: a unifying theme for cohesin and gene expression from model organisms to humans. *Current opinion in cell biology*, 25(3):327–333, 2013.

M Doyle, L Badertscher, L Jaskiewicz, S Güttinger, S Jurado, T Hugenschmidt, U Kutay, and W Filipowicz. The double-stranded RNA binding domain of human Dicer functions as a nuclear localization signal. *RNA*, 19(9):1238–1252, 2013.

M Dreyfus and P Régnier. The poly(A) tail of mRNAs: bodyguard in eukaryotes, sca-venger in bacteria. *Cell*, 111(5):611–613, 2002.

IA Drinnenberg, DE Weinberg, KT Xie, JP Mower, KH Wolfe, GR Fink, and DP Bartel. RNAi in budding yeast. *Science*, 326(5952):544–550, 2009.

N Dutrow, DA Nix, D Holt, B Milash, B Dalley, E Westbroek, TJ Parnell, and BR Cairns. Dynamic transcriptome of *Schizosaccharomyces pombe* shown by RNA-DNA hybrid mapping. *Nat Genet*, 40(8):977–986, 2008.

MJ Dye and NJ Proudfoot. Terminal exon definition occurs cotranscriptionally and promotes termination of RNA Polymerase II. *Molecular Cell*, 3(3):371 – 378, 1999.

S Emmerth, H Schober, D Gaidatzis, T Roloff, K Jacobeit, and M Bühler. Nuclear retention of fission yeast Dicer is a prerequisite for RNAi-mediated heterochromatin assembly. *Developmental cell*, 18(1):102–113, 2010.

A Enright, B John, U Gaul, T Tuschl, C Sander, and D Marks. microRNA targets in *Drosophila*. *Genome Biology*, 5(1):R1, 2003.

D Fagegaltier, A Bougé, B Berry, E Poisot, O Sismeiro, J Coppée, L Théodore, O Voinnet, and C Antoniewski. The endogenous siRNA pathway is involved in heterochromatin formation in *Drosophila*. *Proceedings of the National Academy of Sciences*, 106(50): 21258–21263, 2009.

MA Faghihi and C Wahlestedt. Regulatory roles of natural antisense transcripts. *Nature*, 10(9):637–643, 2009.

K Fejes-Toth, V Sotirova, R Sachidanandam, G Assaf, GJ Hannon, P Kapranov, S Foissac, AT Willingham, R Duttagupta, E Dumais, et al. Post-transcriptional processing generates a diversity of 5-modified long and short RNAs. *Nature*, 457 (7232):1028–1032, 2009.

Wenyi Feng, David Collingwood, Max E Boeck, Lindsay A Fox, Gina M Alvino, Walton L Fangman, Mosur K Raghuraman, and Bonita J Brewer. Genomic mapping of single-stranded DNA in hydroxyurea-challenged yeasts identifies origins of replication. *Nature cell biology*, 8(2):148–155, 2006.

SL Forsburg and P Nurse. Cell cycle regulation in the yeasts saccharomyces cerevisiae and schizosaccharomyces pombe. *Annual review of cell biology*, 7(1):227–256, 1991.

MR Friedländer, W Chen, C Adamidi, J Maaskola, R Einspanier, S Knespel, and N Rajewsky. Discovering microRNAs from deep sequencing data using miRDeep. *Nature biotechnology*, 26(4):407–415, 2008.

KT Gagnon, L Li, Y Chu, BA Janowski, and DR Corey. RNAi factors are present and active in human cell nuclei. *Cell reports*, 6(1):211–221, 2014.

MA Garcia, EF Meurs, and M Esteban. The dsRNA protein kinase PKR: virus and cell control. *Biochimie*, 89(6):799–811, 2007.

M Gartenberg. Heterochromatin and the Cohesion of sister chromatids. *Chromosome research*, 17(2):229–238, 2009.

M Gause, Z Misulovin, A Bilyeu, and D Dorsett. Dosage-sensitive regulation of cohesin chromosome binding and dynamics by Nipped-B, Pds5, and Wapl. *Molecular and cellular biology*, 30(20):4940–4951, 2010.

NH Gehring, U Frede, G Neu-Yilik, P Hundsdoerfer, B Vetter, MW Hentze, and AE Kulozik. Increased efficiency of mRNA 3' end formation: a new genetic mechanism contributing to hereditary thrombophilia. *Nature Genetics*, 28(4):389–392, 2001.

D Gerlich, B Koch, F Dupeux, J Peters, and J Ellenberg. Live-cell imaging reveals a stable cohesin-chromatin interaction after but not before DNA replication. *Current biology*, 16(15):1571–1578, 2006.

O Gick, A Kramer, W Keller, and ML Birnstiel. Generation of histone mRNA 3' ends by endonucleolytic cleavage of the pre-mRNA in a snRNP-dependent in vitro reactionl. *EMBO J*, 5(6):1319–1326, 1986.

A Gil and NJ Proudfoot. A sequence downstream of AAUAAA is required for rabbit $\beta - globin$ mRNA 3'-end formation. *Nature*, 312(5993):473–4, 1984.

JA Goodrich and R Tjian. Transcription factors IIE and IIH and ATP hydrolysis direct promoter clearance by RNA polymerase II. *Cell*, 77(1):145–156, 1994.

JH Graber, CR Cantor, SC Mohr, and TF Smith. Genomic detection of new yeast pre-mrna 3'-end-processing signals. *Nucleic Acids Research*, 27(3):888–894, 1999.

JH Graber, GD McAllister, and TF Smith. Probabilistic prediction of *Saccharomyces cerevisiae* mRNA 3'-processing sites. *Nucleic Acids Research*, 30(8):1851–1858, 2002.

N Gromak, S West, and NJ Proudfoot. Pause sites promote transcriptional termination of mammalian RNA Polymerase II. *Molecular and Cellular Biology*, 26(10):3986–3996, 2006.

JA Guerra-Assunção and AJ Enright. MapMi: automated mapping of microRNA loci. *BMC bioinformatics*, 11(1):133, 2010.

M Gullerova and NJ Proudfoot. Cohesin complex promotes transcriptional termination between convergent genes in *S. pombe*. *Cell*, 132(6):983–995, 2008.

M Gullerova and NJ Proudfoot. Convergent transcription induces transcriptional gene silencing in fission yeast and mammalian cells. *Nature structural & molecular biology*, 19(11):1193–1201, 2012.

M Gullerova, D Moazed, and NJ Proudfoot. Autoregulation of convergent RNAi genes in fission yeast. *Genes & development*, 25(6):556–568, 2011.

EJ Gumbel. *Statistics of extremes*. Courier Dover Publications, 2012.

M Hackenberg, M Sturm, D Langenberger, JM Falcon-Perez, and AM Aransay. miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic acids research*, 37(suppl 2):W68–W76, 2009.

M Hafner, M Landthaler, L Burger, M Khorshid, J Hausser, P Berninger, A Rothballer, M Ascano Jr, A Jungkamp, M Munschauer, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1):129–141, 2010.

M Halic and D Moazed. Dicer-independent primal RNAs trigger RNAi and heterochromatin formation. *Cell*, 140(4):504–516, 2010.

GJ Hannon. RNA interference. *Nature*, 418(6894):244–251, 2002.

D Haussecker and NJ Proudfoot. Dicer-dependent turnover of intergenic transcripts from the human β-globin gene cluster. *Molecular and Cellular Biology*, 25(21):9724–9733, 2005.

M Hayashi, Y Katou, T Itoh, M Tazumi, Y Yamada, T Takahashi, T Nakagawa, K Shirahige, and H Masukata. Genome-wide localization of pre-RC sites and identification of replication origins in fission yeast. *The EMBO journal*, 26(5): 1327–1339, 2007.

L He. microRNAs: small RNAs with a big role in gene regulation. *Nature*, 5(7):522–531, 2004.

SS Hébert, K Horré, L Nicolaï, B Bergmans, AS Papadopoulou, A Delacourte, and B De Strooper. microRNA regulation of alzheimer's amyloid precursor protein expression. *Neurobiology of disease*, 33(3):422–428, 2009.

C Heichinger, C J Penkett, J Bähler, and P Nurse. Genome-wide characterization of fission yeast DNA replication origins. *The EMBO journal*, 25(21):5171–5179, 2006.

IL Hofacker, W Fontana, PF Stadler, LS Bonhoeffer, M Tacker, and P Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie/Chemical Monthly*, 125(2):167–188, 1994.

PJ Horn and CL Peterson. Heterochromatin assembly: a new twist on an old model. *Chromosome Research*, 14(1):83–94, 2006.

T Huang, B Fan, M Rothschild, Z Hu, K Li, and S Zhao. MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans. *BMC bioinformatics*, 8(1):341, 2007.

T Hubbard, D Barker, E Birney, G Cameron, Y Chen, L Clark, T Cox, J Cuff, V Curwen, T Down, et al. The ensembl genome database project. *Nucleic acids research*, 30(1): 38–41, 2002.

T Humphrey, P Sadhale, T Platt, and NJ Proudfoot. Homologous mRNA 3' end formation in fission and budding yeast. *EMBO J*, 10:3505–3511, 1991.

T Humphrey, CE Birse, and NJ Proudfoot. RNA 3' end signals of the *S.pombe ura4* gene comprise a site determining and efficiency element. *EMBO J*, 13(10):1441–51, 1994.

MV Iorio, M Ferracin, C Liu, A Veronese, R Spizzo, S Sabbioni, E Magri, M Pedriali, M Fabbri, M Campiglio, et al. microRNA gene expression deregulation in human breast cancer. *Cancer research*, 65(16):7065–7070, 2005.

A Jakymiw, RS Patel, N Deming, I Bhattacharyya, P Shah, RJ Lamont, CM Stewart, DM Cohen, and EKL Chan. Overexpression of Dicer as a result of reduced let-7 microRNA levels contributes to increased cell proliferation of oral cancer cells. *Genes, Chromosomes and Cancer*, 49(6):549–559, 2010.

CH Jan, RC Friedman, JG Ruby, and DP Bartel. Formation, regulation and evolution of *Caenorhabditis elegans* 3' UTRs. *Nature*, 469(7328):97–101, 2011.

BA Janowski, KE Huffman, JC Schwartz, R Ram, D Hardy, DS Shames, JD Minna, and DR Corey. Inhibiting gene expression at transcription start sites in chromosomal DNA with antigene RNAs. *Nature chemical biology*, 1(4):216–222, 2005.

M Jiang, J Anderson, J Gillespie, and M Mayne. uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. *BMC bioinformatics*, 9(1):192, 2008.

B John, AJ Enright, A Aravin, T Tuschl, C Sander, and DS Marks. Human microRNA targets. *PLoS biology*, 2(11):e363, 2004.

D Kaida, MG Berg, I Younis, M Kasim, LN Singh, L Wan, and G Dreyfuss. U1 snRNP protects pre-mRNAs from premature cleavage and polyadenylation. *Nature*, 468(7324):664–668, 2010.

H Kaneko, S Dridi, V Tarallo, BD Gelfand, BJ Fowler, WG Cho, ME Kleinman, SL Ponicsan, WW Hauswirth, VA Chiodo, et al. Dicer1 deficit induces alu RNA toxicity in age-related macular degeneration. *Nature*, 471(7338):325–330, 2011.

C Kanellopoulou, SA Muljo, AL Kung, S Ganesan, R Drapkin, T Jenuwein, DM Livingston, and K Rajewsky. Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. *Genes & development*, 19(4): 489–501, 2005.

T Kawai and S Akira. Toll-like receptors and their crosstalk with other innate receptors in infection and immunity. *Immunity*, 34(5):637–650, 2011.

DH Kim, LM Villeneuve, KV Morris, and JJ Rossi. Argonaute-1 directs siRNA-mediated transcriptional gene silencing in human cells. *Nature structural & molecular biology*, 13(9):793–797, 2006.

DH Kim, P Sætrom, O Snøve, and JJ Rossi. microRNA-directed transcriptional gene silencing in mammalian cells. *Proceedings of the National Academy of Sciences*, 105 (42):16230–16235, 2008.

VN Kim, J Han, and MC Siomi. Biogenesis of small RNAs in animals. *Nature reviews Molecular cell biology*, 10(2):126–139, 2009.

M Kiriakidou, PT Nelson, A Kouranov, P Fitziev, C Bouyioukos, Z Mourelatos, and A Hatzigeorgiou. A combined computational-experimental approach predicts human microRNA targets. *Genes & development*, 18(10):1165–1178, 2004.

T Kivioja, A Vähärautio, K Karlsson, M Bonke, M Enge, S Linnarsson, and J Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nature methods*, 9(1):72–74, 2012.

N Kotaja, SN Bhattacharyya, L Jaskiewicz, S Kimmins, M Parvinen, W Filipowicz, and P Sassone-Corsi. The chromatoid body of male germ cells: Similarity with processing bodies and presence of Dicer and microRNA pathway components. *Proceedings of the National Academy of Sciences of the United States of America*, 103(8):2647–2652, 2006.

A Kozomara and S Griffiths-Jones. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic acids research*, 42(D1):D68–D73, 2014.

A Krek, D Grün, MN Poy, R Wolf, L Rosenberg, EJ Epstein, P MacMenamin, I da Piedade, KC Gunsalus, M Stoffel, et al. Combinatorial microRNA target predictions. *Nature genetics*, 37(5):495–500, 2005.

L Kuai, F Fang, JS Butler, and F Sherman. Polyadenylation of rRNA in *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences of the United States of America*, 101(23):8581–8586, 2004.

S Kueng, B Hegemann, B Peters, J Lipp, A Schleiffer, K Mechtler, and J Peters. Wapl controls the dynamic association of cohesin with chromatin. *Cell*, 127(5):955–967, 2006.

Y Kurihara and Y Watanabe. *Arabidopsis* microRNA biogenesis through Dicer-like 1 protein functions. *Proceedings of the National Academy of Sciences of the United States of America*, 101(34):12753–12758, 2004.

E Ladewig, K Okamura, AS Flynt, JO Westholm, and EC Lai. Discovery of hundreds of mirtrons in mouse and human small RNA data. *Genome research*, 22(9):1634–1645, 2012.

H Lange, FM Sement, J Canaday, and D Gagliardi. Polyadenylation-assisted RNA degradation processes in plants. *Trends in plant science*, 14(9):497–504, 2009.

B Langmead, C Trapnell, M Pop, and SL Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

RC Lee, RL Feinbaum, and V Ambros. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell*, 75(5):843–854, 1993.

M Legendre and D Gautheret. Sequence determinants in human polyadenylation site selection. *BMC Genomics*, 4(1):7, 2003.

A Lengronne, Y Katou, S Mori, S Yokobayashi, GP Kelly, T Itoh, Y Watanabe, K Shirahige, and F Uhlmann. Cohesin relocation from sites of chromosomal loading to places of convergent transcription. *Nature*, 430(6999):573–578, 2004.

AKL Leung, AG Young, A Bhutkar, GX Zheng, AD Bosson, CB Nielsen, and PA Sharp. Genome-wide identification of Ago2 binding sites from mouse embryonic stem cells with and without mature microRNAs. *Nature structural & molecular biology*, 18(2): 237–244, 2011.

BP Lewis, I Shih, MW Jones-Rhoades, DP Bartel, and CB Burge. Prediction of mammalian microRNA targets. *Cell*, 115(7):787–798, 2003.

BP Lewis, CB Burge, and DP Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120 (1):15–20, 2005.

L Li, J Xu, D Yang, X Tan, and H Wang. Computational approaches for microRNA studies: a review. *Mammalian genome*, 21(1-2):1–12, 2010.

LP Lim, NC Lau, EG Weinstein, A Abdelhakim, S Yekta, MW Rhoades, CB Burge, and DP Bartel. The microRNAs of *Caenorhabditis elegans*. *Genes & development*, 17(8): 991–1008, 2003.

B Liu, J Li, and MJ Cairns. Identifying miRNAs, targets and functions. *Briefings in bioinformatics*, 15(1):1–19, 2012.

AR Lohe, AJ Hilliker, and PA Roberts. Mapping simple repeated DNA sequences in heterochromatin of *Drosophila melanogaste*. *Genetics*, 134(4):1149–74, 1993.

JC Loke, EA Stahlberg, DG Strenski, BJ Haas, PC Wood, and QQ Li. Compilation of mRNA polyadenylation signals in *Arabidopsis* revealed a new signal element and potential secondary structures. *Plant Physiol.*, 138(3):1457–1468, 2005.

A Losada. Cohesin regulation: fashionable ways to wear a ring. *Chromosoma*, 116(4): 321–329, 2007.

A Losada, M Hirano, and T Hirano. Identification of xenopus Smc protein complexes required for sister chromatid cohesion. *Genes & Development*, 12(13):1986–1997, 1998.

A Losada, Ti Yokochi, and T Hirano. Functional contribution of Pds5 to cohesin-mediated cohesion in human cells and xenopus egg extracts. *Journal of cell science*, 118(10): 2133–2141, 2005.

R Lyne, G Burns, J Mata, CJ Penkett, G Rustici, D Chen, C Langford, D Vetrie, and J Bähler. Whole-genome microarrays of fission yeast: characteristics, accuracy, reproducibility, and processing of array data. *BMC genomics*, 4(1):27, 2003.

SS Mandal, C Chu, T Wada, H Handa, AJ Shatkin, and D Reinberg. Functional interactions of RNA-capping enzyme with factors that positively and negatively regulate promoter escape by RNA polymerase II. *Proceedings of the National Academy of Sciences of the United States of America*, 101(20):7572–7577, 2004.

M Mangone, AP Manoharan, D Thierry-Mieg, J Thierry-Mieg, T Han, SD Mackowiak, E Mis, C Zegar, MR Gutwein, V Khivansara, O Attie, K Chen, K Salehi-Ashtiani, M Vidal, TT Harkins, P Bouffard, Y Suzuki, S Sugano, Y Kohara, N Rajewsky, F Piano, KC Gunsalus, and JK Kim. The landscape of *C. elegans* 3'UTRs. *Science*, 329(5990): 432–5, 2010.

S Marguerat and J Bähler. RNA-seq: from technology to biology. *Cellular and Molecular Life Sciences*, 67(4):569–579, 2010.

S Marguerat, A Schmidt, S Codlin, W Chen, R Aebersold, and J Bähler. Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. *Cell*, 151(3):671–683, 2012.

JHA Martens, RJ O'Sullivan, U Braunschweig, S Opravil, M Radolf, P Steinlein, and T Jenuwein. The profile of repeat-associated histone lysine methylation states in the mouse epigenome. *The EMBO journal*, 24(4):800–812, 2005.

J Mata. Genome-wide mapping of polyadenylation sites in fission yeast reveals widespread alternative polyadenylation. *RNA biology*, 10(8):1407–1414, 2013.

ML Mayer, I Pot, M Chang, H Xu, V Aneliunas, T Kwok, R Newitt, R Aebersold, C Boone, GW Brown, et al. Identification of protein complexes required for efficient sister chromatid cohesion. *Molecular biology of the cell*, 15(4):1736–1745, 2004.

C Mayr and DP Bartel. Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, 138(4):673–684, 2009.

P Maziere and AJ Enright. Prediction of microRNA targets. *Drug discovery today*, 12 (11):452–458, 2007.

J McLauchlan, D Gaffney, JL Whitton, and JB Clements. The consensus sequence YGTGTTYY located downstream from the AATAAA signal is required for efficient formation of mrna 3' termini. *Nucleic Acids Research*, 13(4):1347–1368, 1985.

AJ McNairn and JL Gerton. Intersection of ChIP and FLIP, genomic methods to study the dynamics of the cohesin proteins. *Chromosome research*, 17(2):155–163, 2009.

CC Mello and D Conte. Revealing the world of RNA interference. *Nature*, 431(7006): 338–342, 2004.

H Mitsuzawa and A Ishihama. RNA Polymerase II transcription apparatus in *Schizosaccharomyces pombe*. *Current Genetics*, 44(6):287–294, 2004. ISSN 0172-8083.

F Miura, N Kawaguchi, J Sese, A Toyoda, M Hattori, S Morishita, and T Ito. A large-scale full-length cDNA analysis to explore the budding yeast transcriptome. *Proceedings of the National Academy of Sciences*, 103(47):17846–17851, 2006.

A Molnár, F Schwach, DJ Studholme, EC Thuenemann, and DC Baulcombe. miRNAs control gene expression in the single-cell alga *Chlamydomonas reinhardtii*. *Nature*, 447(7148):1126–1129, 2007.

DP Morris, GA Michelotti, and DA Schwinn. Evidence that phosphorylation of the RNA Polymerase II carboxyl-terminal repeats is similar in yeast and humans. *Journal of Biological Chemistry*, 280(36):31368–31377, 2005.

KV Morris, S Chan, SE Jacobsen, and DJ Looney. Small interfering RNA-induced transcriptional gene silencing in human cells. *Science*, 305(5688):1289–1292, 2004.

EP Murchison, JF Partridge, OH Tam, S Cheloufi, and GJ Hannon. Characterization of

Dicer-deficient murine embryonic stem cells. *Proceedings of the National Academy of Sciences of the United States of America*, 102(34):12135–12140, 2005.

EP Murchison, P Stein, Z Xuan, H Pan, MQ Zhang, RM Schultz, and GJ Hannon. Critical roles for Dicer in the female germline. *Genes & development*, 21(6):682–693, 2007.

U Nagalakshmi, Z Wang, K Waern, C Shou, D Raha, M Gerstein, and M Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881):1344–1349, 2008.

SH Nagaraj, RB Gasser, and S Ranganathan. A hitchhiker's guide to expressed sequence tag (EST) analysis. *Briefings in Bioinformatics*, 8(1):6–21, 2007.

K Nasmyth and CH Haering. Cohesin: its roles and mechanisms. *Annual review of genetics*, 43:525–558, 2009.

H Neil, C Malabat, Y d/'Aubenton Carafa, Z Xu, LM Steinmetz, and A Jacquier. Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature*, 457(7232):1038–1042, 2009.

KC Neuman, EA Abbondanzieri, R Landick, J Gelles, and SM Block. Ubiquitous transcriptional pausing is independent of RNA Polymerase backtracking. *Cell*, 115 (4):437–447, 2003.

N Nonaka, T Kitajima, S Yokobayashi, G Xiao, M Yamamoto, SIS Grewal, and Y Watanabe. Recruitment of cohesin to heterochromatic regions by Swi6/HP1 in fission yeast. *Nature Cell Biology*, 4(1):89–93, 2002.

NM Nunes, W Li, B Tian, and A Furger. A functional human Poly(A) site requires only a potent DSE and an A-rich upstream sequence. *The EMBO Journal*, 29(9):1523–1536, 2010.

C Okada, E Yamashita, SJ Lee, S Shibata, J Katahira, A Nakagawa, Y Yoneda, and T Tsukihara. A high-resolution structure of the pre-microRNA nuclear export machinery. *Science*, 326(5957):1275–1279, 2009.

K Okamura, JW Hagen, H Duan, DM Tyler, and EC Lai. The mirtron pathway generates microRNA-class regulatory RNAs in *Drosophila*. *Cell*, 130(1):89–100, 2007.

I Onn and D Koshland. *In vitro* assembly of physiological Cohesin/DNA complexes. *Proceedings of the National Academy of Sciences*, 108(30):12198–12205, 2011.

F Ozsolak, AR Platt, DR Jones, JG Reifenberger, LE Sass, P McInerney, JF Thompson, J Bowers, M Jarosz, and PM Milos. Direct RNA sequencing. *Nature*, 461(7265): 814–818, 2009.

F Ozsolak, P Kapranov, S Foissac, W Kim, E Fishilevich, AP Monaghan, B John, and PM Milos. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell*, 143(6):1018–1029, 2010.

V Pancaldi, F Schubert, and J Bähler. Meta-analysis of genome regulation and expression variability across hundreds of environmental and genetic perturbations in fission yeast. *Molecular BioSystems*, 6(3):543–552, 2010.

MY Park, G Wu, A Gonzalez-Sulser, H Vaucheret, and RS Poethig. Nuclear processing and export of microRNAs in *Arabidopsis*. *Proceedings of the National Academy of Sciences of the United States of America*, 102(10):3691–3696, 2005.

J Parkinson and M Blaxter. *Expressed Sequence Tags: An Overview*, volume 533 of *Methods in Molecular Biology*, book section 1, pages 1–12. Humana Press, 2009.

AE Pasquinelli. microRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. *Nature Reviews Genetics*, 13(4):271–282, 2012.

AE Pasquinelli, BJ Reinhart, F Slack, MQ Martindale, MI Kuroda, B Maller, DC Hayward, EE Ball, B Degnan, P Müller, et al. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, 408(6808):86–89, 2000.

N Passon, A Gerometta, C Puppin, E Lavarone, F Puglisi, G Tell, C Di Loreto, and G Damante. Expression of Dicer and Drosha in triple-negative breast cancer. *Journal of clinical pathology*, 65(4):320–326, 2012.

V Pelechano, W Wei, and LM Steinmetz. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature*, 497(7447):127–131, 2013.

J Peters, A Tedeschi, and J Schmitz. The Cohesin complex and its roles in chromosome biology. *Genes & development*, 22(22):3089–3114, 2008.

M Petronczki, B Chwalla, MF Siomos, S Yokobayashi, W Helmhart, AM Deutschbauer, RW Davis, Y Watanabe, and K Nasmyth. Sister-chromatid cohesion mediated by the alternative RF-CCtf18/Dcc1/Ctf8, the helicase Chl1 and the polymerase-α-associated protein Ctf4 is essential for chromatid disjunction during meiosis II. *Journal of cell science*, 117(16):3547–3559, 2004.

KP Porkka, MJ Pfeiffer, KK Waltering, RL Vessella, TLJ Tammela, and T Visakorpi. microRNA expression profiling in prostate cancer. *Cancer research*, 67(13): 6130–6135, 2007.

NJ Proudfoot. Sequence analysis of the 3' non-coding regions of rabbit α- and β-globin messenger RNAs. *EMBO J*, 107(4):419–525, 1976.

NJ Proudfoot. How RNA Polymerase II terminates transcription in higher eukaryotes. *Trends Biochem Sci*, 14:105–110, 1989.

NJ Proudfoot. Ending the message: poly(A) signals then and now. *Genes and Development*, 25(17):1770–1782, 2011.

P Provost, D Dishart, J Doucet, D Frendewey, B Samuelsson, and O Rådmark. Ribonuclease activity and RNA binding of recombinant human Dicer. *The EMBO Journal*, 21(21):5864–5874, 2002.

M Rehmsmeier, P Steffen, M Höchsmann, and R Giegerich. Fast and effective prediction of microRNA/target duplexes. *RNA*, 10(10):1507–1517, 2004.

BJ Reinhart, FJ Slack, M Basson, AE Pasquinelli, JC Bettinger, AE Rougvie, HR Horvitz, and G Ruvkun. The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature*, 403(6772):901–906, 2000.

D Retelska, C Iseli, P Bucher, CV Jongeneel, and F Naef. Similarities and differences of polyadenylation signals in human and fly. *BMC Genomics*, 7(1):176, 2006.

M Reuter, P Berninger, S Chuma, H Shah, M Hosokawa, C Funaya, C Antony, R Sachidanandam, and RS Pillai. Miwi catalysis is required for piRNA amplification-independent LINE1 transposon silencing. *Nature*, 480(7376):264–267, 2011.

N Rhind, Z Chen, M Yassour, DA Thompson, BJ Haas, N Habib, I Wapinski, S Roy, MF Lin, DI Heiman, et al. Comparative functional genomics of the fission yeasts. *Science*, 332(6032):930–936, 2011.

A Rodriguez, S Griffiths-Jones, J. Ashurst, and A Bradley. Identification of mammalian microRNA host genes and transcription units. *Genome Research*, 14(10a):1902–1910, 2004.

MF Rojas-Duran and WV Gilbert. Alternative transcription start site selection leads to large differences in translation activity in yeast. *RNA*, 18(12):2299–2305, 2012.

HM Rothnie, J Reid, and T Hohn. The contribution of AAUAAA and the upstream element UUUGUA to the efficiency of mRNA 3'-end formation in plants. *The EMBO journal*, 13(9):2200, 1994.

CH Ruby, JGand Jan and DP Bartel. Intronic microRNA precursors that bypass Drosha processing. *Nature*, 448(7149):83–86, 2007.

S Rudra and RV Skibbens. Cohesin codes–interpreting chromatin architecture and the many facets of cohesin function. *Journal of cell science*, 126(1):31–41, 2013.

V Rusinov, V Baev, IN Minkov, and M Tabler. MicroInspector: a web tool for detection of miRNA binding sites in an RNA sequence. *Nucleic acids research*, 33(suppl 2): W696–W700, 2005.

V Saint-André, E Batsché, C Rachez, and C Muchardt. Histone h3 lysine 9 trimethylation and HP1γ favor inclusion of alternative exons. *Nature structural & molecular biology*, 18(3):337–344, 2011.

MP Samanta, W Tongprasit, H Sethi, C Chin, and V Stolc. Global identification of noncoding RNAs in *Saccharomyces cerevisiae* by modulating an essential RNA processing pathway. *Proceedings of the National Academy of Sciences of the United States of America*, 103(11):4192–4197, 2006.

CE Samuel. Antiviral actions of interferons. *Clinical microbiology reviews*, 14(4): 778–809, 2001.

M Schena, RA Heller, T Theriault, K Konrad, E Lachenmeier, and RW Davis. Microarrays: biotechnology's discovery platform for functional genomics. *Trends in Biotechnology*, 16(7):301–306, 1998.

M Schlackow, S Marguerat, NJ Proudfoot, J Bähler, R Erban, and M Gullerova. Genome-wide analysis of poly(A) site selection in *Schizosaccharomyces pombe*. *RNA*, 19(12):1617–1631, 2013.

M Schmid and Tk Jensen. The exosome: a multipurpose RNA-decay machine. *Trends in biochemical sciences*, 33(10):501–510, 2008.

C Schmidt, N Brookes, and F Uhlmann. Conserved features of cohesin binding along fission yeast chromosomes. *Genome Biology*, 10(5):R52, 2009.

D Schmidt, PC Schwalie, CS Ross-Innes, A Hurtado, GD Brown, JS Carroll, P Flicek, and DT Odom. A CTCF-independent role for cohesin in tissue-specific transcription. *Genome research*, 20(5):578–588, 2010.

D Schmitter, J Filkowski, A Sewer, RS Pillai, EJ Oakeley, M Zavolan, P Svoboda, and W Filipowicz. Effects of Dicer and Argonaute down-regulation on mRNA levels in human HEK293 cells. *Nucleic acids research*, 34(17):4801–4815, 2006.

D Schübeler, D Scalzo, C Kooperberg, B van Steensel, J Delrow, and M Groudine. Genome-wide DNA replication profile for *Drosophila melanogaster*: a link between transcription and replication timing. *Nature genetics*, 32(3):438–442, 2002.

M Segurado, A de Luis, and F Antequera. Genome-wide distribution of DNA replication origins at A+T-rich islands in *Schizosaccharomyces pombe*. *EMBO reports*, 4(11): 1048–1053, 2003.

A Sehgal, BT Hughes, and PJ Espenshade. Oxygen-dependent, alternative promoter controls translation of *tco1+* in fission yeast. *Nucleic Acids Research*, 36(6): 2024–2031, 2008.

M Selbach, B Schwanhäusser, N Thierfelder, Z Fang, R Khanin, and N Rajewsky. Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455(7209): 58–63, 2008.

P Sethupathy, M Megraw, and AG Hatzigeorgiou. A guide through present computational approaches for the identification of mammalian microRNA targets. *Nature methods*, 3 (11):881–886, 2006.

SA Shabalina and EV Koonin. Origins and evolution of eukaryotic RNA interference. *Trends in ecology & evolution*, 23(10):578–587, 2008.

D Shalon, SJ Smith, and PO Brown. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Research*, 6 (7):639–645, 1996.

Z Shan, Q Lin, C Deng, J Zhu, L Mai, J Liu, Y Fu, X Liu, Y Li, Y Zhang, et al. miR-1/miR-206 regulate Hsp60 expression contributing to glucose-mediated apoptosis in cardiomyocytes. *FEBS letters*, 584(16):3592–3600, 2010.

Y Shen, G Ji, BJ Haas, X Wu, J Zheng, GJ Reese, and QQ Li. Genome level analysis of rice mRNA 3'-end processing signals and alternative polyadenylation. *Nucleic Acids Res.*, 36(9):3150–3161, 2008a.

Y Shen, Y Liu, L Liu, C Liang, and QQ Li. Unique features of nuclear mRNA poly(A) signals and alternative polyadenylation in *Chlamydomonas reinhardtii*. *Genetics*, 179 (1):167–176, 2008b.

A Sherstnev, C Duc, C Cole, V Zacharaki, C Hornyik, F Ozsolak, PM Milos, GJ Barton, and GG Simpson. Direct sequencing of *Arabidopsis thaliana* RNA reveals patterns of cleavage and polyadenylation. *Nature structural & molecular biology*, 19(8):845–852, 2012.

T Shiraki, S Kondo, S Katayama, K Waki, T Kasukawa, H Kawaji, R Kodzius, A Watahiki, M Nakamura, T Arakawa, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proceedings of the National Academy of Sciences*, 100(26):15776–15781, 2003.

F Simmer, A Buscaino, IC Kos-Braun, A Kagansky, A Boukaba, T Urano, ARW Kerr, and RC Allshire. Hairpin RNA induces secondary small interfering RNA synthesis and silencing in trans in fission yeast. *EMBO reports*, 11(2):112–118, 2010.

CC Siow, SR Nieduszynska, CA Müller, and CA Nieduszynski. OriDB, the DNA

replication origin database updated and extended. *Nucleic acids research*, 40(D1): D682–D686, 2012.

RV Skibbens. Chl1p, a DNA helicase-like protein in budding yeast, functions in sister-chromatid cohesion. *Genetics*, 166(1):33–42, 2004.

RV Skibbens, LB Corson, D Koshland, and P Hieter. Ctf7p is essential for sister chromatid cohesion and links mitotic chromosome structure to the DNA replication machinery. *Genes & Development*, 13(3):307–319, 1999.

A Stark, J Brennecke, RB Russell, and SM Cohen. Identification of *Drosophila* microRNA targets. *PLoS Biology*, 1(3):e60, 2003.

DA Steege. Emerging features of mRNA decay in bacteria. *RNA*, 6(8):1079–1090, 2000.

DB Stetson and R Medzhitov. Type I interferons in host defense. *Immunity*, 25(3): 373–381, 2006.

I Sumara, E Vorlaufer, C Gieffers, B Peters, and J Peters. Characterization of vertebrate cohesin complexes and their regulation in prophase. *The Journal of cell biology*, 151 (4):749–762, 2000.

M Tanaka, Y Sakai, O Yamada, T Shintani, and K Gomi. In silico analysis of 3'-end-processing signals in *Aspergillus oryzae* using expressed sequence tags and genomic sequencing data. *DNA Res.*, 18(3):189–200, 2011.

F Tang, M Kaneda, D OCarroll, P Hajkova, SC Barton, YA Sun, C Lee, A Tarakhovsky, K Lao, and MA Surani. Maternal microRNAs are essential for mouse zygotic development. *Genes & development*, 21(6):644–648, 2007.

R Thadani and MT Tammi. MicroTar: predicting microRNA targets from RNA duplexes. *BMC bioinformatics*, 7(Suppl 5):S20, 2006.

DM Thompson and R Parker. Cytoplasmic decay of intergenic transcripts in *Saccharomyces cerevisiae*. *Molecular and cellular biology*, 27(1):92–101, 2007.

B Tian, J Hu, H Zhang, and CS Lutz. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Research*, 33(1):201–212, 2005.

A Tóth, R Ciosk, F Uhlmann, M Galova, A Schleiffer, and K Nasmyth. Yeast cohesin complex requires a conserved protein, Eco1p (Ctf7), to establish cohesion between sister chromatids during DNA replication. *Genes & development*, 13(3):320–333, 1999.

S Tuduri, L Crabbé, C Conti, H Tourrière, H Holtgreve-Grez, A Jauch, V Pantesco, J De Vos, A Thomas, C Theillet, et al. Topoisomerase I suppresses genomic instability by preventing interference between replication and transcription. *Nature cell biology*, 11(11):1315–1324, 2009.

S van Dongen, C Abreu-Goodger, and AJ Enright. Detecting microRNA binding and siRNA off-target effects from expression data. *Nature methods*, 5(12):1023–1025, 2008.

S Vasudevan, Y Tong, and JA Steitz. Switching from repression to activation: microRNAs can up-regulate translation. *Science*, 318(5858):1931–1934, 2007.

K Venkataraman, KM Brown, and GM Gilmartin. Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition. *Genes & development*, 19(11):1315–1327, 2005.

TA Vickers and ST Crooke. siRNAs targeted to certain polyadenylation sites promote specific, RISC-independent degradation of messenger RNAs. *Nucleic acids research*, 40(13):6223–6234, 2012.

T Volpe, V Schramke, GL Hamilton, SA White, G Teng, RA Martienssen, and RC Allshire. RNA interference is required for normal centromere function infission yeast. *Chromosome Research*, 11(2):137–146, 2002.

Xi Wang, J Zhang, F Li, J Gu, T He, X Zhang, and Y Li. microRNA identification based on sequence and structure alignment. *Bioinformatics*, 21(18):3610–3614, 2005.

T Watanabe, K Miyashita, TT Saito, T Yoneki, Y Kakihara, K Nabeshima, YA Kishi, C Shimoda, and H Nojima. Comprehensive isolation of meiosis-specific genes identifies novel proteins and unusual non-coding transcripts in *Schizosaccharomyces pombe*. *Nucleic acids research*, 29(11):2327–2337, 2001.

KS Wendt, K Yoshida, T Itoh, M Bando, B Koch, E Schirghuber, S Tsutsumi, G Nagae, K Ishihara, T Mishiro, et al. Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature*, 451(7180):796–801, 2008.

E White, M Schlackow, K Kamieniarz-Gdula, NJ Proudfoot, and M Gullerova. Human nuclearDicer restricts the deleterious accumulation of endogenous double-stranded RNA. *Nature Structural & Molecular Biology*, 2014.

EJ White, O Emanuelsson, D Scalzo, T Royce, S Kosak, EJ Oakeley, S Weissman, M Gerstein, M Groudine, M Snyder, et al. DNA replication-timing analysis of human chromosome 22 at high resolution and different developmental states. *Proceedings of the National Academy of Sciences of the United States of America*, 101 (51):17771–17776, 2004.

B Wightman, I Ha, and G Ruvkun. Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. *Cell*, 75(5): 855–862, 1993.

BT Wilhelm, S Marguerat, S Watt, F Schubert, V Wood, I Goodhead, CJ Penkett, J Rogers, and J Bähler. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, 453(7199):1239–1243, 2008.

JE Wilusz and DL Spector. An unexpected ending: Noncanonical 3' end processing mechanisms. *RNA*, 16(2):259–266, 2010.

MA Winters and M J Edmonds. A poly(A) Polymerase from calf thymus. characterization of the reaction product and the primer requirement. *J Biol Chem*, 248:4763–4768, 1973a.

MA Winters and MJ Edmonds. A poly(A) Polymerase from calf thymus. Purification and properities of the enzyme. *J Biol Chem*, 248:4756–4762, 1973b.

K Woodfine, H Fiegler, DM Beare, JE Collins, OT McCann, BD Young, S Debernardi, R Mott, I Dunham, and NP Carter. Replication timing of the human genome. *Human molecular genetics*, 13(2):191–202, 2004.

KJ Woolcock, D Gaidatzis, T Punga, and M Bühler. Dicer associates with chromatin to repress genome activity in *Schizosaccharomyces pombe*. *Nature structural & molecular biology*, 18(1):94–99, 2011.

S Wuchty, W Fontana, IL Hofacker, P Schuster, et al. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49(2):145–165, 1999.

J Xu, Y Yanagisawa, AM Tsankov, C Hart, K Aoki, N Kommajosyula, KE Steinmann, J Bochicchio, C Russ, A Regev, et al. Genome-wide identification and characterization of replication origins by deep sequencing. *Genome Biol*, 13(4):R27, 2012.

Z Xu, W Wei, J Gagneur, F Perocchi, S Clauder-Münster, J Camblong, E Guffanti, F Stutz, W Huber, and LM Steinmetz. Bidirectional promoters generate pervasive transcription in yeast. *Nature*, 457(7232):1033–1037, 2009.

S Yamanaka, S Mehta, FE Reyes-Turcu, F Zhuang, RT Fuchs, Y Rong, GB Robb, and SIS Grewal. RNAi triggered by specialized machinery silences developmental genes and retrotransposons. *Nature*, 493(7433):557–560, 2013.

J Yan and TG Marr. Computational analysis of 3'-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat. *Genome Research*, 15(3): 369–375, 2005.

N Yanaihara, N Caplen, E Bowman, M Seike, K Kumamoto, M Yi, RM Stephens, A Okamoto, J Yokota, T Tanaka, et al. Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. *Cancer cell*, 9(3):189–198, 2006.

H Yang, W Kong, L He, J Zhao, JD O'Donnell, J Wang, RM Wenham, D Coppola, PA Kruk, SV Nicosia, et al. microRNA expression profiling in human ovarian cancer: miR-214 induces cell survival and cisplatin resistance by targeting PTEN. *Cancer research*, 68(2):425–433, 2008.

M Yassour, T Kaplan, HB Fraser, JZ Levin, J Pfiffner, X Adiconis, G Schroth, S Luo, I Khrebtukova, A Gnirke, C Nusbaum, D Thompson, N Friedman, and A Regev. *Ab initio* construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proceedings of the National Academy of Sciences*, 106(9):3264–3269, 2009.

A Zehir, LL Hua, EL Maska, Y Morikawa, and P Cserjesi. Dicer is required for survival of differentiating neural crest cells. *Developmental biology*, 340(2):459–467, 2010.

L Zhang, D Hou, X Chen, D Li, L Zhu, Yu Zhang, J Li, Z Bian, X Liang, Xing Cai, et al. Exogenous plant MIR168a specifically targets mammalian LDLRAP1: evidence of cross-kingdom regulation by microRNA. *Cell research*, 22(1):107–126, 2011.

X Zhang, M Cairns, B Rose, C O'Brien, K Shannon, J Clark, J Gamble, and N Tran. Alterations in miRNA processing and expression in pleomorphic adenomas of the salivary gland. *International Journal of Cancer*, 124(12):2855–2863, 2009.

Zhihong Zhang and Fred S Dietrich. Mapping of transcription start sites in *Saccharomyces cerevisiae* using 5' SAGE. *Nucleic acids research*, 33(9):2838–2851, 2005.

J Zhao, L Hyman, and C Moore. Formation of mRNA 3' ends in eukaryotes: Mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiology and Molecular Biology Reviews*, 63(2):405–445, 1999.

DG Zisoulis, MT Lovci, ML Wilbert, KR Hutt, TY Liang, AE Pasquinelli, and GW Yeo. Comprehensive discovery of endogenous argonaute binding sites in *Caenorhabditis elegans*. *Nature structural & molecular biology*, 17(2):173–179, 2010.

Appendix A

3'RACE protocol

A.1 RNA-purification

1. Spin down cells
    - 3000 rpm for 3-4 minus
    - discard supernatant
    - get pellet and add 0.5 ml of water, re-suspend
    - Replace into 2 thick eppendorf tubes
2. - Centrifuge 1 min at 13 K
    - discard supernatant
    - Re-suspend in ~250 µl buffer (Buffer: 50 mM Tris-HCL pH 7.5, 10 mM EDTA pH 8, 100 mM NaCl, 1% SDS).
    - add phenol: chloroform ("ACID, MONI "), same amount (~250 µl)
    - add washed beads, same amount (~250 µl).
    - Keep on ice
3. - In homogeniser: 1 min spin, 5 minutes on ice. Repeat 3 times. (Alternatively vortex for 15 minutes, 5 minutes on ice. Repeat 3 times)

- Centrifuge at 4°C (cold centrifuge) at 14 K

4.
- Take upper phase without the beads (clear liquid, should be ~250 μl)
- add same amount as liquid of isopropanol to both tubes
- add 1/10 of the volume of salt NaAc 3 M ph 7.5
- mix carefully (gets cloudy)

5.
- Centrifuge at 4°C for 15 minutes at 13 K
- get pellet, take off supernatant
- Add 1/2 ml of 70% ethanol and remove ("wash ")
- leave pellet to dry
- re-suspend in 50-100 μl of water (say 50 μl).
- AT THIS POINT SAMPLE IS FREEZABLE

6.
- Add DNase 1(enzyme) (1/10th of the Volume, i.e. 5 μl of buffer, 2 μl of enzyme)
- incubate at 37°C for one hour.
- AT THIS POINT SAMPLE IS FREEZABLE

7.
- Add 15 μl of water
- 200 μl of phenol: chloroform
- Vortex
- Centrifuge at 4°C at 13 K for 8 minutes (5-10 minutes)
- Refill aquaphase into fresh tubes
- Add same amount of isopropanol
- Add 1/10th of the volume of salt (20 μl)
- Centrifuge at 4°C for 15 minutes at 13 K

- Suck off supernatant, wash with 70% ethanol
- re-suspend pellet in ~50 µl water
- AT THIS POINT SAMPLE IS FREEZABLE

8. This is now the final RNA sample. The concentration can be measured by nano-drop.

A.2 RT-PCR

The purified RNA sample will now be reverse transcribed into double- stranded cDNA. The negative control does not contain reverse transcriptase, hence the cDNA should not be created and amplified.

1. Use the previous purified RNA samples. Depending on how many dilutions are wanted to be used later on, create a multiple of 2 tubes as the input and the negative control.

   Create a mix for the input and the negative control as indicated in the Table A.1.

Add	**input**	**negative control**	**Comment**
RNA-sample	100 ng of RNA	100 ng of RNA	100 ng=x µl depending on sample concentration
water	11µl − x µl	11 µl − x µl + 1 µl	negative has more water to replace the reverse transcriptase
DNTP 10 mM	1 µl	1 µl	dilute 1 M of DNTP by 1:10 with water
Phased d(N)T	1 µl	1 µl	oligo: stretch of Ts with A, C, G in front, acts as a primer for the poly(A) tail.
RnaseOUT	1 µl	1 µl	Ribonuclease inhibitor
0.1 M DTT	1 µl	1 µl	DNase protecting agent
5×first strand buffer	4 µl	4 µl	Do not dilute, 5×refers to final concentration
SSIII RT	1 µl	0 µl	Reverse transcriptase, was substituted with water in negative control
SUM	20 µl	20 µl	

Table A.1: Mix for input and negative control for RT-PCR.

2. Vortex solutions or mix with pipettes

3. Put into small holes thermocycler, Program 88 (annealing step at 48°C for 1 hour, elongation period at 75°C 5min, see Section A.3 for precise description of the PCR procedure), close tight

4. create templates for PCR: for different dilutions, add eg. 80 µl or 180 µl of water

A.3 PCR

Corresponding to the number of dilutions obtained, create 2 tubes per gene for each input and negative control.

1. Create Mastermix (MM) as given in Table A.2. An alternative Recipe for the mastermix is given in Table A.3.

Add	**amount per tube (μl)**	**MM for 18 genes**(40 μl)	Mastermix is for all genes suppose 2 tubes for input/negative control per gene to have MM extra, multiply column 2 by 40 (column 3)
10×Thermo buffer	5	200	
25 mM MgCl	3	120	Aids Taq polymerase as a cofactor, speeding up the reaction
dNTPs	2	80	collection of free bases A, T, C, G
dT linker oligont.	1	40	Necessary as a starting point for the Taq polymerase
Taq polymerase	2	80	thermostable bacterial polymerase adds the dNTPs to annealed primer and copies the DNA
water	35	1400	

Table A.2: Mastermix recipe for the PCR. Note that there are more ingredients added separately to the tubes for gene.

Add	**amount per tube (μl)**	**MM for 18 genes**(40 μl)
5×buffer	10	400
dNTPs	1	40
dT linker oligont.	1	40
Taq polymerase	1	40
water	35	1400

Table A.3: Alternative mastermix recipe for the PCR. Note that there are more ingredients added separately to the tubes for gene (same as for the previous mastermix). This recipe was used once the previous did not yield satisfactory results. It is likely to need further optimisation.

2. Now create 2 tubes for input and negative control for each gene. Add 48 μl of MM into each of them. For each gene add (1 μl) of the corresponding oligonucleotide into the 2 tubes (again necessary as a starting point for the Taq polymerase). Add (1 μl) of input template into the input tube and (1 μl) of negative control template into the negative control tube (note that the templates were the ones previously obtained by the RT-PCR in the previous section).

Run the tubes in the PCR block, Program 66, which uses the following settings: 3 minutes at 95°C for the heat activation of the polymerase. Then there were 35 cycles, 30 seconds at each temperature, of 95°C (denaturation step to separate the DNA strands by disruption of hydrogen bonds), then 55°C (annealing step: primers associate with the complementary sequence in the single-stranded DNA) and then 72 °C (elongation step: 72°C is optimal for the Taq polymerase to add dNTPs to the primer and synthesize a new DNA strand corresponding to each probe). The middle step of each cycle should be a different temperature for the optimum of each oligonucleotide, according to its T_m, but for simplicity I chose 55°C for each of them (as it turns out it gave me the desired results). To ensure that any single-stranded DNA is fully extended after the cycling, the samples are kept at 72°C for 5 minutes. Finally the samples are held for 4°C to store it until loading the gel. The gel (for its preparation see below Section A.3.1) is loaded with 10 µl-20 µl of each sample and 15 µl of DNA ladder. The bands corresponding to each probe are observed after gel electrophoresis.

To run the gel always run it from - (black electrode) to plus (red electrode) and leave space on the gel at the plus-end to observe the bands. If the present samples are to be reused, make droplets if 5 µl of orange dye and 5 µl of sample on a parafilm, mix with pipette without creating bubbles and add 10 µl to the lanes on the gel. Alternatively if the samples are not to be reused the dye can be added directly to the tubes. If the samples are to be reused in PCR green dye does not actually impede the PCR action. It might also be useful to add 20 µl of sample to the lanes. Connect electrodes and run the gel on 120bV.

If the above only gives very weak bands the PCR can be run again by using the now higher concentrated samples as the template. Then repeat as follows: Create the mastermix as above, but use the now after-PCR tubes as the templates for input and negative control. Add 5 µl of these to the new tubes. Run again on Program 66 and

then proceed to gel-electrophoresis.

A.3.1 Preparation of Agarose gel

To prepare 500 ml of Agarose gel

- Create 1×E buffer from the available 50×E buffer
- Add 2% Agarose Low EEO in 1×E buffer (10 g in 500 ml)
- Microwave bottle until the liquid is completely clear, i.e. there are no crystals (needs to be watched to not overboil)
- Add Ethidium-bromide to have final concentration of 0.1% (here ~50 µl).
- Set up the Gel-tank: Add the same 1×E buffer as for the gel to the tank

Appendix B

Yeast Transformations

B.1 PCR purification - using Quiagen Purification Kit

PCR purification is a way to clean DNA. Use purification columns

- Use purification columns
- BINDING STEP: add 200 µl PCR product and 5×this Volume (here 1000 µl) to columns (possibly spin twice and columns hold 600 µl)
- spin column at 13K, 1min (possibly spin twice and columns hold 600 µl)
- discard supernatant
- WASH STEP: add 700 µl of PE buffer
- spin at 13 K, 1 min
- discard supernatant
- DRY STEP: spin at 13 K, 1 min
- Put filter from column to Eppendorf tube
- add 50 µl EB buffer

- spin at 13 K, 2 min
- Results in ~45 µl PCR insert.

B.2 Plasmid restriction

Cut the plasmid and the promoter-*ura4*-ORF with restriction enzymes *PstI* and *XhoI* as described below. All other terminator inserts are to be cut with *XmaI* and *SacI* and/or *BamHI* and *SacI*.

B.2.1 Restriction of Plasmid pJR1-3XH

- 5 µl plasmid
- 2 µl (3) Buffer (optimised Buffer number for the two restriction enzymes used)
- 2 µl *PstI*
- 2 µl *XhoI*
- 9 µl water

put to 37°C for 1 h. Add 2 µl of CIP and put to oven for 20 minutes. CIP phosphorylates the open ends, so that the plasmid doesn't close again and remains a linear DNA molecule.

Run to check on the gel: should have one band corresponding to the cut plasmid with missing promoter-*nmt1*-ORF (~8 kb) and one band corresponding to the cut out promoter-*nmt1*-ORF (~1 kb).

B.3 Restriction PCR amplified promoter-*ura4*-ORF

After PCR purification,

- 45 µl PCR product

- 5 µl (3) Buffer (optimised Buffer number for the two restriction enzymes used)
- 2 µl *PstI*
- 2 µl *XhoI*

put to 37°C for 1 h.

B.3.1 PCR amplified terminator inserts

after PCR purification:

- 45 µl PCR product
- 2 µl (4) Buffer (optimised Buffer number for the two restriction enzymes used)
- 2 µl *XmaI*
- 2 µl *SacI*

put to 37°C for 1 h.

B.4 Ligation

This step concerns the ligation of promoter-*ura4*-ORF to Plasmid with cut out promoter-*nmt1*-ORF or the ligation of any terminator insert to the Plasmid with inserted promoter-*ura4*- ORF and cut out *nmt1* terminator.

- 10 µl of 2x ligation buffer
- 1 µl T4 DNA ligase
- 9 µl of insert and cut plasmid, ratio to be determined by PCR (usually there is a higher concentration of plasmid than insert). 7 µl of insert and 2 µl of plasmid should be reasonable. For the negative control use water instead of insert.
- leave 30min on the bench

B.5 Bacterial transformation

When transforming bacteria only bacteria with the closed plasmid will grow, as the plasmid contains an Ampicillin Resistance Gene.

- Warm LB Amp plates in 37°C oven, leave them open to dry
- Use competent bacterial cells (e.g.XL1), keep them on ice to melt, but not overheat.
- In one tube: 100 µl of competent cells, 20 µl ligation mixture (with negative control). Mix slowly with the pipette
- incubate on ice for 10mins
- Heat shock on the thermo-block for 60 s, 42°C
- back on ice for 1min
- Add ~300 µl of YT medium to all samples
- Put to 37°C shaker for 30 min in order to recover the cells
- quick, light spin to get more cells on the bottom
- plate 80 µl on the warmed plates. Work close to the fire for sterility. Use glass spreader by dipping it in ethanol, burning it with fire, holding it on the side of the plate and letting it cool down, spread the cells.
- plates to 37°C oven overnight to let bacterial cultures grow

Ideally the negative control gets only few colonies, while the positive plates grow many. This is an indicator of how well the preparation has worked and if it should be repeated or weather one can proceed to harvest the colonies (they will be double checked again anyway).

B.5.0.1 Bacterial Minipreps

B.6 LiOAc Transformations of *S. pombe*

Adapted from Bähler et al. (1998)

- Inoculate overnight culture of *ura4* deficient strain (*D-18*) from a fresh plate (YEA or EMM medium).
- Next day dilute into 100 ml of medium to an OD600 ~0.08 (about 5 transformations)
- Grow to an OD600 of 0.2-0.5 (~4 to 6h)
- Spin cells down at 3300 rpm for 4 minutes
- Wash with 1×ddH_2O
- Wash with 20 ml LiOAc mix ("Lite" - 1ml 10×TE pH 7.5, 1 ml 1 M LiOAc pH 7.0, 8 ml ddH_2O, pellet again, resuspend in 100 µl of LiOAc mix (for every 20 ml of culture)
- Let stand on bench for 10 min, meanwhile boil salmon sperm for 10 min at 95°C
- For each transformation add into an Eppendorf tube: 10 µl of salmon sperm, 10 µl (1-5 µg) of DNA (desired plasmid), 100 µl of cells
- Incubate at room temperature for 10 mins. Add 260 µl of PEG mix (10×TE, 0.1 < Tris-Cl pH 7.5, 10 mM EDTA). Pipette up and down or vortex to mix well
- Incubate at 30°C for 30-60 mins
- Add 43 µl DMSO. Mix.
- Heat shock at 42°C for 5 min

- Spin down for 10 s. Aspirate off supernatant
- wash in 1 µl of water
- add 100 µl of water to ~300 µl of sample
- plate on YEA plates (-His plates in our case - only cells with the plasmid should survive, as it contains his3). 250 µl per plate.
- Incubate at 30°C for several days
- once colonies are visible, replate on new -His plates (can also check success via colony-PCR)

B.7 Real Time quantitative PCR

After RT-PCR, the qPCR is performed. Dilute each RT sample with 20 µl of water

- Control: Probe Actin (in the genome, all cells should have the same levels of Actin transcription more or less).
- per sample: 7.8 µl of Sensimix (Cybergreen mix), 4 µl of water, 0.6 µl of forward and 0.6 µl of reverse Oligo.
- Pipette 13 µl in each tube
- 2 µl of RT-PCR sample
- can be stored in the fridge in the dark

Appendix C

S. pombe polyadenylation additional figures and tables

Data Set	Cycling. Strand-specific	Cycling, Non strand-specific	Meiotic	Quiescent (24 hours nitrogen depletion)	Quiescent (7 days nitrogen depletion)
Number of sequences in data set	58517960	"26858469	18513752	31394347	28801407
Number of sequences mapped back before elimination of heterogeneity of cleavage site	6612	26102	2000	9582	8236
Number of sequences mapped back after elimination of heterogeneity of cleavage site	3093	12385	988	4618	3215
Number of genes with identified cleavage sites after the gene (total in genome)	1964 (5153)	4198 (5153)	812 (5153)	2627 (5153)	2024 (5153)
Number of tandem genes with identified cleavage sites (total in genome)	843 (2332)	1840 (2332)	351 (2332)	1120 (2332)	868 (2332)
Number of convergent genes with identified cleavage sites (total in genome)	1121 (2816)	2358 (2816)	461 (2816)	1504 (2816)	1155 (2816)
Length of RNA-seq reads (nt)	51	30	50 (trimmed)	51	76

Table App.C.1 Transcript pool composition of all analysed datasets from varying physiological conditions

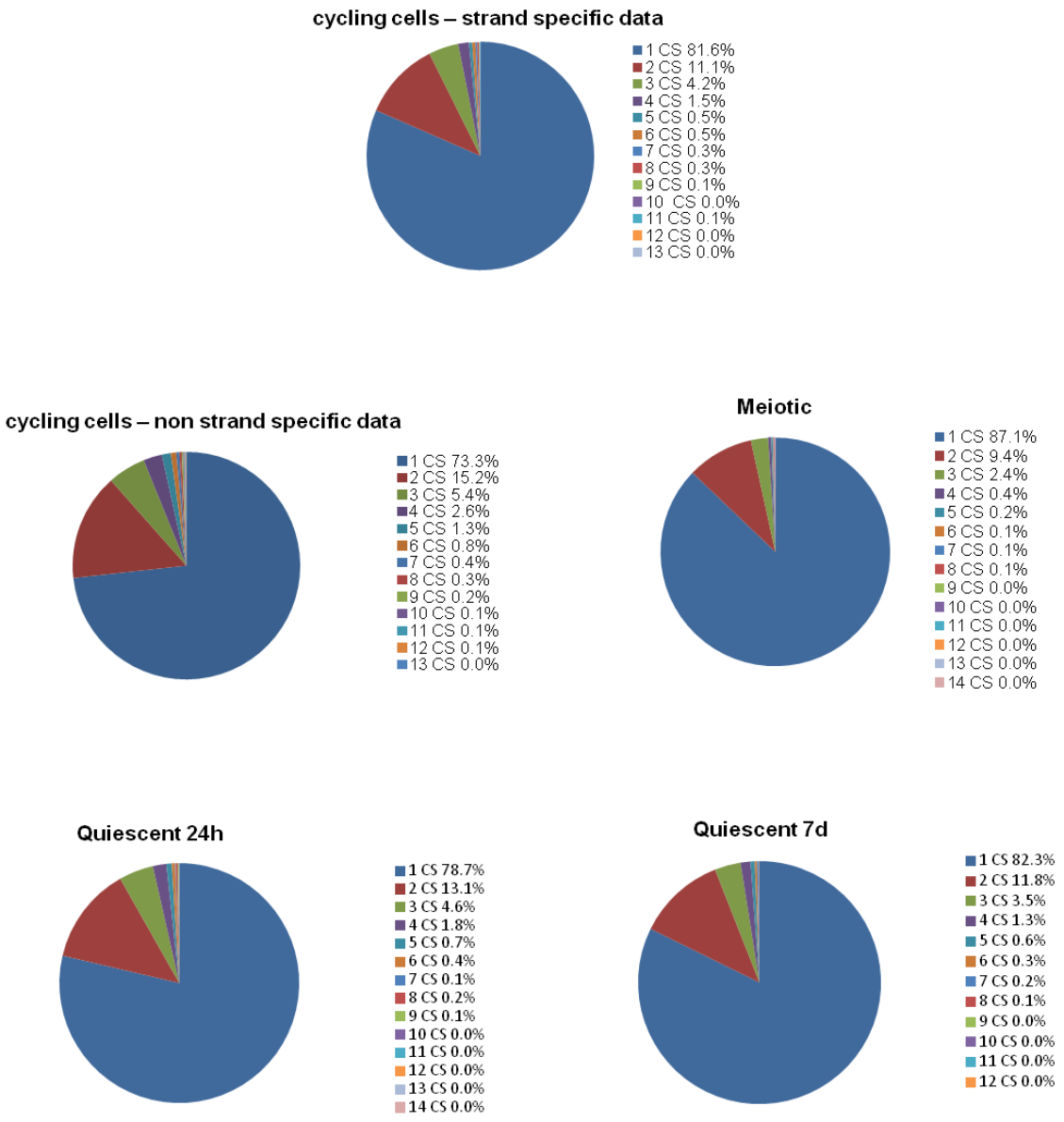

Figure C.1: Heterogeneity: the proportion of all CS identified, which are grouped into one representative CS if an individual separation of 1- 6 nt is observed, is shown for all analysed data sets.

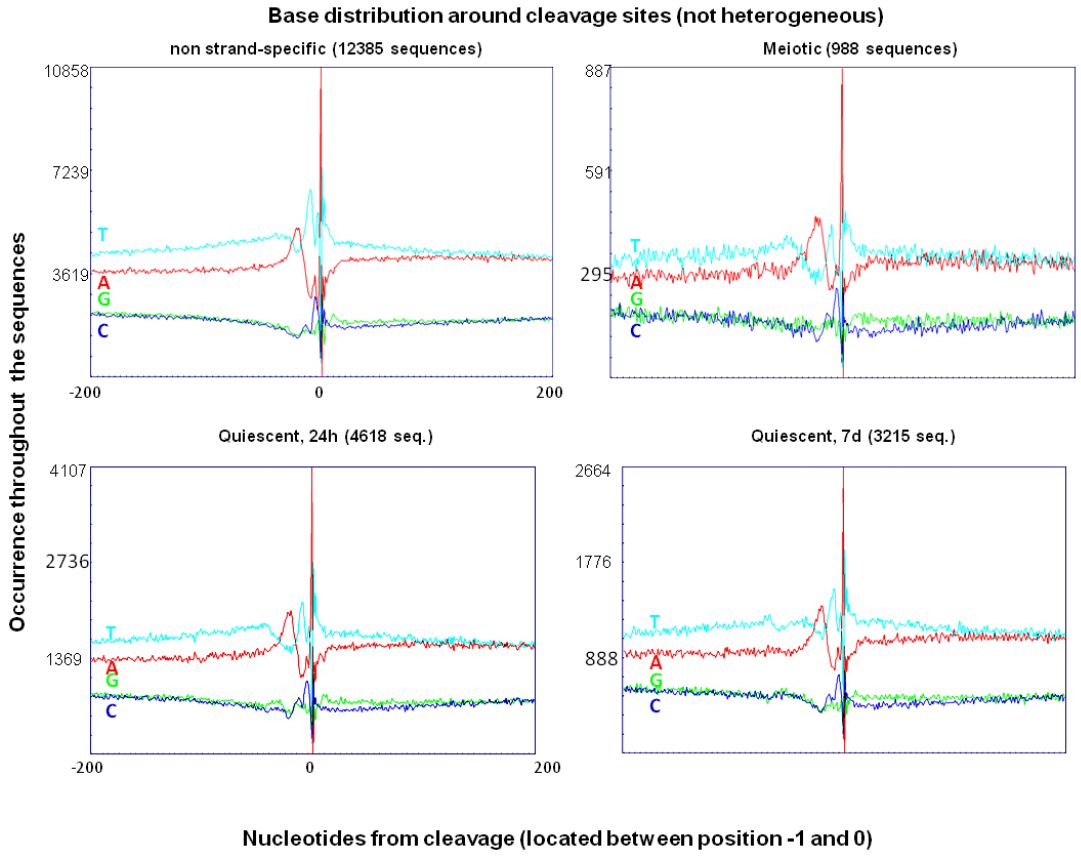

Figure C.2: Nucleotide composition around mapped CS is shown for all data sets. CS are processed as described in Figure 2.16 (see Materials and Methods).

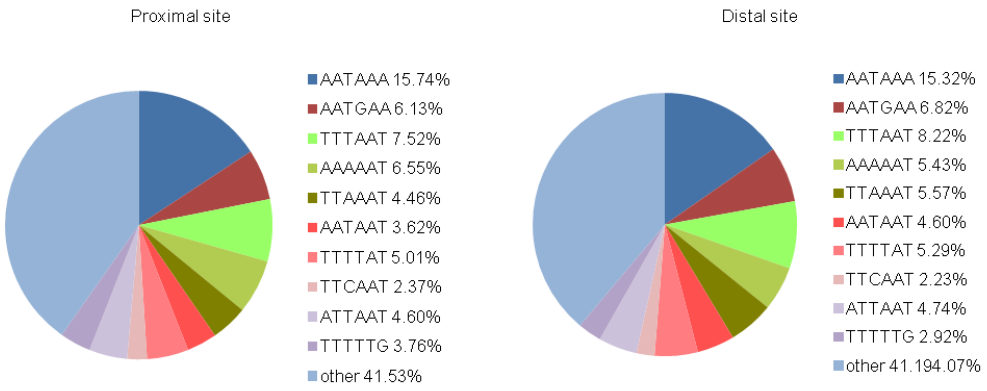

Figure C.3: The proportion of all alternatively polyadenylated genes, which have the identified motifs either at the most proximal site (left) or at the most distal site (right). Genes with ranking motifs were excluded for calculation of the proportions with lower ranking motifs.

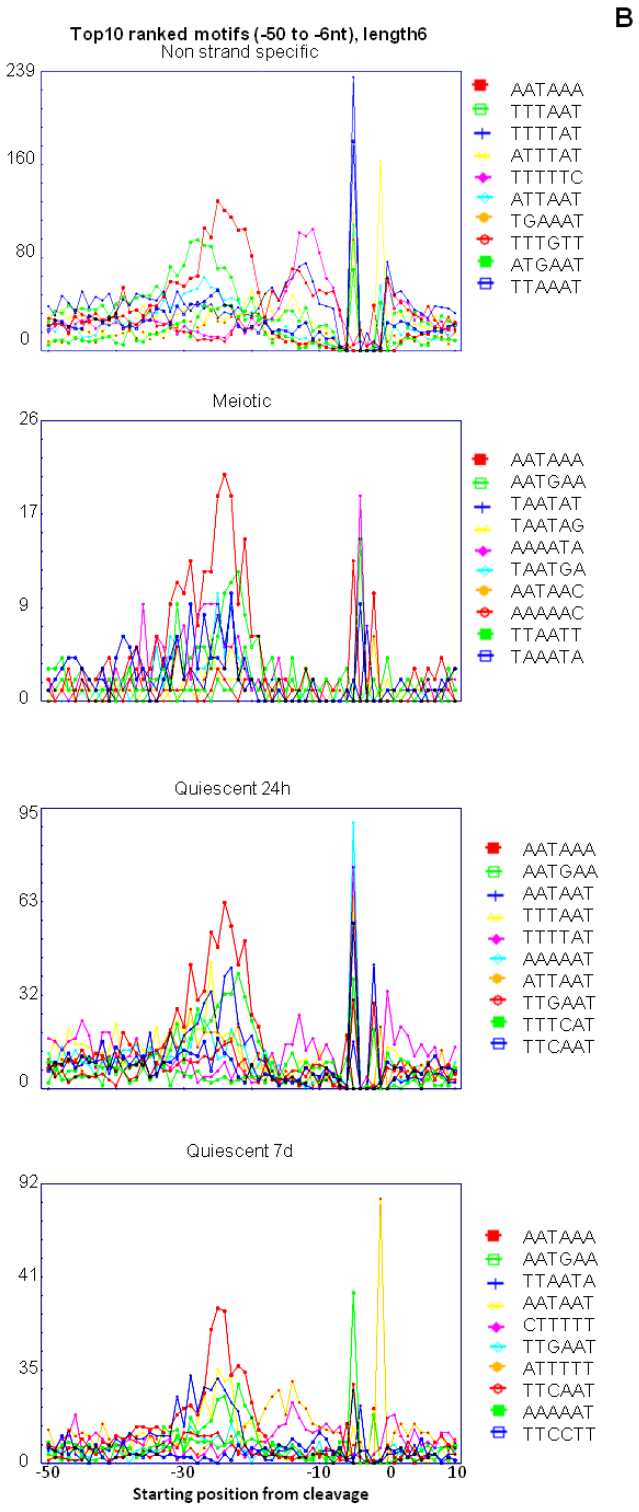

Figure C.4: As in Figure 2.21, possible PASs were analysed for all other available data sets. The positions of the Top 10 motifs are shown.

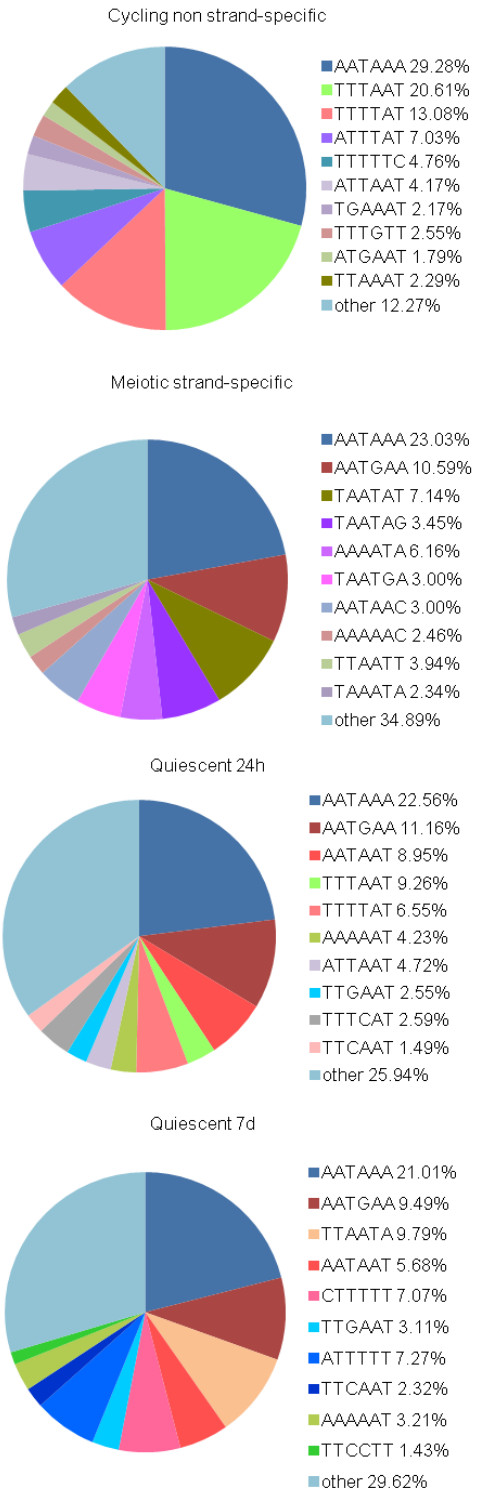

Figure C.5: The proportions of genes with the detected possible PAS is shown for all data sets, procedures are as for Figure 2.23

Improbizer Outcomes

Non strand specific cycling						
AATAAA						
A	0.990	0.988	0.003	0.458	0.466	0.485
C	0.003	0.004	0.041	0.110	0.085	0.138
G	0.003	0.005	0.148	0.236	0.144	0.113
T	0.003	0.003	0.808	0.196	0.305	0.265

Meiotic							
TAATAAA							
A	0.285	0.978	0.981	0.003	0.508	0.534	0.573
C	0.138	0.011	0.003	0.111	0.089	0.132	0.115
G	0.221	0.008	0.011	0.030	0.290	0.147	0.158
T	0.356	0.003	0.005	0.855	0.114	0.187	0.153

Quiescent (24h)						
AAATAA						
A	0.318	0.985	0.990	0.003	0.499	0.452
C	0.145	0.009	0.003	0.105	0.111	0.128
G	0.260	0.003	0.003	0.058	0.225	0.161
T	0.277	0.003	0.003	0.834	0.165	0.260

Quiescent (7d)						
AAATAA						
A	0.279	0.879	0.964	0.004	0.547	0.458
C	0.191	0.058	0.003	0.068	0.057	0.170
G	0.262	0.059	0.029	0.003	0.213	0.137
T	0.268	0.003	0.003	0.924	0.183	0.236

Figure C.6: Improbizer result for cycling non strand specific, meiotic, quiescent (24 h) and quiescent (7 d) data.

	Cycling strand-specific					Meiotic				
Gene Primary Name	CS position	Distance from stop codon	Number of hits	Hit score	p-value Fisher, left	CS position	Distance from stop codon	Number of Hits	Hit score	p-value Fisher, right
SPAC343.09	1655775	260	2	3.417754139		1655775	260	1	5.40139027	
	1655818	303	1	1.70887707		1655818	303	3	16.2041708	0.041336218
	1655830	315	13	22.2154019		1655830	315	1	5.40139027	
SPAC22H10.09	2389853	181	2	3.417754139		2389873	201	4	21.6055611	0.028940293
	2389873	201	2	3.417754139						
SPAC3F10.19	2831368	160	1	1.70887707		2831595	387	3	16.2041708	0.012510582
	2831539	331	2	3.417754139		2831608	400	1	5.40139027	
	2831608	400	1	1.70887707						
SPAC1142.06	3638382	292	1	1.70887707		3638382	292	1	5.40139027	
	3638421	331	2	3.417754139						
	3638564	474	1	1.70887707						
SPAC27F1.07	4330648	193	1	1.70887707		4330560	105	1	5.40139027	
	4330678	223	5	8.544385348		4330648	193	1	5.40139027	
SPAC17C9.11c	4481999	70	1	1.70887707		4482101	172	1	5.40139027	
	4482331	402	1	1.70887707		4482544	615	7	37.8097319	0.000230521
	4482346	417	1	1.70887707						
	4482479	550	2	3.417754139						
	4482544	615	1	1.70887707						
SPAC14C4.14	5259083	255	3	5.126631209		5259129	301	1	5.40139027	
	5259109	281	1	1.70887707		5259182	354	1	5.40139027	
	5259129	301	1	1.70887707						
	5259182	354	3	5.126631209						
	5259198	370	1	1.70887707						
SPAC11E3.15	5315376	60	2	3.417754139		5315429	113	2	10.8027805	
	5315429	113	1	1.70887707						
SPBC660.11	216700	205	3	5.126631209		216748	253	1	5.40139027	
	216748	253	1	1.70887707						
SPBC800.09	269114	161	2	3.417754139		269473	520	1	5.40139027	
	269310	357	1	1.70887707						
	269412	459	1	1.70887707						
	269476	523	1	1.70887707						
	269529	576	1	1.70887707						
	269652	699	3	5.126631209						
SPBC428.11	467150	229	2	3.417754139		466992	71	1	5.40139027	
	467215	294	1	1.70887707		467211	290	1	5.40139027	
SPBC1685.09	516765	31	1	1.70887707		516769	35	1	5.40139027	
	516780	46	8	13.67101656		516780	46	1	5.40139027	
	516814	80	1	1.70887707						
SPBC18H10.02	1772516	268	1	1.70887707		1772522	274	1	5.40139027	
	1772645	397	1	1.70887707		1772632	384	1	5.40139027	
	1772659	411	2	3.417754139						
SPBC2G5.05	2588636	59	1	1.70887707		2588651	74	2	10.8027805	
	2588651	74	3	5.126631209		2588666	89	1	5.40139027	
	2588666	89	1	1.70887707		2588682	105	6	32.4083416	
	2588682	105	21	35.88641846						
	2588719	142	2	3.417754139						
SPBC1539.06	4373407	171	1	1.70887707		4373407	171	1	5.40139027	
	4373428	192	8	13.67101656						
SPCC1020.06c	774655	190	5	8.544385348		774655	190	5	27.0069514	0.058969449
	774782	317	2	3.417754139						
	774821	356	1	1.70887707						
SPCC1393.03	798785	43	1	1.70887707		798785	43	1	5.40139027	
	798816	74	7	11.96213949		798893	151	1	5.40139027	
	798837	95	2	3.417754139						
	798854	112	6	10.25326242						
	798893	151	11	18.79764776						
	798911	169	6	10.25326242						
SPCC1322.04	1293777	80	1	1.70887707		1293785	88	1	5.40139027	
	1293785	88	1	1.70887707						
	1293814	117	1	1.70887707						
	1293822	125	2	3.417754139						
	1293881	184	1	1.70887707						
	1293940	243	1	1.70887707						
SPCC576.11	2100518	35	3	5.126631209		2100515	32	2	10.8027805	0.053911967
	2100538	55	22	37.59529553	0.018130665	2100538	55	1	5.40139027	
	2100555	72	1	1.70887707						

Table App.C.2 part 1: Candidate alternatively polyadenylated genes between cycling and meiotic cells. The *p*-values are computed according to the right Fisher Exact Test for cycling cells and according to the left Fisher Exact Test for meiotic cells

	Cycling strand-specific					Meiotic				
Gene Primary Name	CS position	Distance from stop codon	Number of hits	Hit score	p-value Fisher, left	CS position	Distance from stop codon	Number of Hits	Hit score	p-value Fisher, right
SPAC3A11.13	3441428	258	1	1.70887707		3441428	258	1	5.40139027	
	3441610	76	4	6.835508278						
SPAC24C9.06c	3048142	266	1	1.70887707		3048142	266	1	5.40139027	
	3048218	190	3	5.126631209						
SPAC6F12.13c	1338293	158	1	1.70887707		1338293	158	1	5.40139027	
	1338347	104	1	1.70887707		1338365	82	1	5.40139027	
	1338367	84	38	64.93732864	4.25E-05					
SPAC139.01c	1016241	124	1	1.70887707		1016241	124	1	5.40139027	
	1016273	92	2	3.417754139						
SPBC26H8.11c	3974094	271	2	3.417754139		3974281	84	1	5.40139027	
	3974281	84	1	1.70887707						
SPBC21D10.12	2420330	459	1	1.70887707		2420534	255	1	5.40139027	
	2420391	398	4	6.835508278						
	2420518	271	1	1.70887707						
	2420537	252	1	1.70887707						
SPBC19G7.03c	2348981	96	168	287.0913477	6.37E-06	2348981	96	20	108.027805	
	2348990	87	18	30.75978725	0.008546363	2349025	52	1	5.40139027	
	2349028	49	1	1.70887707						
SPBC1D7.04	1743966	146	2	3.417754139		1744023	89	1	5.40139027	
	1744015	97	2	3.417754139						
	1744023	89	1	1.70887707						
SPBC1306.01c	1177865	214	2	3.417754139		1177929	150	1	5.40139027	
	1177929	150	1	1.70887707						
SPCC1840.06	2268642	241	5	8.544385348		2268700	183	1	5.40139027	
	2268700	183	1	1.70887707						
SPCC31H12.04c	626914	297	1	1.70887707		627016	195	1	5.40139027	
	626946	265	1	1.70887707						
	627022	189	1	1.70887707						
	627043	168	13	22.2154019	0.032120595					
	627062	149	3	5.126631209						
	627124	87	1	1.70887707						
	627165	46	2	3.417754139						
	627172	39	3	5.126631209	0.06804865					

Table App.C.2 part 2.

	Cycling strand-specific					Quiescent 24h				
Gene Primary Name	CS position	Distance from stop codon	Number of hits	Hit score	p-value Fisher, left	CS position	Distance from stop codon	Number of Hits	Hit score	p-value Fisher, right
SPAC17C9.11c	4481999	70	1	1.70887707		4482544	615	4	12.7411473	
	4482331	402	1	1.70887707		4482554	625	1	3.18528683	
	4482346	417	1	1.70887707						
	4482479	550	2	3.417754139						
	4482544	615	1	1.70887707						
SPAC3G6.02	5383701	248	1	1.70887707		5383560	107	1	3.18528683	
	5383806	353	2	3.417754139		5383701	248	6	19.111721	
SPBC18H10.13	1792936	134	6	10.25326242	0.004626889	1793010	208	1	3.18528683	
						1793055	253	3	9.55586049	
SPBC1711.06	2144273	46	18	30.75978725		2144273	46	23	73.2615971	
	2144551	324	2	3.417754139		2144347	120	1	3.18528683	
SPBC14F5.06	4167195	319	1	1.70887707		4167133	257	2	6.37057366	
	4167243	367	2	3.417754139		4167195	319	5	15.9264341	
SPCC14G10.04	751146	383	4	6.835508278		751146	383	5	15.9264341	
	751161	398	1	1.70887707		751153	390	1	3.18528683	
	751604	841	2	3.417754139		751161	398	4	12.7411473	
						751469	706	1	3.18528683	
SPCC1322.04	1293777	80	1	1.70887707		1293761	64	1	3.18528683	
	1293785	88	1	1.70887707		1293775	78	1	3.18528683	
	1293814	117	1	1.70887707		1293940	243	2	6.37057366	
	1293822	125	2	3.417754139						
	1293881	184	1	1.70887707						
	1293940	243	1	1.70887707						
SPAC26F1.06	5173808	163	1	1.70887707		5173862	109	1	3.18528683	
	5173815	156	2	3.417754139		5173910	61	41	130.59676	
	5173910	61	26	44.43080381						
	5173959	12	1	1.70887707						
SPAC1952.08c	4979165	973	2	3.417754139		4979165	973	1	3.18528683	
	4979208	930	2	3.417754139		4979998	140	2	6.37057366	
	4979595	543	1	1.70887707		4980069	69	1	3.18528683	
	4979854	284	1	1.70887707						
SPAC16.02c	4424659	590	1	1.70887707		4424733	516	5	15.9264341	
	4424726	523	1	1.70887707		4424799	450	1	3.18528683	
	4424733	516	1	1.70887707		4424845	404	3	9.55586049	
	4425123	126	1	1.70887707		4424859	390	1	3.18528683	
	4425141	108	2	3.417754139		4424867	382	2	6.37057366	
	4425169	80	1	1.70887707		4425167	82	1	3.18528683	
SPAC27F1.06c	4327201	81	4	6.835508278		4327201	81	5	15.9264341	
	4327230	52	2	3.417754139						
SPAC2C4.11c	4278459	269	3	5.126631209	0.06804865	4278489	239	3	9.55586049	
	4278527	201	1	1.70887707						
SPAC31G5.17c	3019354	297	2	3.417754139		3019403	248	3	9.55586049	
	3019403	248	7	11.96213949	0.061131654	3019492	159	6	19.111721	0.042887874
	3019497	154	3	5.126631209	0.06804865	3019501	150	1	3.18528683	
	3019509	142	1	1.70887707						
SPBC14F5.04c	4157753	418	7	11.96213949	0.001888141	4157756	415	1	3.18528683	
	4157819	352	22	37.59529553		4157807	364	1	3.18528683	
	4157836	335	1	1.70887707		4157819	352	51	162.449628	0.038914418
	4157878	293	1	1.70887707		4157826	345	6	19.111721	0.042887874
	4157928	243	1	1.70887707		4157836	335	3	9.55586049	
	4158090	81	1	1.70887707		4157909	262	2	6.37057366	
SPBC21C3.09c	3810915	458	1	1.70887707		3810915	458	2	6.37057366	
	3811054	319	3	5.126631209	0.06804865	3811033	340	1	3.18528683	
SPBC21H7.07c	2266529	136	1	1.70887707		2265831	834	2	6.37057366	
	2266563	102	6	10.25326242		2266563	102	1	3.18528683	
SPBC17A3.05c	1410806	555	1	1.70887707		1410748	613	2	6.37057366	
	1411138	223	2	3.417754139		1410803	558	5	15.9264341	0.072497838
SPBC1709.15c	1127496	472	2	3.417754139		1127457	511	1	3.18528683	
	1127888	80	2	3.417754139		1127888	80	3	9.55586049	
SPBC646.10c	940694	130	2	3.417754139		940777	47	7	22.2970078	
	940758	66	1	1.70887707						
	940777	47	4	6.835508278						
SPBC1271.12	341980	428	8	13.67101656	0.000770444	341976	432	7	22.2970078	0.025370292
						341984	424	1	3.18528683	
						342135	273	2	6.37057366	

Table App.3 part 1: Candidate alternatively polyadenylated genes between cycling and quiescent (24 h) cells. The *p*-values are computed according to the right Fisher Exact Test for cycling cells and according to the left Fisher Exact Test for quiescent cells.

	Cycling strand-specific					Quiescent 24h				
Gene Primary Name	CS position	Distance from stop codon	Number of hits	Hit score	p-value Fisher, left	CS position	Distance from stop codon	Number of Hits	Hit score	p-value Fisher, right
SPBC1683.10c	164005	424	2	3.417754139		164026	403	1	3.18528683	
	164026	403	4	6.835508278	0.09353511	164265	164	2	6.37057366	
	164196	233	1	1.70887707						
SPCC16C4.18c	699876	288	1	1.70887707		699856	308	3	9.55586049	
	699970	194	2	3.417754139		699950	214	1	3.18528683	
						700082	82	2	6.37057366	
SPCC31H12.04c	626914	297	1	1.70887707		627013	198	1	3.18528683	
	626946	265	1	1.70887707		627034	177	4	12.7411473	
	627022	189	1	1.70887707		627043	168	5	15.9264341	
	627043	168	13	22.2154019	0.000770444	627060	151	5	15.9264341	0.072497838
	627062	149	3	5.126631209	0.06804865					
	627124	87	1	1.70887707						
	627165	46	2	3.417754139						
	627172	39	3	5.126631209	0.06804865					
SPCC794.12c	275213	300	2	3.417754139		275167	346	1	3.18528683	
	275511	2	1	1.70887707		275261	252	1	3.18528683	
						275280	233	1	3.18528683	
						275297	216	1	3.18528683	
						275381	132	3	9.55586049	
						275503	10	1	3.18528683	
						275511	2	2	6.37057366	
SPCC757.15	59471	956	2	3.417754139		59931	496	3	9.55586049	
	59934	493	1	1.70887707		60210	217	1	3.18528683	

Table App.3 part 2.

	Cycling strand-specific					Quiescent 7d				
Gene Primary Name	CS position	Distance from stop codon	Number of hits	Hit score	p-value Fisher, left	CS position	Distance from stop codon	Number of Hits	Hit score	p-value Fisher, right
SPAC1D4.04	646432	56	1	1.70887707		646432	56	1	3.4720526	
	646479	103	3	5.126631209	0.088284678					
SPAC1296.02	712258	109	8	13.67101656	0.00154276	712233	84	2	6.9441052	
	712265	116	3	5.126631209	0.088284678	712259	110	1	3.4720526	
SPAC9.09	1481042	122	2	3.417754139		1481172	252	5	17.360263	0.052481593
	1481159	239	1	1.70887707						
	1481169	249	2	3.417754139						
	1481200	280	1	1.70887707						
	1481221	301	3	5.126631209	0.088284678					
SPAC57A7.04c	1545065	240	1	1.70887707		1544849	24	2	6.9441052	
	1545202	377	63	107.6592554	5.11E-19	1545202	377	3	10.4161578	
SPAC6B12.13	2436940	353	12	20.50652483	6.05E-05	2436848	261	4	13.8882104	0.094635176
	2437006	419	1	1.70887707		2436938	351	2	6.9441052	
SPAC2F3.02	3925455	125	1	1.70887707		3925404	74	1	3.4720526	
	3925780	450	2	3.417754139		3925455	125	1	3.4720526	
SPAC1F7.04	4225659	541	6	10.25326242	0.084079269	4225351	233	1	3.4720526	
	4225671	553	46	78.6083452	8.63E-12	4225420	302	1	3.4720526	
	4225692	574	1	1.70887707		4225659	541	2	6.9441052	
	4225786	668	4	6.835508278	0.039304374	4225671	553	5	17.360263	
						4225695	577	13	45.1366838	0.000468593
						4225784	666	2	6.9441052	
SPAC27F1.07	4330648	193	1	1.70887707		4330648	193	4	13.8882104	
	4330678	223	5	8.544385348	0.017496854	4330863	408	1	3.4720526	
SPAC17C9.11c	4481999	70	1	1.70887707		4482346	417	1	3.4720526	
	4482331	402	1	1.70887707		4482544	615	2	6.9441052	
	4482346	417	1	1.70887707						
	4482479	550	2	3.417754139						
	4482544	615	1	1.70887707						
SPAC1782.11	4774605	100	6	10.25326242		4774605	100	12	41.6646312	
	4774638	133	2	3.417754139		4774613	108	2	6.9441052	
SPAC1006.07	5087016	276	30	51.26631209	2.82E-08	5086850	110	1	3.4720526	
	5087032	292	2	3.417754139		5086912	172	1	3.4720526	
	5087088	348	2	3.417754139		5087016	276	3	10.4161578	
	5087108	368	1	1.70887707		5087032	292	3	10.4161578	
	5087118	378	1	1.70887707		5087081	341	2	6.9441052	
	5087210	470	13	22.2154019		5087141	401	1	3.4720526	
						5087210	470	20	69.441052	
SPAC29A4.04c	5141969	139	1	1.70887707		5142095	265	16	55.5528416	0.008572065
	5142079	249	9	15.37989363	0.000686549					
	5142095	265	3	5.126631209						
	5142197	367	3	5.126631209	0.088284678					
SPAC14C4.14	5259083	255	3	5.126631209		5258922	94	4	13.8882104	0.094635176
	5259109	281	1	1.70887707		5259083	255	1	3.4720526	
	5259129	301	1	1.70887707		5259129	301	1	3.4720526	
	5259182	354	3	5.126631209	0.088284678	5259278	450	1	3.4720526	
	5259198	370	1	1.70887707						
SPAC3G6.02	5383701	248	1	1.70887707		5383701	248	2	6.9441052	
	5383806	353	2	3.417754139						
SPBC800.09	269114	161	2	3.417754139		269393	440	2	6.9441052	
	269310	357	1	1.70887707		269412	459	1	3.4720526	
	269412	459	1	1.70887707						
	269476	523	1	1.70887707						
	269529	576	1	1.70887707						
	269652	699	3	5.126631209	0.088284678					
SPBC947.15c	643635	295	1	1.70887707		643379	39	1	3.4720526	
	643718	378	2	3.417754139		643532	192	8	27.7764208	0.008948067
						643635	295	1	3.4720526	

Table App.C.4 part 1: Candidate alternatively polyadenylated genes between cycling and quiescent (7 d) cells. The *p*-values are computed according to the right Fisher Exact Test for cycling cells and according to the left Fisher Exact Test for quiescent cells.

	Cycling strand-specific					Quiescent 7d				
Gene Primary Name	CS position	Distance from stop codon	Number of hits	Hit score	p-value Fisher, left	CS position	Distance from stop codon	Number of Hits	Hit score	p-value Fisher, right
SPBC1709.05	1108382	86	1	1.70887707		1108419	123	1	3.4720526	
	1108405	109	185	316.1422579	2.37E-66	1108472	176	1	3.4720526	
	1108424	128	2	3.417754139		1108500	204	548	1902.68482	
	1108500	204	1306	2231.793453	2.97E-128					
SPBC409.13	1166412	166	5	8.544385348	0.066039633	1166300	54	2	6.9441052	
						1166314	68	9	31.2484734	0.004961233
						1166412	166	1	3.4720526	
SPBC83.10	1528717	243	9	15.37989363		1528714	240	1	3.4720526	
	1528831	357	1	1.70887707	0.000686549	1528831	357	1	3.4720526	
SPBC28F2.11	1590687	205	1	1.70887707		1590684	202	2	6.9441052	
	1590717	235	1	1.70887707						
	1590852	370	2	3.417754139						
SPBC19C2.07	1689489	47	1	1.70887707		1689478	36	1	3.4720526	
	1689543	101	200	341.7754139	9.94E-72	1689507	65	2	6.9441052	
	1689579	137	10	17.0887707		1689520	78	2	6.9441052	
	1689599	157	1	1.70887707		1689546	104	16	55.5528416	7.98E-05
						1689579	137	26	90.2733675	0.029768046
						1689652	210	1	3.4720526	
SPBC18H10.14	1794627	88	162	276.8380853	3.68E-56	1794605	66	1	3.4720526	
	1794699	160	28	47.84855795	5.41E-07	1794616	77	1	3.4720526	
	1794706	167	1	1.70887707		1794627	88	1	3.4720526	
	1794715	176	1	1.70887707		1794699	160	4	13.8882104	
						1794755	216	2	6.9441052	
SPBC19G7.17										
	2384376	100	1	1.70887707		2384567	291	22	76.3851572	2.31E-06
	2384530	254	6	10.25326242	0.007788299					
	2384570	294	1	1.70887707						
SPBC36B7.03	2389367	330	6	10.25326242	0.007788299	2389323	286	4	13.8882104	0.094635176
						2389369	332	1	3.4720526	
SPBC2G5.05	2588636	59	1	1.70887707		2588607	30	1	3.4720526	
	2588651	74	3	5.126631209	0.088284678	2588636	59	1	3.4720526	
	2588666	89	1	1.70887707		2588674	97	1	3.4720526	
	2588682	105	21	35.88641846	1.59E-05	2588682	105	3	10.4161578	
	2588719	142	2	3.417754139		2588713	136	5	17.360263	0.052481593
SPBC4F6.04	2691564	41	9	15.37989363	0.004116009	2691564	41	1	3.4720526	
	2691576	53	5	8.544385348	0.066039633	2691576	53	1	3.4720526	
	2691683	160	1	1.70887707		2691683	160	1	3.4720526	
SPBC405.07	3153913	38	4	6.835508278	0.039304374	3153904	29	1	3.4720526	
	3154155	280	1	1.70887707		3153917	42	807	2801.94645	2.50E-215
	3154168	293	19	32.46866432	0.000206866	3154147	272	1	3.4720526	
						3154155	280	1	3.4720526	
						3154168	293	4	13.8882104	
SPBC776.01	3172626	66	3	5.126631209	0.088284678	3172674	114	7	24.3043682	
	3172674	114	70	119.6213949	9.04E-18	3172696	136	31	107.633631	1.29E-06
	3172696	136	2	3.417754139						
SPBC776.11	3196971	35	1	1.70887707		3196985	49	1	3.4720526	
	3196985	49	38	64.93732864	9.30E-13	3196993	57	4	13.8882104	
	3196993	57	3	5.126631209	0.039304374					
SPBC19F8.08	3245954	24	12	20.50652483	0.000463333	3245954	24	1	3.4720526	
	3245993	63	4	6.835508278		3245973	43	15	52.080789	
	3246030	100	1	1.70887707		3246030	100	7	24.3043682	0.066466229
SPBC11C11.07	3362577	146	4	6.835508278	0.039304374	3362620	189	3	10.4161578	
	3362606	175	6	10.25326242	0.007788299	3362632	201	1	3.4720526	
	3362620	189	3	5.126631209						
	3362632	201	1	1.70887707						
SPBPB7E8.01	3496385	160	3	5.126631209	0.088284678	3496468	243	1	3.4720526	
	3496468	243	1	1.70887707		3496516	291	1	3.4720526	
	3496487	262	9	15.37989363	0.000686549	3496524	299	1	3.4720526	
	3496498	273	4	6.835508278	0.039304374	3496552	327	3	10.4161578	
	3496514	289	2	3.417754139						
	3496534	309	1	1.70887707						
SPBP8B7.06	3643653	70	11	18.79764776	0.000965915	3643653	70	1	3.4720526	
						3643754	171	2	6.9441052	

Table App.C.4 part 2

	Cycling strand-specific					Quiescent 7d				
Gene Primary Name	CS position	Distance from stop codon	Number of hits	Hit score	p-value Fisher, left	CS position	Distance from stop codon	Number of Hits	Hit score	p-value Fisher, right
SPCC1235.01	176375	217	2	3.417754139						
	176467	309	1	1.70887707						
SPCC1235.06	190213	122	2	3.417754139		190266	175	6	20.8323156	0.029103011
						190370	279	1	3.4720526	
SPCC1682.14	401425	78	2	3.417754139		401349	2	1	3.4720526	
	401445	98	2	3.417754139		401377	30	7	24.3043682	0.016137837
						401445	98	4	13.8882104	
SPCC14G10.04	751146	383	4	6.835508278		751146	383	1	3.4720526	
	751161	398	1	1.70887707		751161	398	1	3.4720526	
	751604	841	2	3.417754139						
SPCC1020.06c	774655	190	5	8.544385348	0.066039633	774655	190	1	3.4720526	
	774782	317	2	3.417754139		774721	256	1	3.4720526	
	774821	356	1	1.70887707		774782	317	3	10.4161578	
SPCC584.04	1537639	116	3	5.126631209		1537639	116	2	6.9441052	
	1537766	243	1	1.70887707		1537801	278	3	10.4161578	
	1537781	258	2	3.417754139		1537857	334	2	6.9441052	
	1537899	376	3	5.126631209	0.088284678					
SPAP8A3.07c	5326542	177	22	37.59529553	1.83E-08	5326595	124	1	3.4720526	
	5326609	110	1	1.70887707		5326609	110	1	3.4720526	
SPAC29A4.15	5117274	251	1	1.70887707		5117278	247	1	3.4720526	
	5117372	153	1	1.70887707		5117383	142	2	6.9441052	
	5117395	130	7	11.96213949	0.00346648	5117391	134	3	10.4161578	
	5117471	54	13	22.2154019	2.69E-05	5117475	50	1	3.4720526	
SPAC27F1.02c	4319255	404	1	1.70887707		4319272	387	1	3.4720526	
	4319476	183	2	3.417754139		4319335	324	2	6.9441052	
	4319493	166	1	1.70887707						
SPAC2C4.11c	4278459	269	3	5.126631209	0.088284678	4278487	241	3	10.4161578	
	4278527	201	1	1.70887707						
SPAC25B8.12c	4176631	389	7	11.96213949	0.00346648	4176597	423	1	3.4720526	
	4176749	271	7	11.96213949	0.00346648	4176637	383	1	3.4720526	
	4176757	263	2	3.417754139		4176659	361	6	20.8323156	0.029103011
SPAC26H5.10c	4141262	477	1	1.70887707		4141267	472	2	6.9441052	
	4141300	439	24	41.01304967	2.00E-06	4141300	439	3	10.4161578	
	4141318	421	6	10.25326242	0.007788299	4141339	400	1	3.4720526	
	4141339	400	6	10.25326242	0.033719209	4141426	313	1	3.4720526	
	4141382	357	2	3.417754139		4141450	289	1	3.4720526	
						4141486	253	1	3.4720526	
						4141523	216	1	3.4720526	
						4141533	206	1	3.4720526	
						4141657	82	1	3.4720526	
SPAC6G9.09c	3261386	84	2	3.417754139		3261425	45	20	69.441052	7.52E-06
	3261431	39	1	1.70887707						
SPAC24C9.12c	3069318	295	38	64.93732864		3069318	295	104	361.09347	8.75E-06
	3069434	179	2	3.417754139		3069400	213	5	17.360263	0.052481593
	3069506	107	1	1.70887707		3069452	161	2	6.9441052	
	3069538	75	21	35.88641846	4.11E-08	3069534	79	2	6.9441052	
SPAC1805.11c	2792651	365	1	1.70887707		2792825	191	1	3.4720526	
	2792721	295	1	1.70887707		2792833	183	1	3.4720526	
	2792749	267	1	1.70887707		2792843	173	1	3.4720526	
	2792837	179	39	66.64620571	1.87E-14	2792883	133	1	3.4720526	
	2792858	158	2	3.417754139		2793005	11	1	3.4720526	
	2792886	130	1	1.70887707						
	2792901	115	1	1.70887707						
	2792924	92	1	1.70887707						
	2792986	30	1	1.70887707						
SPAC4G9.06c	2261006	956	5	8.544385348	0.017496854	2261185	777	11	38.1925786	0.001524895
	2261014	948	4	6.835508278	0.039304374					
SPAC1002.09c	1811844	288	2	3.417754139		1811973	159	11	38.1925786	0.001524895
	1811889	243	1	1.70887707						
	1811969	163	1	1.70887707						
	1812046	86	4	6.835508278	0.039304374					
SPAP27G11.10c	1624751	324	3	5.126631209	0.088284678					
	1624853	222	7	24.30436819	0.00346648					
SPAC10F6.13c	1230604	183	2	3.417754139		1230564	223	1	3.4720526	
	1230714	73	3	5.126631209	0.088284678	1230604	183	2	6.9441052	
	1230731	56	2	3.417754139		1230614	173	1	3.4720526	

Table App.C.4 part 3

	Cycling strand-specific					Quiescent 7d				
Gene Primary Name	CS position	Distance from stop codon	Number of hits	Hit score	p-value Fisher, left	CS position	Distance from stop codon	Number of Hits	Hit score	p-value Fisher, right
SPAC30D11.12	1096761	128	1	1.70887707		1096729	160	5	17.360263	0.052481593
	1096808	81	1	1.70887707		1096761	128	1	3.4720526	
	1096823	66	2	3.417754139		1096857	32	3	10.4161578	
	1096834	55	3	5.126631209	0.088284678					
	1096847	42	1	1.70887707						
	1096861	28	9	15.37989363	0.000686549					
SPAC23C4.09c	1042977	242	2	3.417754139		1043176	43	3	10.4161578	
SPAC821.10c	1002806	192	2	3.417754139		1002834	164	2	6.9441052	
	1002825	173	1	1.70887707		1002898	100	10	34.720526	0.002750594
	1002875	123	1	1.70887707		1002908	90	5	17.360263	0.052481593
						1002946	52	1	3.4720526	
						1002965	33	1	3.4720526	
SPAC222.04c	950073	224	1	1.70887707		950073	224	1	3.4720526	
	950127	170	2	3.417754139						
SPAC2F7.13c	558978	338	3	5.126631209	0.088284678	559123	193	4	13.8882104	0.094635176
	559055	261	2	3.417754139						
	559119	197	1	1.70887707						
SPBC1289.05c	4392037	53	2	3.417754139		4391989	101	3	10.4161578	
	4392046	44	2	3.417754139						
SPBC1539.07c	4373296	378	3	5.126631209	0.088284678	4373369	305	4	13.8882104	0.094635176
	4373632	42	1	1.70887707		4373393	281	1	3.4720526	
						4373408	266	4	13.8882104	0.094635176
SPBC14F5.04c	4157753	418	7	11.96213949	0.00346648	4157836	335	2	6.9441052	
	4157819	352	22	37.59529553	1.83E-08	4157932	239	1	3.4720526	
	4157836	335	1	1.70887707		4158045	126	3	10.4161578	
	4157878	293	1	1.70887707		4158113	58	1	3.4720526	
	4157928	243	1	1.70887707						
	4158090	81	1	1.70887707						
SPBC56F2.02	4123072	66	1	1.70887707		4123072	66	6	20.8323156	
	4123084	54	9	15.37989363		4123084	54	8	27.7764208	
	4123099	39	1	1.70887707		4123093	45	1	3.4720526	
	4123113	25	12	20.50652483	0.001916924	4123113	25	2	6.9441052	
SPBC215.09c	4045834	144	2	3.417754139		4045812	166	2	6.9441052	
	4045844	134	26	44.43080381	7.14E-10	4045834	144	1	3.4720526	
						4045842	136	2	6.9441052	
SPBC13G1.06c	3736601	478	2	3.417754139		3736646	433	17	59.0248942	4.42E-05
	3736654	425	1	1.70887707		3736660	419	1	3.4720526	
SPBC2G2.03c	3440963	158	1	1.70887707		3441013	108	12	41.6646312	
	3441013	108	10	17.0887707						
	3441033	88	1	1.70887707						
	3441054	67	10	17.0887707	0.000305498					
	3441074	47	3	5.126631209	0.088284678					
SPBC4F6.18c	2726367	935	2	3.417754139		2726352	950	219	760.379519	2.39E-57
	2726382	920	9	15.37989363	0.000686549	2726577	725	2	6.9441052	
	2727202	100	5	8.544385348		2726643	659	1	3.4720526	
						2727202	100	8	27.7764208	
						2727224	78	23	79.8572098	1.28E-06
SPBC19G7.03c	2348981	96	168	287.0913477	1.03E-31	2348981	96	31	107.633631	
	2348990	87	18	30.75978725		2348990	87	20	69.441052	
	2349028	49	1	1.70887707						
SPBC17G9.11c	2190726	147	8	13.67101656	0.00154276	2190641	232	2	6.9441052	
	2190807	66	1	1.70887707						
SPBC1306.01c	1177865	214	2	3.417754139		1177929	150	5	17.360263	
	1177929	150	1	1.70887707		1177938	141	1	3.4720526	
SPBC839.05c	603533	170	2	3.417754139		603579	124	1	3.4720526	
	603663	40	8	13.67101656	0.00154276	603672	31	55	190.962893	
	603672	31	47	80.31722227						
SPBC106.07c	386248	257	1	1.70887707		385991	514	2	6.9441052	
	386256	249	2	3.417754139						
SPBC1271.12	341980	428	8	13.67101656	0.00154276	341976	432	11	38.1925786	0.001524895
						341984	424	1	3.4720526	
SPCC70.05c	2358531	546	5	8.544385348	0.066039633	2358490	587	2	6.9441052	
						2358531	546	1	3.4720526	
						2358775	302	1	3.4720526	

Table App.C.4 part 4

	Cycling strand-specific					Quiescent 7d				
Gene Primary Name	CS position	Distance from stop codon	Number of hits	Hit score	p-value Fisher, left	CS position	Distance from stop codon	Number of Hits	Hit score	p-value Fisher, right
	2350384	343	2	3.417754139		2350523	204	1	3.4720526	
	2350428	299	1	1.70887707						
SPCC70.03c	2350475	252	1	1.70887707						
	2136098	102	8	13.67101656	0.008392449	2136098	102	1	3.4720526	
SPCC126.11c	2136114	86	3	5.126631209	0.088284678					
	2129816	224	2	3.417754139		2129874	166	1	3.4720526	
SPCC126.08c	2129874	166	1	1.70887707						
	2114482	833	1	1.70887707		2114482	833	1	3.4720526	
SPCC126.01c	2115142	173	3	5.126631209	0.088284678					
	1419119	337	3	5.126631209	0.088284678	1419207	249	2	6.9441052	
SPCC622.12c	1419438	18	1	1.70887707		1419323	133	2	6.9441052	
SPCC338.14	1349597	464	4	6.835508278	0.039304374	1349844	217	2	6.9441052	
	1349615	446	1	1.70887707		1349966	95	1	3.4720526	
	1349844	217	1	1.70887707						
	626914	297	1	1.70887707		626939	272	1	3.4720526	
	626946	265	1	1.70887707		626946	265	1	3.4720526	
	627022	189	1	1.70887707		627022	189	1	3.4720526	
	627043	168	13	22.2154019	0.000221068	627034	177	2	6.9441052	
	627062	149	3	5.126631209	0.088284678	627043	168	1	3.4720526	
	627124	87	1	1.70887707						
	627165	46	2	3.417754139						
SPCC31H12.04c	627172	39	3	5.126631209	0.088284678					
	611829	70	1	1.70887707		611832	67	1	3.4720526	
	611843	56	123	210.1918796	3.48E-32	611843	56	10	34.720526	
SPCC1183.08c						611850	49	2	6.9441052	
	331218	806	1	1.70887707		331596	428	1	3.4720526	
	331596	428	1	1.70887707		331635	389	1	3.4720526	
SPCC736.10c	331635	389	5	8.544385348	0.066039633					
SPCC794.12c	275213	300	2	3.417754139		275507	6	1	3.4720526	
	275511	2	1	1.70887707						

Table App.C.4 part 5

Appendix D

miRNA Supplementary Material

D.1 All miRNA and target candidates

In the following section a complete list of potential miRNA genes and targets is provided, with the RNA sequence of the predicted pre-miRNA hairpin. Localisation of the potential mature miRNA is indicated by lower case letters of the sequence. The predicted folding by *RNAfold* (Hofacker et al., 1994) and MapMi (Guerra-Assunção and Enright, 2010) is presented in "[(." notation. The predicted negative folding energy is presented in brackets after the folding structure.

D.1.1 miRNA targets identified from APA of meiotic cells compared to cycling cells

D.1.1.1 sRNA data for Yamanaka et al. (2013)

potential miRNA gene = SPCC330.11 potential miRNA targets = SPBC365.06 SPCC16C4.16c

Ucggaaaagagcggacaaaaacu CCGUAUCAACCAAAAAGAUCGAUACGGACGCACCGUGCUGCACAUCGCCGUUUCUGA

(((((((.(.((((.....((((((((.....)))))))).(((.....)))...)))))..))))))) (-20.40)

potential miRNA gene = SPBC1734.01c potential miRNA targets = SPBC365.06

UUGGAACAAACACAAGACGGUAAGCCUGAGCUUAAAAUCAAAAAGCGUAAAGCUGAAAAAGgagaucaacgucaagaacuugaUCGCAUUGUCAAAUCUAUUAAACGUUC

UGG

..(((((.....((...(((((....)))))...)).....((.....))(((.....(.(((((((.((.....)))))))).).....))).....))))).. (-15.20)

potential miRNA gene = SPAC1527.02 potential miRNA targets = SPBC31F10.08 SPCC126.11c

UUAACACCGACGUGCUUUGGGACUUGAAUUAGUAAACCAAAGCUACAUUAUCUUUAGAAAAAUggaaggaguacggucugaUUUUUAUUGUUAUAGAAUUGAAAUGUUAA

((((((.....(((((((((.(((.....))).....)))))))).)).(((..((((....(((((((((.((.....)).)))))))))....)))))..))))..)))))) (-16.50)

potential miRNA gene = SPBC405.07 potential miRNA targets = SPAC806.06c

GCGCAGUAAGUUUUCCCCUACACCAAUCUCACAauggcaccuggucugguuGUUGGUUUAAACAAAGGAAAGGUUUUGACCAAGCGUCAACUUCCUGAGC

((.(((.((((...........(((.........))).((((..(((..(((((......))))).))).))))..((((.....)))))))).))).)) (-18.90)

potential miRNA gene = SPCC1235.16 potential miRNA targets = SPAC4F10.18 SPAC13G6.07c

UUGCAUUCGUGAUUUAAAAUGGAAAGGaaaucccaaguuucagauACAAAUAAUAACUCUAUACCUACGUAAGUUUAUUAAUUAUUAGUUUAUUUCUUUACGCCGGA

CUGCGA

((((((((((........)))))...(((((.....)))))......................((..((((((.....(((........)))....))))))..))..))))) (-10.90)

potential miRNA gene = SPBC16E9.11c potential miRNA targets = SPAC29B12.07

UGAUUAUCCUUCAAAAGAUACACUGCAUGAAAAAUUGAGCCUAGCCGUUGAAAACACGGUaggguuuggaaacgaguaAUGACAACAAUCUUGAAAUUUCAUAAUCA

(((((((.......(((((....(((((......((((((((.(((((.......))))).))))))))).........))).))...)))))........))))))) (-19.30)

potential miRNA gene = SPAC26A3.14c potential miRNA targets = SPBC1921.07c

CAGCUGUCGAAAAGUAUCUAAACUUUGGugaaugggccuuugaucaAUCUUCUGCCUUCUUAGGUUCUUCAACAUGGUUCAAAAAGGUAUAAAGACAUUG

(((.((((.....((((((.((((.((.((((.((((((..((..............))..)))))))))).)).)))).....))))))...))))))) (-20.20)

potential miRNA gene = SPCC1322.10 potential miRNA targets = SPBC31F10.08

CACUUGCUGCUUAUGCCGUCUCUGUCACUUCUCCUACUCGugacaccaccuggcaaucCGGCCAAGUGAACACCGUCACCUGGUCUUCUGUCAGCACUGAUCCCGA

GGAGAUGGUUAUUAUGUUGGUG

((((.((.......(((((((((((((...........))))))....((((((....(((((.((((......)))).)))))...))))))............))))))))......)).)))) (-38.20)

potential miRNA gene = SPAC3G6.09c potential miRNA targets = SPCC1223.06

AUUAUCACCUUCCUGCUccucgugauuuaggagaacUUUUGCAUAACCUAUGAAAUAACCGGGAUAAU

..........((((((((((........)))))........((((...))))).......))))).... (-9.50)

potential miRNA gene = SPAC17H9.10c potential miRNA targets = SPBC31F10.08

UGGUGGUCAUGAGGGUACAAAGCUUGAUAUUUCUGUUCAAGAUCUAAAAAGUAUUAUUGAAaaucuugagaaacuacauccUUAACGUUUUCUAUCAAUUCUGUUU

UUGCCGCUA

(((((((...((.((.......(((((.........)))))................((((((.(((((..........))))).))))))........)).))...))))))) (-17.50)

potential miRNA gene = SPBC1709.06 potential miRNA targets = SPCC645.10

AGUAUGUCUAGCUCCCAugguuaggauuggggaacuuCCCAUGAGGUUAUUAGCACUUCGAUAUGGCGCUAAUCUAGUAUGGGGGCCAGAAAUUGUGGAUAAAGCU

......(((.((((((((.(..(((((((.(......(((((((((.......)))))...))))).)))))))).))))))))).)))................ (-30.20)

potential miRNA gene = SPAP27G11.12 potential miRNA targets = SPCC1442.06

UUUUUUUGUCCAAUUUCUCAACCugccguagguuuggggaaGAACCAAUACAUUAUUAACCAAAUUACAGGUUAGACUUUAGGAACGCGGUUUUACGACUACACGAAGAAA

(((((((((....(((((((((((...)))).))))))))(((((......(((((((((........))))))....)))......)))))).........))))))))) (-19.80)

potential miRNA gene = SPBC1289.09 potential miRNA targets = SPCC1235.01 SPAC1805.11c

GCCUUUAUGCCGUUUAGCCUCGUCACUAUAUACAUCUUCUUUGCAUCUUUUUGUAUCAACUAGAUGAUUcguacuauucaaaggcagagcgcUACAAGGU

(((((...((.((((.((((............(((((....((((......))))......)))))..............)))).))))))...))))) (-18.30)

potential miRNA gene = SPAC869.02c potential miRNA targets = SPAC6G9.11

UUCUUGAUUGAUUUAAUAUAUAUAUGGCAUAACCGUUUGGAAUUAGAGAUUCCAAAGUAAAAUGUCCAGUguugagguuaaucgugaa

(((.(((((((((((((((.....((((((.....(((((((((...)))))))))).....))).)))))))))).)))))))).))) (-22.30)

potential miRNA gene = SPAP7G5.04c potential miRNA targets = SPBC16G5.11c

GAAGCUACUUUGGAGAAGCUUGCAUCGGCUGGUggucguggaggcgcuccaACUCAUUAGAUUUAUUAAGCAGUUAAAUUAUCCAAAAUGAAAAAUUUAGAAUCUAU

GUAGACUUC

((((((((.((((((..((((.((.(((((...))))))).)))).)))))).(((((.(((...((((....))))...)))...)))))................)))).)))) (-27.50)

potential miRNA gene = SPAC8E11.03c potential miRNA targets = SPBC839.05c

GUUCAAUAUAAAAAUAAAGUGCAAUGGAAGAAUUCgcagaggggaaugaugaugaaC

(((((.(((..........((((((......))).)))..........))).))))) (-6.80)

potential miRNA gene = SPAC1002.12c potential miRNA targets = SPBC365.06

GGUGGAGUCAAAGAAAGUGGAUUUGGCCGUGAAGGCAGCAAGUACGGUAUUUCUGAAUAUCUCGAUAucaaguccuuaacuaucucGACCCUCU

((.((.(((..(((.(((((((((((.((.(((..(((.(((((...))))))))..).)).))...))))))))...))).))).)))))).)) (-22.10)

potential miRNA gene = SPBC1271.09 potential miRNA targets = SPCC306.04c

GCAUUGAGUAUAUCAAGCUUAUUAAUAAUUUCUUAGCCCUuuccugagugcgaccuuuUCUUUUCAUCAUAUCAUUUUAUGCUAAUCUUUUCUUUUUUAUUGUACCCU

UAUUAAAUACAAAUCAUGC

(((((.((........(((...............)))......)).)))))................((((......)))).................(((((...........)))))....... (-9.20)

potential miRNA gene = SPCC1442.05c potential miRNA targets = SPAC1805.11c

GCAACGAAUGUUGGUUUUGGAACAAACUCAAAAGAAUAUACAAAAUGCCCAGAAGUCCAUGACUUCUGcaguugagggaguccaaaAGCAUUAUGAAAACGC

((.....(((.((.(((((((.....(((((..................((((((((...))))))))...)))))....))))))).)).)))......)) (-23.60)

potential miRNA gene = SPAC4D7.08c potential miRNA targets = SPBC11G11.01

AGCUUUGCAGAAGAUGAGgaacgugaagcucccgaagacaUUUCUCUCCAUAACACACAUUCAGAUGUUACUUUUGAUUUUGUUUAAUUCCCCAAUUUAUUCAAUGCU

(((.(((...(((.((.((((....((((...((((((((((...................)))))))...)))).....))))..)))).)).)))...))).))) (-11.50)

potential miRNA gene = SPAC1039.04 potential miRNA targets = SPAC12G12.04

AUCUAUAUCAuccauuucaaaggcgucaAUGAACUCAGAUGAAAAGUUAGACAAGAAGGAAUGGGAUCAGCAGCUGCCUAUUGAUUUUGGAGAAGGU
.....((.(((((..((((((((((...(((.((((.....)))).)))...)).)))).)))))...)))))).... (-18.40)
potential miRNA gene = SPCC757.02c potential miRNA targets = SPAC5H10.10
CAGAAAGAGAAAAACUCAAUGCUCAUGCUUUAGAAGUUGggacuugggcguuuugugaUUUUUUGUUCGGCAGGACGUACAACGUUAUAUCCUCAAUGAGCAAAGC
CAGGAAGCUUGGUUACACUG
.....(((((..((..(((((((.((((.....))).).)))))))..))..)))))(((((((..(((((.(((.....))).)))))...)))))).(((((((...)))))))..... (-29.30)
potential miRNA gene = SPBC839.12 potential miRNA targets = SPCC306.04c
GUUAAACUCUCUGUCACACUAGUUCGAACGUUUGAACCCCAUUCUUACAAUCAUCACAAUGCGACguggagcucguggaucuacaaaUUUGGGAGGAAUGACAUGGCA
AGAAGUACUUGAGUUUAAU
(((((((((((((((.....((((((....)))))))..(((((.....(.((((.........)))).)(((.(((((.....))))).)))))))...)))))..)))).....))))))))) (-23.60)
potential miRNA gene = SPBC1348.07 potential miRNA targets = SPBC11G11.01
AUGAGUUAUUUAUGGUAAGAGAUACUCGUAAUGUGGACcuggagcggggacuugaaUUGUGUAAGCCUGAAAAGGUAAACAAACAAAAUCUCUUUACCAACAUCAU
.....((((((((((....((..(((..((((....((((..(((.....)))))))))...))))..)))..))...)))))).)))))..... (-23.50)
potential miRNA gene = SPCC364.01 potential miRNA targets = SPAC19B12.05c
GGGGAACUAAAUGAUGUUGAGUUAAAAAUCCCUGCAAUUUCUCAUUCUUUAAAAUUAAaggaugguagcauucuuguuCUCC
(((((((.....(((((((.....((((((((((....)))))))))))))))...))))))) (-18.20)
potential miRNA gene = SPAC1A6.01c potential miRNA targets = SPAC23C4.08 SPBC557.05
ACACUCUAGUCAUGUCGCCGUGCCUGAAUCUAUUCUUAAUAAAAUAAAUCAGAAGUGGUCUAAaguucaggaugaugauggCAUGCCAUCGAUGUUAUAAGUAUGU
(((..((.((.((((((.(((((...(((.((((((.(((....((.((((....)))).)).)).)))))))))...)))))))....)))))).)).))..))) (-20.10)
potential miRNA gene = SPBPB8B6.03 potential miRNA targets = SPBC582.03
UGUUGGCUCUUUUUUUCCAACUUGGAGUAAAGAACCCUUGAUUCAUUAUUUUAAAGUAUUAUGGAUACuccgacuugggaaguggugCCGCA
(((.((((.((.(((((((.(((((((.....(((((.((.....)).)))))))))))))).)))))))).)).))))))) (-24.90)
potential miRNA gene = SPCC830.10 potential miRNA targets = SPCC1223.02
GACAGGAUAUCAAGAUGACAAGUAAGUCAAUUUUGGAAGGAUUCUUCAUAAUUAAUUAAUGUgcgguaucugcaucuucaauAUUUUUCAGCGGAAUGAUAACUAA
CAAGAUUGCUGUC
(((((.....(((((((((.....))).)))))))(((....)))).....((((((...))))))......((((.(((....))).)))).....)))))) (-21.50)
potential miRNA gene = SPBC1348.07 potential miRNA targets = SPBC11G11.01
AUGAGUUAUUUAUGGUAAGAGAUACUCGUAAUGUGGACcuggagcggggacuugaaUUGUGUAAGCCUGAAAAGGUAAACAAACAAAAUCUCUUUACCAACAUCGU
.....((((((((((....((..(((..((((....((((..(((.....)))))))))...))))..)))..))...)))))).)))))..... (-23.50)
potential miRNA gene = SPBC342.04 potential miRNA targets = SPBC3B8.05
GUUCCAGGAACAAAGCUUUUACGUGCUGACCCAGAAAAaggguaugcucgcauauuaUGUAUGAUAGCUAACAAAUCACAGUUAUAUAGUCAUGAAUCGGGAC
(((((.................((.((..((((.....))))..)).))....((((((.(((((((((.....)))))).)).)))))...))))) (-20.30)
potential miRNA gene = SPAC4H3.03c potential miRNA targets = SPBC21B10.03c
UUCUGAAGAGUGUUCACUCUCAAGUGAAUCAUUGGGuaacaguccgcaagcguuuAGCUCUAUCGCUGCAAUUGCUGCUGCUCAUAUUUUGGACAAGGCCCUUUCUA
UUUUCAAUGAA
...(((((((((((((....)))))))..))).(((((....(((((((((((((((.....)))))).)).)))))..)))).........)))))).....)))))..... (-25.60)
potential miRNA gene = SPAC27D7.06 potential miRNA targets = SPAPB17E12.13 SPBP4G3.02
CAACGAAGUAUUCUAACAUGCUAAGGACUACuagccggacauuuucuagggcUUUGUCUUCUUCUAACUUUAAAAUAAAUUGUGGAAGAAGGCAUUGGUUUAGUGUU
UUGACGUUG
(((((.(((((.....)))))..((.........))(((((....(((((.(..((((((((((...((.....)).....)))))))))))..).)))))))))...))))) (-27.50)
potential miRNA gene = SPBC1711.18 potential miRNA targets = SPAC806.06c
GAUUGUUUGAGAAggcgggcagguugucuguAAGGUCACAAACUGUGCAAACCUUUCAACAAAGGAUC
((((.((((.((((((.(.((((..(((((....)).)))..)))).)...))))))...)))))).)))) (-18.20)
potential miRNA gene = SPCC613.06 potential miRNA targets = SPAC4F10.18
ACAUGGGCcgcgauauuuacaaggacgaGACAUUGACGAUCCCUGAGGGCGUUUCAGUUGACAUCAAGGCUCGUUUGGUGACUGUAAAAGGCCCUCGU
((..(((((...(((.((((..(((((((.(.((((.(.((.(((((.....))))))..))).)))).)))))))).)))).)))....))))))..)) (-29.70)
potential miRNA gene = SPBC2G2.17c potential miRNA targets = SPAC1834.01
CCAUGGUUCCACCGCUgacagcagugguuguacagucuCUGUCACCGGAGGCAAUGCCUACUUUGAGUUGUACUAAAUUGCGUGUUAAAAGUGGAUGCUUGG
(((.(..(((((...((((((((((....((((((.(((.........((((...))))).....)))))))))...))))))..))))))...))))..).))) (-32.20)
potential miRNA gene = SPBC1348.07 potential miRNA targets = SPBC11G11.01
AUGAGUUAUUUAUGGUAAGAGAUACUCGUAAUGUGGACcuggagcggggacuugaaUUGUGUAAGCCUGAAAAGGUAAACAAACAAAAUCUCUUUACCAACAUCAU
.....((((((((((....((..(((..((((....((((..(((.....)))))))))...))))..)))..))...)))))).)))))..... (-23.50)
potential miRNA gene = SPBC1773.02c potential miRNA targets = SPBC31F10.08
GUUAGGCGCUGAGAAGCCUGGUGGUGGAAAGCuguuuagaagccauuggaucuuugAGAAAGGUACUGGCAAAUGCAUAGUGAAGGAAAUUGACAUAUCACCUCUUG
UCAGUGUUGAC
(((((.(((((((.(((...(((((((....(((((.((...(((((....((((((...)))))).)))))...)).)))))).....)))))))).)))))))))))) (-32.00)

potential miRNA gene = SPBC16E9.11c potential miRNA targets = SPAC29B12.07

UUUCUUCGAAAAUUCUAAAGUAGGAGAAGUAUAAGGAGACGAAUAUAUAGAUGCAGAAUCUGAUGAAUUUaggguuuggaaacgagcaUGCUUGGGAAAUCGUGUA

GGGGAAGGAGUGCGAGAAGAAA

(((((((....(((((....((.(.((....((((.(.........(((((.....)))))).........(((((....)))))).).))))).....)).).))......))))))....))))))) (-18.30)

potential miRNA gene = SPAC1B3.08 potential miRNA targets = SPBC557.05

CCAUUAAUUCAGCUGUUGCGCGAUCUAAUAGUGUCCUACUAGCGAAAAAUUUAUCaauuccuuguggggaaaacgCUAACCUCAGUGCUGAACUUUGAUAUUCCUGG

(((..(((((((..(((((((.......(((((((((.((((.(((.(.....))).)))))))...)))))).....))))..))).)))).)))..))) (-17.70)

potential miRNA gene = SPBC1683.08 potential miRNA targets = SPAC31G5.16c

GUUAUGUUAAUUUUCGUCUCCAUGGCCGGUUGGAUGUUuggugcugauaccgguucCAUUGGUGGUGUUACUAGUAUGCGCGAUUUUCGUGAGCGUUAUGCUGAUA

GAUAUGAC

(((((((.............(((.(((((((.(((...(((((((....)))))).)))))))))).)))....((((((((((.....)))).)))))))...)))))))) (-28.90)

potential miRNA gene = SPBC691.01 potential miRNA targets = SPBC365.06

AAUAUACUAAACGCUUUGUUGGAGUUCAUGUCUACAAAuuuagaagagaaacagaggAGAUUUUCUAAGCUUCAAAAGACUUUGGGUGUU

.((((.(((((..((((..(((((((....(((((.....))))(((((((.(.....).)))))))).)))))))))..)))))))). (-13.60)

potential miRNA gene = SPAC24B11.10c potential miRNA targets = SPAC20G8.08c

UGGGCGCAUAAAGCGGCUUGCAAAGGUUUGGCUAAAGCUCAGUAUGCUGUUGGCUUUAUGAUGGAACAAGgcauugguguugcugcugAUCCAUCCUCUGCUCA

(((((((((((((((.(((...(((((.....))))))..))).)))....)))))))))(((((.((.(((((.......))))..)).))))....)))))) (-29.90)

potential miRNA gene = SPBC1709.05 potential miRNA targets = SPCC1235.02

CAAUAACCUACUUCUCUCAAAAAAAUCGAGGGUCAGUUUUAACUUCUUGUUGCUUGAAAaugagugaaguuuaugaagguGCUAUUG

((((((((((((.(((((.....(((((.(.(((...........))).).)))))...)))))..))))).....))))..)))))) (-16.10)

potential miRNA gene = SPBC16G5.09 potential miRNA targets = SPCC191.07

CAUGAUUUCGAAGUUACUGAAAAUAGUCCUUUUAacugcagguguuauuggcaACACACUUGCUGAUUCUAGAAUUCAUG

(((((.(((((((..(((((....)))))..))))...(.(((((((((.....)))))))))..)).....))).)))))) (-14.90)

potential miRNA gene = SPAC4G9.09c potential miRNA targets = SPBC16A3.19

GACAAUUUACUUAAAGGUGCUGCUACUCAAGCUCUCCAGAAUCUCAAUCUGUCCUGUGGUUACGAUGAauaugccgguauccauuuggauUGAUUAUAUUUUUUGUA

GCCCAUUCUUUUGUU

(((((.....((.(((.....))))).)).((((.....))))...((.(((((((((.(((((((.....((((....))))....)))))))).)))))))))).....)))))) (-24.40)

potential miRNA gene = SPCC645.13 potential miRNA targets = SPBC365.06

GUUGUUCGGUGUGUUUGUAAAUCGCAAGAAGACAUUggcgauacguggguucaauGUGAUGGGUGUGAUUGCUGGCAACAUGCCAGUUGCGUUGGAUUAGCGGAUAAG

GAUAUACCUGAAAGC

....((((((((((....(((((.....)))))).(((.((((((((((((.((((((.(((....)))))))))..)))))))))))))).)))).....))))))))).)))... (-37.10)

potential miRNA gene = SPAC4G8.04 potential miRNA targets = SPBC365.06

UCUUCCCCUUAUUCAUUUGUUUUGGAaugcaaccaucuguuucugaCUCUUCAGUUUACCUUUCUUCCGCUCCCACGAAGCCCAUUGCUAGUGGAAGUGUCGCUACA

GGGAUAGA

(((..(((((....(((((.....)))))).....((((....)))).....((((.((((..(((((.....))).)))))..)))))))))).....))))..))) (-23.00)

potential miRNA gene = SPBC405.07 potential miRNA targets = SPAC806.06c

AUGCUACCAAAAGUUUUUUCCAGACGGUAGGUACAGUGUAGACUUUACACGauggcaccuggacuggguuGUUGGUUUAAACAAGGGAAAGACUUUGACUAAGCGU

(((((....((((((((((...(((((((((.(((((((.....))))))..))..))))).))))).(((((.....)))))))))))))))).....))))) (-28.10)

potential miRNA gene = SPAC13F5.06c potential miRNA targets = SPBC577.12

GCAGUCAUGCAAGACAAGCGUAAAGCAGGUUUUAGUGAAAUGAACAUUCAUGAAGUUCUCCGUCUacguuuugauuugccugaaaACUUGAAGUUGCAACUGAACAA

AGAAGAAGCUUUAUUGC

....(((.(((((.(((((((((..(.(((.(((.(((((.....)))))).))).)).)...)))).)))).)))))).))).....(((((((....((.....))....)))))))).... (-21.50)

potential miRNA gene = SPBC1861.01c potential miRNA targets = SPBC1815.01

CAUGCGACCGACACCUUGGAAAAUAAAAGAaggggaauaggggauuuucCAAACGAACGAUGACAAUACUAAUAGUGUGUUAUGGAUUUCGAUGUACAAAAGUCA

UUUAGUCGUUAUG

(((((((((.....((((.............))))(((((.........)))))...(((((.(((((((.((((...))))))))))).)..)))))(((((.((....)))))...)))))..)))) (-22.50)

potential miRNA gene = SPBC713.10 potential miRNA targets = SPBC557.05

UUUGUUCAAAAUUACUAUAUUACAUUAAAAAAUGUCGcuuccaagggcugucggucGAUUUAUAAUUGUUGGUUCGCAAGGUAUGUUUGUGGCCAUUCUCCUAUCAAGGA

UUUAUUCAAA

.....(((((.(((.(((((((..((...)).)))).)).)))).)))).)))(((((.(((((((.....)))))))))).....((((....)))).....)))).)))) (-14.50)

potential miRNA gene = SPCC1795.03 potential miRNA targets = SPBC31F10.08 SPAC1B3.06c

AAGAGGCACUUACAUUCCAAUGACAACCCAAGaugcugcugcuaaggacguuGAUCAUAAGCAUUAAUUUAAGAUGUUCUGUAAUUCUCAAAGAAUGUAUUCUUUUU

.(((((....(((((((..(((((((((.....((....))).....)))).)))).)))))(((((.....)))))).............)))))))..)))))). (-16.40)

potential miRNA gene = SPAC2E12.02 potential miRNA targets = SPCC645.10

GGACAUAUAUAUUUUUGUUAAAUUAACGGCACCCAUAAUAAAUUGAUUUUUAAUGAuucagacugcaucgacaaUUUCAAGCAAUGGCGGAACGAAUCAGGGAAUG

GAGUCC

((((..........((((((...))))))).............((.((((((..((((((...(((((.....(........).....))))).....)))))))))))).)))))) (-14.10)

potential miRNA gene = SPBC1683.08 potential miRNA targets = SPAC31G5.16c

UCGUGGUCUUCAUUUCUAUGGCCGGAUGGCUAGGaggugcugauaccgguucCAUUAGUGGUAUCUUGGGAAUGCGUGAUUUUCAAUCUCGUUUUGCUGACCGG

UAUAAUCCUAUCACGA

(((((((...(((((((.((((....))))))))))))...(((((((((((....)))).....((.((.(((.((((...)))).))).)).)))))))))).....))))))) (-36.80)

potential miRNA gene = SPCC1235.16 potential miRNA targets = SPAC4F10.18 SPAC13G6.07c

GUUGGUACUUCUCAUAUCCGCCUCGCGUGGGGUCAUAAUACGCUUCCUGUAUCUGCCUUGUCUCAAUCUcaaucccaaguuucagauGAGGGCUUCGACAGAACUC

UUUCCGCUAAC

((((((.............(((((....)))))).............((((....(((((((((.((.((........)).)).)))))..)))))....)))).........)))))) (-21.00)

D.1.1.2 sRNA data for Halic and Moazed (2010)

potential miRNA gene = SPAC31A2.16 potential miRNA targets = SPCC126.15c SPCC126.11c

AAGGAGAAGAUUGAGGGGGAUGAGCUGUACGAUAAUAUUUugaaggaaacuuuuaaugaagaaUUACUCUCUCAUUACCCACCAAAUAUUAUUUAUGCGACCUUUCAA

AAAUAUUUGAGCUCCUU

.(((((....(((((((.....((.(((.((((((((((.(..((.......(((((((((.....))).)))))))))..).)))))))))))))..)))))))).....))))). (-24.90)

potential miRNA gene = SPAC29A4.14c potential miRNA targets = SPAC23C4.08

UGUAUAAGGAAUGCUUUUAUGGAAUUUAUGGAAACACGAAAUUCCAAGUCCAAguaaagugucuucuauuacuuaacGGAAUUGAAGCUAAUUAGAUGCA

(((((.......(((((..(((((((.((....)).)))))))..(((((((((..(....)..)))))))...)))...)))))).....))))) (-19.70)

potential miRNA gene = SPAC806.05 potential miRNA targets = SPAC806.06c

GUAACGGAUUUCGCCAAAACAAUUGCUCCUUUUAUUCAaaggaaaaaguugaaaauacaGUUAUAUUUUUGGAAAUGGCCGCGAGCAGGAAAUCUGUAGAACACUUAAU

ACUAAAAAAUCGGUUGC

(((((.(((((.((((...(((((..(((((((.....)))))))...)))))(((((.....)))))).....))))).....(((.....)))).....................)))))).)))))) (-20.80)

potential miRNA gene = SPCC645.11c potential miRNA targets = SPCC645.10

AUUUGCAAUGUUCCGCUUGUGGAACUCGUUUUUUUAAuaaugauaaugaaaacacaagaUGCUAUGUCAAAUUGAACUACUGUUAUCCAGACUGUACCGAAAUUACACCA

AGCACCAGCGAAU

((((((.........(((((((....(((((.((........)).)))))...)))))))..((((..((..(((((....(((.(((.....))).)))....)))))..))..))))...)))))) (-18.60)

potential miRNA gene = SPBC2G2.02 potential miRNA targets = SPCC191.02c

AGUGCUGACCCAGCCAAACGUCGUGAAUGGGAGUCAUGUGUCAAACGCUUAUUAAACGGUAAAUGuacaucuggaccuggauauGUUCAAUUACGUUCUGGACAACU

GGUUGGUACAGCACU

(((((((..(((((((...(((..(((((..(...(((((((.....((.....((((.....)))).....))).....)))))))....)..)))))..)))...)))))))..))))))) (-35.10)

potential miRNA gene = SPCP1E11.09c potential miRNA targets = SPAC1093.02 SPAC4F10.18 SPBC409.18 SPCC613.06

UGCCGAAGAAGAAAAGAAGGAAGAAGCCAAGGAggaggaagagucugaugaggacAUGGGUUUUGGCUUGUUUGACUAGAUCAAUUAUACGAAUGAGUAUCUUUGU

GCCUUAAUUUACGUCGCA

(((.((.....(((..(((((((.(((((((((.....(.((((.....)))))).)...)))))))))).))).....(((..(((((.....))))..))))..))))...))))) (-20.10)

potential miRNA gene = SPCC1902.02 potential miRNA targets = SPBC365.06

GGUGGUGUCUCUUGUGACuuucuuuccaaaauucagaaaaAGCAAUGGGAAACAGGGGUUGGUCAUAUGUGUUUCUACCCAUUGUCUAUUAUAAAUGACGAACC

(((..((((..(((((((.(((((.....)))))..(((((((((((((.(.....).)))))...)))))))))...)))))..)))).))) (-20.50)

potential miRNA gene = SPAC56E4.06c potential miRNA targets = SPAC1F7.12

CCAUCUUCUUAAAGAUCCAAUuuggcagccuauccuuaugccaaaUGGGAAAGUACUUAGGGCUGGCGAUAAAAUGUUUCGUCCGGCAUAUGCUAAGACUUUGG

...((((....))))(((.(((((((.....)))))))).)))((((.(((((.(((((((.....))).)))))).....)))))))... (-29.40)

potential miRNA gene = SPCC576.16c potential miRNA targets = SPBC21H7.07c

GCCUAAUGAUUUUUAUUUCUUUAUGGUguaagugacacgauaggacuauUCUUACGUUCAUUCACUAACUCAUUAGGC

(((((((((.............((((((.(((((((((....)))))))))))))))))..))))))))) (-20.90)

potential miRNA gene = SPCC5E4.06 potential miRNA targets = SPAC23C4.08

GAUAUUGAUGGAAAGAUAUCUUCAAAAGAAGAAGugauugggagagcuaaaggugaaACUGACACUACAAAGUCUAAAUUUGAAGAUAUUGUGAAAACAUUUG

AUGGGUAUC

((((((.((.(((.(((((((((((...(((((((.(((((....(((...)))....)))))))).....)))...))))))))))))(((....)).))).)).)))))) (-22.60)

potential miRNA gene = SPBC3B9.17 potential miRNA targets = SPAC806.06c

GUGAUGUGGUUAAGAAGACUAUAAGUUUUAAUUAAGuauuuccuggguuauugacuaaaauAUCUAGGGUAUUUGUGCGUGGAAAAGCCAGGGUUGUUGCUGACAACAUUUC

AUUAC

(((((((.((((.((((((...))))))..))))).))...(((((((.((.((((((((((...)))))))))...)).)..)))))))(((((...)))))....)))))) (-25.70)

potential miRNA gene = SPCC576.16c potential miRNA targets = SPBC21H7.07c

GGUGUGCUAGUUGCCUAAUGAUUUUUAUUUCUUUAUGGUguaagugacacgauaggacuauUCUUACGUUCAGUCACUAACUCAUUAGACUUCUAUGAAACUUGGAUC

AAGGCAGUCAAAGUAACU

(((.((((.(((((((..(((((((...(((((.....(((.((((((((((.((((((....)))))))))...)))))))).....)).)))).....)))))...)))))))))))))).)..)))))) (-28.70)

potential miRNA gene = SPAC17A5.02c potential miRNA targets = SPCC777.15

AGAAGUUGUCAAAACGUAAGCAUUGAAUUACUUACAUCCUCUGACGUUUGAUAGGUUAAAAAGAAAAGAGgaaauaugguugaugaaaacuUGGUUAAAAAAGACAA GUUAGGAAUUCCUAGGAACUUUU

.((((((.(......(((((..........))))).((((((....(((((....)))))......))))))....((((........((((((...........))))))........))))).)))))). (-21.20)

potential miRNA gene = SPBC1734.06 potential miRNA targets = SPBP16F5.04

CCUGUUUCAAAAAGAAAUUUAAUUCGCCAACUUAUAGAUUGGGAGCGCGUUCAGUCGAAAUCAAUuggaguagaaaaagaaaaaCUAGGAGGGGGCGAUUGGGAGAAGG CAUACGCUGAGG

((((((((.....)))))).....(((((..(((.(((.((.....(...((((.((.((.....)).)).).)))...)...))))).)))..)))))..)))....(((....))).... (-19.40)

potential miRNA gene = SPAC11G7.01 potential miRNA targets = SPAC20G8.08c

CAUGUCGGCGUAGUUGUUGGUUGUUCUGUUGCUAUuccuguugguguaguucuaauuuUGAUUGGUUUAGGAAUUUUUCUUUGGAAAAGGCAUCAACGGUCCAAACGAAU UAAGGCUGAGCGCAUG

(((((..((.((((((((.(((...((((((((((.((....)).))))((((((...........)))))).((((((....))))))....))))))....))).)))...))))).))))))) (-29.70)

potential miRNA gene = SPAC22F8.10c potential miRNA targets = SPAC31G5.16c

GUCCAAUUCCGACUGGAGCGCAGUGGGGAUUUCAUCCugguguuggggaaaaccUCCUGUGGAC

((((((((((.((((.....)))).)))))........((.((((......)))).))..))))) (-22.60)

potential miRNA gene = SPBC1734.04 potential miRNA targets = SPAC4F10.18 SPBC557.05

UUGUUAAUCUCUUCCACAAUGAAAGCAAAUAGAGACUUuggaagagaugguggcccuuuuaGUAUUACGCCGAAUCGAUUCCAGCCUAAAUCCAGCGGAAAUCCCAU AUUCCGACAG

......((((((((((......................))))))))))((((...............)))).......................(((((........))))).... (-21.10)

D.1.2 miRNA targets identified from APA of quiescent (24 hours nitrogen depletion) cells compared to cycling cells

D.1.2.1 sRNA data for Yamanaka et al. (2013)

potential miRNA gene = SPBC3H7.18 potential miRNA targets = SPAC630.11

UCUACAAAUAUACAAUGAAUaggacuucagaaagugugGAACCUCAACAAAAUGAAAAAACAGCAGUGCAUUGGAGCAGAGAAUGGGUUCCUGUUGUUGUAGA

(((((((..((((..((((......))))....))))(((((((((......))).....((.(..(((......)))..)..))))))))).....))))))) (-16.90)

potential miRNA gene = SPCC330.11 potential miRNA targets = SPCC16C4.16c

UcggaaaagagcggacaaaaacuCCGUAUCAACCAAAAAGAUCGAUACGGACGCACCGUGCUGCACAUCGCCGUUUCUGA

(((((((.(.((((........((((((((.............)))))))).(((......)))...))))).))))))) (-20.40)

potential miRNA gene = SPBP16F5.05c potential miRNA targets = SPCC757.15

UCGAUUACUCCAUAAGAAAUUUUCUCUCAAAUCAAUUAAAACGUUCCAGUUUUUUAACUUGACUGUUGCugaagaaaaaauggaagaAGAGUAGCAGUGGAUG CUGGUAGUUGA

(((((((((.....((((...)))).............(((((......))))).....((.(((((((((...................))))))))).))....))))))))) (-20.30)

potential miRNA gene = SPCC4G3.16 potential miRNA targets = SPAC4G8.02c

CgugaagcaguucgugguCACCCCCCAAAGCACUAAUUUUAUUAANNNNNNNNAUUCUUUAAACAUUUAGACGACGUCGCUGGUAUAUG

((((.(.((((.(((.(((...))).)))..)))).).)))) (-11.65)

potential miRNA gene = SPAC1527.02 potential miRNA targets = SPCC663.02

UUAACACCGACGUGCUUUGGGACUUGAAUUAGUAAACCAAAGCUACAUUAUCUUUAGAAAAAUggaaggaguacggucugaUUUUUAUUGUUAUAGAAUUGAAAUG UUAA

((((((.....(((((((((.(((......)))...))))))).)).(((..((((....((((((((.((......))..))))))))....))))..)))..)))))) (-16.50)

potential miRNA gene = SPBC405.07 potential miRNA targets = SPAC17G6.12 SPAC806.06c

GCGCAGUAAGUUUUCCCCUACACCAAUCUCACAauggcaccuggucugguuGUUGGUUUAAACAAAGGAAAGGUUUUGACCAAGCGUCAACUUCCUGAGC

((.(((.((((...........(((.........))).((((..(((..(((((......))))).))).))))..((((.....)))))))).))).)) (-18.90)

potential miRNA gene = SPBC30D10.03c potential miRNA targets = SPAC17G6.12

AAGGUUUAGGCAAUGGCUUCUAGAUAUCGUGUUGAGUAUGCGUUAaagcaacaucggaaggacGAUUUUAUUGAAUGGAUUAAAGGUUUGUUGUCAGUUCCUU

.(((....((((((((..(((....(((..(((.(((((((......)))..((((......))))..)))).))).)))...))).))))))))....))). (-17.30)

potential miRNA gene = SPBC19C7.02 potential miRNA targets = SPCC1259.01c

AAAGUUAUUAUUUGUAGCAUGCCAUGUUACAAGACGAGucuagcagaucguugccuAGAAGCGCAUUGAUUCGGAGACGGCUUU

..........((((((((((...))))))))))....((((..(.(((((.(((((...)).))).))))).).))))...... (-19.00)

potential miRNA gene = SPBC405.04c potential miRNA targets = SPCC645.03c

AACAGUGGCUAAGUUAGCUCUGGAGAACAUGGAUUCUGACGAUAUUGCGGCAGACUUUACAGAUCCUAUUcauuuagauauggagagucAAAAAACUUCAUGUUACUGUU

.((((((((.(((((.((((.(((.....((((.((((.((......)).)))).))))....)))..((((((......)))))))))).....)))))...)))))))). (-22.70)

potential miRNA gene = SPAC9E9.13 potential miRNA targets = SPCC663.02 SPCC757.15
GGUGAAGAUGCUAGUGACAGCGAACCUGAACUCGAAGAGgaagaggaaguuggcaguaacgAAAAGAAAGAGUAAUAAUGUUAUCAUGUAAUCAGUCUUAAAAUGCAUCC
AUUUUAUC
(((((((((((.(.((((...(((...(((...))).....))))..)))).)).))).....((((..((.....(((....))).....))..))))).....))))...))))))) (-18.90)
potential miRNA gene = SPAC17C9.16c potential miRNA targets = SPBC31A8.02
UCCACUGAGAACAAGAACUUGCAUUGAGAACCGAUUGGAUGUGGUUUUUAAGAGAAACAAUuuuuggugaauuagugga
(((((((...((((((.(((..((((((((((.....)))))))))).....)))))))))....))))))) (-21.10)
potential miRNA gene = SPBC16C6.05 potential miRNA targets = SPAC1F3.05 SPBC31A8.01c
CAGUAUUauacugcgauguuuguacauugccCGUGGAAGUACGUAAUAAACAACUUGUUUUAUAAUUAUAUUAACUAUGAAUGCGUUGUAGUACUG
(((((((....(((((((....)))))))..((((....))))........((((....((((((.(((...))).)))))))...)))))))))))) (-19.20)
potential miRNA gene = SPCC550.04c potential miRNA targets = SPBPB7E8.01
ACGCUGACAUGCUUAUGACCCUCGUAUUGAUCAUAAGUGCAUAUAUUGCUUUGAAUGCUGUUgcaucuguuguuauugcgu
(((((((((.(((((((.(........).)))))))))((.....)))).....((((....))))....))))..)))) (-17.00)
potential miRNA gene = SPBP16F5.05c potential miRNA targets = SPCC757.15
GAAGAAAACACUGGUGAugaagaaaaauuggaagaUGGAAUAUGAUCGUUUUAUUUGAUAAUUUCUGAUUUUUAUUUUGUAAAAAAAUUUACCUCUUUCUUU
(((((((.....(((((((((((((((((((((.((((.........)))))).))..)))))))))))).)))))).....))))..))))))) (-15.20)
potential miRNA gene = SPBC1709.06 potential miRNA targets = SPCC645.10
AGUAUGUCUAGCUCCCAugguuaggauuggggaacuuCCCAUGAGGUUAUUAGCACUUCGAUAUGGCGCUAAUCUAGUAUGGGGGCCAGAAAUUGUGGAUAAAGCU
......(((.(((((((((.(..(((((((.(......(((((((((.......)))))...))))))..)))))))).)))))))))).))).................. (-30.20)
potential miRNA gene = SPAC977.12 potential miRNA targets = SPBC211.08c
AAGUAAGGAGAUGUGGGGAUUCAUAguaacuugugggauuuuucuaguGUUACUAUGCCAAUUGCGUUUACUGAGUAAACGCAAAUCAAAAUCCACUCCUUAUAAUGC
CUUGCUU
.(((((((...(((((((..(((((((((...(((((...)))))..)))))))))....((((((((((...))))))))))..........))))))))...))))))). (-39.30)
potential miRNA gene = SPBC1604.25 potential miRNA targets = SPAC3H8.04 SPBC16E9.05 SPAC2G11.11c
AGAAUUAUUGCCAAUGUCUACUGCCUCAAAACUUUGuaucgucggggccgguguCUUUAGUGUUUACAUGGUUUAUUUCGUACAUAAUAAUCAGAUUAUUGAACGAGA
GGUAAUAUUUACU
.....((((((.........((((((...((........)).)))))(.((((..........)))).)....((((((...(((((((...)))))))..)))))))))))..... (-19.40)
potential miRNA gene = SPAP27G11.12 potential miRNA targets = SPAC17G8.05 SPCC757.15
UUUUUUUGUCCAAUUUCUCAACCugccguagguuuggggaaGAACCAAUACAUUAUUAACCAAAUUACAGGUUAGACUUUAGGAACGCGGUUUUACGACUACACGAA
GAAA
((((((((((....((((((((((...)))).))))))))(((((.....(((((((((.......))))))....)))).....)))))).........))))))))) (-19.80)
potential miRNA gene = SPBC1289.09 potential miRNA targets = SPAC8C9.04
GCCUUUAUGCCGUUUAGCCUCGUCACUAUAUACAUCUUCUUUGCAUCUUUUUGUAUCAACUAGAUGAUUcguacuauucaaaggcagagcgcUACAAGGU
((((((...((.((((.((((.............(((((....(((.....)))).....)))))).....)))).)))))...)))))) (-18.30)
potential miRNA gene = SPBC18E5.13 potential miRNA targets = SPAC13D6.02c SPBC31F10.15c
GUUUUucgaaaaugccagagccgGACUUUUAUGAAUUCUCAAUUUGCAAGUGCGAAACUC
...(((((....(((.(((...(((.(((...))).)))...)))))....)))))... (-6.70)
potential miRNA gene = SPAC869.02c potential miRNA targets = SPAC6G9.11
UUCUUGAUUGAUUUAAUAUAUAUAUGGCAUAACCGUUUGGAAUUAGAGAUUCCAAAGUAAAAUGUCCAGUguugagguuaaucgugaa
(((.(((((((((((((.....(((((((.....(((((((((...)))))))))).....))).)))))))))).)))))))).))) (-22.30)
potential miRNA gene = SPBP35G2.07 potential miRNA targets = SPAC3G6.04 SPBC31F10.05
CUGUGCUGGCUCCCCUUAGAAGAUUGCAUacgagggcggcuuucucguCUUAUGGACGAGAAAUUGCCCUUCAAAAGAGAUUUUUAAAUUUAAAUAGUUGCUCAGCU
GUGCGG
((((((((((.....((((((((((.....(((((((((.(((((((((((((((((...)))))))))))))))).....)))))))).....)))..)).)))))...)))) (-34.20)
potential miRNA gene = SPAP7G5.04c potential miRNA targets = SPBC9B6.07 SPBC25H2.12c
GAAGCUACUUUGGAGAAGCUUGCAUCGGCUGGUggucguggaggcgcuccaACUCAUUAGAUUUAUUAAGCAGUUAAAUUAUCCAAAAUGAAAAAUUUAGAAUCUAUG
UAGACUUC
(((((((.((((((..((((.((.((((((...)))))))..))))..)))))))..((((.(((...(((((....)))))...)))...))))).....))))).))))) (-27.50)
potential miRNA gene = SPAC8E11.03c potential miRNA targets = SPBC18H10.03 SPAC22G7.01c SPBC839.05c
GUUCAAUAUAAAAAUAAAGUGCAAUGGAAGAAUUCgcagaggggaaugaugaugaaC
(((((.(((.....((((((.....))).)))).....))).))))) (-6.80)
potential miRNA gene = SPBC1271.09 potential miRNA targets = SPCC23B6.01c SPCC306.04c
GCAUUGAGUAUAUCAAGCUUAUUAAUAAUUUCUUAGCCCUuuccugagugcgaccuuuUCUUUUCAUCAUAUCAUUUUAUGCUAAUCUUUUCUUUUUUAUUGUACCCUU
AUUAAAUACAAAUCAUGC
(((((.((.....(((.....)))).....)).))))).....((((.....))))....((((.....)))))..... (-9.20)
potential miRNA gene = SPAC688.10 potential miRNA targets = SPCC965.09 SPAC25G10.05c

UUCAAUCGAGUAUAUAGAUUGGGAAUUAAGUCCUUGUGAUCCAGCUUAUGAUUUCGUAGGGAAAGAUUUGccgacugcaaaugaagaguUAACCACUGUUCCAGUUAUU
CGAGUUUUUGGAUUGAA
(((((((.........(((((((.....))))...))))...(((((((....))))))).(((((((((..(((((.....(((.(((.....))).)))))))...)))))))))..))))))) (-25.90)
potential miRNA gene = SPCC1442.05c potential miRNA targets = SPAC26A3.01 SPAC1805.11c
GCAACGAAUGUUGGUUUUGGAACAAACUCAAAAGAAUAUACAAAAUGCCCAGAAGUCCAUGACUUCUGcaguugagggaguccaaaAGCAUUAUGAAAACGC
((.....(((.((.(((((((.....(((((.....................(((((((...)))))))...))))))....)))))))).)).)))).....)) (-23.60)
potential miRNA gene = SPBC23G7.13c potential miRNA targets = SPCC663.02
UCGGAAGAAAGGUUAAUGAUAUUGUUGAAGACGAAGGCGUUGUGGUGGAAACCAUAUCGAUUGGCUCUgguucgaaggagaaaguuaAUUUUGNNNNNNNNGACAUUGA
AAGUGUCUAAAAUCUGG
.((((...(((((((((....((.(((((((((.((..((.((((((....)))))).)).).).)))...)))))..))....)))))))))).....((((((...)))))).....)))). (-21.00)
potential miRNA gene = SPAC31A2.09c potential miRNA targets = SPBC1703.07
UGUGCAAUACUUAAGGGUCAGUGAACCUUCUAAUAGCAAAUACAAAAGUAUUAAAUGGGUAAGGUACUCAacuagggcuggaacuugcGAGAUACGCAUUUAAGAAGUAU
UGUGAUACA
...(((((((((.(((((......)))))....(((((.(((((....)))))...(((((((...)))))).....))))).....(((.....))))......)))))))).... (-22.20)
potential miRNA gene = SPBC29A3.15c potential miRNA targets = SPBC30D10.06
GGUUGAUAAUUAUUUGCAAAAUCAUGUCCUUAGUGGUgguaauccgcgaaaguuuUUUGAUGCUUGCACACGUCUAGCAUGAUUGAGAUAUUCAAAAAUUUCUAUCCA
CUACAACU
(((((.........(((.....)))...((((((((((.(((....(((.(((((.(.(((((.((....)).)))))..)..)))))..)))....))).)))..)))))))))) (-19.70)
potential miRNA gene = SPCC74.09 potential miRNA targets = SPAC17G8.05
UAAUGGCUAUUAUGCACAAAAUUGGCUGugccggaagugaaguugaGACCGACCGUGUUCUUGAAAAAGACCAAGCAAACAACGGUCACUUCGUUGAUAGUGAUUG
((((.(((((((.(((((........)))))......(((((((.....(((((((((((((.....)))))..))))..)))))))))))))).)))))))).)))) (-27.60)
potential miRNA gene = SPBC9B6.09c potential miRNA targets = SPBC31F10.15c
CCGAGCUGAUCAAAUAAUUGUUGUUGGAGACGGGAAaguuuuggagcaagggucuUUUGAACGAUUGUCUCGUCCUGGUACCAACUUUUACAAGUUAAUGAGAUGGCAG
CUUGG
(((((((.(((((..(((.(((((.(((((.....))))))..))))))..))))).....((((((((.((.(((.....))).)).)...)))))))))))))) (-32.20)
potential miRNA gene = SPCC757.02c potential miRNA targets = SPAC5H10.10
CAGAAAGAGAAAAACUCAAUGCUCAUGCUUUAGAAGUUGggacuugggcguuuugugaUUUUUUGUUCGGCAGGACGUACAACGUUAUAUCCUCAAUGAGCAAAGCCAG
GAAGCUUGGUUACACUG
......(((((..((..(((((((((.(((.....))).).)))))))..))..)))))((((((..((((.(((.....))).)))))...)))))).(((((((...)))))))..... (-29.30)
potential miRNA gene = SPBC839.12 potential miRNA targets = SPAC10F6.12c SPCC306.04c
GUUAAACUCUCUGUCACACUAGUUCGAACGUUUGAACCCCAUUCUUACAAUCAUCACAAUGCGACguggagcucguggaucuacaaaUUUGGGAGGAAUGACAUGGCA
AGAAGUACUUGAGUUUAAU
(((((((((((((((.....((((((....)))))))..(((((.......(.((((.........)))).)(((.((((.....)))))).)))))))...))))).)))).....))))))))) (-23.60)
potential miRNA gene = SPAC1A6.01c potential miRNA targets = SPAC17G8.05
ACACUCUAGUCAUGUCGCCGUGCCUGAAUCUAUUCUUAAUAAAAUAAAUCAGAAGUGGUCUAAaguucaggaugaugauggCAUGCCAUCGAUGUUAUAAGUAUGU
(((..((.((.(((((((.((((((...(((.(((((.(((.....((.((((.....)))))).))..)))).)))))))...))))))....)))))))).)).)..))) (-20.10)
potential miRNA gene = SPAC1F5.04c potential miRNA targets = SPAC11D3.07c
UGAAGUAGAGAUgcgaaaauucgucaaagggACAGGACCCUAAUUUUUCUUAUGACUCUAUUCUGUCUACUCCA
...((((((.(...(((....(((((.(((((.(((.....))).))))))....)))))..)))))))... (-12.40)
potential miRNA gene = SPBPB8B6.03 potential miRNA targets = SPAC1296.02 SPBC31F10.05 SPBC21B10.11
UGUUGGCUCUUUUUUUCCAACUUGGAGUAAAGAACCCUUGAUUCAUUAUUUUAAAGUAUUAUGGAUACuccgacuugggaaguggguggCCGCA
(((.((((.((.(((((((.((((((((.....(((((.((.....)).)))))))))))).)))))))).)).)))))))) (-24.90)
potential miRNA gene = SPAC31G5.03 potential miRNA targets = SPBC31F10.05
Gauggccacagaacuugu cGUGCAAAGUGAGCGUGCAUUCCAAAAGCAACCUCAUAUCUUCCAAAACGCCAAGAAGGGCGCUGGCCGUC
(((((((.....(((...))))).(((((..(((.....)))..)))))).....(((.....))))).)))))) (-24.60)
potential miRNA gene = SPAC521.05 potential miRNA targets = SPAC4G8.02c
UCUCUUUCUGCAAAAGGAUCAGGaugggguauuacgagagauUCCCGCCACAAGCGCUCUGCUACCGGUGCUAAGCGUGCUCAAUACCGUAAGAAGAGA
(((((((.(((....((....((.((((.(((.....))).)))))))....(((((..((.....))..)))))).....)))))).))))))) (-23.80)
potential miRNA gene = SPBC106.07c potential miRNA targets = SPAC20H4.04
GGACGGAACUGGCCUUUGCUUACGGGAUUCAUAAAGCUCUUGUUGUUGCUAGAGUUCCUCUUACUGCAGCaauuguaccucccuuggcaAAGCGCUUCCGUGGUCC
((((.(....((((((((((...(((((...(((((.....(((((((.(..((((...)))))..).)))))))))))...)))))..)))))))).)))....)))) (-34.20)
potential miRNA gene = SPAC977.12 potential miRNA targets = SPBC211.08c
AAGUAAGGAGAUGUGGGGAUUCAUAguaacuugugggauuuuucuaguGUUACUAUGCCAAUUGCGUUUACUGAGUAAACGCAAAUCAAAAUCCACUCCUUAUAAU
GCCUUGCUU
.(((((((...(((((((((..((((((((((...(((((...)))))..)))))))))....((((((((((...))))))))))...........)))))))))...))))))). (-39.30)
potential miRNA gene = SPBC2D10.04 potential miRNA targets = SPBC14C8.12

UCUCCCUCCUCCCGCCAAACUCCAUAugaggggcgaauugaauaUUCCAGCGCUUAGUCGCUUUGGGAAAUCGAUAUCGGCUUCUUUACACCACCAAAAUGGUUCUU
CUCGAGA
((((.((((((.................))))))(((((((...(((((.((......)).)))))..)))))..))))..)))..........((((......))))........)))) (-23.10)
potential miRNA gene = SPAC27D7.06 potential miRNA targets = SPBP4G3.02
CAACGAAGUAUUCUAACAUGCUAAGGACUACuagccggacauuuucuagggcUUUGUCUUCUUCUAACUUUAAAAUAAAUUGUGGAAGAAGGCAUUGGUUUAGUGU
UUUGACGUUG
(((((.(((((.......)))))..((.........)(((((....(((((.(..(((((((((((...((......))...)))))))))))..).)))))))))...))))) (-27.50)
potential miRNA gene = SPBC1711.18 potential miRNA targets = SPAC806.06c
GAUUGUUUGAGAAggcgggcagguugucuguAAGGUCACAAACUGUGCAAACCUUUCAACAAAGGAUC
((((.((((.((((((.(.((((..(((((....))..)))).))))).)...))))))...)))))).)))) (-18.20)
potential miRNA gene = SPBC106.04 potential miRNA targets = SPBC30D10.06
GGGAUUCUUUUUUAUUCUAAUUUUUAGAGCUGAAAUGAACauggaacaagaagaugaucaAGUCCCAGCUGUUGCAGCUGAAACUGUUCCUUUAAAACGAUAUGUAAC
GAAUCCU
(((((((..((((((((((...)))))).)))))...(((((.....((((..((.((.((..(((((((...)))))))..)))))..)))))).)))))).....))))...)))))))) (-22.30)
potential miRNA gene = SPBC21D10.11c potential miRNA targets = SPBC16D10.03
AUAAAGAAUUAGCAAAAUUUCCUCUAUUUUGGAAACAACCAUGUCCAAAAUGGGAAGUUUGAUGCUUUCUagagcaagaugccuucuugaacUACGAAGUGCUUUAUU
UUUUGU
(((((((((.(((((((((((..(((((((.....)))))))))))))))))....((((.(((..((((((.....))))))..))).)))))))..)))))))) (-30.80)
potential miRNA gene = SPBC2G2.17c potential miRNA targets = SPAC22A12.11 SPAC4G9.17c
CCAUGGUUCCACCGCUgacagcaguggguuguacagucuCUGUCACCGGAGGCAAUGCCUACUUUGAGUUGUACUAAAUUGCGUGUUAAAAGUGGAUGCUUGG
(((.(..((((((...((((((((((....((((((.(((.....((((...))))).....)))))))))...)))))).)))))...))))))..).))) (-32.20)
potential miRNA gene = SPCC1322.11 potential miRNA targets = SPBC1734.06 SPAC22G7.01c
AACUUGCGGGUUCGACGUUCCUUCUACAACCAAACACCAGCAGAAAAUUAUCACAAAAUGUCUCguggacgcggugcugccucuGGUACUAAGUAUAGAAUGACCUUGGG
UCUUCCAGUU
.(((.(.(((((((((.(((..((((((..((.....(((((((.....(((((.....)))))..)))))))...))))).....)).)))))).)))).)))))))). (-20.20)
potential miRNA gene = SPBC30D10.10c potential miRNA targets = SPCC645.03c
GGAGUAUUUUAGUGAUCUAAAAAACAAAAAUGAAAGCAUUcaacuagcagccgcugaucAAUUAAAAGAAUUUGUACAUAGUUCUACUAAGGAAUUAUCAGGUGAAUCC
(((((.((((((....)))))).))).....((((....)))))..((((....)))).....((((((...(((((((((.....)))))))))))))...))) (-14.00)
potential miRNA gene = SPBC428.01c potential miRNA targets = SPBC27B12.10c
UUGAAUAUGUUUCUUAUUUAGGAAAGGCCGGAGAAUGuagccuuucaaccccuaaUGGGCUUUUUUCUUUAUAAGUGUCAAUGUAUAAGAGAUUCAAUUCGA
((((((..(((((((((....(((((((.....)))))))...(((....)))).....................)))))))))...)))))) (-20.90)
potential miRNA gene = SPBP16F5.05c potential miRNA targets = SPCC757.15
ACUCAAUUAGAUUCCACUAUUCCUACUCAAAUUGAUUCUGUUCAACGUAACAAGGAUCAAGAUGAugaagaaauauuggaagaAAGACGAGAGUCAAGAAGGGACUCCA
AAACAUUUUUGAGU
((((((.....(((((.((((.((.((((..(((((((((((.....))).)))))))))..)))).).).).)))).)))))).....((((.....)))))..........)))))) (-27.70)
potential miRNA gene = SPBC1683.08 potential miRNA targets = SPAC17G8.05
GUUAUGUUAAUUUUCGUCUCCAUGGCCGGUUGGAUGUUuggugcugauaccgguucCAUUGGUGGUGUUACUAGUAUGCGCGAUUUUCGUGAGCGUUAUGCUGAUAGA
UAUGAC
(((((((.............(((.(((((((...(((((((....)))))).)))))))))).)))....((((((((((.....))).)))))))...))))))) (-28.90)
potential miRNA gene = SPBC691.01 potential miRNA targets = SPBC31A8.01c
AAUAUACUAAACGCUUUGUUGGAGUUCAUGUCUACAAAuuuagaagagaaacagaggAGAUUUUCUAAGCUUCAAAAGACUUUGGGUGUU
.((((.((((((..(((((((....(((((.....))))(((((((.(.....).)))))))).)))))))))..)))))))). (-13.60)
potential miRNA gene = SPAC22F3.09c potential miRNA targets = SPAC1F3.05 SPBC16D10.03
UCCGCGCUACGCAUUUUAACUGCCUUUUGCUGUAAUGGCUCCACGUUCUUCCGCAGUACAUGUUGCCGUUuauucugggguugaagucuAUGAAUGCUUUAUUAAAG
GUGUCUCUGUGAUGCGCAGA
((.((((..(((((((((((.((.........((((((..((.((.((.....)).)).))..))))))......)))))))))).....(((((((....)))))))....)))..)))))) (-26.10)
potential miRNA gene = SPAC24B11.10c potential miRNA targets = SPAC23H3.07c
UGGGCGCAUAAAGCGGCUUGCAAAGGUUUGGCUAAAGCUCAGUAUGCUGUUGGCUUUAUGAUGGAACAAGgcauuggguugcugcugAUCCAUCCUCUGCUCA
(((((((((((((((.(((...(((((.....)))))..))).)))....)))))))(((((.((.(((((.......)))).)).)))))....))))))) (-29.90)
potential miRNA gene = SPBC16G5.09 potential miRNA targets = SPCC191.07
CAUGAUUUCGAAGUUACUGAAAAUAGUCCUUUUAacugcagguguuauuggcaACACACUUGCUGAUUCUAGAAUUCAUG
(((((.(((((((..(((((....)))).)))))...(.(((((((.....)))))))).)......))).)))) (-14.90)
potential miRNA gene = SPCC825.02 potential miRNA targets = SPBC29A10.01 SPCC663.02
CUGAGAAGAUGAAGUUCAGUCAAUGGUACacguugacugcaccuuuguugaucAG
((((...(((((((.((((((((.....)))))))))).).)))))))..)))) (-20.30)
potential miRNA gene = SPAC869.11 potential miRNA targets = SPAC17G8.05

UGAGAAUCUCGAUGAAGAAGAAGAUGAGCCCACcacugcuucuaagaggauuuUCA

((((((((((...((((.((..(.((....)))..)).)))...)).)))))))) (-12.00)

potential miRNA gene = SPAC4G8.04 potential miRNA targets = SPBC839.09c

UCUUCCCCUUAUUCAUUUGUUUUGGAaugcaaccaucuguuucugaCUCUUCAGUUUACCUUUCUUCCGCUCCCACGAAGCCCAUUGCUAGUGGAAGUGUCGCUAC

AGGGAUAGA

(((..((((....(((((......)))))).....((((....)))).....((((.((((..(((.....))).)))).)))).....)))))..))) (-23.00)

potential miRNA gene = SPBC405.07 potential miRNA targets = SPAC17G6.12 SPAC806.06c

AUGCUACCAAAAGUUUUUUCCAGACGGUAGGUACAGUGUAGACUUUACACGauggcaccuggacuggguuGUUGGUUUAAACAAGGGAAAGACUUUGACUAAGCGU

((((((....((((((((((...(((((((.(((((((.....)))))..))..)))))..)))))).((((((.....))))))))))))))))).....)))))) (-28.10)

potential miRNA gene = SPAC688.06c potential miRNA targets = SPBC83.13 SPAC1B3.01c

GCUGAAGCGGAAAAAUUUAAAUGUUUCGCUAGAUCUUAUAAAGUCUUGGUGCGAUAAGUAUGGAGUACUCaugaaagggaguuggcauuAGCUAGCUAAUCAGU

((((((((((.....)))))))...((((....(((((((.(((((.....))))).)))..)))..)))))).(((((((....)))))))..)))))) (-26.10)

potential miRNA gene = SPCC1223.15c potential miRNA targets = SPAC24C9.03 SPBC13G1.11

GCAUUUCAACGAUUGGAAGACUCUAUUUCAAGCUUACAGAGACAAGAGGAUUCGCUUAAAACAAAAUACGagcuucaggaagcucgguUGGAUAUGC

((((((((((.....(((..((((.(((((.....)))))..)))))..))).....((((((((....)))))))))))))).)))) (-23.90)

potential miRNA gene = SPAC13F5.06c potential miRNA targets = SPBC577.12

GCAGUCAUGCAAGACAAGCGUAAAGCAGGUUUUAGUGAAAUGAACAUUCAUGAAGUUCUCCGUCUacguuuugauuugccugaaaACUUGAAGUUGCAACUGAACA

AAGAAGAAGCUUUAUUGC

....(((.((((((.(((((((..(.(((.(((.(((((.....)))))).)))).)...)))).)))).)))))).)))).....(((((((....((.....)).....)))))))).... (-21.50)

potential miRNA gene = SPBC713.10 potential miRNA targets = SPAC17G8.05 SPCC622.02 SPCC965.09

UUUGUUCAAAAUUACUAUAUUACAUUAAAAAAUGUCGcuuccaagggcugucggucGAUUUAUAAUUGUUGGUUCGCAAGGUAUGUUUGUGGCCAUUCUCCUAUCA

AGGAUUUAUUCAAA

.....((((((.(((.((((((..((...)).).)))).)).)))).)))((((.((((((.....))))))))))...((((....)))).......... (-14.50)

potential miRNA gene = SPCC1795.03 potential miRNA targets = SPAC17C9.12 SPCC1259.01c

AAGAGGCACUUACAUUCCAAUGACAACCCAAGaugcugcugcuaaggacguuGAUCAUAAGCAUUAAUUUAAGAUGUUCUGUAAUUCUCAAAGAAUGUAUUCUUUUU

.(((((....(((((((..(((((((.....((....)).....)))).)))).((((((.....))))))..............)))))))..))))). (-16.40)

potential miRNA gene = SPAC17G6.09 potential miRNA targets = SPBP23A10.05 SPAC4G9.09c

AGGAUUUUGCCUAUGGAUUCUUCUAAUGUACcaguuuugaaggaugaagacaAGUGCAAGUUUUCGAUGCGCUUUACAAACUUCCU

((((.((((((...))(((((((.(((.....))).)))))))).....(((((((.(....)..)))))))..)))))..)))) (-15.20)

potential miRNA gene = SPAC2E12.02 potential miRNA targets = SPCC645.10

GGACAUAUAUAUUUUUGUUAAAUUAACGGCACCCAUAAUAAAUUGAUUUUUAAUGAuucagacugcaucgacaaUUUCAAGCAAUGGCGGAACGAAUCAGGGAAUGGAGUCC

((((..........((((((...))))))..............((.(((((((..(((((((...(((((.....(........).....))))...)))))))))))).)))))) (-14.10)

potential miRNA gene = SPBC1921.06c potential miRNA targets = SPBP23A10.05

AGUGCCUUGauagaaacugaacgacuuuCGUGUCUUCAGGCAACAAUUCCUUUACCUUCUUUGGACGUUUAAUCGUUUCUCCUCUAUCGUUAUUCU

.......(((((((..(((((.(((.....)))))))).................((.....))(((.....))).....)))))))..... (-16.50)

potential miRNA gene = SPBC14C8.05c potential miRNA targets = SPBC56F2.12

GAUGGUCACCaaaaaggugcaagagauuuGACAUGGAGUUAUUCAUCCUUAUUAAAUGCCAUC

((((((((((.....)))).....(((((((...(((.........)))...)))))))))))) (-14.70)

potential miRNA gene = SPAC1002.07c potential miRNA targets = SPAC4G8.02c

GAGCACAAAAAGUUGGAGGAUGGAGUAUGAUGAGAGUCACcggcgagaaccuaaaggcucUUGCUGAUAAACUUCCCGGCAACUUUUAAACUCAUACCCAAUACUUAAC

GCUC

((((.......((((.((..(((.(((((((((((((..((((.(((.........(((....)))).....)))))))..)))))))))...)))))))))) (-31.40)

potential miRNA gene = SPBC1683.08 potential miRNA targets = SPAC17G8.05

UCGUGGUCUUCAUUUCUAUGGCCGGAUGGCUAGGaggugcugauaccgguucCAUUAGUGGUAUCUUGGGAAUGCGUGAUUUUCAAUCUCGUUUUGCUGACCGGUAUA

AUCCUAUCACGA

(((((((...(((((((.((((((....)))))))))))...(((((((((((....)))).....((.((.(((.((((...)))).))).)).)))))))))).....))))))) (-36.80)

potential miRNA gene = SPBC2A9.10 potential miRNA targets = SPAC22F3.03c SPAC1A6.09c

GGGUGAUUUGAAAAACCUCCUCGACUUGGGAUGUCGAAUUUUCAACAUGGAAAUUAUCACAGUUAUUACagcaugaggggaggaaccUCAAUUAUCGAUCC

((((((((((....((((((((((.....))).(((((((...)))))))).(((..(((.....))).)))))))))....)))))..))).)))) (-21.60)

potential miRNA gene = SPAC13A11.05 potential miRNA targets = SPCC70.03c

UGAAGgguuuaggucuaaguacgagaaCCUUUAAUUGGAGCUCACUUUCUUCCAUUCUUUUACCAAGAAUUCCUUUA

(((((((....((...((((..(((..(((.....)).)))))....))(((((.....)))))))))))) (-12.90)

D.1.2.2 sRNA data for Halic and Moazed (2010)

potential miRNA gene = SPAC1687.15 potential miRNA targets = SPBP4H10.13 SPAC328.01c SPBC23E6.04c

CAUCAUCCAUAAUGAAUCAUGGUACUAAAAGUAAGUUGGCuaacuucgauagugcuguuguUCUAACCGUAUAAAUAGUUCCUGUGGACCCCUUUCGUAUUAUAA

AGGAAACCGCUCGUGAUG

((((((.....((((.(((((((((.(((((.((....)).)))))...)))))))))))))).....................((((...(((((.....)))))...)))))..)))))) (-23.30)

potential miRNA gene = SPAC30D11.01c potential miRNA targets = SPCC4B3.18

UAGGAAAUAACUUGACUAAAACUUUUUGGGCAACAGGUUAuagcgauucaccagaggcuaaUAUGUAUGGCUCACAUCCUUUCUAUAUGGAGCAACGCUAUAUUCCC

AUAGGAACUACCAAUACCUA

((((...(((((((.((((((....))))))....)))))))((((.............))))....(((((((.....(((.........)))).....)))))))(((....))).........)))) (-22.30)

potential miRNA gene = SPAC1142.02c potential miRNA targets = SPBC646.13 SPAC23H3.07c

UGACCCUAAUAUUGCCAACAUGGCCCGUAACUUUCAaucuggaggagguaugccugacUUGAGUUCUCUUGCCAAUAACCCCCAACUGCAAAAUUUGGCUCGCAAUUUUA

UGAACA

.....(((.((((((...((.(((..(((.(((((.........)))))))))))))....(((....)))(((((.....................))))))..)))))).)))..... (-15.00)

potential miRNA gene = SPAC806.05 potential miRNA targets = SPAC806.06c

GUAACGGAUUUCGCCAAAACAAUUGCUCCUUUUAUUCAaaggaaaaaguugaaaauacaGUUAUAUUUUUGGAAAUGGCCGCGAGCAGGAAAUCUGUAGAACACUUA

AUACUAAAAAAUCGGUUGC

(((((.((((((.((((...(((((..(((((((.....)))))))...)))))(((((((.....)))))).....))))).....(((.....)))).....................)))))).))))) (-20.80)

potential miRNA gene = SPBC26H8.08c potential miRNA targets = SPAC23H3.07c SPAC4G9.09c

GAGGCUGCACAAGCGUCUAACCCAGAUGAUGAGGAAgaguuugugauggaagaagacaACCUUGGAGAGGCCCCUUUAUUAGUUGACUCCGAGUCUUAUGAGGCAUCA

GUGAAGGCAGAUACCUC

(((((((.(((((.(((..((((.....)).)).)))))))(((((....(((((....(((((((.(((.....)))..)))))))))))).....)))))).....))))...)))) (-31.10)

potential miRNA gene = SPBC1683.08 potential miRNA targets = SPAC959.04c

GGUGCUAuuaauuuugguacaacguuugguGCUUUAUACACAAUUGAUAACUUAGGACGUCGUAACCCACUAAUUUUUGGAGCU

(((.(((..(((((.((((((.((((((((((...(((.........)))..)))..)))))))..)))..)))..)))))..)))..))).))) (-14.20)

potential miRNA gene = SPCC645.11c potential miRNA targets = SPCC645.10

AUUUGCAAUGUUCCGCUUGUGGAACUCGUUUUUUUAAuaaugauaaugaaaacacaagaUGCUAUGUCAAAUUGAACUACUGUUAUCCAGACUGUACCGAAAUUACACCA

AGCACCAGCGAAU

((((((.........((((((....(((((.((........)).))))))...)))))).((((..((..(((((....(((.(((.....))).)))....)))))..))..)))))...)))))) (-18.60)

potential miRNA gene = SPCC1682.03c potential miRNA targets = SPCC4B3.18

GUUGAUUUAGCUGGUCGUGCCUCACCUCGUUAUCAAGCUAAugauucauggauuucuaaaaagaGAAAUGCCCCCUUGUCAUCUAUAAGUCAAU

((((((((((((((((.(......).)).)))..)))..((((((....(((((((.....)))))))..))).....))))))...)))))))) (-16.00)

potential miRNA gene = SPAC4A8.11c potential miRNA targets = SPBP8B7.19

CGACAUUUCUAAAACCAACGAAACUACACGUACUCUACAAAGCCUAACAACCUCAUUGACCAACuccaaugaaaauguuggcguggauGUUG

.................(((.........)))(((((((...(((.(((((...((((((.........)))))))...)))))))))))).))).. (-17.20)

potential miRNA gene = SPBC2G2.02 potential miRNA targets = SPCC191.02c

AGUGCUGACCCAGCCAAACGUCGUGAAUGGGAGUCAUGUGUCAAACGCUUAUUAAACGGUAAAUGuacaucuggaccuggauauGUUCAAUUACGUUCUGGACAACUG

GUUGGUACAGCACU

(((((((..(((((((...(((..(((((..(...(((((((.....((.....(((.....)))).....)).....)))))))....)..))))))..)))).....)))))))..))))))) (-35.10)

potential miRNA gene = SPAC9.09 potential miRNA targets = SPBC21B10.04c

AGCUUACUGAGGCUGGUGCAGAGUACAUCCAAAUUGAUGagccuaucccuuacucuugacuugcCCCAAGAGAUCCUUGCUAGCUACAAGGAAGCUUAUGAGACUCUUGGUA

AGAUCGGUAAGCU

((((((((((((.(((..((((((((((.....)))).........)))))))..))).)))(((((((...(((((..(((((.....)))))..)))).)))))))).....)))))))))) (-41.70)

potential miRNA gene = SPAC110.04c potential miRNA targets = SPAC4D7.14

GAUUGGUUCUGUUAGUGAUGUUGUUUAUUUCUAUCCcugcuugguuuacugauauucaacGCCGUGCUUUAUUAGAAGCCGCCAACAUCGCUGGUCUCAAUC

((((((..((...((((((((((...((((((.....(((.((((....(((...)))..))))..)))).)).....))))))....))))))))))..)))))) (-26.70)

potential miRNA gene = SPCP1E11.09c potential miRNA targets = SPAC8C9.04 SPBC409.18 SPBC4F6.10 SPAC1B3.01c SPAC23H3.07c

UGCCGAAGAAGAAAAGAAGGAAGAAGCCAAGGAggaggaagagucugaugaggacAUGGGUUUUGGCUUGUUUGACUAGAUCAAUUAUACGAAUGAGUAUCUUUGUGC

CUUAAUUUACGUCGCA

(((.((.....(((..((((((.((((((((((.......(.((((.....)))))..)...)))))))))).))).........(((..((((.....))))..)))..))))...))))) (-20.10)

potential miRNA gene = SPBC530.02 potential miRNA targets = SPAC2F3.11

GUCUGGCUUAGGAGCGCAAAUUAAGCGAUAAGAAAGAAAUGGAUCUGUAUCUUUUAGCAAGCGGCAGAAAguuuggguuguaaagucUGAC

(((.((((((.((.(.((((((..((.....((((((.((.....)).)))))).....))..........)))))).).)).).).)))))).))) (-18.30)

potential miRNA gene = SPCC965.08c potential miRNA targets = SPCC569.08c

CUCCAAUUGUUGGUAGGGUUUCCAUGGACAUGCUUACUGUUGAUCUUUCAGACAUUCCUGAUGUCAAgccaggggaugaaguaaauauUUUGGGGAACUCCAGAAUUGACAG

UGGCAG

.......((((.((.((....)))).))))(((((.(((((((((.........(((((((...(...))..)))))))).........((((((....))))))))))))))))). (-27.70)

potential miRNA gene = SPAPB2B4.05 potential miRNA targets = SPCC576.03c

GAAAGAUUAUUUUAUGGUCCAAGACAGUAAGUCUUUAUUUUAAuaaauuuaugacugacuuuugccUCUUUAACAGAAUCACCUCUGGAAUAUGUUUCUUCUUUU

((((((.........(((..(((.((((.....((((....)))).....)))).)))..)))).....((((.....))))(((.....))).)))))) (-12.30)

potential miRNA gene = SPAC56E4.06c potential miRNA targets = SPAC30D11.05

CCAUCUUCUUAAAGAUCCAAUuuggcagccuauccuuaugccaaaUGGGAAAGUACUUAGGGCUGGCGAUAAAAUGUUUCGUCCGGCAUAUGCUAAGACUUUGG

...((((....))))(((.(((((((((.....)))))))).)))(((((.(((((((((.....))).)))))).....)))))))... (-29.40)

potential miRNA gene = SPCC1682.11c potential miRNA targets = SPAC4F10.20 SPBC56F2.12

UAUUCCAUUCAUGAAAAUGGUGUCUUUUAUGGCUACAUUGUAGGACUUUUAGGCGGAUUCCUUGCUUGGCuuaucauuggcgccauagagGGCGGCCUCAGCAUGAUG

UGAGUA

(((((((.((((.....(((((((((((((..((.((((((.((...((((((.....)))))))))))).)).))..)))))))))))).))))....)))))))).)))) (-39.10)

potential miRNA gene = SPCC576.16c potential miRNA targets = SPBC21H7.07c

GCCUAAUGAUUUUUAUUUCUUUAUGGUguaagugacacgauaggacuauUCUUACGUUCAUUCACUAACUCAUUAGGC

(((((((((.............((((((..((((((.(((.((((((....)))))))))))))))))..)))))))))) (-20.90)

potential miRNA gene = SPBC30D10.15 potential miRNA targets = SPBC31F10.05

GGUUucugguaaucaagccuugccuggCGAAGCAAAUUUCCAGUCACU

(((..((((.(((...((.((((...)))))).))..)))).)))))..))) (-10.60)

potential miRNA gene = SPAC1610.04 potential miRNA targets = SPCC364.03

UGCCUGAGUCAAAAGaauugaaauaugccauggaucACAAUGGACUAAACCCACUUUAUAUCUCGUUUGGCA

(((((((((...((((.............(((.........)))).........))))....)))))...)))) (-10.30)

potential miRNA gene = SPBC3B9.17 potential miRNA targets = SPAC806.06c

GUGAUGUGGUUAAGAAGACUAUAAGUUUUAAUUAAGuauuuccugguuauugacuaaaauAUCUAGGGUAUUUGUGCGUGGAAAAGCCAGGGUUGUUGCUGACAACAUUU

CAUUAC

(((((((.((((.(((((((...)))))))..)))))..))...(((((((.((.((((((((((...))))))))...)).))..)))))))(((((...)))))))....)))))) (-25.70)

potential miRNA gene = SPBP4G3.02 potential miRNA targets = SPAC17C9.13c

UUUUGAUGGUAAAUUUGAUUUCAAGGAACAUUUGACuucccguagccccuaucauaaaacCCUACUUUUACGGCCCAAGCAUUGACUUUCCCACCACUUGUAAGAUAAAGC

AGGUCCAUACUCUUCAAAG

.(((((.((((...(((....)))((((.....((.(((.(((((((.....................))))))...)))))......))))....((((((.........))))))...)))))..)))))). (-15.70)

potential miRNA gene = SPCC576.16c potential miRNA targets = SPBC21H7.07c

GGUGUGCUAGUUGCCUAAUGAUUUUUAUUUCUUUAUGGUguaagugacacgauaggacuauUCUUACGUUCAGUCACUAACUCAUUAGACUUCUAUGAAACUUGGAUCAA

GGCAGUCAAAGUAACU

(((.((((.(((((((..(((((((...(((((.....(((.((((((((((((.((((((....))))))))...)))))))........)).)))).....)))))...)))))))))))))).)..))))))) (-28.70)

potential miRNA gene = SPBC1685.13 potential miRNA targets = SPBC1734.06 SPBC31F10.05 SPBC1677.03c

CGUCGGUCGCUUCCUUCCCACCAUCUUGGGUAAUccuugguugauugccuuuuaCGACUUUGUCAACUGGGUCUUUGCACUCACCGGUGGUUGCUGUAUCGCCGUCGCC

AUUCGUGUCCAUG

....((.(((......((((.....))))..((((..(((((((.............)))))))))).....((((((.....)))))).....))).))... (-25.60)

potential miRNA gene = SPBC23G7.06c potential miRNA targets = SPBP8B7.19

UGGUUUAAaguucgaaucucuuucuaacGGUGGUGAAUUAAUCGUAAGCGCCGAUAUUUGGUACGAAGGCGAUUGCAGUUUUAAGGCA

((.((((((((..((((.(((((((.(((((.(.............).))))))..))).....))))))..))))..)))))))..)) (-17.40)

potential miRNA gene = SPAC9E9.09c potential miRNA targets = SPAC3G6.04

GCGAAUUGGGCUCUUACGGUCUUACCAAUUAUACUCAAacaaaagcuguucacaucaaCCUGGGAAUGGAUUCUCCCAUUUAAGGAAAUCAGACGC

(((((((((.............))))))).....((((.....)))).....(((..(((((((.....)))))))..)))).....))) (-15.40)

potential miRNA gene = SPAC23H4.03c potential miRNA targets = SPBC1734.06

UGuuuagaucccuuucuuucuuuGAAAAAGAAAUCCUUUAUGAGUUUAGCUUUCUGACUGACAUUUAAAUGUAGGUGCUAUAGAGUCUGAACA

(((((((((...(((((((.....)))))))).....((((...))))(((((((.((((....)))).)))....)))))))))))))) (-13.60)

potential miRNA gene = SPBC14C8.03 potential miRNA targets = SPBP8B7.19

ACUGACGAagagaagcgcgcuuuggaccGACAAAAUUUUGAUCAGUAUAAUGAUUUGCGUCGUGCUGCUGAAGUACAUCGCCAAGCUCGUCAGU

(((((((.............((((.....)))).....(((((.....))))).(((..(((((.....)))))..))).....)))))))) (-24.20)

potential miRNA gene = SPAC17A5.02c potential miRNA targets = SPCC777.15 SPAC323.07c

AGAAGUUGUCAAAACGUAAGCAUUGAAUUACUUACAUCCUCUGACGUUUGAUAGGUUAAAAAGAAAAGAGgaaauauggguugaugaaaacuUGGUUAAAAAAGA

CAAGUUAGGAAUUCCUAGGAACUUUU

.((((((.(.....(((((.........)))))).((((((....(((((....)))))).....))))))....(((.........((((((.........))))))....))))).))))))). (-21.20)

potential miRNA gene = SPBC1734.06 potential miRNA targets = SPBP16F5.04

CCUGUUUCAAAAAGAAAUUUAAUUCGCCAACUUAUAGAUUGGGAGCGCGUUCAGUCGAAAUCAAUuggaguagaaaaagaaaaaCUAGGAGGGGGCGAUUGGGAGAAGG

CAUACGCUGAGG

(((((((((.....)))))).....(((((..(((.(((.(((.....(...(((((.((.((.....)).).).)))).)...)))))..)))..))))))...)))).... (-19.40)

potential miRNA gene = SPBP16F5.02 potential miRNA targets = SPAC1296.02

CCGCUCUUUCCUUACCAACUCAUAUUCGCGuauguuuagacauggucggauuuCCUUCAAUUACUUGGAAACUGUACUCAAUAUAAACAUUCCGACUUUACUCACAAA
AUUUUAGUCCACGG
.....(((.(.(((((((.....((((((.(((((.....)))))))).)))).....)))))))..)))..... (-16.60)
potential miRNA gene = SPAC9G1.06c potential miRNA targets = SPAPB17E12.11 SPCC23B6.01c
GGUACAUUGCAUucugguguugaauacaaaUUUAAUGUUUCGGCCUAUGCCUGCCAGCCAUCUACCUCCAUUAGCAAUACCAGAUUAGCAAUACAGACC
(((..(((((((.((((((((((.....(((.(.(((.....))))).))))..........)))))))))).).)))))....))) (-22.70)
potential miRNA gene = SPAC11G7.01 potential miRNA targets = SPAPB17E12.11 SPBC19C2.03
CAUGUCGGCGUAGUUGUUGGUUGUUCUGUUGCUAUuccuguugguguaguucuaauuuUGAUUGGUUUAGGAAUUUUUCUUUGGAAAAGGCAUCAACGGUCCAAAC
GAAUUAAGGCUGAGCGCAUG
(((((..((.(((((((((.(((...((((((((((.((....)).)))))((((((.....)))))).((((((....))))))....))))))....)))).)))...))))))..)))))))) (-29.70)
potential miRNA gene = SPBC29A10.13 potential miRNA targets = SPBC1734.06
GAUUGUUGAAACAAUGGUUACCAAGGGCAAAUGGGUCAUCCCAGGCUACCGUGAAAAGUUUGGUGAUuugaguauuauguaguuugcucCAUUAAUGAAACAUUUUA
UAACUAGCAAUC
(((((((.....(((..((.....((((....)))))))..)))((((((.((((.((.....(((((.....)))))).....)).)))).)))))))...))))))) (-23.50)
potential miRNA gene = SPBC11B10.09 potential miRNA targets = SPCP25A2.02c
UGucacguauacugacuaaugcuAGGAACCUAUGGCGUUGUUUAUAAAGCAAGACA
(((((..((.....(((.(((((((.....)))))))))).....)).)))) (-10.80)
potential miRNA gene = SPAC167.07c potential miRNA targets = SPAC1A6.09c
GUUUGAUGUUCAAAAUGUCUUAAAAAAUGAUUCCGGCUACUGGUUUCAACUACAAUUACUAAUUGAUguuuacucacgcaugcuuuuuaCCAUGAUCGACGAUGAG
UUUCAUAACGAUAAGC
.(((((...))))).....(((..((((...)))))..))).....((.((((.(((..((((((((((.....)))).....)).)))))...)))).)))). (-16.50)
potential miRNA gene = SPAC6B12.18 potential miRNA targets = SPAC630.11
AUGCAAGAUCAUUGCAAUGAAGAGUUUACUAAGAAAAUGcaaacugaaggaaagaaugauGUCCCAGAGGAAAAACCCAUAAUUUUAUAAGCGAUUAAUUAAUAAGGA
CAGAUCGUGUCGAACUGCAU
.(((((.....)))))...................(((((..(((..(((.(.....).)))))).((.....)).....((((((..((.....))..))))))..)))))) (-14.80)
potential miRNA gene = SPAC22F8.10c potential miRNA targets = SPAC3H8.03 SPBC9B6.07
GUCCAAUUCCGACUGGAGCGCAGUGGGGAUUUCAUCCugguggguuggggaaaaccUCCUGUGGAC
((((((((((.((((.....)))).)))))).....((.((((.....)))).))..)))))) (-22.60)
potential miRNA gene = SPAC821.06 potential miRNA targets = SPAC17C9.13c
UCACGUUACAAUGUUGACUUGGCCAAAACCAACAAAUUCCAAAUUCAUUCAAUUCGCAACUUGCAAUGGaaguuccuucugcaguuacCCUAAAUAAUUAUGUGGG
UAUGAAAACAAAAGAAUGUUGA
..(((((....((((...((((.....)))).....((((.(((..(((...)))...)))))))(((((.....)))))....)))))....)))))... (-15.90)
potential miRNA gene = SPCC16A11.01 potential miRNA targets = SPBC1683.10c
GUCGUAGGUAUCAUUUGCGCCAUUGCAUUCGGUGGUCuucaucacaauuaccguggucgUUCAGCACGCUUUGAAUGGGUUGUGUGUUUUUUAUGGUCCAUUUACA
UUUGCUUAUUAUGUCUUGAC
(((((.(((((((.((..(((((((((....)))))))).....((((((((..(((((((.....)))))))....))))).))))........))..)).)))))))))) (-23.70)
potential miRNA gene = SPBC1683.08 potential miRNA targets = SPAC959.04c
GGUGCUAuuaauuuugguacaacguuugguGCUUUAUACACAAUUGAUAACUUAGGACGUCGUAACCCACUAAUUUUUGGAGCU
(((.(((..((((.(((((((.(((((((((...(((.......)))..)))..)))))))..)))..)))))..))).))) (-14.20)
potential miRNA gene = SPBC1734.04 potential miRNA targets = SPBC31F10.05 SPBC1677.03c
UUGUUAAUCUCUUCCACAAUGAAAGCAAAUAGAGACUUuggaagagaugguggcccuuuuaGUAUUACGCCGAAUCGAUUCCAGCCUAAAUCCAGCGGAAAUCCCAU
AUUCCGACAG
.....((((((((((.....)))))))))(((((.....)))))..........(((((.....)))))).... (-21.10)

D.1.3 miRNA targets identified from APA of quiescent (7 days nitrogen depletion) cells compared to cycling cells

D.1.3.1 sRNA data for Yamanaka et al. (2013)

potential miRNA gene = SPBC3H7.08c potential miRNA targets = SPBC30D10.04
GCUGAUCAGUUUGCUGUUAGAUAACUUGGUGUGUCAUGACUUAUACCUUUGGAGUugcugguuauaugcuugguAUGUUAAAAAUUAGUAGAAAAUUAGAAAUCAGC
(((((((((....)))).....((((((((((.....)))))))....)))))(((((((((((((...)))))))....)))))))).....)))))) (-19.80)
potential miRNA gene = SPAC1D4.14 potential miRNA targets = SPAPB1A10.13
GUGUUCAUGACUUCGUUGCCAgaaaaggaucagcaaggcgAAGUUAGUGUUUCAGGUAAGAAGUUCGAGGUAAUACCGAUUAAUUUUUUAUUAACAGAAUAU

((((((.((((((((((..((.....)).))))..)))))))))).((((...((((((((((...(((...))).....))))))))))))))))))) (-21.90)

potential miRNA gene = SPCC330.11 potential miRNA targets = SPCC16C4.16c

UcggaaaagagcggacaaaaacuCCGUAUCAACCAAAAAGAUCGAUACGGACGCACCGUGCUGCACAUCGCCGUUUCUGA

(((((((.(.((((.....((((((((.............)))))))).(((.....))))...)))))).))))))) (-20.40)

potential miRNA gene = SPAC1687.04 potential miRNA targets = SPAC4H3.10c

UUUCUUCAAAGAUUUCAGGAUGCAUUGUUUGCAGGAUCUAAGCCCGAUUUACAAGGCACuuuaggcaucgaugaagaaG

.(((((((..(((.((((((.(((.((((.....((.....)).....)))))..)))))))..)).))))..))))))). (-18.10)

potential miRNA gene = SPCC4G3.16 potential miRNA targets = SPAC4G8.02c

CgugaagcaguucgugguccACCCCCCAAAGCACUAAUUUUAUUAANNNNNNNNAUUCUUUAAACAUUUAGACGACGUCGCUGGUAUAUG

((((.(.((((.(((.(((.....................................))))..)))))..).)))) (-11.65)

potential miRNA gene = SPAC1527.02 potential miRNA targets = SPAC4H3.05 SPCC613.10 SPCC663.02

UUAACACCGACGUGCUUUGGGACUUGAAUUAGUAAACCAAAGCUACAUUAUCUUUAGAAAAAUggaaggaguacggucugaUUUUUAUUGUUAUAGAAUUGAAAUGUUAA

((((((.....(((((((((.(((.....)))...)))))))).)).(((..((((....(((((((.((.....)).)))))))....)))))..))))..)))))) (-16.50)

potential miRNA gene = SPBC405.07 potential miRNA targets = SPBC23E6.03c

GCGCAGUAAGUUUUCCCCUACACCAAUCUCACAauggcaccuggucugguuGUUGGUUUAAACAAAGGAAAGGUUUUGACCAAGCGUCAACUUCCUGAGC

((.(((.((((.....((((.....))))..((((..(((..((((((.....)))))).))).)))))..((((.....)))))))).))).)) (-18.90)

potential miRNA gene = SPCC1235.16 potential miRNA targets = SPAC4F10.18

UUGCAUUCGUGAUUUAAAAUGGAAAGGaaaucccaaguuucagauACAAAUAAUAACUCUAUACCUACGUAAGUUUAUUAAUUAUUAGUUUAUUUCUUUACGCCGGA

CUGCGA

((((((((((.....))))))...(((((.....)))))).....................((..((((((.....(((.....))))....)))))))..))..)))))) (-10.90)

potential miRNA gene = SPBC19C7.02 potential miRNA targets = SPAC4G8.02c

AAAGUUAUUAUUUGUAGCAUGCCAUGUUACAAGACGAGucuagcagaucguugccuAGAAGCGCAUUGAUUCGGAGACGGCUUU

.....((((((((((...))))))))))....((((..(.((((((.(((((...)).))))..)))))).).))))..... (-19.00)

potential miRNA gene = SPBC405.04c potential miRNA targets = SPCC645.03c

AACAGUGGCUAAGUUAGCUCUGGAGAACAUGGAUUCUGACGAUAUUGCGGCAGACUUUACAGAUCCUAUUcauuuagauauggagagucAAAAAACUUCAUGUUACUGUU

.(((((((.(((((.(((((.....(((((.(((((.((.....)).)))))..))))).....)))..((((((.....)))))))))).....))))))...)))))))). (-22.70)

potential miRNA gene = SPAC9E9.13 potential miRNA targets = SPCC663.02

GGUGAAGAUGCUAGUGACAGCGAACCUGAACUCGAAGAGgaagaggaaguuggcaguaacgAAAAGAAAGAGUAAUAAUGUUAUCAUGUAAUCAGUCUUAAAAUGCAUC

CAUUUUAUC

((((((((((((.((.((((...(((...(((...))).....))))..)))))).)).))).....((((..((.....(((.....))).....))..)))).....))))))...)))))) (-18.90)

potential miRNA gene = SPBC405.01 potential miRNA targets = SPAC23H3.09c SPAC23H4.12

UGCAAUGGUAGCUUGCUCGCUGUAACGAAAGCAUGGAACCCAUUAUUGCGUUGCUGAUUggcaauggcggucgugaacauaccAUCGCA

(((.(((((((.....))(((.((((..((((((...)))).....(((((((....)))))))))..))))..)))))))))))))) (-24.20)

potential miRNA gene = SPBC16C6.05 potential miRNA targets = SPBC16E9.15

CAGUAUUauacugcgauguuuguacauugccCGUGGAAGUACGUAAUAAACAACUUGUUUUAUAAUUAUAUUAACUAUGAAUGCGUUGUAGUACUG

(((((((((....(((((((....)))))))..((((....))))........((((....(((((((.(((...))).)))))))...)))))))))))) (-19.20)

potential miRNA gene = SPBC2F12.07c potential miRNA targets = SPCC965.11c

GUACACCAAUCUGUUACAUACUCUUUCAACAAACUAACAAugggacgugugaucagagcUCAAAGAAAGAGUGGUGGAAUUUUCCAAGCUCAUACCCGUCUUCGUAAGGG

UGCUGCUCAGCUCCGUAC

((((.....((((((((((.(((.....................)))..)))))))..))))((((.......))))..((((....))))((((((((((.....)))))).)))...))))..)))) (-27.00)

potential miRNA gene = SPBC16E9.11c potential miRNA targets = SPAC29B12.07

UGAUUAUCCUUCAAAAGAUACACUGCAUGAAAAAUUGAGCCUAGCCGUUGAAAACACGGUaggguuuggaaacgaguaAUGACAACAAUCUUGAAAUUUCAUAAUCA

(((((((.......((((((....((((((.....(((((((.(((((.....)))))..)))))))).....)))..)))...))))).....)))))))) (-19.30)

potential miRNA gene = SPBC28F2.12 potential miRNA targets = SPAC31A2.09c

CCGACACCACCUACGACUUAAUCUUUGCAACUCAGGUAGAaugagcgggauucagguaUGUUGAACGUAUUUACCAUUACUAAUACAUAGUUUUCACCUUCUUCUG

UCCCCUUACGG

(((.......((((.....)))).....((((.(((((..(((...(((((((((.....)))))...)))))..)))...)))..)))))))))))) (-15.70)

potential miRNA gene = SPAC3G6.09c potential miRNA targets = SPAC25B8.14

AUUAUCACCUUCCUGCUccucgugauuuaggagaacUUUUGCAUAACCUAUGAAAUAACCGGGAUAAU

.....((((((((((.....))))).....((((...)))).....)))))).... (-9.50)

potential miRNA gene = SPBC27B12.01c potential miRNA targets = SPAC1A6.09c

UCGUUGUGUUUCUCCUCAGUUUCAACAGUAAGUGAGAGUAUGAAUUGAUUCAUACUAACUAAUucccacaggaucgcaauuccCAACCUAUGGCCAACCUCUGCAAAAGAA

GGACAUGCGA

(((((((.(((((.((((.(.....)).))))(((((((....)))))))).....(((.(((.....))).)))..........)))))).)))) (-25.10)

potential miRNA gene = SPBC1709.06 potential miRNA targets = SPAC1093.03

AGUAUGUCUAGCUCCCAugguuaggauuggggaacuuCCCAUGAGGUUAUUAGCACUUCGAUAUGGCGCUAAUCUAGUAUGGGGGCCAGAAAUUGUGGAUAAAGCU

.....(((.((((((((((.(..(((((((.(.....(((((((((.....))))))...))))))).)))))))).))))))))).))).... (-30.20)

potential miRNA gene = SPAC17G8.03c potential miRNA targets = SPAC12B10.01c

GAAGAACCUACUGGUGAGGAAACUGAAGCAGAUCACGAGGAGAAAgccucuguugaucaugguAAUUUUUCUGACAAGACAACUAGUGAAGCCUCUUC

((((((..(((((((..((((((....(.(((((((((((.....)))))).).)))))).)).)....)))))).........)))))..))..)))))) (-22.10)

potential miRNA gene = SPCC330.01c potential miRNA targets = SPBC409.04c

GUUGCUUUGAACCUUACAGAGGCAUCUCAAGUGUUUAUGAUGGAUCCUUGGUGGAAUGGAGCUGUACaguggcaggcaauggaucGAAUUCAUCGUAUUGGACAAAAAAG

GCCGAUAAAAGUAAU

(((((((((((((...(((....)))))))..)))))(((((((..(((((...((....(((.....))))...))..)))))))))))))(((((((.(.....).)))))))..)))))))) (-24.70)

potential miRNA gene = SPBC1604.25 potential miRNA targets = SPCC1322.02

AGAAUUAUUGCCAAUGUCUACUGCCUCAAAACUUUGuaucgucggggccgguguCUUUAGUGUUUACAUGGUUUAUUUCGUACAUAAUAAUCAGAUUAUUGAACGAGAGGU

AAUAUUUACU

.....(((((((.........((((((...((.........)).)))))(.((((.............)))).)....(((((((...(((((((...)))))))..)))))))))))))..... (-19.40)

potential miRNA gene = SPBC1289.09 potential miRNA targets = SPCC1235.01 SPAC1805.11c SPBC1778.06c

GCCUUUAUGCCGUUUAGCCUCGUCACUAUAUACAUCUUCUUUGCAUCUUUUUGUAUCAACUAGAUGAUUcguacuauucaaaggcagagcgcUACAAGGU

((((((...((.((((.((((.............(((((....(((((.....)))).....)))))).....)))))..............)))).)))))...)))))) (-18.30)

potential miRNA gene = SPBC18E5.13 potential miRNA targets = SPAC13D6.02c

GUUUUucgaaaaugccagagccgGACUUUUAUGAAUUCUCAAUUUGCAAGUGCGAAACUC

...(((((....(((.(((...(((.(((...))).)))...)))))))....)))))... (-6.70)

potential miRNA gene = SPAP7G5.04c potential miRNA targets = SPAC17G8.15 SPBC9B6.07 SPBC16G5.11c SPBC3D6.11c

GAAGCUACUUUGGAGAAGCUUGCAUCGGCUGGUggucguggaggcgcuccaACUCAUUAGAUUUAUUAAGCAGUUAAAUUAUCCAAAAUGAAAAAUUUAGAAUCUAU

GUAGACUUC

(((((((((.((((((..((((.((.((((((...)))))))..)))).)))))))).(((((.(((...(((((....)))))...)))...)))))..............)))).)))) (-27.50)

potential miRNA gene = SPBC16A3.03c potential miRNA targets = SPBC16A3.07c

CUUGuugugcggucacuaaaccAGCUUAGCGUUCAUCAUGAAAAGGUAACUGGGUGCAGGCCUCGACAAG

(((((((...(((((.....(((((.(((.(.(((((...)))))..).)))))))).....)))))..)))))))) (-19.90)

potential miRNA gene = SPAC8E11.03c potential miRNA targets = SPAC1071.01c

GUUCAAUAUAAAAAUAAAGUGCAAUGGAAGAAUUCgcagaggggaaugaugaugaaC

(((((.(((..........((((((.....))).)))..........))).))))) (-6.80)

potential miRNA gene = SPAC688.10 potential miRNA targets = SPAC25G10.05c

UUCAAUCGAGUAUAUAGAUUGGGAAUUAAGUCCUUGUGAUCCAGCUUAUGAUUUCGUAGGGAAAGAUUUGccgacugcaaaugaagaguUAACCACUGUUCCAGUUAU

UCGAGUUUUUGGAUUGAA

(((((((.........(((((((.....)))))...)))))...(((((((....))))))).((((((((..(((((.....(((.(((.....))).)))))))...)))))))))..)))))))) (-25.90)

potential miRNA gene = SPBC23G7.13c potential miRNA targets = SPCC663.02

UCGGAAGAAAGGUUAAUGAUAUUGUUGAAGACGAAGGCGUUGUGGUGGAAACCAUAUCGAUUGGCUCUgguucgaaggagaaaguuaAUUUUGNNNNNNNNGACAUU

GAAAGUGUCUAAAAUCUGG

.((((...((((((((((....((.((((((((.((..((.(((((((....))))))).).).).)))...)))))..))....))))))))).........(((((((...)))))))).....)))). (-21.00)

potential miRNA gene = SPCC16A11.06c potential miRNA targets = SPAC343.17c

AGCGGUUUUAUAGCUUAUACAGCUAUCUUGACUCCCgaggguugaaguaugaagaAGUACAGAGGUAUUAUAACUCCUUGAUUCCUGAAAGUCGAGAGAGAGCUGGAG

CUCUGUU

.((((......((((((...(((((.(((((((((..(.((((.(.(((........((((((....))))).....)))).).)))))..)))))))))...)))))))))))))). (-30.10)

potential miRNA gene = SPAC1039.04 potential miRNA targets = SPAC12G12.04

AUCUAUAUCAuccauuucaaaggcgucaAUGAACUCAGAUGAAAAGUUAGACAAGAAGGAAUGGGAUCAGCAGCUGCCUAUUGAUUUUGGAGAAGGU

.......((.(((((..((((((((((...(((.((((.........................))))).)))...)).)))).)))))...)))))).... (-18.40)

potential miRNA gene = SPAC212.07c potential miRNA targets = SPAC25B8.14

CAGAAUACUCCUUCUGuauccgaagggcugugucgAUGUAGCCGGAUACGAAAGGCAAUUUUUG

((((((....((((.(((((((....(((((........)))))))))))).))))....)))))) (-19.40)

potential miRNA gene = SPBC839.12 potential miRNA targets = SPAC343.01c

GUUAAACUCUCUGUCACACUAGUUCGAACGUUUGAACCCCAUUCUUACAAUCAUCACAAUGCGACguggagcucguggaucuacaaaUUUGGGAGGAAUGACAUG

GCAAGAAGUACUUGAGUUUAAU

(((((((((((((((.....((((((....)))))))..(((((.......(.((((.........)))).)(((.((((((.....)))))).)))))))...))))).)))).....))))))))) (-23.60)

potential miRNA gene = SPBC1348.07 potential miRNA targets = SPBC16A3.07c SPBC2G2.04c

AUGAGUUAUUUAUGGUAAGAGAUACUCGUAAUGUGGACcuggagcggggacuugaaUUGUGUAAGCCUGAAAAGGUAAACAAACAAAAUCUCUUUACCAACAUCAU

.............((((((((((....((..(((..((((....((((..(((((.......)))))))))...))))..)))..))...))))))).)))))..... (-23.50)

potential miRNA gene = SPCC364.01 potential miRNA targets = SPCC584.01c

GGGGAACUAAAUGAUGUUGAGUUAAAAAUCCCUGCAAUUUCUCAUUCUUUAAAAUUAAaggaugguagcauucuuguuCUCC

(((((((.....(((((((.....((((((((((....))))))))))))))))...))))))) (-18.20)

potential miRNA gene = SPAC1A6.01c potential miRNA targets = SPAC23C4.08

ACACUCUAGUCAUGUCGCCGUGCCUGAAUCUAUUCUUAAUAAAAUAAAUCAGAAGUGGUCUAAaguucaggaugaugauggCAUGCCAUCGAUGUUAUAAGUAUGU

(((..((.((.(((((((.(((((((...(((.(((((.(((....((.((((....)))))..))..))).)))))))...)))))))....)))))))..))).))..))) (-20.10)

potential miRNA gene = SPCC16A11.11 potential miRNA targets = SPAC17G8.03c SPAC630.06c SPBC30D10.04

UUCUUCCUCUCCUUUUGAUUUUAUACGAAAAAUAUGCUGGGUGCAUUUAACUCGACUUUGGCUCgcuuuggcggucuuguucauaaAGUACCUUGGAGAUUAUCCCAG

CGAAGAA

((((((((((.....(((((((.(((.....((((.....)))).....))).....(((.(((....)))))).....)))))))).....)))))).....).)))))) (-22.90)

potential miRNA gene = SPBPB8B6.03 potential miRNA targets = SPAC17C9.13c SPBC582.03 SPAC630.06c SPBC428.03c SPCC736.04c

UGUUGGCUCUUUUUUUCCAACUUGGAGUAAAGAACCCUUGAUUCAUUAUUUUAAAGUAUUAUGGAUACuccgacuugggaagugguggCCGCA

(((.((((.((.(((((((((.(((((((.............(((((.((........)).)))))))))))).)))))))).)).)))))))) (-24.90)

potential miRNA gene = SPAC57A7.10c potential miRNA targets = SPAC630.06c

AGUUAUUCUAAAAAAGAUGAUGAUggugguauguugauuguuuGGAAUUAUGAGCUGACCUUUUAACAAUAGACGAAUCGAUAUUUGCCAAUGUCAAUCAAGU

CACAGUAACU

(((((((.....(((((((.((((((.(((((((((((...(((.(((.....))).)))...)))))).)))))))))))))))))).)))).....))))))) (-23.40)

potential miRNA gene = SPBC1348.07 potential miRNA targets = SPBC16A3.07c SPBC2G2.04c

AUGAGUUAUUUAUGGUAAGAGAUACUCGUAAUGUGGACcuggagcggggacuugaaUUGUGUAAGCCUGAAAAGGUAAACAAACAAAAUCUCUUUACCAACAUCGU

.....(((((((((((....((..(((..((((....((((..(((((.....)))))))))...))))..)))..)))...)))))).)))))..... (-23.50)

potential miRNA gene = SPAC521.05 potential miRNA targets = SPAC4G8.02c

UCUCUUUCUGCAAAAGGAUCAGGaugggguauuacgagagauUCCCGCCACAAGCGCUCUGCUACCGGUGCUAAGCGUGCUCAAUACCGUAAGAAGAGA

(((((((.(((....((....((.((((.(((.....))).)))))....((((((..((.....))..)))))).....)))))).))))))) (-23.80)

potential miRNA gene = SPAC27E2.09 potential miRNA targets = SPAC22F8.11

GUUCAAAUUCUCAAUGAGCUUguacaagucauuggacguUGCAAUUGAUUAUGCAAUUUCUCAGUUGGGAGAGUUCCAGUUCCAACCAAUUCGUACGCAAUCGAAC

.....((((((.(((....)))))))))((.((((((.....)))))).))...(((((((.....)))))))...((((.....)))). (-22.00)

potential miRNA gene = SPBC106.07c potential miRNA targets = SPAC630.06c

GGACGGAACUGGCCUUUGCUUACGGGAUUCAUAAAGCUCUUGUUGUUGCUAGAGUUCCUCUUACUGCAGCaauuguaccucccuuggcaAAGCGCUUCCGUGGUCC

((((.(....((((((((((...((((...(((.....(((((((.(..((((...)))))..).)))))))))...))))..)))))))).)))....)))) (-34.20)

potential miRNA gene = SPBC2D10.04 potential miRNA targets = SPBC14C8.12

UCUCCCUCCUCCCGCCAAACUCCAUAugaggggcgaauugaauaUUCCAGCGCUUAGUCGCUUUGGGAAAUCGAUAUCGGCUUCUUUACACCACCAAAAUGGUUCUUCU

CGAGA

((((.((((((.................))))))(((((((((...(((((((.((......)).)))))))..)))))).)))..........(((((.....))))).....)))) (-23.10)

potential miRNA gene = SPBC1348.09 potential miRNA targets = SPAC17G8.03c

UCCGUCUAUUCUCCACUUGAAAGUACUACCGAAGAACAAAUUCAUAACAUUUUCAAUACAAAUGUGUuuggcgcuuuggagguuauCAAAGCAAUAACUCCUAUCU

UCCGCUCACAACAUAAUGGA

(((((...........(((((((.....(((.....)))).....)))))))).....((((((((((...(((.((((((.....)))))).)))).....))).)).)))))))))) (-18.40)

potential miRNA gene = SPAC9E9.08 potential miRNA targets = SPAC13C5.01c

AGUUUUGacuuggagaguuuggguuCGGAUGAAAUAUUUGAAGGCGUGAAUUUAGAUGAAUUAGAGCAACAAGCUCAAACCCAGGUUCAAGCU

(((((.(((((((.(((((((.((((.(((....(((((((.....)))))))).))).)))))..)))))))....)))))))..))))) (-25.40)

potential miRNA gene = SPAC630.05 potential miRNA targets = SPCC736.04c

ACUGAUUGUUUUCUAUUCCUGUAAUGacaacgcuggaaucucugGAUAUGCCAGAGAUGACAGAGGUCGAAGAUGAUACAGU

((((((((((((.............(((....(((..(((((((((.....)))))))))..)))..)))))))))))).)))) (-27.30)

potential miRNA gene = SPCC613.06 potential miRNA targets = SPAC4F10.18

ACAUGGGCcgcgauauuuacaaggacgaGACAUUGACGAUCCCUGAGGGCGUUUCAGUUGACAUCAAGGCUCGUUUGGUGACUGUAAAAGGCCCUCGU

((..(((((...(((.((((..(((((((.(.((((.(.(.(.(((((.....)))))..)))).)))).)))))))).)))).)))....))))..)) (-29.70)

potential miRNA gene = SPAC1D4.12 potential miRNA targets = SPBC646.13

UGUGUGAUUUGAAACAUUCUUUGGAugcgggaggcauugcgcuCUUAGAAAUGCCUUCGGGUACUGGAAAGACCAUUUCUCUUUUGUCCCUUAUUGUUUCAUA

UCAACA

(((.((((.(((((((.....(((((...(((((((((.(.......).)))))))))).(((.((....)))))).....))))).....)))))))..)))))) (-28.80)

potential miRNA gene = SPBC2G2.17c potential miRNA targets = SPAC1834.01 SPAC4G9.17c

CCAUGGUUCCACCGCUgacagcaguggguuguacagucuCUGUCACCGGAGGCAAUGCCUACUUUGAGUUGUACUAAAUUGCGUGUUAAAAGUGGAUGCUUGG

(((.(..(((((...((((((((((....((((((.(((.........((((...))))).....)))))))))...))))))..)))))...)))))..).))) (-32.20)

potential miRNA gene = SPBC1348.13 potential miRNA targets = SPAC3A11.06

AGUUUUAACGAUAUAUUCCUUGAUAUGUUUgauuuugguauuguacguaaaaUUAAAGAUAAAAAUGCUAUUCUUUUUAUAAUAGCUAACUCUAAAGGAAUUGA

AAGAGUAAUAACU

((((...((.....(((((((..(((.((((((((((((...))).)))))))))).)))).....((((((........))))))........)))))))..)))) (-17.90)

potential miRNA gene = SPCC1322.11 potential miRNA targets = SPBC1734.06

AACUUGCGGGUUCGACGUUCCUUCUACAACCAAACACCAGCAGAAAAUUAUCACAAAAUGUCUCguggacgcggugcugccucuGGUACUAAGUAUAGAAUGACCUU

GGGUCUUCCAGUU

.(((.(.(((((((((.(((..(((((..((.....(((((((.........(((.....(((((...)))))..)))))))...))))).....)).)))))).))).)))))))).).))). (-20.20)

potential miRNA gene = SPBC1348.07 potential miRNA targets = SPBC16A3.07c SPBC2G2.04c

AUGAGUUAUUUAUGGUAAGAGAUACUCGUAAUGUGGACcuggagcggggacuugaaUUGUGUAAGCCUGAAAAGGUAAACAAACAAAAUCUCUUUACCAACAUCAU

.............((((((((((.....((..(((..(((((....(((((..(((((.........)))))))))...)))))..)))..))...))))))).)))))..... (-23.50)

potential miRNA gene = SPBC30D10.10c potential miRNA targets = SPCC645.03c

GGAGUAUUUUAGUGAUCUAAAAAACAAAAAUGAAAGCAUUcaacuagcagccgcugaucAAUUAAAAGAAUUUGUACAUAGUUCUACUAAGGAAUUAUCAGGUGAAUCC

(((((.((((((....)))))).))).....((((....)))))..((((....))))..............((((((...(((((((((.....)))))))))))))...))) (-14.00)

potential miRNA gene = SPBC428.01c potential miRNA targets = SPBC27B12.10c

UUGAAUAUGUUUCUUAUUUAGGAAAGGCCGGAGAAUGuagccuuucaaccccuaaUGGGCUUUUUUCUUUAUAAGUGUCAAUGUAUAAGAGAUUCAAUUCGA

((((((..(((((((((.....(((((((.............)))))))...(((....)))).....................)))))))))...)))))) (-20.90)

potential miRNA gene = SPBC16E9.11c potential miRNA targets = SPAC29B12.07

UUUCUUCGAAAAUUCUAAAGUAGGAGAAGUAUAAGGAGACGAAUAUAUAGAUGCAGAAUCUGAUGAAUUUaggguuuggaaacgagcaUGCUUGGGAAAUCGUGUA

GGGGAAGGAGUGCGAGAAGAAA

(((((((....(((((....((.(.((....((((.(.........((((.....)))))...........(((((....)))))).).))))).....)).).))......))))))....))))))) (-18.30)

potential miRNA gene = SPAC6F12.16c potential miRNA targets = SPAC1F5.11c SPBC409.04c

CAAAAAAUGGAAGACACUAUCGCAUGCAUACAUcgggauauuguuuuuagugCAUCACUAUAUUUGUAGAAUUUAUUCAGAUUUUGCUUAAUAUUAAUUUUUG

((((((...(((((((.((((.((((....))).).)))))..)))))))(((((..))))(((((.((((((((((....)))))))))..))))))...)))))) (-18.70)

potential miRNA gene = SPCC1442.02 potential miRNA targets = SPAC57A10.09c

GUAAAUUAUCUUUCGUCCGUUUACCUCACACCUUAAUAUUCAUAAGAAGAAUGGACUGGAAAGAAgcgauuccaggugccgccAAGGUAGCCAUUGAAACCAAAGAUUAU

AUUAC

((((((.((((((.(((((((((..((.....................))..)))))))))(((((........)))))(.((((.....)))).).........)))))).)).)))) (-19.90)

potential miRNA gene = SPCC1919.04 potential miRNA targets = SPAC17G8.03c

CACUUACAUCCAGAGUGCUAAUUAUUAAGAGCACUAAUUUAAAACUUUUUUGUUUAAGACAuggaucucucguuuauugagGGCUUUGAAACUAUAUGGACUGUAGU

CAGAGCUGUAGUG

((((.(((((((.((((((..........))))))......((((......))))......))))))..(((((.....))))(((((((..((((((.....))))))))))))))))) (-25.40)

potential miRNA gene = SPAC6G10.08 potential miRNA targets = SPBC16H5.14c SPAC25G10.05c

CUGACACAAUauguuggaaguacgugcugccGUAAAGGCACCUUUCAAGUUGGCUGCCGCAGGCCGUGGUUUCAUGAAUAUGCGUAUG

.......((((((((((.((((..(((((.....((((((.........)).)))))))..)))))..)))))).....)))))). (-24.30)

potential miRNA gene = SPAC24B11.10c potential miRNA targets = SPAC23H3.07c

UGGGCGCAUAAAGCGGCUUGCAAAGGUUUGGCUAAAGCUCAGUAUGCUGUUGGCUUUAUGAUGGAACAAGgcauugguguugcugcugAUCCAUCCUCUGCUCA

((((((((((((((((.(((...((((((.....))))))..)))).)))....))))))))((((((.((.((((.......))))..)).)))))....)))))) (-29.90)

potential miRNA gene = SPBC1709.05 potential miRNA targets = SPCC1235.02

CAAUAACCUACUUCUCUCAAAAAAAUCGAGGGUCAGUUUUAACUUCUUGUUGCUUGAAAaugagugaaguuuaugaagguGCUAUUG

((((((((((((((.((((.....((((((.(.(((..........))).).))))))...)))))..)))))).....))))..)))))) (-16.10)

potential miRNA gene = SPAC17H9.14c potential miRNA targets = SPCC320.12

GAUUUUAAAUGAUGAAAAUUGGGUACAUAAAGAAGCUAAGCGUUUGUCUAAACUUUUACGCCAGAAGAgcaucgcuuuggccagugcuGAUGAUUUUAAAACUCGCCUUAA

CAUUCUAAAUU

(((((..((((.....((((.(.(.....).).))))((((..((....))....))))....(((((...)))))(((.(((..(((.....))))..))).)))....)))))..)))))) (-16.20)

potential miRNA gene = SPAC4G9.09c potential miRNA targets = SPBC11C11.06c

GACAAUUUACUUAAAGGUGCUGCUACUCAAGCUCUCCAGAAUCUCAAUCUGUCCUGUGGUUACGAUGAauaugccgguauccauuuggauUGAUUAUAUUUUUUGUAGCCCA

UUCUUUUGUU

(((((.........((.(((........))).)).((((.......)))))...((.(((((((((.(((((((.....((((....))))....)))))))).)))))))))).....))))) (-24.40)

potential miRNA gene = SPAPB2B4.07 potential miRNA targets = SPAC17G8.15

GAUUGCCCAAACUGUACACACCCAAUCCUcugcugacaaaagaagaaguCGAGGUACGUCGAAAUGAGUUUUGGGAAACCUGCUGGGCUUAUGGUGGCUCAAAAGA

AAUAUGGGAUGUU

(((..((((.........((((((..(((...((((.....................))))....)))).....)))))....((..((((((.....))))))..))).....)))))..))) (-24.60)

potential miRNA gene = SPAC4G8.04 potential miRNA targets = SPBC839.09c

UCUUCCCCUUAUUCAUUUGUUUUGGAaugcaaccaucuguuucugaCUCUUCAGUUUACCUUUCUUCCGCUCCCACGAAGCCCAUUGCUAGUGGAAGUGUCGCUACAGG

GAUAGA

(((..((((....((((((.....)))))).............((((....)))).............((((.((((..(((.....))).)))))))).....)))))..))) (-23.00)

potential miRNA gene = SPBC1348.09 potential miRNA targets = SPAC17G8.03c

UCCGUCUAUUCUCCACUUGAAAGUACUACCGAAGAACAAAUUCAUAACAUUUUCAAUACAAAUGUGUuuggcgcuuuggagguuauCAAAGCAAUAACUCCUAUCUUCC

GCUCACAACAUAAUGGA

(((((.....((((((((.....(((.....)))).....)))))))).....((((((((((...(((.((((((.....)))))).))).....)))).)).)))))))))) (-18.40)

potential miRNA gene = SPBC405.07 potential miRNA targets = SPBC23E6.03c

AUGCUACCAAAAGUUUUUUCCAGACGGUAGGUACAGUGUAGACUUUACACGauggcaccuggacuggguuGUUGGUUUAAACAAGGGAAAGACUUUGACUAAGCGU
(((((....((((((((((...(((((((((.(((((((.....)))))..))..))))..)))).(((((.....)))))))))))))))).....))))) (-28.10)
potential miRNA gene = SPAC343.17c potential miRNA targets = SPBC16A3.12c
AAGCACGGCGUCGGCGCAAUGCAGUACGCAAAGAUCCUAAGCUUUCUCGGCAGCCUCAAGUUGgucgcuugcuagaagaaaaCUCUGUCGAUGACAUUCCUUUGGCAAC
CAUGCUU
.((((.((...(((((.(((((..((...((..(((.....)))..))...))).....)))))))).(((((((((.(((.....((((...)))))))..)))))))))).)))). (-25.90)
potential miRNA gene = SPCC1223.15c potential miRNA targets = SPBC17D1.17
GCAUUUCAACGAUUGGAAGACUCUAUUUCAAGCUUACAGAGACAAGAGGAUUCGCUUAAAACAAAAUACGagcuucaggaagcucgguUGGAUAUGC
((((((((((.....(((..((((.(((.....)))))..))))..))).....((((((((....))))))))))))))).)))) (-23.90)
potential miRNA gene = SPAC13F5.06c potential miRNA targets = SPBC577.12
GCAGUCAUGCAAGACAAGCGUAAAGCAGGUUUUAGUGAAAUGAACAUUCAUGAAGUUCUCCGUCUacguuuugauuugccugaaaACUUGAAGUUGCAACUGAACAAAG
AAGAA
GCUUUAUUGC
....(((.(((((.(((((((..(.(((.(((.(((((.....)))))).))).)).)...)))).)))).)))))..))).....(((((((....((.....)).....)))))))).... (-21.50)
potential miRNA gene = SPAC17D4.01 potential miRNA targets = SPBC646.13
UGUCAUUGAGGACUCCUGGGUUUCAAGGUAAGCUAaauauaagggaagcgaagcaagaca
((((.(((..(..((((..((((..((.....)))))))...)))))..)....)))))) (-8.90)
potential miRNA gene = SPBC1861.01c potential miRNA targets = SPBC1815.01
CAUGCGACCGACACCUUGGAAAAUAAAAGAaggggaauaggggauuuucCAAACGAACGAUGACAAUACUAAUAGUGUGUUAUGGAUUUCGAUGUACAAAAGUCAUUU
AGUCGUUAUG
(((((((((.....((((.............))))(((.........))))...(((((.(((((((.((((...)))))))))).)..))))(((.(.((....)))))...))))).)))) (-22.50)
potential miRNA gene = SPBC713.10 potential miRNA targets = SPCC622.02
UUUGUUCAAAAUUACUAUAUUACAUUAAAAAAUGUCGcuuccaagggcugucggucGAUUUAUAAUUGUUGGUUCGCAAGGUAUGUUUGUGGCCAUUCUCCUAUCAA
GGAUUUAUUCAAA
.....(((((((.(((.(((((((..((...))..).))).)).)).))).)))).)))(((((.((((((.....)))))))))).....((((....)))).......... (-14.50)
potential miRNA gene = SPCC1795.03 potential miRNA targets = SPAC17C9.12
AAGAGGCACUUACAUUCCAAUGACAACCCAAGaugcugcugcuaaggacguuGAUCAUAAGCAUUAAUUUAAGAUGUUCUGUAAUUCUCAAAGAAUGUAUUCUUUUU
.(((((....(((((((..((((((((.......((....)).......)))).)))).((((((.......))))))..............)))))))..))))). (-16.40)
potential miRNA gene = SPBC1289.16c potential miRNA targets = SPAC3F10.09
GGUCAUGUUUAUGGAGGAUAUCAUAAGCAUUCUGGUAGUGgaggccacuuggucgacaUGAUGAAAAAUAUCAGUGAUGUUACUCAUGAUUUUGCUAUGGGAAAU
UUUAGAUAUCAUAAGUAUGAUU
.((((((((((....((((((.((((..((((.(((((((((((....)))))...((.((((..(((((((...)))))))..)))))..))))))))))..)))))))))))))))))). (-32.70)
potential miRNA gene = SPBC14C8.05c potential miRNA targets = SPBC56F2.12
GAUGGUCACCaaaaaggugcaagagauuuGACAUGGAGUUAUUCAUCCUUAUUAAAUGCCAUC
((((((((((.....))))......(((((((...(((.........)))...)))))))))))) (-14.70)
potential miRNA gene = SPAC1002.07c potential miRNA targets = SPAC23H3.09c SPAC4G8.02c
GAGCACAAAAAGUUGGAGGAUGGAGUAUGAUGAGAGUCACcggcgagaaccuaaaggcucUUGCUGAUAAACUUCCCGGCAACUUUUAAACUCAUACCCAAUACU
UAACGCUC
((((.......((((.((..(((.(((((((((((((..((((.(((.........(((....)))).......)))))))..)))))))))...)))))))))).)))) (-31.40)
potential miRNA gene = SPCC1235.16 potential miRNA targets = SPAC4F10.18
GUUGGUACUUCUCAUAUCCGCCUCGCGUGGGGUCAUAAUACGCUUCCUGUAUCUGCCUUGUCUCAAUCUcaaucccaaguuucagauGAGGGCUUCGACAGAACUCUUU
CCGCUAAC
((((((.............(((((....)))))).............((((....(((((((((.((.((.........)).)).)))))..))))))....)))).........)))))) (-21.00)
potential miRNA gene = SPBC2A9.10 potential miRNA targets = SPAC1A6.09c
GGGUGAUUUGAAAAACCUCCUCGACUUGGGAUGUCGAAUUUUCAACAUGGAAAUUAUCACAGUUAUUACagcaugaggggaggaaccUCAAUUAUCGAUCC
((((((((((....((((((((((.......)).(((((((...)))))))))..(((..(((.....))).)))))))))....)))))..)))).)))) (-21.60)

D.1.3.2 sRNA data for Halic and Moazed (2010)

potential miRNA gene = SPAC31A2.16 potential miRNA targets = SPCC126.15c
AAGGAGAAGAUUGAGGGGGAUGAGCUGUACGAUAAUAUUUugaaggaaacuuuuaaugaagaaUUACUCUCUCAUUACCCACCAAAUAUUAUUUAUGCGACCUUU
CAAAAAUAUUUGAGCUCCUU
.(((((....(((((((.....((.(((.((((((((((.(..((.....((((((((((.....))).)))))))))..).)))))))))))))..)))))))).....))))). (-24.90)
potential miRNA gene = SPBC582.07c potential miRNA targets = SPBC28F2.11
CGGAUUUUUUCGAACCUUGGCCGAUGUUGAAGUUAAUCAUCUUAAAUGUGAUCAAUUUCUUGUCGCCCAUuauagauacuaugugcgugaaauGAGAAGACGUG
CUUACGCCCAGUUACUUGAAUCUUACCG

.((((((.....(((..(((.(((((..((((((.(((((.......))))).)))))).))))).)))............((.((((((.(((......)))..)))))).)))))....))))))..... (-25.20)

potential miRNA gene = SPAC29A4.14c potential miRNA targets = SPAC23C4.08 SPBC1685.03

UGUAUAAGGAAUGCUUUUAUGGAAUUUAUGGAAACACGAAAUUCCAAGUCCAAguaagugucuucuauuacuuaacGGAAUUGAAGCUAAUUAGAUGCA

(((((.......(((((..((((((((.((....))..))))))))..(((((((((..(....)..))))))...)))...))))).......))))) (-19.70)

potential miRNA gene = SPAC1687.15 potential miRNA targets = SPBC23E6.04c

CAUCAUCCAUAAUGAAUCAUGGUACUAAAAGUAAGUUGGCuaacuucgauagugcuguuguUCUAACCGUAUAAAUAGUUCCUGUGGACCCCUUUCGUAUUAUAAAGG
AAACCGCUCGUGAUG

((((((.......((((.(((((((((.((((.((....)).))))...))))))))).))))...................((((...(((((........)))))...))))..)))))) (-23.30)

potential miRNA gene = SPAC1142.02c potential miRNA targets = SPAC23H3.07c SPBC19G7.07c

UGACCCUAAUAUUGCCAACAUGGCCCGUAACUUUCAaucuggaggagguaugccugacUUGAGUUCUCUUGCCAAUAACCCCCAACUGCAAAAUUUGGCUCGCAAUUUUAUG
AACA

......(((.(((((...((.(((..(((.(((((........)))))))))))))....(((....)))(((((..................)))))..))))).)))...... (-15.00)

potential miRNA gene = SPCC1450.02 potential miRNA targets = SPAC17G8.03c

GUGGCAGUGGAAUGCCUCCGCCUcaacaaaaauauuguuuggcaauUGUAAGACAAUUAAAAAGAACCAAAAAUUCCGCACCAU

((((..(((((((......(((..(((((.....))))).)))((((((...))))))............))))))).)))) (-17.00)

potential miRNA gene = SPBC26H8.08c potential miRNA targets = SPAC23H3.07c

GAGGCUGCACAAGCGUCUAACCCAGAUGAUGAGGAAgaguuugugauggaagaagacaACCUUGGAGAGGCCCCUUUAUUAGUUGACUCCGAGUCUUAUGAGGCAUCAGU
GAAGGCAGAUACCUC

((((((((.(((((.(((..((((.....)).)).))))))))))(((((....(((((....((((((.(((..........)))..))))))))))......)))))......))))...)))) (-31.10)

potential miRNA gene = SPAC4A8.11c potential miRNA targets = SPBC30D10.04

CGACAUUUCUAAAACCAACGAAACUACACGUACUCUACAAAGCCUAACAACCUCAUUGACCAACuccaaugaaaauguuggcguggauGUUG

................(((........)))(((((((...(((.((((...((((((........))))))...)))))))))))).)).. (-17.20)

potential miRNA gene = SPBC2G2.03c potential miRNA targets = SPBC2G2.02

UGAAAUUAUACACUGAUGAGGCUAGUGGAUUUAAAGUguaagccguuaaacucaaucauUAUUAAGGUUUUUCUAACGUUAAUUAAGUGACCCCGUUGUAGUGAUGGUAUU
AAGCGUUGGUUUCA

((((((((.((.((..((((..((((((.((((.....)))))))))))..)))).((((((((((.(((.....))).)))))..)))))..(((((.....)))))......)).)))))))))) (-23.30)

potential miRNA gene = SPAC1A6.04c potential miRNA targets = SPCC736.04c

GUGCCGUUGUUAAGCGUACuuuggagcguaacaacaucaCUGCUAGCCCCGAGUGCCAACAAUGCUACUAUAACUACUGCUGGAGCGGUCUUUAUGAUGACAGCGC

((((((((((..(.((((((.((.((..................)))).))))))))))))))....(((((..(((((....)))))..)))))......)))) (-24.40)

potential miRNA gene = SPAC9.09 potential miRNA targets = SPBC21B10.04c

AGCUUACUGAGGCUGGUGCAGAGUACAUCCAAAUUGAUGagccuauccuuacucuugacuugcCCCAAGAGAUCCUUGCUAGCUACAAGGAAGCUUAUGAGACUCUUG
GUAAGAUCGGUAAGCU

(((((((((((((.(((..((((((((((......))))..........))))))..))).)))(((((((...((((..((((.......))))..)))).)))))))).....)))))))))) (-41.70)

potential miRNA gene = SPCP1E11.09c potential miRNA targets = SPAC1093.02 SPAC4F10.18 SPBC4F6.10 SPCC285.08 SPAC23H3.07c SPAC630.06c SPBC17D1.17
SPCC1620.06c

UGCCGAAGAAGAAAAGAAGGAAGAAGCCAAGGAggaggaagagucugaugaggacAUGGGUUUUGGCUUGUUUGACUAGAUCAAUUAUACGAAUGAGUAUCUUUGUGCC
UUAAUUUACGUCGCA

(((.((.....(((..((((((.(((((((((........(.((((.....)))).)...))))))))).)).........(((..((((......))))..)))..))))..)))...))))) (-20.10)

potential miRNA gene = SPBC530.02 potential miRNA targets = SPAC2F3.11

GUCUGGCUUAGGAGCGCAAAUUAAGCGAUAAGAAAGAAAUGGAUCUGUAUCUUUUAGCAAGCGGCAGAAAguuugggguuguaaagucUGAC

(((.((((((.((.(.((((((..((.....((((((.((......)).))))))).....)).......)))))).).)).).))))).))) (-18.30)

potential miRNA gene = SPAC56E4.06c potential miRNA targets = SPAC30D11.05 SPAC4G8.02c

CCAUCUUCUUAAAGAUCCAAUuuggcagccuauccuuaugccaaaUGGGAAAGUACUUAGGGCUGGCGAUAAAAUGUUUCGUCCGGCAUAUGCUAAGACUUUGG

...((((....))))(((.((((((((...........)))))))).)))((((.(((((.((((((((.........))).))))).....)))))))))... (-29.40)

potential miRNA gene = SPCC1682.11c potential miRNA targets = SPAC4F10.20 SPBC56F2.12

UAUUCCAUUCAUGAAAAUGGUGUCUUUUAUGGCUACAUUGUAGGACUUUUAGGCGGAUUCCUUGCUUGGCuuaucauuggcgccauagagGGCGGCCUCAGCAUGAU
GUGAGUA

((((((((.((((.....((((((((((((((..((.((((((.((...((((((.....))))))))))))).)).))..)))))))))))).)))....))))))).))))) (-39.10)

potential miRNA gene = SPBC1709.02c potential miRNA targets = SPBC115.01c

AACUUUUGucaauugguuucuaauaauugaaGGAUUCUUCCGCACCUGGUUCUGGUGAGCCUAGGCCUAAAACCGAGAAAGAAUGUGAGUU

.((((...((..(((((((............(((....)))...((((((((....))))).)))....)))))))....))....)))). (-19.90)

potential miRNA gene = SPCC576.16c potential miRNA targets = SPBC21H7.07c

GCCUAAUGAUUUUUAUUUCUUUAUGGUguaagugacacgauaggacuauUCUUACGUUCAUUCACUAACUCAUUAGGC

(((((((((..............(((((..(((((.(((.(((((....))))))))))))))))))..))))))))) (-20.90)

potential miRNA gene = SPBC30D10.15 potential miRNA targets = SPBC16G5.06

GGUUucugguaaucaagccuugccuggCGAAGCAAAUUUCCAGUCACU

(((..((((.(((...((.((((...)))).))..))).))))..))) (-10.60)

potential miRNA gene = SPCC5E4.06 potential miRNA targets = SPAC23C4.08

GAUAUUGAUGGAAAGAUAUCUUCAAAAGAAGAAGugauugggagagcuaaaggugaaACUGACACUACAAAGUCUAAAUUUGAAGAUAUUGUGAAAACAUUUGAUGG

GUAUC

((((((.((.(((.((((((((((((...(((((((.((((....(((...)))....))))))))......)))...))))))))))))((....)).))).)).)))))) (-22.60)

potential miRNA gene = SPAC1610.04 potential miRNA targets = SPCC364.03

UGCCUGAGUCAAAAGaauugaaauaugccauggaucACAAUGGACUAAACCCACUUUAUAUCUCGUUUGGCA

((((((((...((((............((((........))))..........))))....))))...)))) (-10.30)

potential miRNA gene = SPAC3A12.10 potential miRNA targets = SPAC57A10.09c

CAAGCCUCUUAAGGCCAAGGUCUUUGGUAUCUGGAUUCGUUAUGACUCUCGUUCCGGUACUCACAACAUGuacaaagaguuccgcgacacuACUCGUGUUGGUGCUGUCG

AGGCUAUG

..((((((..(((((....))))).((((.(((((..((..........)).)))))))))....(((.((((...........((((((.....)))))))))))).))))))... (-29.90)

potential miRNA gene = SPAPB1A10.10c potential miRNA targets = SPAC4C5.01

ACUAGCUCAAAUAUGGUUAACGAUCUUUACAUCAGCUucucuaugaguauugcuuuggCAAUUUGUUAACAUCGUCAUUUUAGUUCGU

((.((((.((((((((((((((......((..(((..(((...)))....))).))......))))))))..))).)))).)))).)) (-13.80)

potential miRNA gene = SPCC576.16c potential miRNA targets = SPBC21H7.07c

GGUGUGCUAGUUGCCUAAUGAUUUUUAUUUCUUUAUGGUguaagugacacgauaggacuauUCUUACGUUCAGUCACUAACUCAUUAGACUUCUAUGAAACUUGGAUCAA

GGCAGUCAAAGUAACU

(((.((((.(((((((..((((((...((((.....(((.(((((((((((.(((((....))))))))...))))))).......)).)))......))))...)))))))))))).)..))))))) (-28.70)

potential miRNA gene = SPBC1685.13 potential miRNA targets = SPBC1734.06 SPAPB1E7.01c SPAC343.01c

CGUCGGUCGCUUCCUUCCCACCAUCUUGGGUAAUccuugguugauugccuuuuaCGACUUUGUCAACUGGGUCUUUGCACUCACCGGUGGUUGCUGUAUCGCCGUCGC

CAUUCGUGUCCAUG

....((.(((......((((......))))..((((..(((((((................)))))))))))............(((((((......)))))))..........))).))... (-25.60)

potential miRNA gene = SPAC25G10.09c potential miRNA targets = SPBC19G7.07c

UUCGCUAAAAUGCAACCACCAGCUCCUUCUCCUCUUCAACCUAGUGGUCACGAUUCUGACAAUUGGAGucaacauggugaugaagaAGAAGAAGAUAGCGAA

(((((((....((........))..((((((.((((((.((....))((((.((..((((.......))))..)).)))))))))).).))))).))))))) (-26.10)

potential miRNA gene = SPBC23G7.06c potential miRNA targets = SPCC584.01c

UGGUUUAAaguucgaaucucuuucuaacGGUGGUGAAUUAAUCGUAAGCGCCGAUAUUUGGUACGAAGGCGAUUGCAGUUUUAAGGCA

((.(((((((((..((((.(((((((.(((((.(............).))))).)).........))))).))))..))))))))).)) (-17.40)

potential miRNA gene = SPAC23H4.03c potential miRNA targets = SPBC1734.06

UGuuuagaucccuuucuuucuuuGAAAAAGAAAUCCUUUAUGAGUUUAGCUUUCUGACUGACAUUUAAAUGUAGGUGCUAUAGAGUCUGAACA

(((((((((...(((((((.......)))))))........((((...))))(((((((.((((....)))).)))....)))))))))))))) (-13.60)

potential miRNA gene = SPAC17A5.02c potential miRNA targets = SPCC777.15 SPAC323.07c

AGAAGUUGUCAAAACGUAAGCAUUGAAUUACUUACAUCCUCUGACGUUUGAUAGGUUAAAAAGAAAAGAGgaaauaugguugaugaaaacuUGGUUAAAAAAGACAAG

UUAGGAAUUCCUAGGAACUUUU

.((((((.(......(((((..........))))).((((((....(((((....)))))......))))))....((((........((((((...........))))))........))))).)))))). (-21.20)

potential miRNA gene = SPBC1734.06 potential miRNA targets = SPBP16F5.04

CCUGUUUCAAAAAGAAAUUUAAUUCGCCAACUUAUAGAUUGGGAGCGCGUUCAGUCGAAAUCAAUuggaguagaaaaagaaaaaCUAGGAGGGGGCGAUUGGGAGAAGGC

AUACGCUGAGG

((((((((.....)))))).....(((((..(((.(((.((.....(...((((.((.((.....)).)).).)))...)...))))).)))..))))).)))....(((....))).... (-19.40)

potential miRNA gene = SPAC11G7.01 potential miRNA targets = SPBC17D1.17

CAUGUCGGCGUAGUUGUUGGUUGUUCUGUUGCUAUuccuguugguguaguucuaauuuUGAUUGGUUUAGGAAUUUUUCUUUGGAAAAGGCAUCAACGGUCCAAACGAAUU

AAGGCUGAGCGCAUG

(((((..((.((((((((.(((...((((((((((.((....)).))))((((((...........)))))).((((((....))))))....))))))....))).)))...))))).))))))) (-29.70)

potential miRNA gene = SPBC29A10.13 potential miRNA targets = SPBC1734.06 SPBC9B6.09c

GAUUGUUGAAACAAUGGUUACCAAGGGCAAAUGGGUCAUCCCAGGCUACCGUGAAAAGUUUGGUGAUuugaguauuauguaguuugcucCAUUAAUGAAACAUUUUAUAA

CUAGCAAUC

((((((((.......(((..((.........((((....))))))..)))((((((.((((.((.....(((((..........)))))......)).)))).))))))...)))))))) (-23.50)

potential miRNA gene = SPBC365.14c potential miRNA targets = SPCC16C4.19

GUCAUUCAUUUUGCUGGUCUCAAAGCAGUUGGUGAAUCUGUACAGGUUCCUUUGAGUUAUUACAAAAAUaacauuuccgguaccauuaauuUAAUAGAGUGCAUGAA

GAAGUAUAAUGUACGUGAC

(((((.((((.((((..(((((..(((.((..(((((..((...(((.((...((((((((.....)))))).))...)).)))))..)))))...)).))).)).))))))).))))...))))) (-22.30)

potential miRNA gene = SPAC167.07c potential miRNA targets = SPAC1A6.09c

GUUUGAUGUUCAAAAUGUCUUAAAAAAUGAUUCCGGCUACUGGUUUCAACUACAAUUACUAAUUGAUguuuacucacgcaugcuuuuuaCCAUGAUCGACGAUGAGU

UUCAUAACGAUAAGC

.(((((...)))))............(((..((((...))))..)))..........((.((((.(((..(((((((((((........))))).....)).)))))...))).)))).)). (-16.50)

potential miRNA gene = SPAC22F8.10c potential miRNA targets = SPBC9B6.07

GUCCAAUUCCGACUGGAGCGCAGUGGGGAUUUCAUCCugguggguuggggaaaaccUCCUGUGGAC

((((((((((.((((.....)))))..))))))).....((.((((.....)))))..))..)))))) (-22.60)

potential miRNA gene = SPAC4F10.20 potential miRNA targets = SPAC17G8.03c

AGAAAAUUAAGGCUCAGGUAUACCAGAUUGACCuuaugaacaacggugaugagauuCAAUGUAAGUUUUUAUUGCAUCUGUUGUUUUGCUUGAUGCUAGUUUAGAUCU

...(((((..((((((((((.....................((((((((((.....)))(((((.....)))))))))))))))))))))).)))))))..... (-20.20)

potential miRNA gene = SPCC1322.10 potential miRNA targets = SPCC24B10.18

GAUCCCGAGGAGAUGGUUAUUAUGUUGGUGAACAAugcucauuaucccaaccaaaACUUUAAUCUCGGAACUGUCCAAUCCUCCGCUGGAUCUUUGGAUACCG

ACAUCUCCAUCUCUUCUGAUC

((((..((((((((((.....((((((.((.............)))))))).....((((...((((((((((.....))))...)))))))))).....)))))))))).)))) (-35.60)

potential miRNA gene = SPAC631.01c potential miRNA targets = SPCC1906.01

GCAGGCGCCGUUCUCAUAAAAAAGGGUAUUUUUAUUuucugucaaauuucauuauauaaauUCUAACUUCAUUUAGCUUCUACUUCUAACUCAAGCGGAUGGGAUUCU

AUUCACGUGUUUGAAUGUCUGC

(((((((((.((.........)).)))..................................((((.....)))).............((((((((((....))))))....)))))..)))))) (-17.50)

potential miRNA gene = SPCC16A11.01 potential miRNA targets = SPAC6F12.04

GUCGUAGGUAUCAUUUGCGCCAUUGCAUUCGGUGGUCuucaucacaauuaccguggucgUUCAGCACGCUUUGAAUGGGUUGUGUGUUUUUUAUGGUCCAUUUACAU

UUGCUUAUUAUGUCUUGAC

((((.(((((((.((..(((((((((....)))))))).................((((((((..(((((((.............)))))))....)))))..))))..)))..)))..)))))))))) (-23.70)

potential miRNA gene = SPAC22F8.09 potential miRNA targets = SPBC16G5.06

GACUacaagcuuugguguauaaaaaUAAAGAUGGUUUAACCGCAACUUCUGCUGCUAGUGAGCUCAAAAAUACCAAACAACCAAAGGAAAGUU

((((.....(((((((((.....................((((((...(((.(.....).)))...))))))).............)).)))))))...)))) (-14.20)

potential miRNA gene = SPCC320.09 potential miRNA targets = SPBC1A4.10c

GAAUAUGAUUCUCGUUCCUAUUGCUUUUACUUcugaucacauugagacucugaaggAGUUAGAAGAUUACAUUGAAGAUGCAAAGCAAAAGGGAAUUACCGGAGUUAAAC

GUGUUU

.....((((((.(((((((.((((((((((((((((...((..((((((...))))))..))))))))....)))))).)).)))))))).)))))).....)))))))..... (-25.70)

potential miRNA gene = SPBC1734.04 potential miRNA targets = SPCC16C4.19 SPAC3A11.13

UUGUUAAUCUCUUCCACAAUGAAAGCAAAUAGAGACUUuggaagagaugguggcccuuuuaGUAUUACGCCGAAUCGAUUCCAGCCUAAAUCCAGCGGAAAUCCC

AUAUUCCGACAG

......((((((((((.....................))))))))))(((((.............)))))..........................(((((........)))))).... (-21.10)

potential miRNA gene = SPAC806.06c potential miRNA targets = SPAC57A10.09c SPCC553.02

CAGGAGCAGACUAAUAUGGAAUUAGUUGCUGGUUACUUUUCUCCGGUUAAUGACCACuacaaaaaagaaggccuugcaccaGCUUAUCAUCGUG

((.(((((((.((((((.....))))))))).(((((((.((..(.....)..)))))))).........(((..((.......))..)))....)).)) (-13.80)

D.2 miRNA candidates for experimental verification

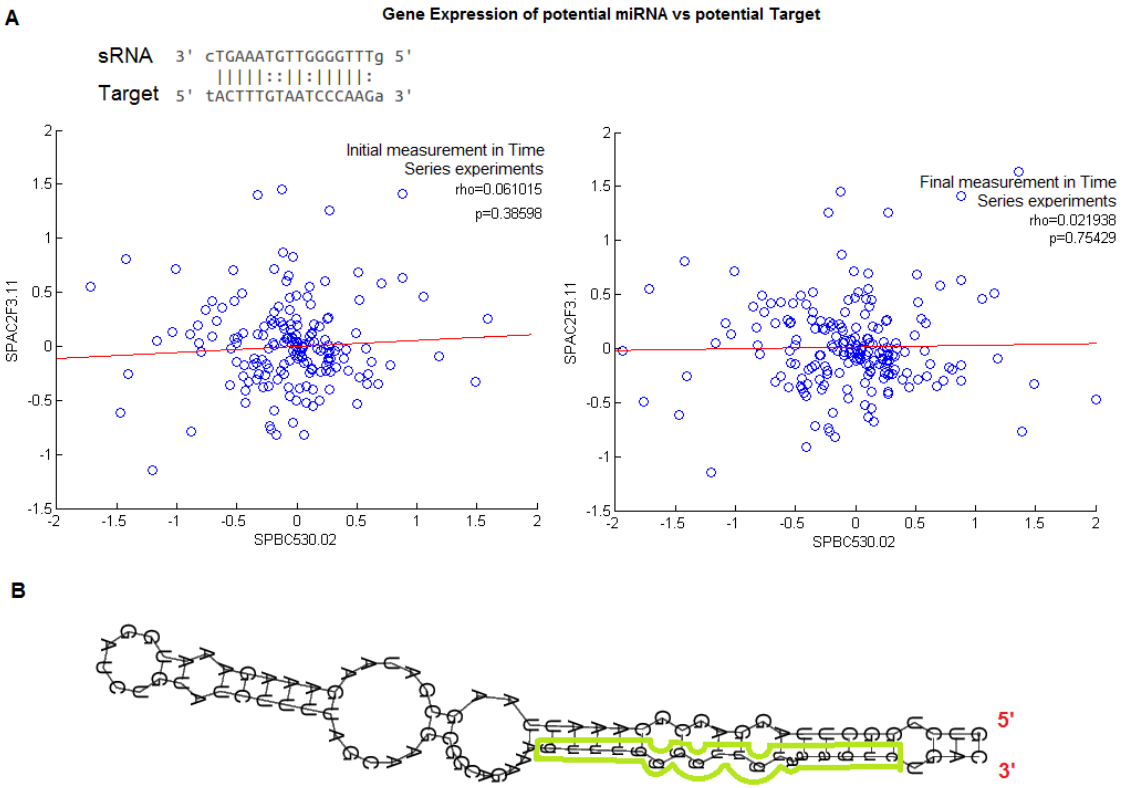

Figure D.1: Correlation of gene expression between the predicted miRNA *SPAC530.02* and its potential target *SPAC2F3.11*. The sRNA (potential miRNA) to target binding profile predicted by miRanda is presented above the left scatter plot. Left scatter plot presents gene expression data based on the first measurement of each time series experiment, while right scatter plot is based on the final measurement of each time series experiment. **B** The predicted hairpin structure of *SPAC530.02* with the predicted mature miRNA outlined in green.

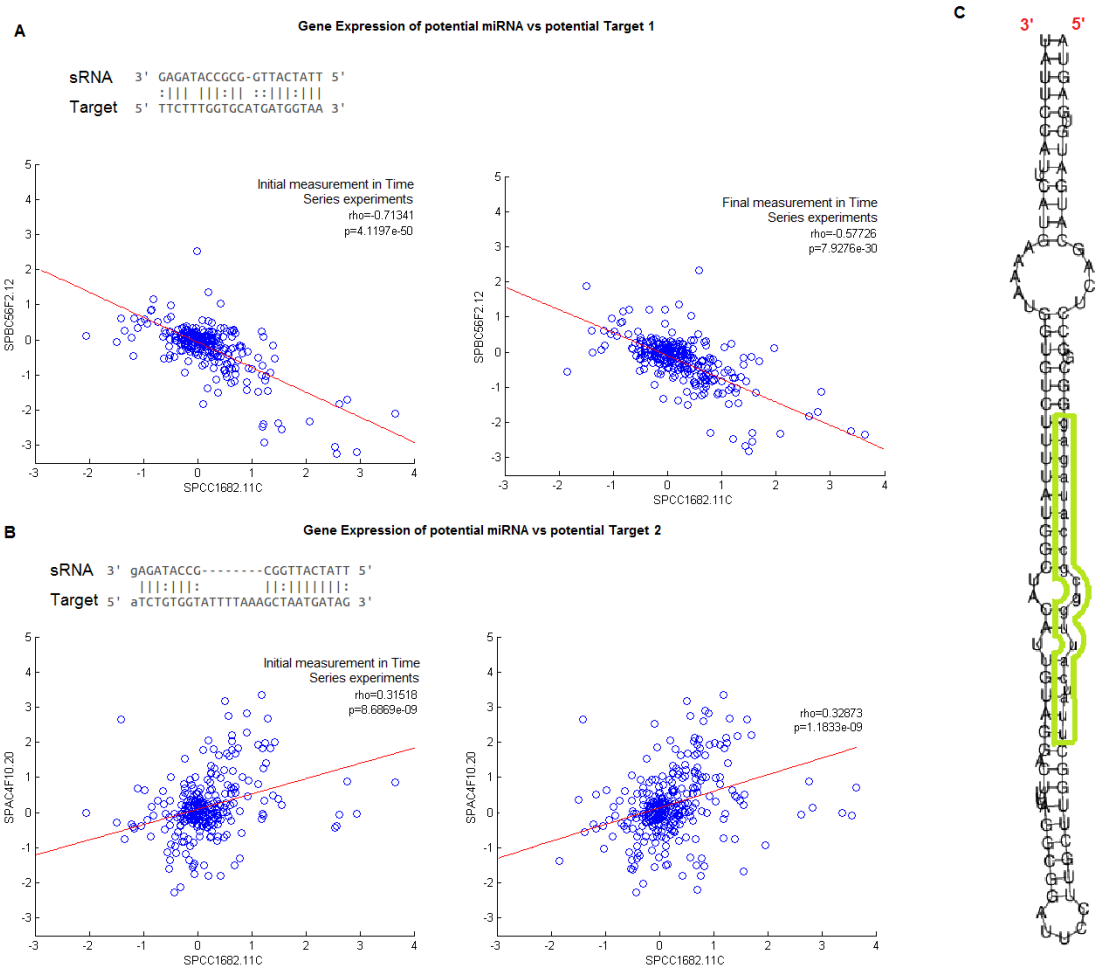

Figure D.2: Correlation of gene expression between the predicted miRNA *SPAC1682.11c* and its potential targets **A** *SPBC56F2.12* and **B** *SPAC4F10.20*. The sRNA (potential miRNA) to target binding profiles predicted by miRanda are presented above the left scatter plots. Left scatter plots present gene expression data based on the first measurement of each time series experiment, while right scatter plots are based on the final measurement of each time series experiment. Notice the strong negative correlation in gene expression between *SPAC1682.11c* and *SPBC56F2.12* in A. **C** The predicted hairpin structure of *SPAC1682.11c* with the predicted mature miRNA outlined in green.

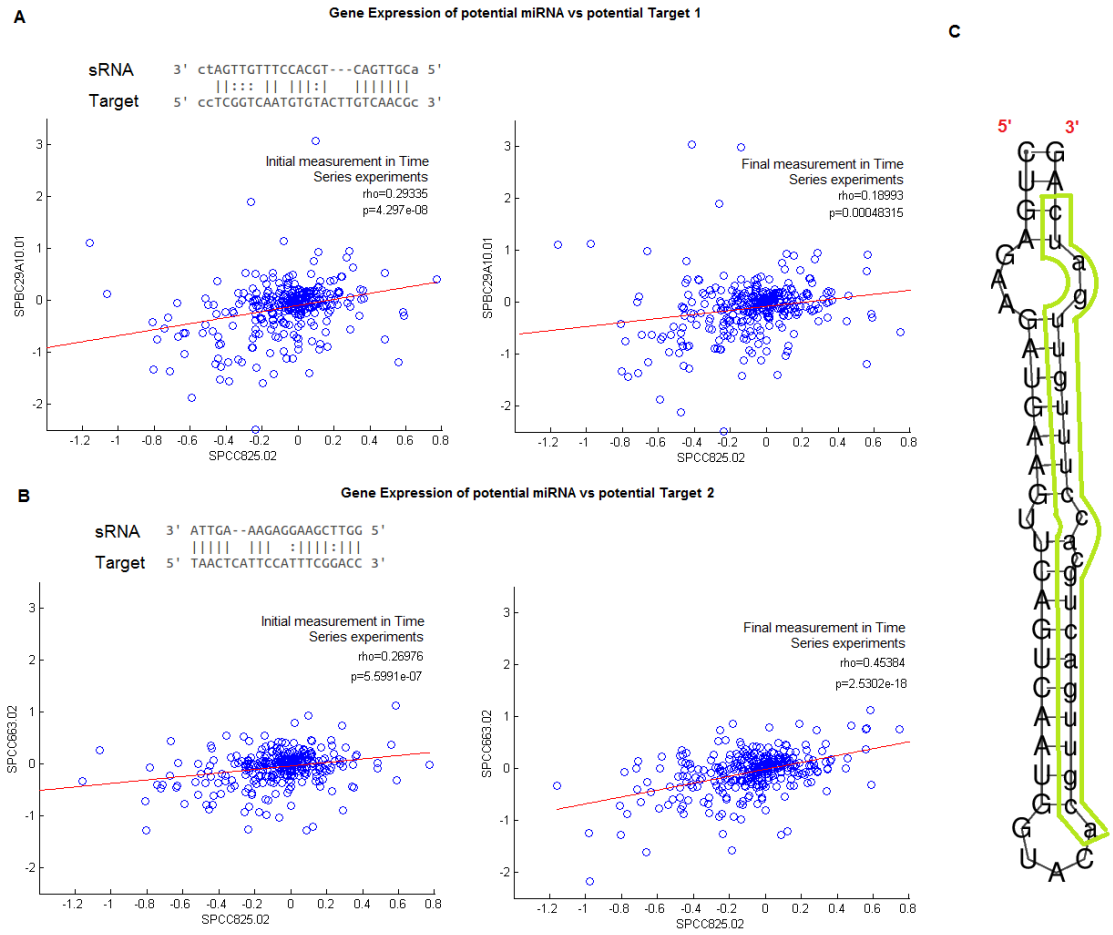

Figure D.3: Correlation of gene expression between the predicted miRNA *SPCC825.02* and its potential targets **A** *SPBC29A10.01* and **B** *SPCC663.02*. The sRNA (potential miRNA) to target binding profiles predicted by miRanda are presented above the left scatter plots. Left scatter plots present gene expression data based on the first measurement of each time series experiment, while right scatter plots are based on the final measurement of each time series experiment. **C** The predicted hairpin structure of *SPCC825.02* with the predicted mature miRNA outlined in green.

Appendix E

Dicer localisation Supplementary Materials and Methods

This section contains supplementary information on experimental methodology, which was implemented by Eleanor White and the analysis performed by Kinga Kamienarz-Gdula. The experimental procedures below were formulated by Eleanor White.

E.1 Experimental Methods

E.1.1 Tissue culture

Human HEK293, Dicer-KD/2b2 (Schmitter et al., 2006) and mouse Dicer knockout ESC29 cell lines were maintained under standard conditions. All cell lines were tested for mycoplasma contamination. Induction of shRNA was carried out by addition of 10 g/ml of doxycycline for 7, 10 or 14 days. Unless otherwise stated, cells were treated with doxycycline for 10 days. Transfections of plasmids expressing shRNA directed against Drosha or PKR 50 were performed using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions every 72 hours for a total of 10 days. Transcription

inhibition was carried out by addition of 2.5 g/ml α-amanitin for 48 hours. Stress induction was carried out by addition of 2 g/ml acivicin for 48 hours. Transient transfections were carried out using Lipofectamin 2000 (Invitrogen).

E.1.2 Immunofluorescence and Microscopy

Dicer nuclear localisation was analysed by immunofluorescence and FRAP experiments, using anti-Dicer 13D6 (Abcam), according to standard protocol using a confocal microscope (Olympus). Anti-PKR1 (Abcam), anti-Drosha (Abcam) and J2 antibody (Scicon,10010200) were also used in immunofluorescence experiments. Note that all presented images show signals at one specific z-axis value.

E.1.3 Chromatin analysis

ChIP experiments were performed using formaldehyde-crosslinked chromatin. Antibodies employed were anti- Dicer 13D6 (Abcam), anti-PolII N20X (Santa Cruz Biotechnology), anti-Ago1 (Millipore), anti-H3K9me2 (Abcam) and anti-H3 (Abcam). Immuno-precipitated, non-precipitated and input DNA was analysed by qRT-PCR. For ChIP-Sequencing (seq), the eluted ChIP DNA was used for library preparation and cluster generation using Illumina kits, following the manufacturer's instructions.

E.1.4 RNA analysis

RNA isolation was carried out using TRIzol reagent (Invitrogen) and reverse transcribed with SuperScript III Reverse Trancriptase (Invitrogen) using gene-specific primers. RNA to prepare dsRNA, ssRNA and miRNA was in vitro transcribed from PCR-generated templates of either exon 2 of the human β-globin gene or the mir22 sequence, using T7 or T3. RNA was then gel purified and for dsRNA and miRNA was hybridized by denaturation at 95°Cand then cooling slowly to allow RNAs to anneal, before being used for IP with J2 antibody (Scicon,10010200). The binding % numbers are based on

quantitation of the input and the IP signal from 100% of input (lane 2) and calculated as follows: (IP/ (Input * 10))*100. Digestions with V1, T1 and S1 were carried out according to manufacturer specification (Ambion). For RNA-Seq, total RNA was treated with T1 nuclease according to manufacturer instruction and remaining dsRNA was purified and used for library preparation using NEBNext Small RNA Library Prep Set for Illumina, following manufacturer instruction.

siRNA isolation was carried out using PEG precipitation and separated on a 20% PAGE, transferred using a semi-dry blot apparatus and chemically crosslinked (using EDC52) before being probed with ^{32}P- labelled PCR products of tested gene loci. Probes were labelled using the DECAprime kit (Ambion).

E.1.5 Protein analysis

Immunoprecipitation and immunodepletion experiments were performed on nuclear extracts using specific antibodies (PolII-8W16, Dicer-13D6). Western blot experiments were performed according to standard protocols using the following antibodies: anti-PolII 8W16 (Abcam), anti-PolII N20 (Abcam), anti-Dicer 13D6 (Abcam), anti- actin (Sigma), anti-Spt5 (Millipore), anti-TLR3 (Abcam), anti-PKR1 (Abcam), anti-IFN- (Abcam), anti-OAS1 (Abcam), anti-ADAR1 (Abcam), anti-Drosha (Abcam), anti-tubulin (Sigma), anti-GRp75 (Abcam).

E.1.6 Flow cytometry

Normal and Dicer, Drosha or PKR depleted cells were Annexin V APC (Ebioscience) and 7-AAD (Ebioscience) labeled and then measured by flow cytometry according to the manufacturer's instructions.

E.2 Computational supplementary Methods

E.2.1 Dicer ChIP-Seq

After passing Solexa CHASTITY quality filter, the reads were mapped to the human genome (hg19) using BOWTIE, allowing maximum 2 mismatches. 990,363 uniquely mapped reads were obtained. 1957 enriched regions were identified by peak calling using MACS V1.4.0 with the default p-value threshold of 10^{-5}. For downstream analysis, only 118 high-confidence, top-scoring regions (score>100) were considered, referenced in the Chapter 4 as "Dicer ChIP-Seq peaks".

E.2.2 dsRNA RNA-Seq

RNA reads were mapped to the human genome as described for ChIP-Seq.

Appendix F

Cohesin Supplementary data

F.1 Summary of FISH Experimental Methods

This Experiment was performed and described by Shweta Bhardwaj.

Sites for analysis were randomly selected out of the Mis4/Rad21 peaks and Rad21 only peaks. Roughly 10 kb regions including intergenic regions between the convergent genes were amplified via PCR. Equimolar amounts of PCR products were labelled with the FISH-Tag Kit (Molecular probes, Invitrogen), as described by the manufacturer. Proliferating cells (G2) were prepared for hybridization by fixing them with para-formaldehyde, sphaeroblasting them with zamolyseT100 (100 units) and treated with RNAse A(1 μg/ml). 100g of labeled probe (1ng/ μl) was hybridised to the fixed cells overnight. Cells were then washed, DAPI-stained (0.2 μ/ml) and moved to poly-I-lysine coated slides. The DeltaVision microscope (100X-1.3 NA objective) was employed with the Softworx program to image the cells. Z-sections were taken at 200nm resolution and FISH signals were processed manually. The Fiji/ImageJ program (NIH) was used to obtain the final images representing maximum intensity plots.

F.2 Transcriptional inhibition reduces Cohesin levels

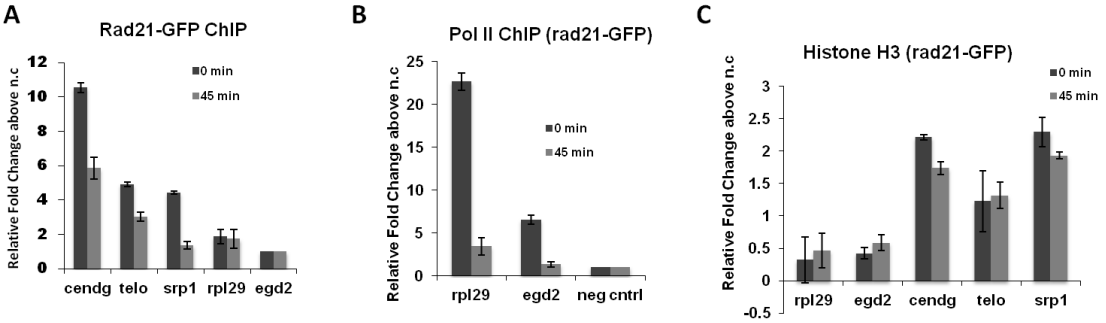

Figure F.1: **A** ChIP Analysis of Rad21-GFP in thiolutin treated cycling cells. Labels correspond to centromeric dg repeats (cendg), sub-telomeric regions (telo), chromosomal arm (srp1) and two promoter regions of rpl29 and edg2. After 45 minutes of thiolutin treatment Rad21 levels at centromeric and telomeric regions dropped one-fold, at the chromosomal arm four-fold and in promoter regions the low levels remained unchanged. **B** Drop in Pol II occupancy confirms effectiveness of thiolutin treatment for 45 minutes. Negative control corresponds to a transcriptionally inactive chromosomal region. **C** H3 levels were measured as a control and remain unchanged between 0 and 45 minutes of thiolutin treatment. Error bars correspond to one standard error of the mean of two biological replicates. This Experiment was performed by Shweta Bhardwaj.

F.3 Called Cohesin Peaks on chromosome II

Rad21 Peaks, chromosome II

Peak Number	Midpoint	Width	Peak Number	Midpoint	Width
1	87552	13500	24	437052	7500
2	108677	7750	25	470427	11750
3	114552	1500	26	477427	1250
4	138677	4750	27	508927	4750
5	148052	2000	28	517927	3250
6	161427	8750	29	521052	1000
7	175927	2250	30	530927	5250
8	182052	3500	31	547302	12000
9	191677	1250	32	583302	4500
10	218177	5250	33	598552	1000
11	235052	5500	34	603302	8000
12	257927	4750	35	623177	1750
13	269052	8000	36	626677	1250
14	284427	5750	37	644302	4000
15	293677	6750	38	653177	4750
16	317052	7500	39	668427	5250
17	321802	1500	40	717052	3500
18	326677	1750	41	719802	1500
19	337177	5750	42	732552	3500
20	350302	4500	43	737927	4250
21	367552	5500	44	756802	8500
22	398552	7500	45	777677	1750
23	424302	6500	46	779552	1500

Table App.F.3.1 - 1 out of 7: Called Rad21 peaks by my peak-calling algorithm on chromosome II.

Peak Number	Midpoint	Width
47	799927	6750
48	809302	5000
49	826802	1500
50	865302	8500
51	873927	4750
52	911177	10750
53	930427	7750
54	935552	1500
55	945677	2750
56	955427	10750
57	969427	5250
58	986927	1250
59	1001927	8250
60	1016052	6000
61	1023177	7250
62	1038052	5000
63	1058177	5750
64	1069927	2750
65	1086802	12500
66	1122302	13000
67	1142677	5250
68	1160177	6250
69	1169552	9000

Peak Number	Midpoint	Width
70	1185052	5500
71	1220677	5750
72	1236052	8000
73	1249927	5250
74	1274177	5750
75	1279927	4250
76	1293427	6250
77	1316302	5500
78	1343052	7000
79	1363802	1000
80	1371552	7000
81	1391927	4750
82	1421052	7500
83	1449552	4500
84	1476052	4500
85	1514052	1000
86	1536677	8750
87	1551677	5250
88	1555802	1000
89	1558802	4500
90	1570052	3000
91	1578552	4500
92	1599677	11250

Table App.F.3.1 - 2 out of 7: Called Rad21 peaks by our peak-calling algorithm on chromosome II.

Peak Number	Midpoint	Width
93	1606427	1750
94	1610927	5750
95	1615427	2750
96	1619052	1500
97	1644552	1500
98	1651427	6250
99	1658427	1750
100	1675802	7000
101	1691677	1250
102	1693052	1000
103	1696901	8000
104	1709901	5500
105	1719651	6000
106	1734026	1250
107	1739901	7500
108	1756901	1000
109	1775276	6250
110	1783901	6500
111	1801651	3000
112	1808151	3500
113	1828526	4750
114	1833901	5000
115	1847526	7750
116	1883526	4750
117	1911526	2250
118	1928651	8000
119	1947151	3500
120	1975026	4750
121	1987651	6000
122	2017901	5000
123	2029401	4500
124	2032401	1000
125	2040901	3000
126	2052776	8250
127	2061276	2250
128	2077026	3250
129	2082401	2000
130	2090401	13000
131	2100651	7000
132	2106776	4750
133	2110026	1250
134	2115276	1250
135	2125276	15750
136	2134901	1500
137	2143151	11500
138	2153276	8250

Table App.F.3.1 - 3 out of 7: Called Rad21 peaks by our peak-calling algorithm on chromosome II.

Peak Number	Midpoint	Width
139	2160776	1750
140	2168901	14000
141	2181401	3500
142	2190526	5250
143	2215776	6750
144	2226026	2250
145	2230651	6500
146	2251651	5500
147	2263526	4750
148	2267151	2000
149	2291901	14000
150	2310276	4750
151	2329526	1250
152	2334526	7750
153	2345276	13250
154	2367651	1000
155	2390401	1500
156	2402776	1750
157	2407526	7250
158	2417401	6000
159	2433151	3500
160	2446276	9750
161	2460151	9000
162	2480026	2250
163	2483401	4000
164	2489651	4500
165	2499026	13250
166	2515526	6250
167	2552151	2000
168	2557151	7500
169	2576151	6500
170	2589151	5000
171	2607026	8750
172	2614401	2500
173	2619401	7000
174	2630276	1750
175	2634401	6000
176	2639026	2750
177	2645526	8750
178	2654401	6500
179	2680026	4250
180	2698651	4000
181	2716776	12250
182	2751651	8000
183	2759901	8000
184	2782901	8500

Table App.F.3.1 - 4 out of 7: Called Rad21 peaks by our peak-calling algorithm on chromosome II.

Peak Number	Midpoint	Width
185	2790276	1750
186	2823901	3500
187	2836651	10000
188	2846276	8250
189	2851151	1000
190	2875526	7750
191	2893026	5250
192	2897276	2750
193	2906776	6250
194	2919526	3250
195	2938776	7750
196	2952276	1250
197	2980526	9250
198	2991026	3250
199	3004026	4750
200	3023526	7250
201	3041901	6500
202	3061651	1500
203	3065026	4750
204	3082776	1250
205	3094276	21250
206	3113901	1500
207	3115901	2000
208	3121151	6500
209	3133276	3750
210	3143276	4250
211	3173651	4500
212	3184151	10500
213	3195776	5250
214	3207026	6250
215	3219901	9000
216	3246151	1500
217	3249026	2750
218	3273026	15750
219	3299026	3750
220	3308026	5250
221	3319526	7250
222	3341151	3000
223	3354276	8250
224	3365276	9750
225	3376151	2500
226	3384776	1250
227	3392026	5750
228	3408276	6250
229	3433526	5250
230	3440901	8000

Table App.F.3.1 - 5 out of 7: Called Rad21 peaks by our peak-calling algorithm on chromosome II.

Peak Number	Midpoint	Width
231	3481901	7000
232	3504776	6750
233	3530151	4500
234	3542401	13000
235	3566901	9000
236	3579901	6000
237	3583651	1000
238	3589651	2500
239	3593401	4000
240	3611276	2750
241	3613901	2000
242	3621651	6500
243	3635776	5250
244	3644651	10000
245	3656401	6000
246	3675276	2750
247	3694401	11500
248	3713026	6250
249	3728026	2250
250	3736026	12250
251	3745526	5250
252	3762276	12750
253	3789401	5000
254	3808651	2500
255	3814901	1500
256	3820276	8750
257	3843901	4000
258	3871026	8750
259	3900526	3750
260	3908401	8000
261	3920651	7000
262	3932151	6500
263	3939526	1750
264	3949776	11250
265	3958401	1500
266	3974776	3750
267	3999401	22500
268	4040651	6000
269	4067026	8750
270	4082651	1000
271	4086151	5500
272	4102151	8000
273	4107651	2000
274	4114651	6000
275	4152526	10250
276	4174151	8000

Table App.F.3.1 - 6 out of 7: Called Rad21 peaks by our peak-calling algorithm on chromosome II.

Peak Number	Midpoint	Width
277	4210026	2250
278	4218026	11750
279	4234776	9750
280	4260276	8750
281	4275901	7500
282	4293151	1000
283	4298651	8500
284	4321026	8750
285	4341651	4500

Peak Number	Midpoint	Width
286	4353151	5500
287	4376901	12500
288	4429151	10000
289	4457651	16000
290	4477901	9000
291	4485276	1750
292	4490901	5000
293	4494276	1250
294	4502276	14250

Table App.F.3.1 - 7 out of 7: Called Rad21 peaks by our peak-calling algorithm on chromosome II.

Mis4 Peaks, chromosome II

Peak Number	Midpoint	Width	Peak Number	Midpoint	Width
1	92427	3750	24	941552	4000
2	103052	1000	25	961802	3000
3	107802	1500	26	978552	2500
4	182552	4500	27	1048302	1000
5	258802	3000	28	1081427	4750
6	285802	2000	29	1107302	4000
7	326552	2000	30	1199802	1500
8	356052	1000	31	1208302	1000
9	412677	5250	32	1237802	4500
10	466677	4750	33	1274552	1500
11	477427	1250	34	1278302	1000
12	498927	1750	35	1297802	1500
13	517677	3250	36	1392802	3000
14	571427	5250	37	1403177	1250
15	625677	6250	38	1407177	1750
16	658302	1000	39	1448927	2250
17	717427	6250	40	1504302	3000
18	736302	4000	41	1514177	3250
19	744802	1000	42	1543302	2000
20	755552	2500	43	1545302	1000
21	780177	1250	44	1558927	2750
22	783052	1000	45	1585802	1000
23	884302	2000	46	1601427	7750

Table App.F.3.2 - 1 out of 4: Called Mis4 peaks by our peak-calling algorithm on chromosome II.

Peak Number	Midpoint	Width
47	1619052	1500
48	1624802	8500
49	1644552	1500
50	1648927	1250
51	1652177	3250
52	1665927	1750
53	1689427	4250
54	1692302	1000
55	1727776	2250
56	1734151	4000
57	1757026	2250
58	1761151	2000
59	1774776	5250
60	1793151	5000
61	1827026	3250
62	1917901	1500
63	1925026	3250
64	1971901	3500
65	2011401	4000
66	2034401	4500
67	2061026	2750
68	2069901	1000
69	2074276	1250
70	2083401	5500
71	2091901	1500
72	2102151	1000
73	2106651	3000
74	2115026	2250
75	2118151	1500
76	2129151	8000
77	2134901	1500
78	2138276	1750
79	2144401	9000
80	2150526	2750
81	2162276	4750
82	2210401	1000
83	2218151	2000
84	2286026	3250
85	2290276	2750
86	2319651	2000
87	2325651	3000
88	2340776	1750
89	2349776	2750
90	2380401	2500
91	2411776	3250
92	2426776	1750

Table App.F.3.2 - 2 out of 4: Called Mis4 peaks by our peak-calling algorithm on chromosome II.

Peak Number	Midpoint	Width
93	2443526	1250
94	2449401	3000
95	2464776	2250
96	2472276	2750
97	2503901	2000
98	2522776	1250
99	2577401	3000
100	2588651	8500
101	2607651	1500
102	2614026	1750
103	2647901	1500
104	2691276	4250
105	2721276	1250
106	2737401	1500
107	2751651	2000
108	2759026	1750
109	2780401	3000
110	2803901	3000
111	2940526	2750
112	3021276	2750
113	3061901	2500
114	3155026	3750
115	3173901	2500
116	3193276	4750
117	3197151	2000
118	3246026	2250
119	3272776	2750
120	3276276	3250
121	3316401	3500
122	3319901	1500
123	3363526	6750
124	3393651	5000
125	3404026	1250
126	3442526	4250
127	3452526	1750
128	3490651	1000
129	3507151	4000
130	3532776	1750
131	3620151	1500
132	3654776	3250
133	3681901	1000
134	3698151	1500
135	3819276	2250
136	3914526	1750
137	3923026	2750
138	3939526	1750

Table App.F.3.2 - 3 out of 4: Called Mis4 peaks by our peak-calling algorithm on chromosome II.

Peak Number	Midpoint	Width
139	3945026	1750
140	3958401	4000
141	3963526	1250
142	3975526	1750
143	3990651	1500
144	4032651	1500
145	4040026	1750
146	4046776	5250
147	4079401	1000
148	4088776	6750
149	4108401	1500

Peak Number	Midpoint	Width
150	4210276	1750
151	4239776	3250
152	4285651	4000
153	4298276	3750
154	4353276	3750
155	4369151	2000
156	4373651	2000
157	4387276	2250
158	4411526	2750
159	4437651	2000
160	4482651	1500
161	4508151	2500

Table App.F.3.2 - 4 out of 4: Called Mis4 peaks by our peak-calling algorithm on chromosome II.

F.4 Cohesin distribution profiles on chromosome II

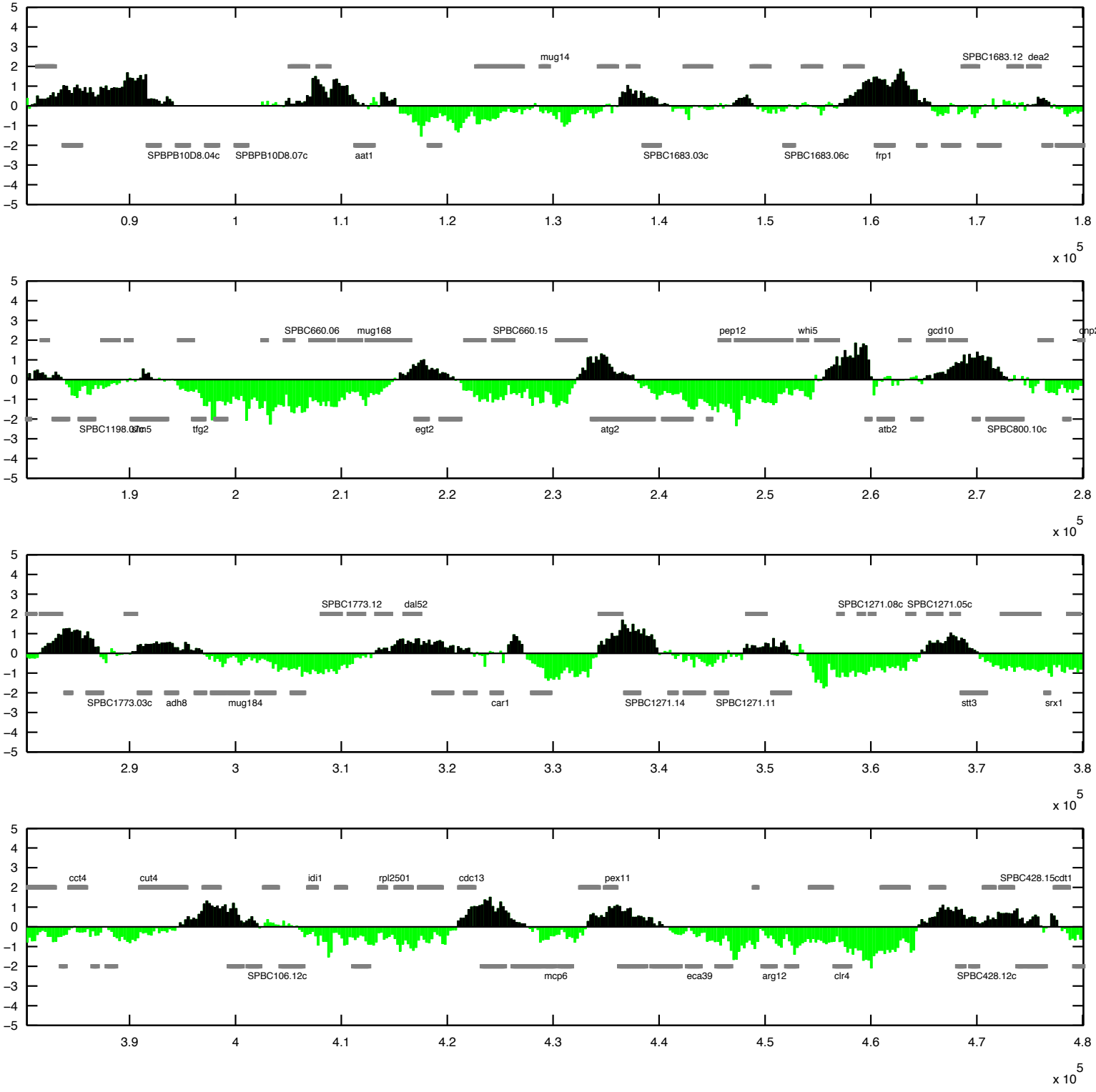

Figure App.F.4.1 Rad21 binding on chromosome II, 1 out of 12. Peaks are based on epitope-tagged Rad21-Pk9 data (Schmidt et al., 2009). Y- axis is in $\log_2$ scale. Called peaks are represented in black, while all other data is plotted in green. Peaks extend from their midpoint to either side until $\log_2$ reaches 0.

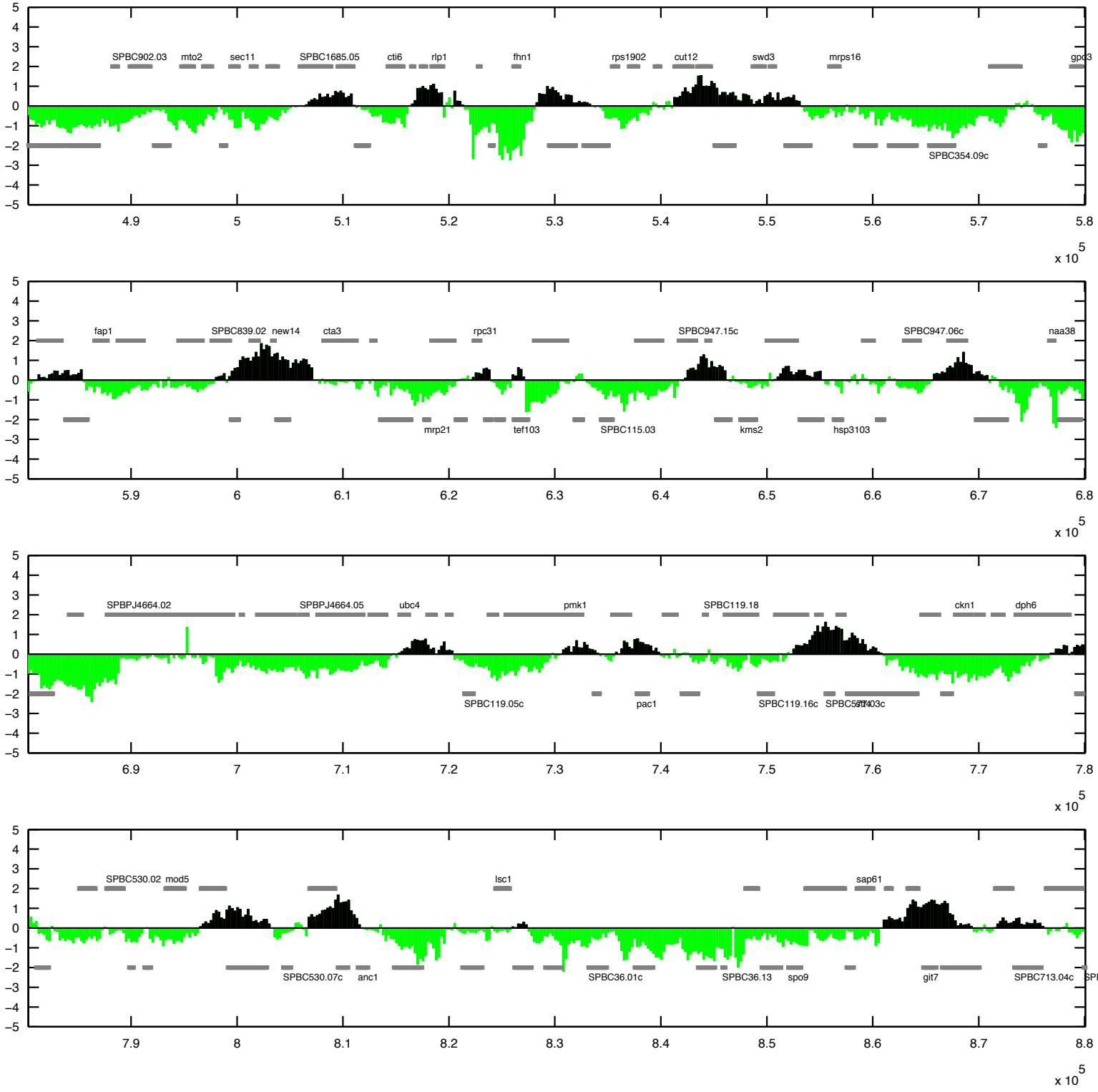

Figure App.F.4.1 Rad21 binding on chromosome II, 2 out of 12.

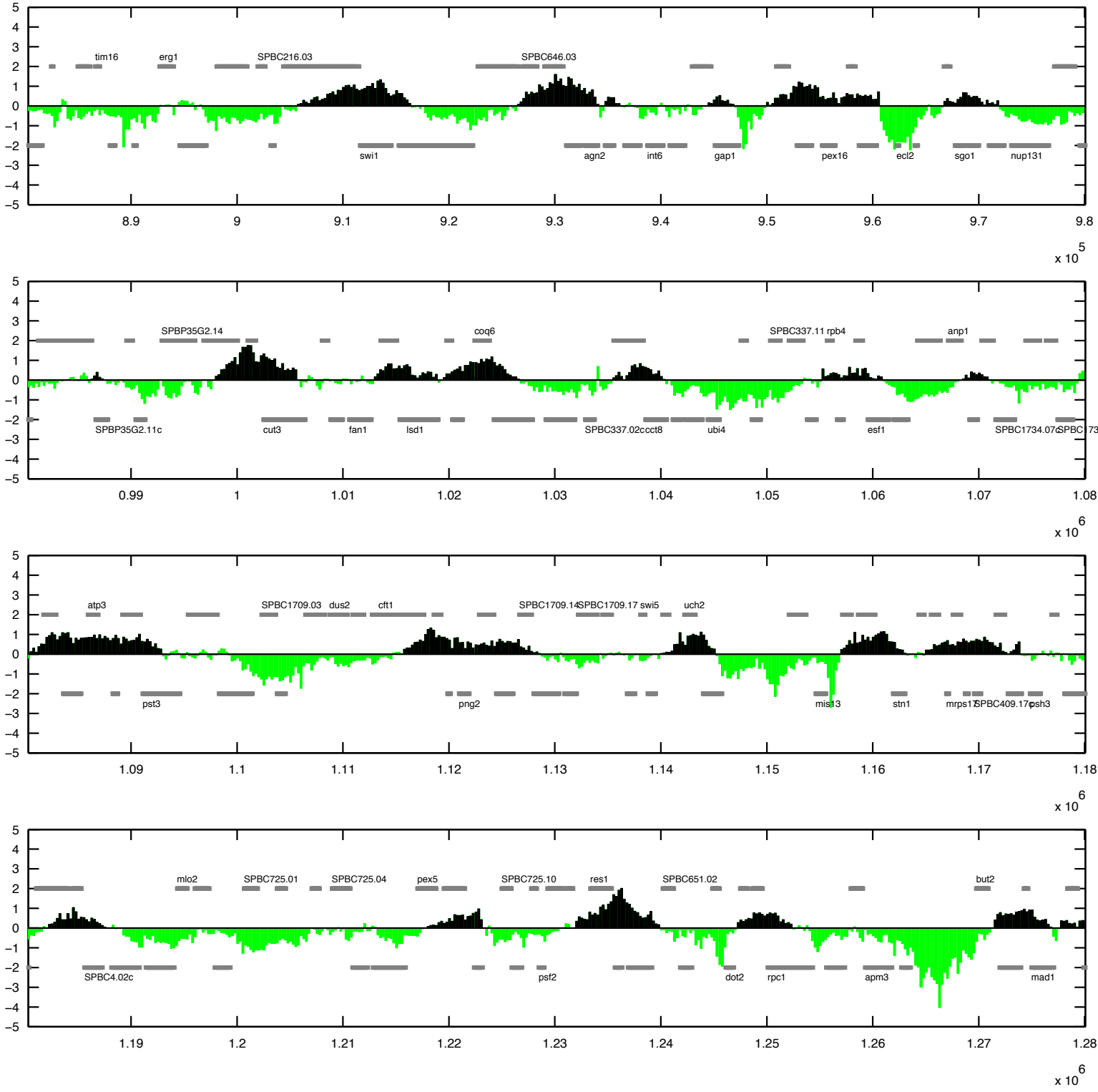

Figure App.F.4.1 Rad21 binding on chromosome II, 3 out of 12.

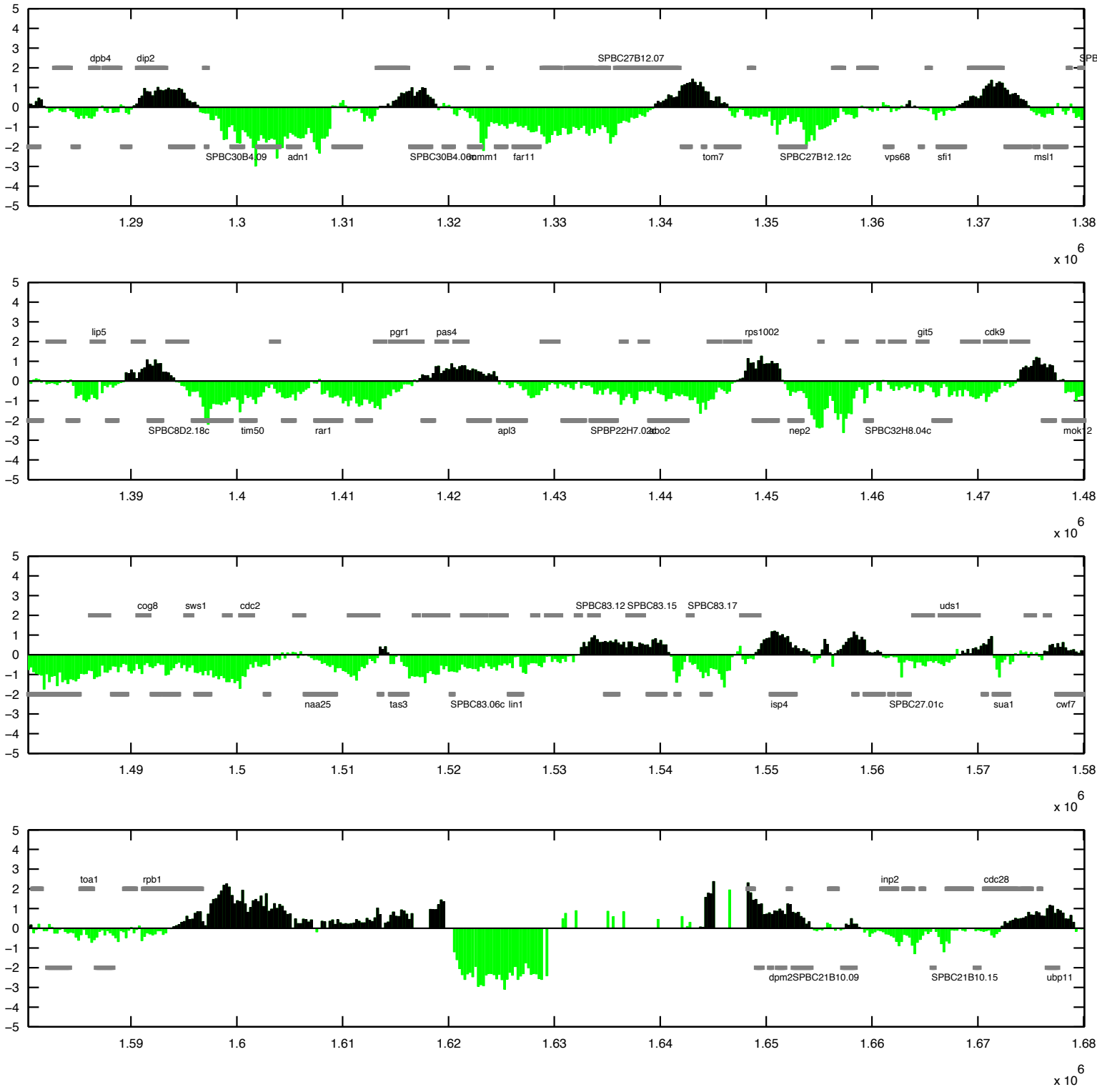

Figure App.F.4.1 Rad21 binding on chromosome II, 4 out of 12.

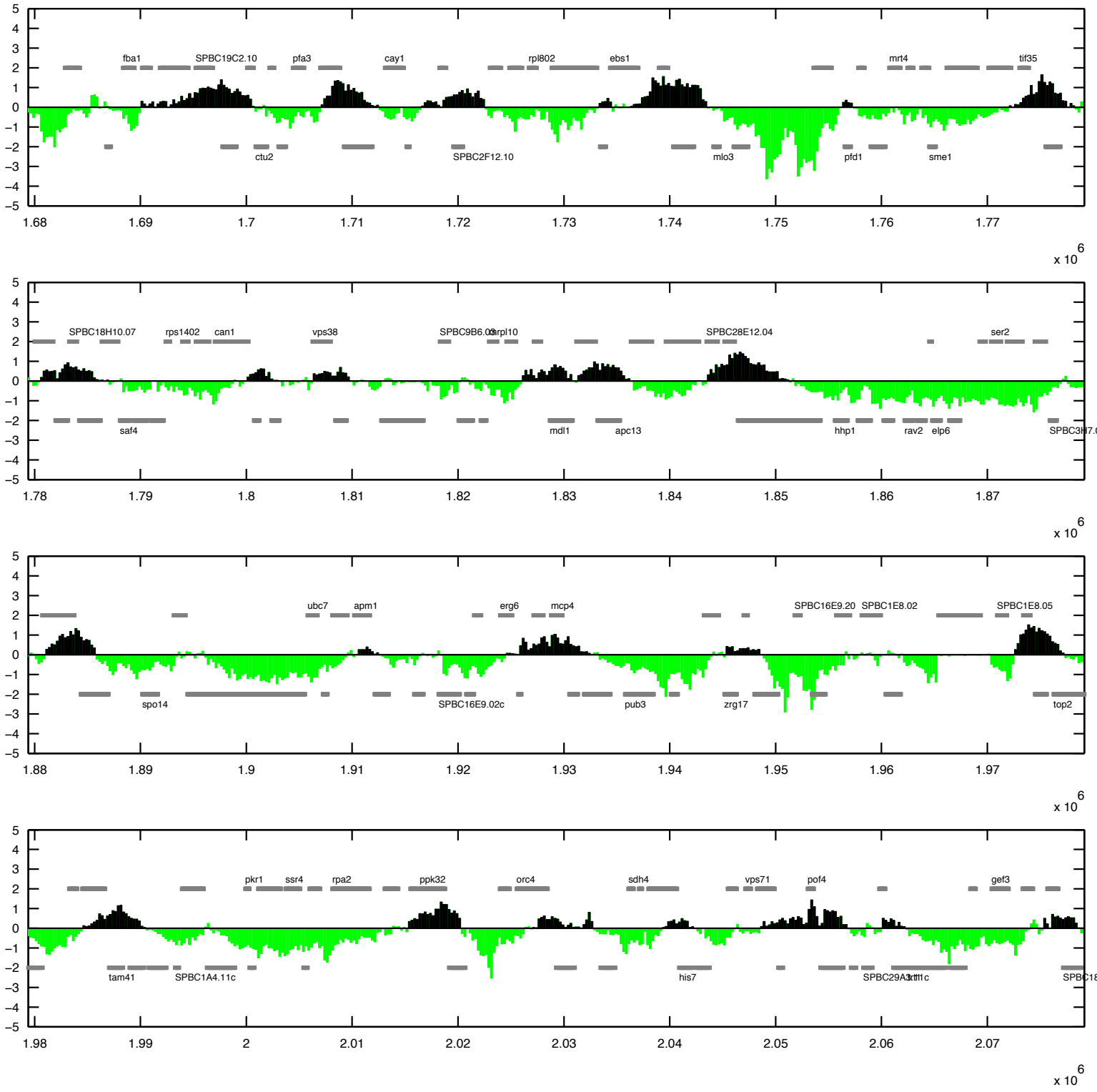

Figure App.F.4.1 Rad21 binding on chromosome II, 5 out of 12.

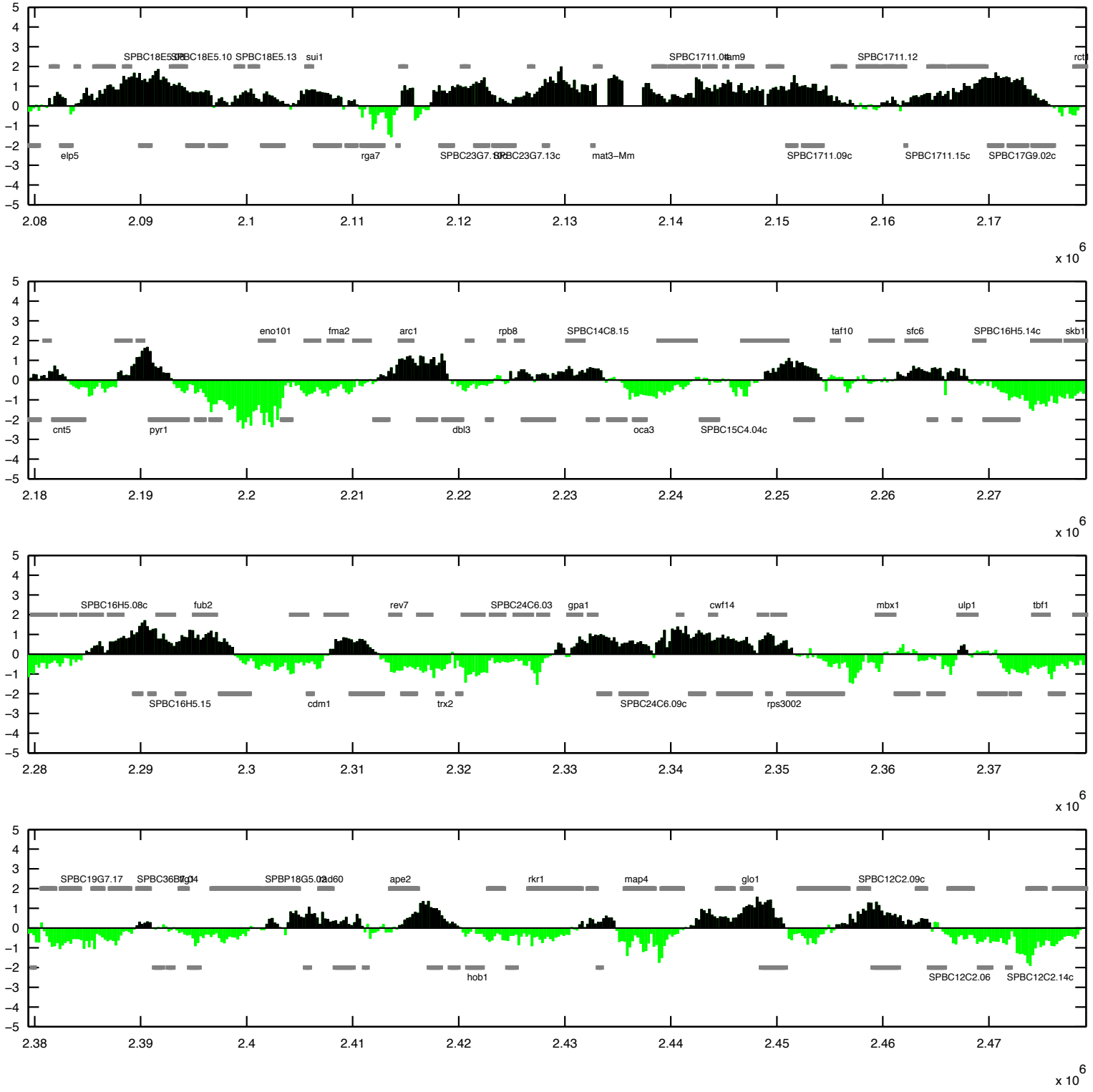

Figure App.F.4.1 Rad21 binding on chromosome II, 6 out of 12.

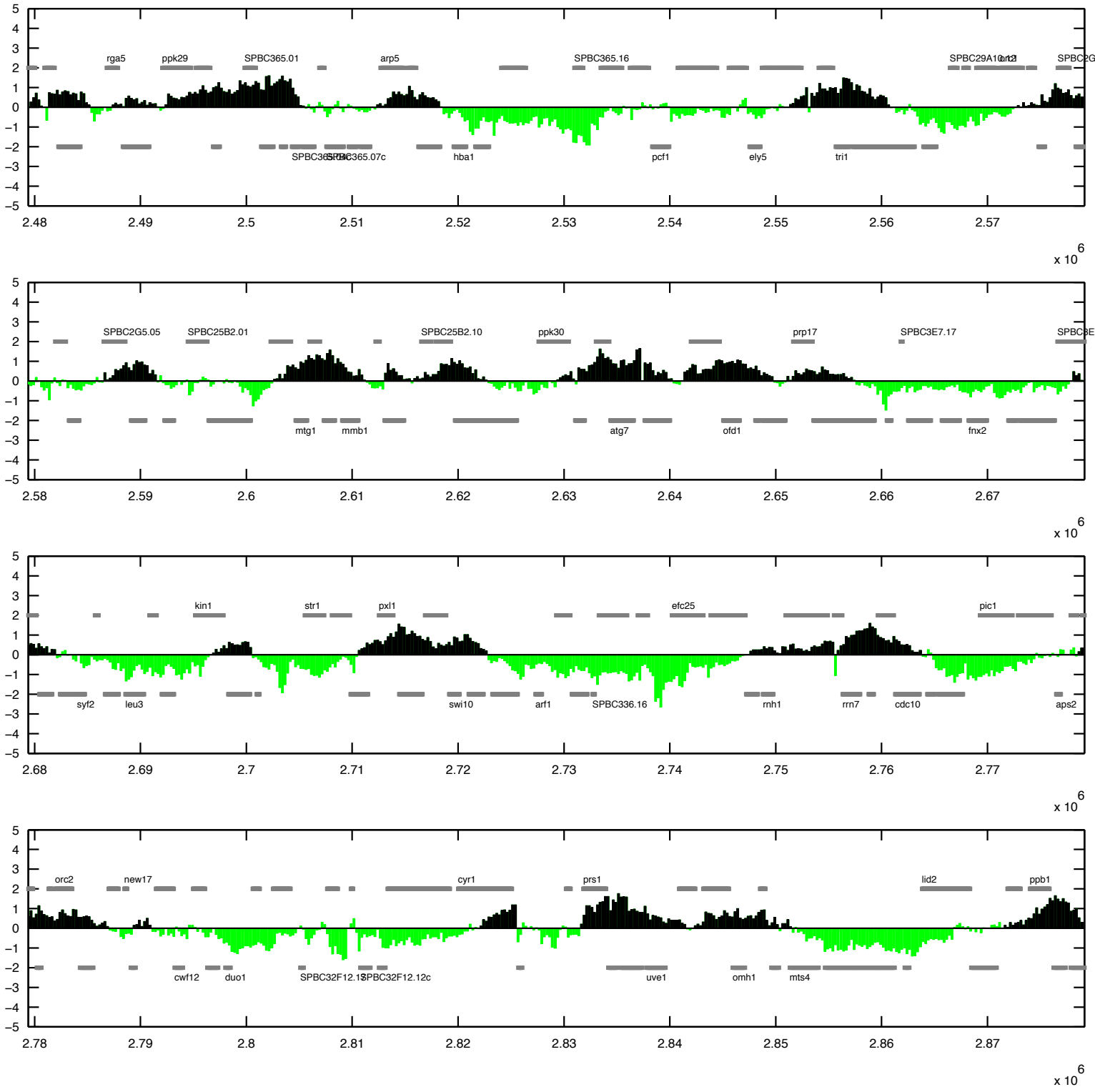

Figure App.F.4.1 Rad21 binding on chromosome II, 7 out of 12.

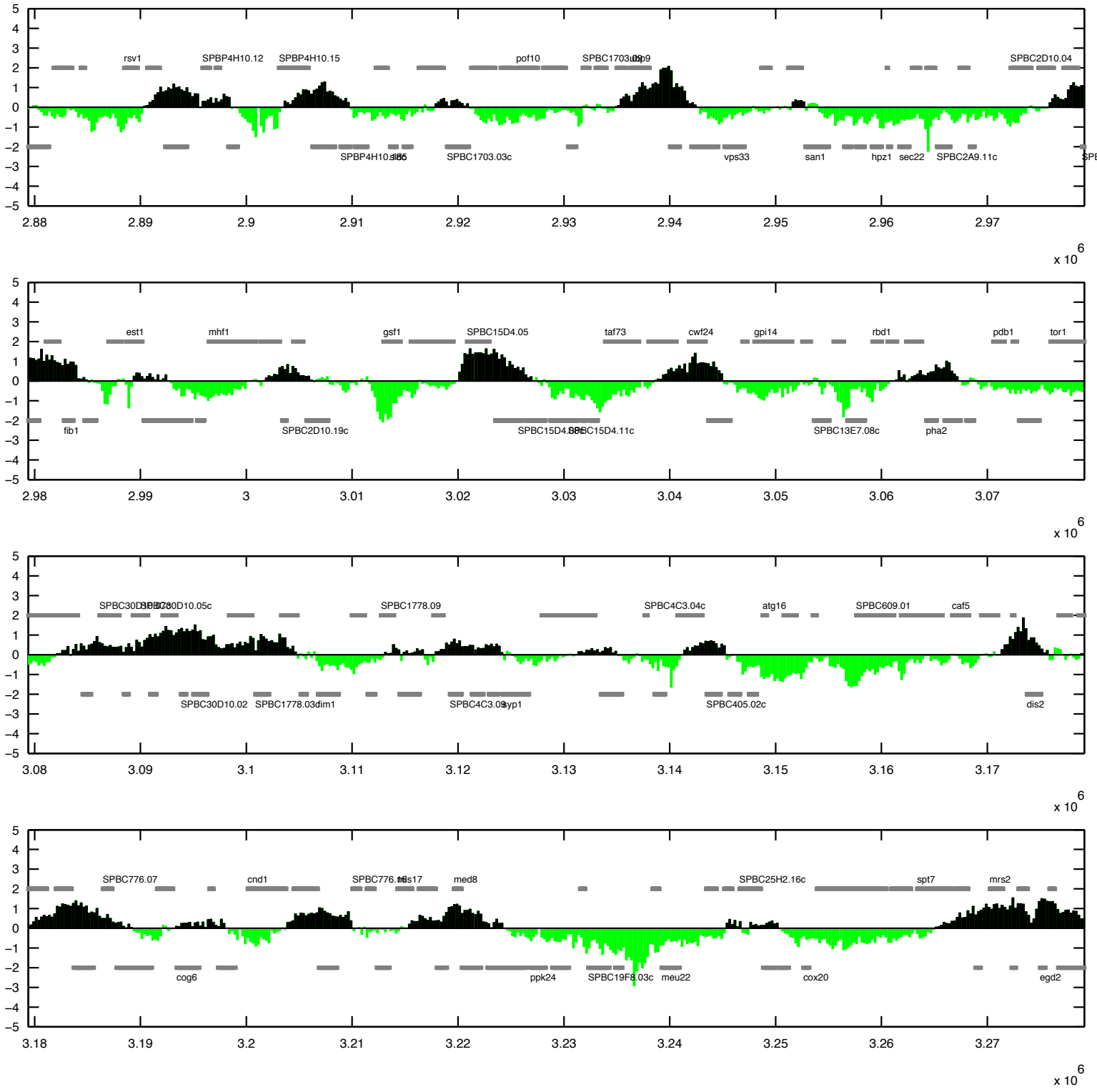

Figure App.F.4.1 Rad21 binding on chromosome II, 8 out of 12.

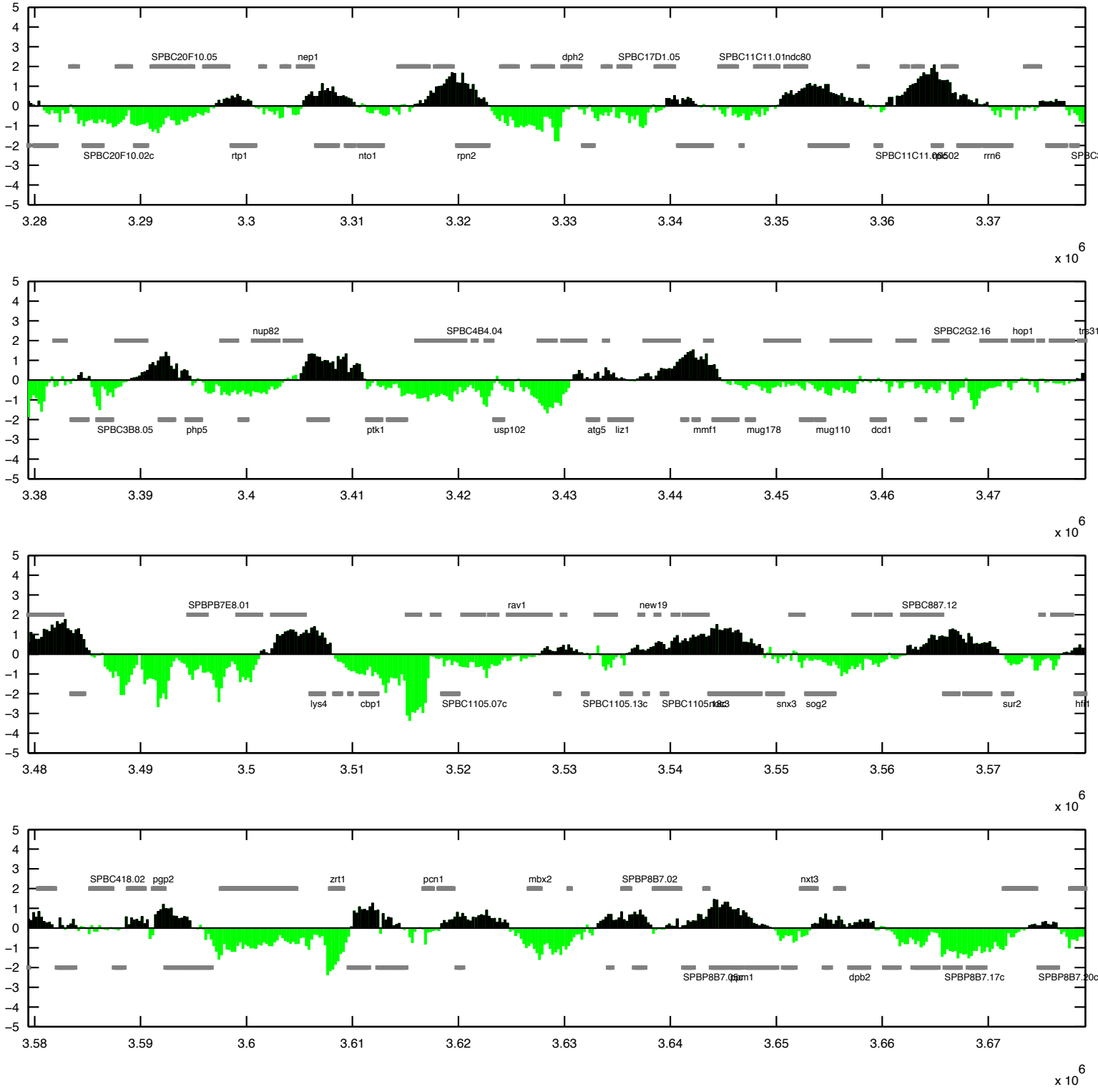

Figure App.F.4.1 Rad21 binding on chromosome II, 9 out of 12.

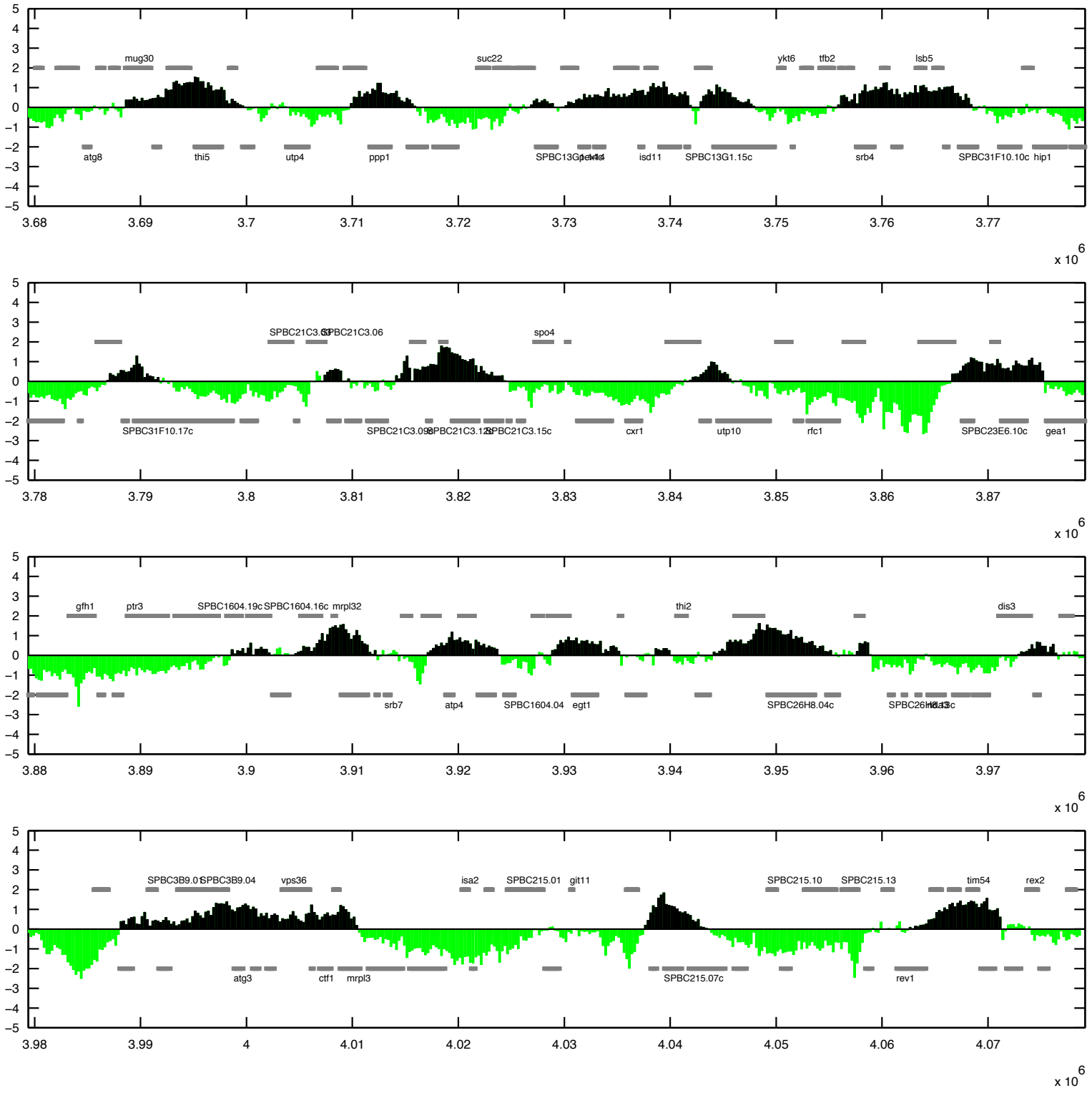

Figure App.F.4.1 Rad21 binding on chromosome II, 10 out of 12.

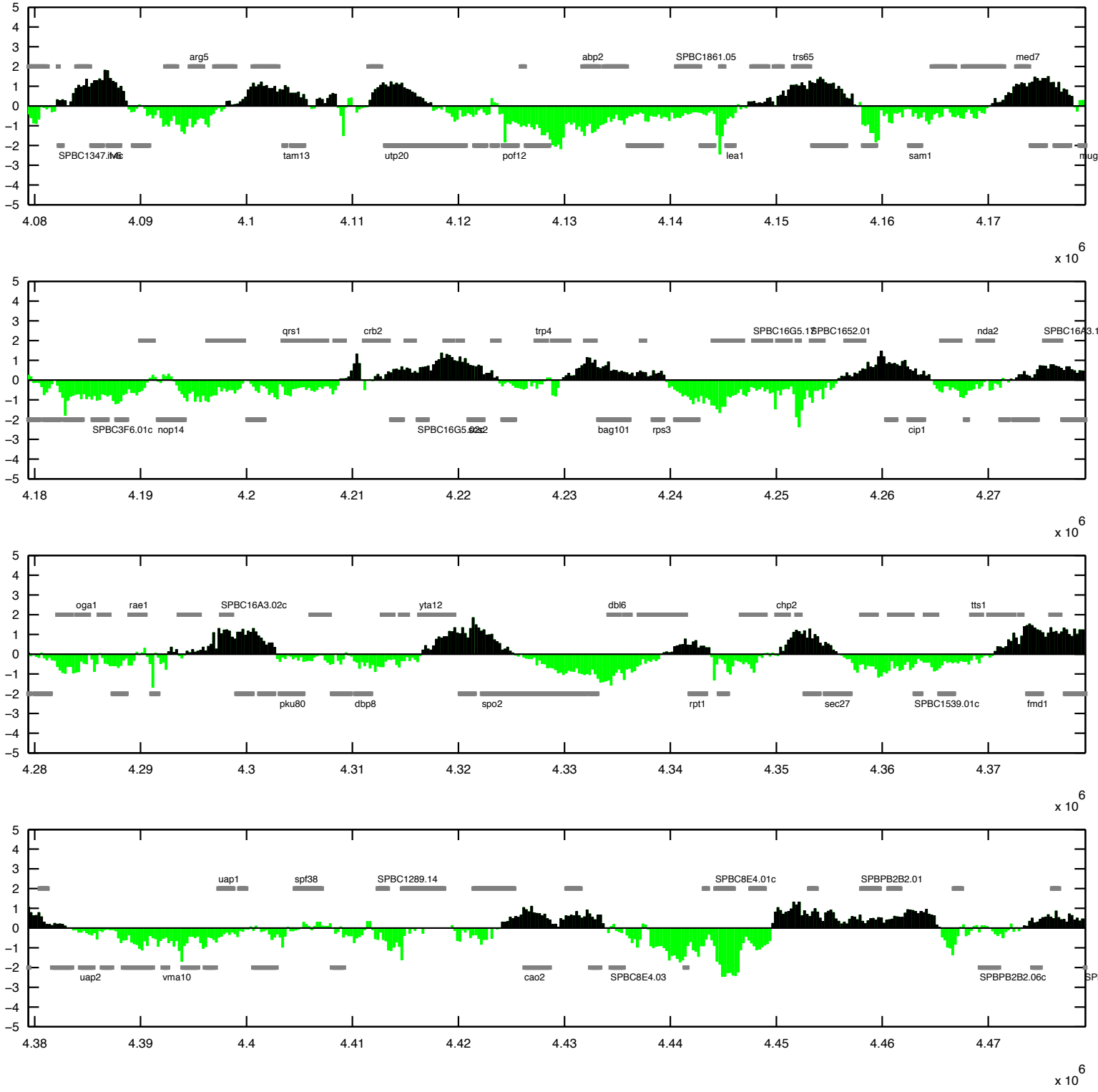

Figure App.F.4.1 Rad21 binding on chromosome II, 11 out of 12.

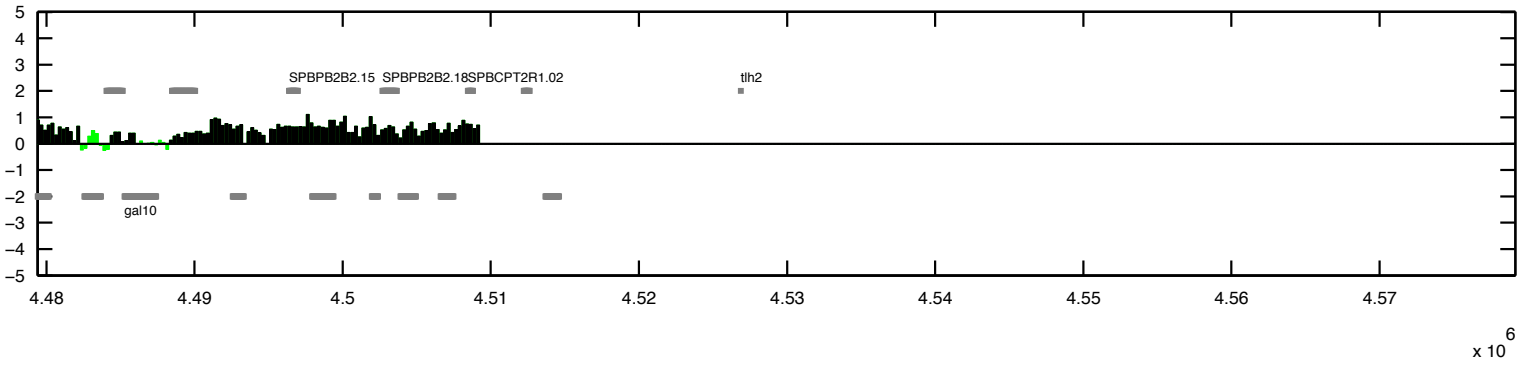

Figure App.F.4.1 Rad21 binding on chromosome II, 12 out of 12.

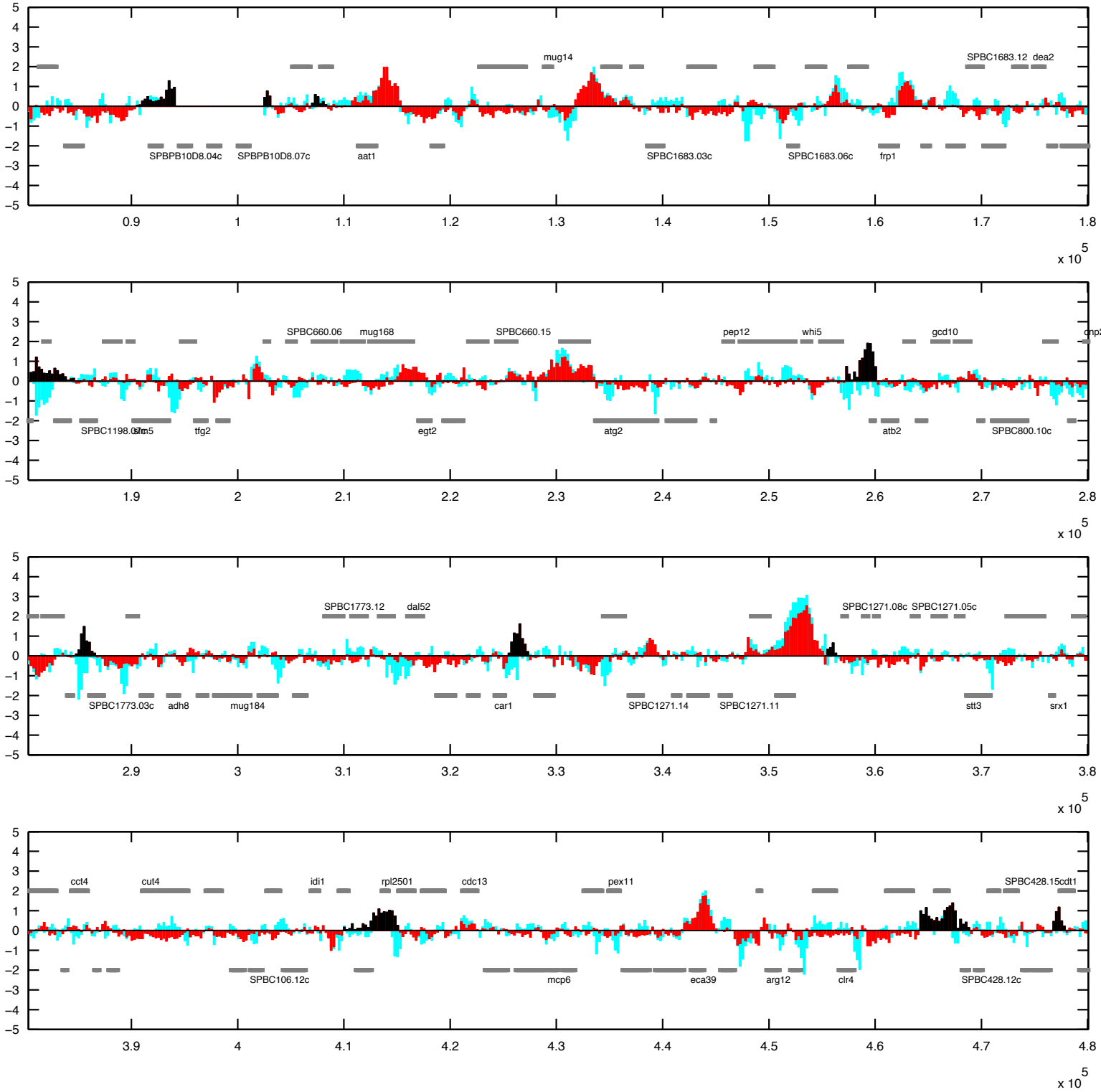

Figure App.F.4.2 Mis4 binding on chromosome II, 1 out of 12. Peaks were called based on comparison of epitope tagged Mis4-Pk9 data to epitope-untagged data Pk9. Y-axis is in $\log_2$ scale. Called peaks are represented in black, while all other data is plotted in red. Note that some regions may appear as peaks, but are not called. This is due to high data of the epitope-untagged data (light blue). Peaks extend from their midpoint to either side until $\log_2$ reaches 0.

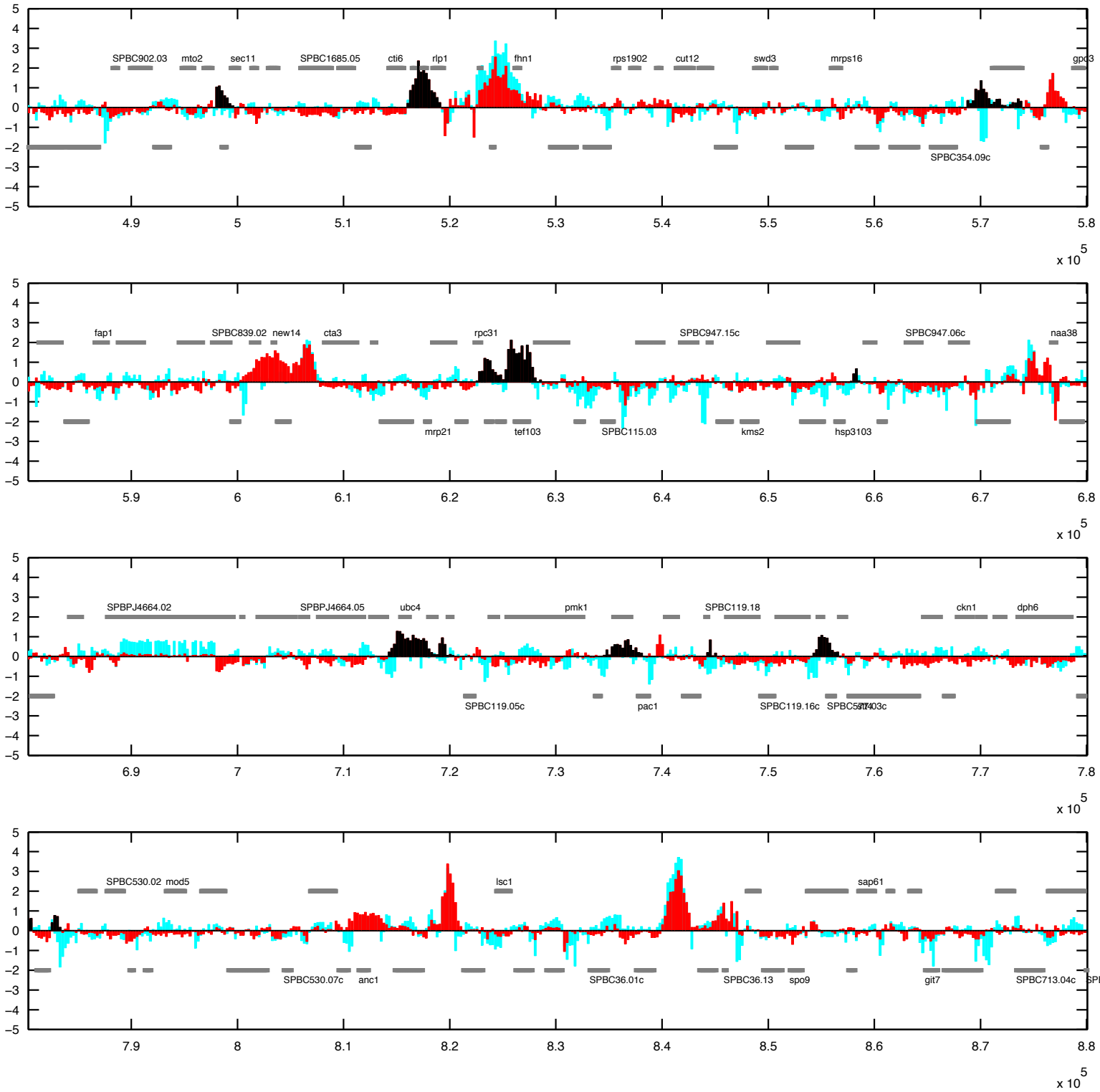

Figure App.F.4.2 Mis4 binding on chromosome II, 2 out of 12

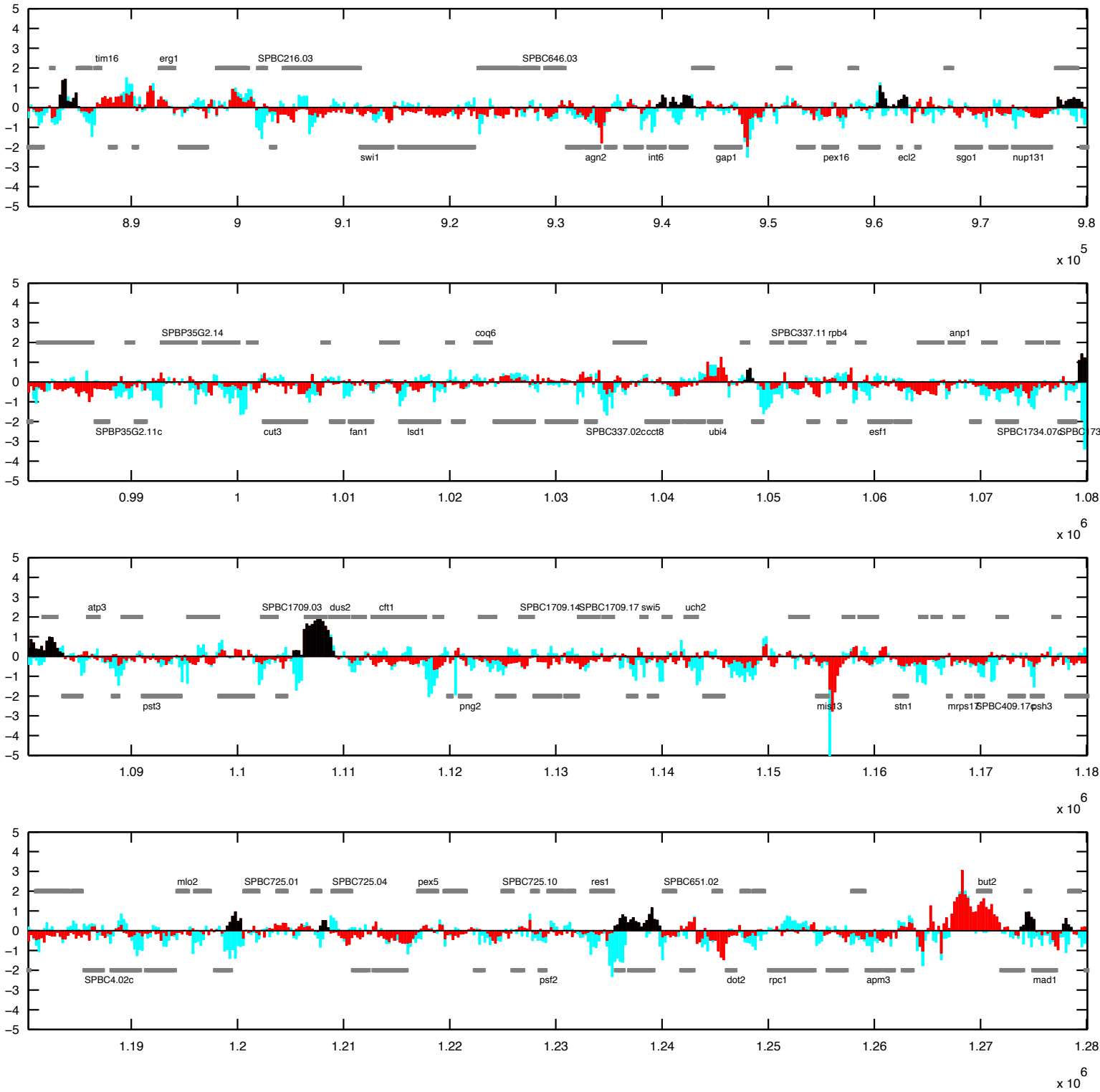

Figure App.F.4.2 Mis4 binding on chromosome II, 3 out of 12

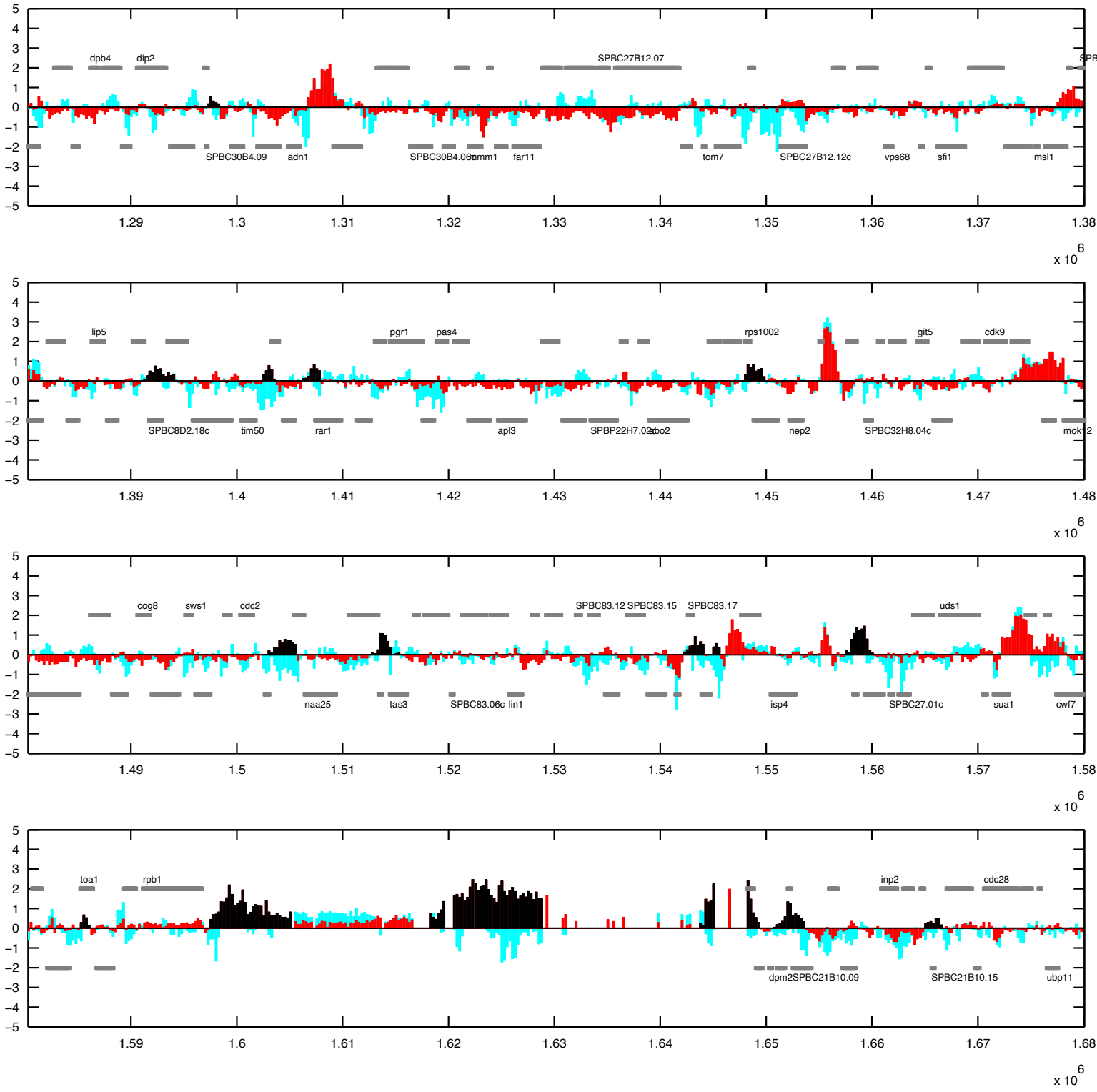

Figure App.F.4.2 Mis4 binding on chromosome II, 4 out of 12

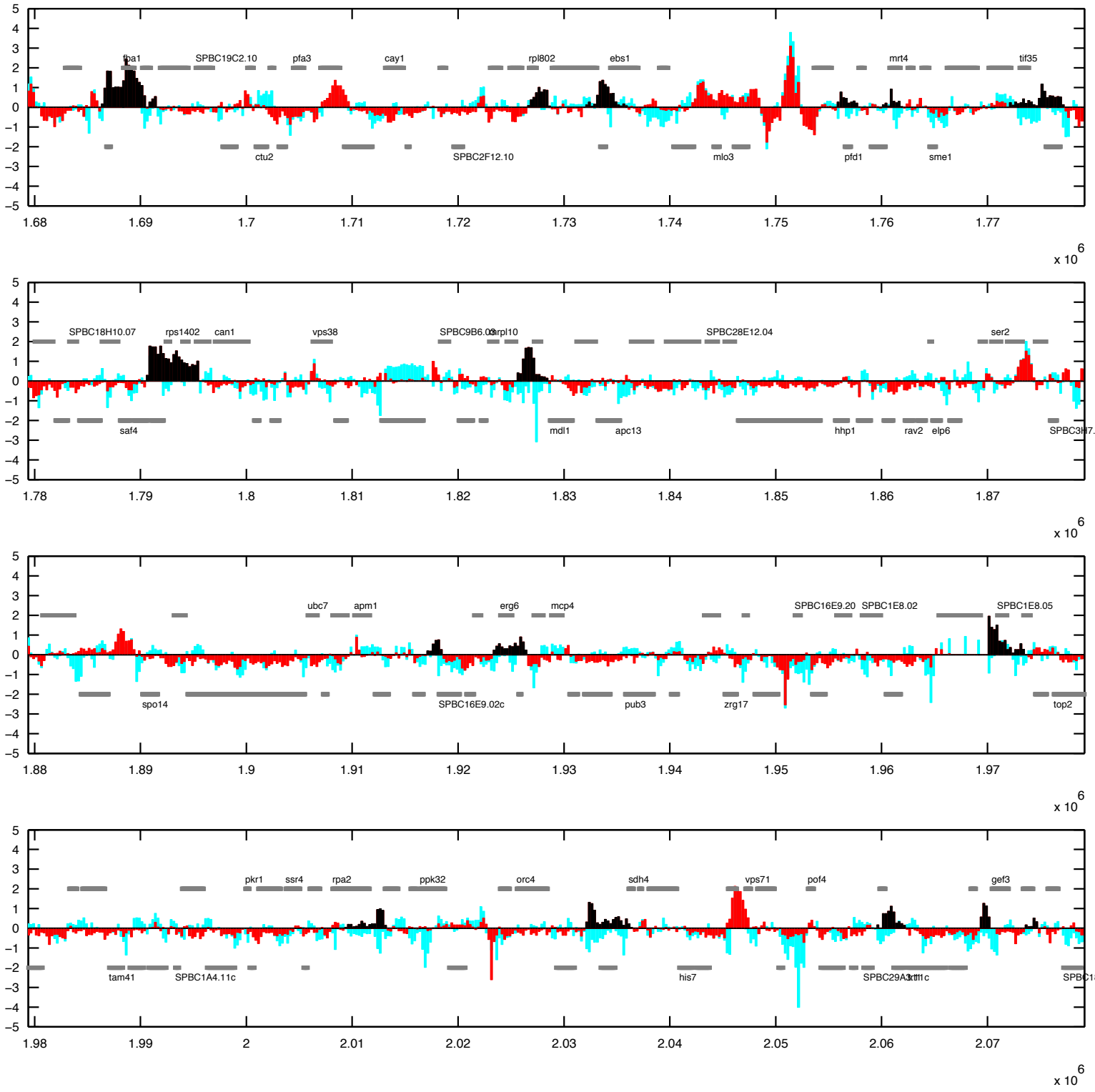

Figure App.F.4.2 Mis4 binding on chromosome II, 5 out of 12

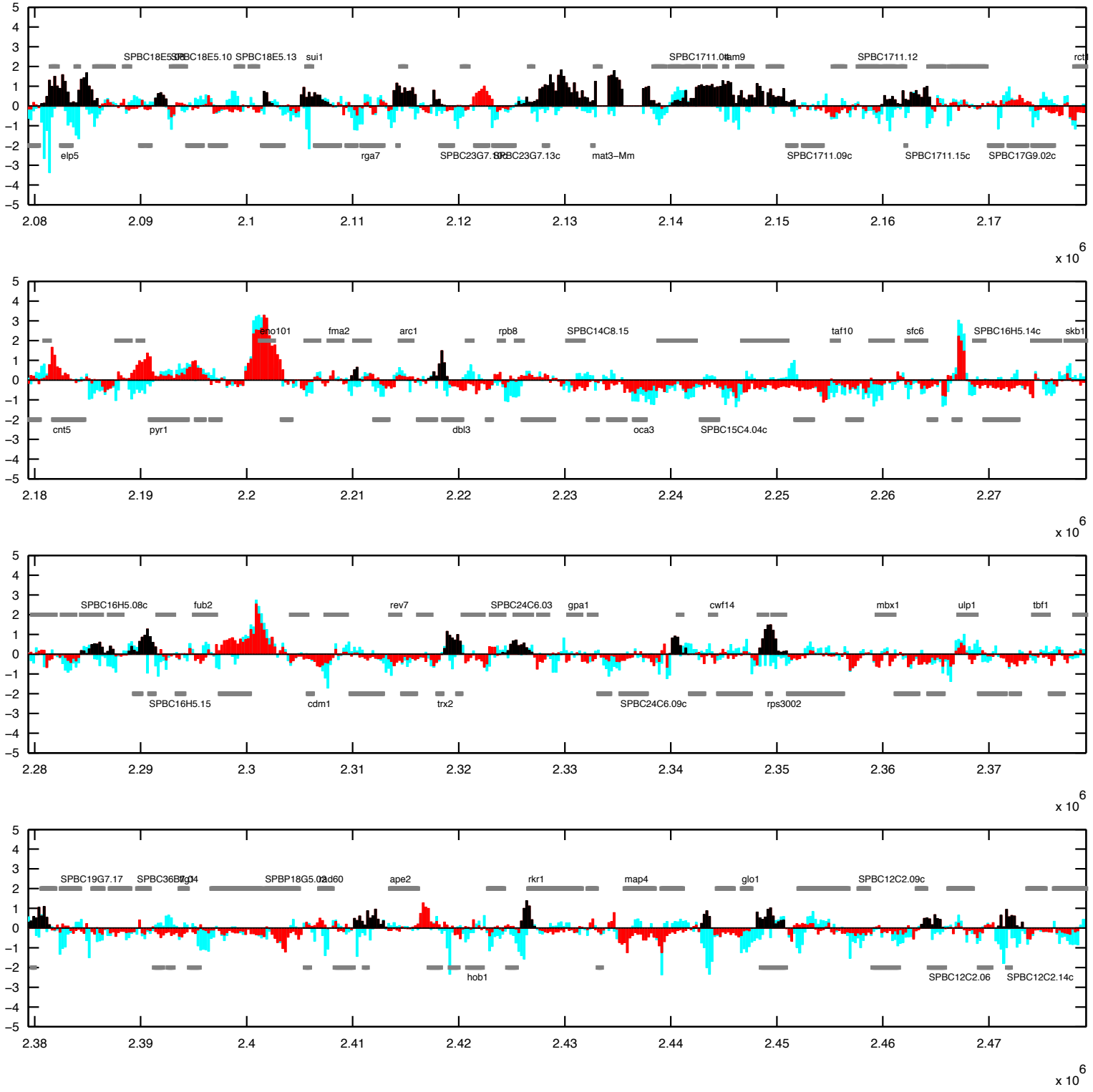

Figure App.F.4.2 Mis4 binding on chromosome II, 6 out of 12

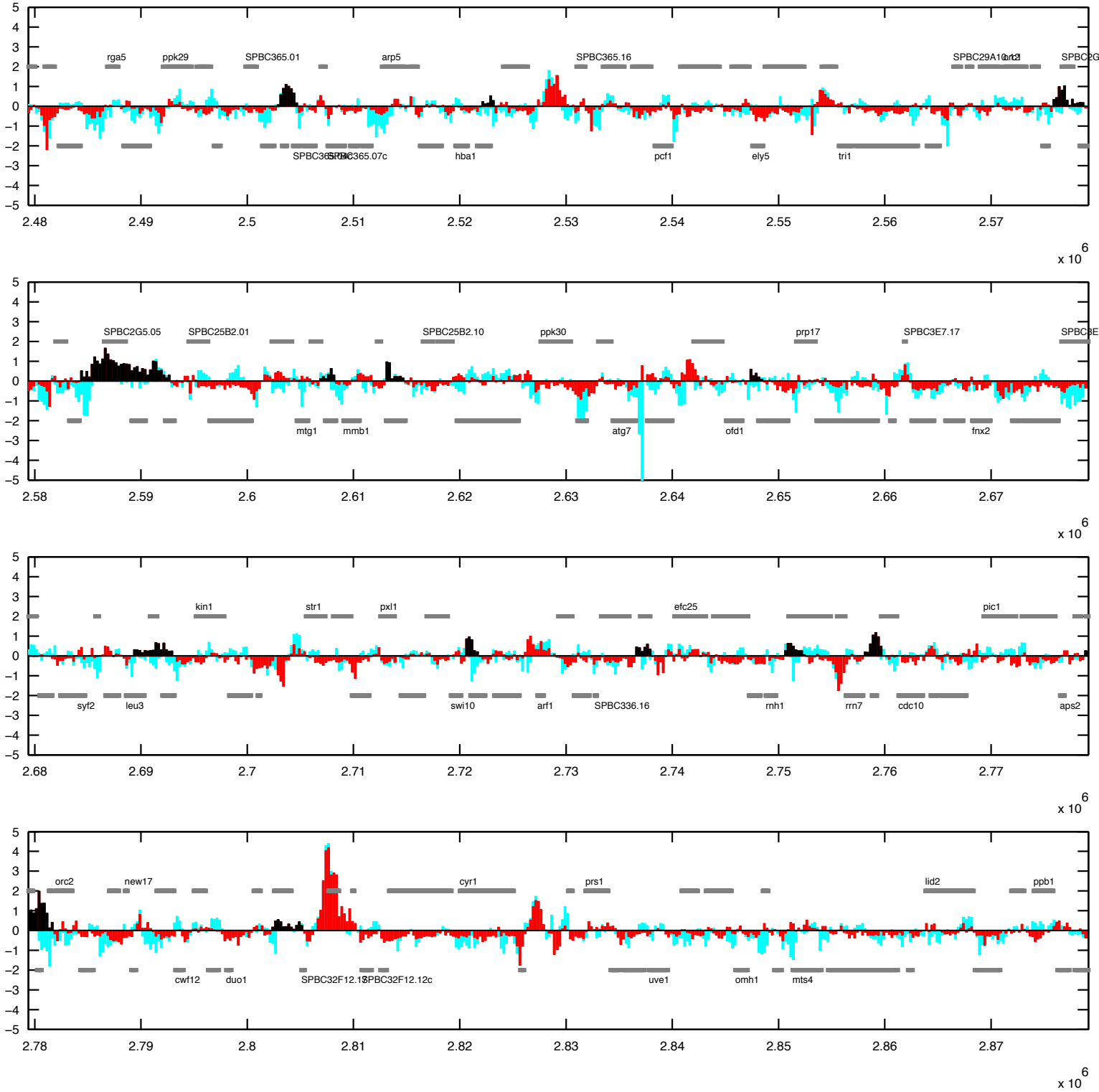

Figure App.F.4.2 Mis4 binding on chromosome II, 7 out of 12

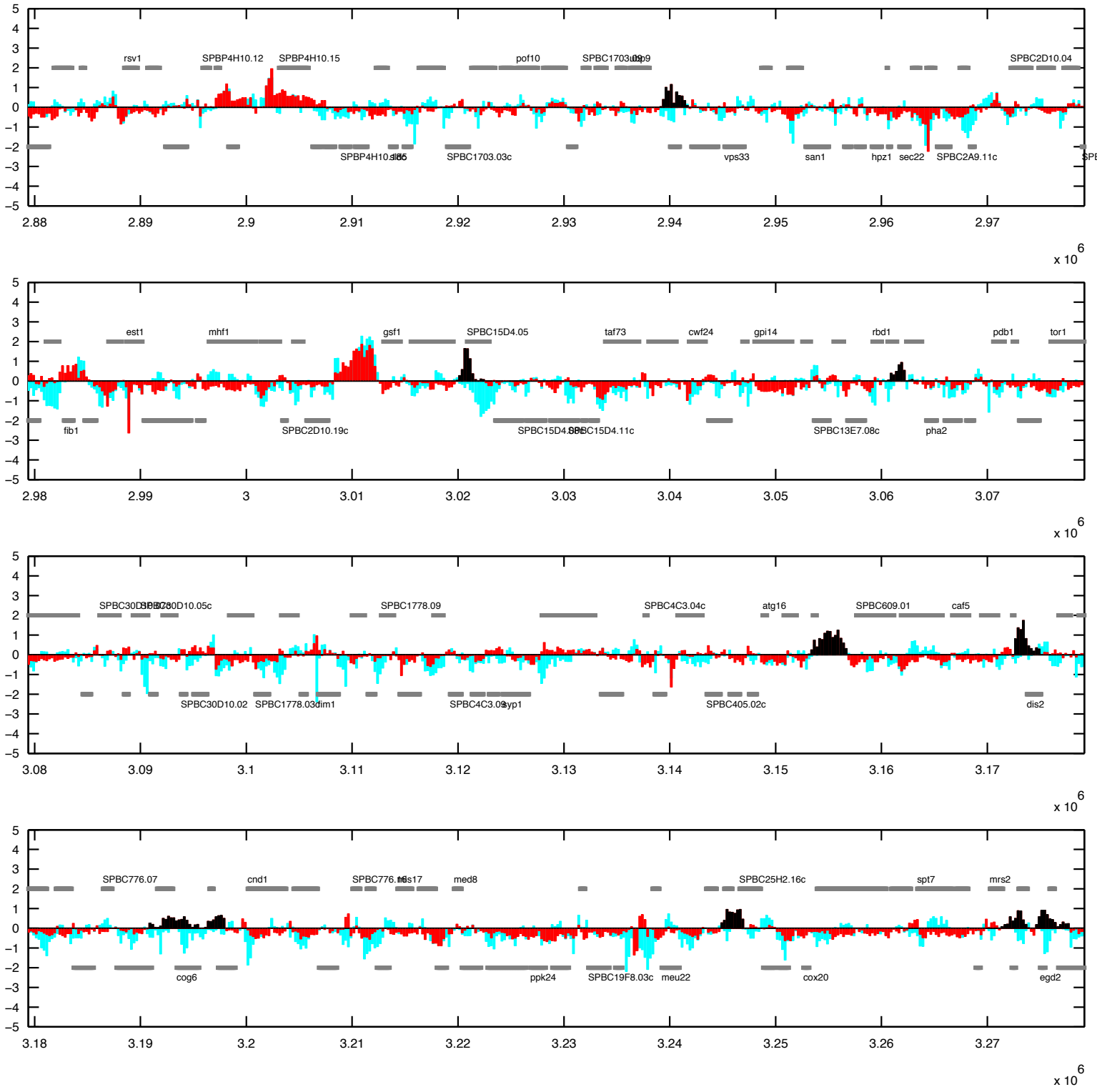

Figure App.F.4.2 Mis4 binding on chromosome II, 8 out of 12

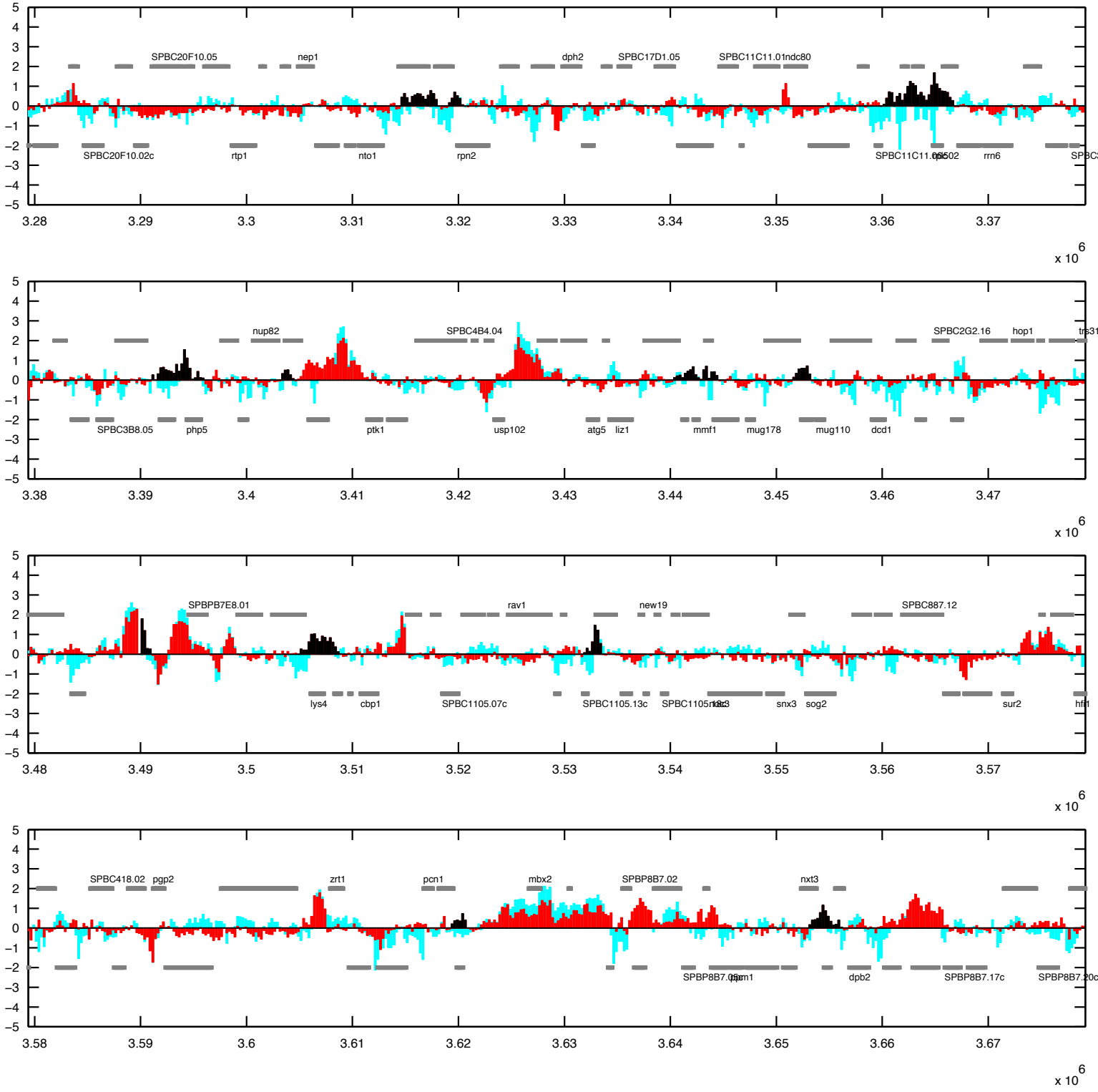

Figure App.F.4.2 Mis4 binding on chromosome II, 9 out of 12

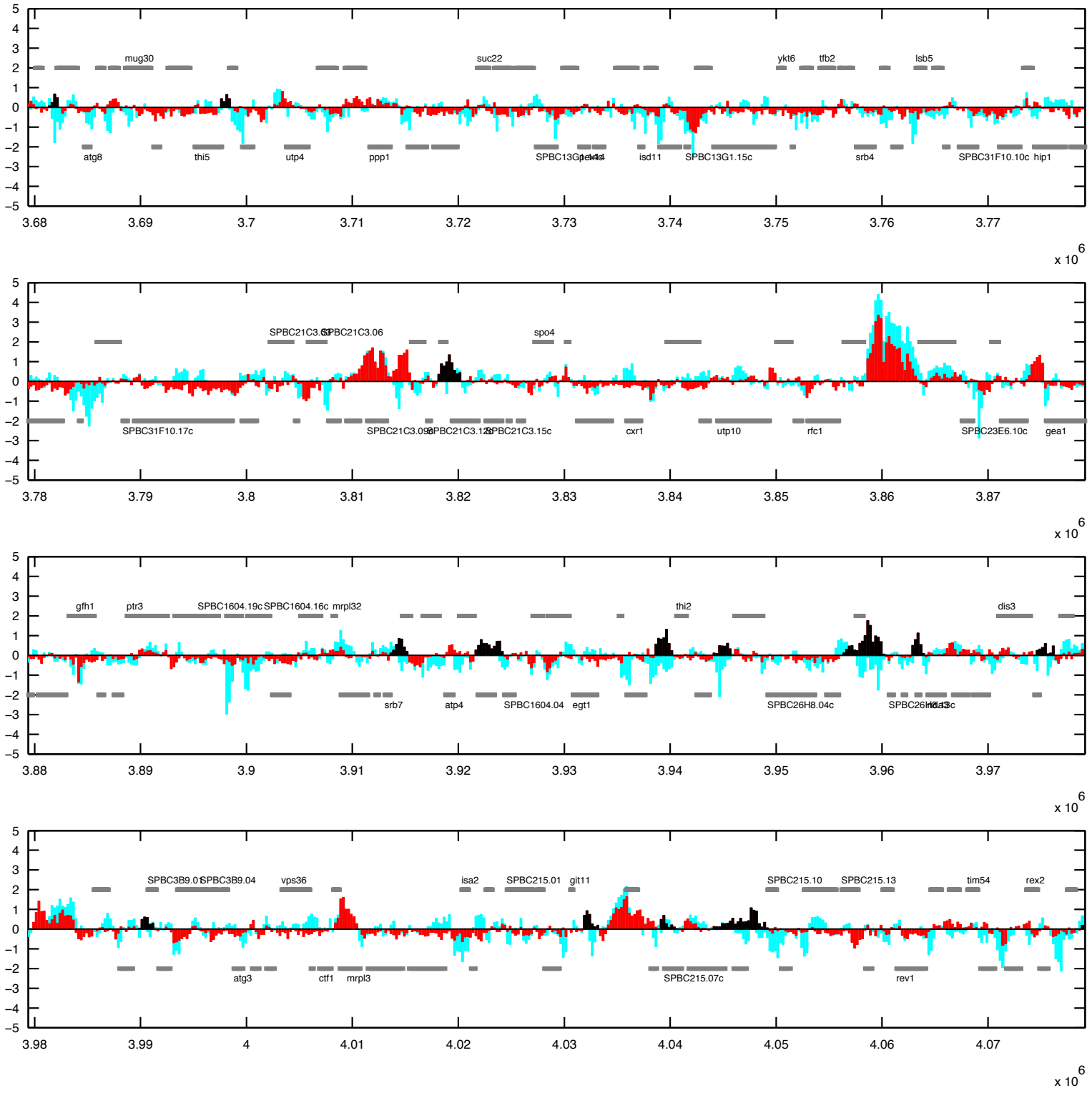

Figure App.F.4.2 Mis4 binding on chromosome II, 10 out of 12

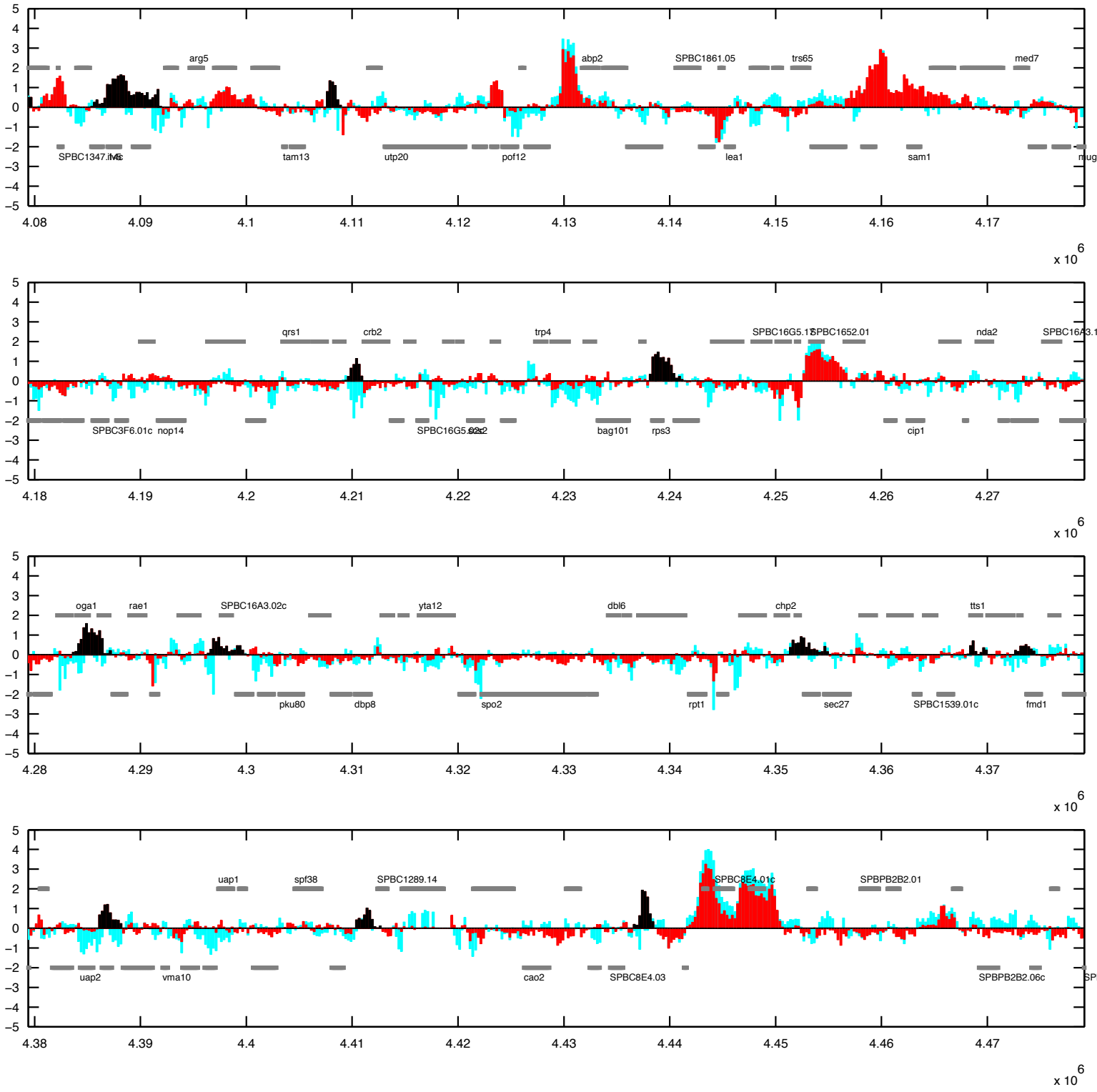

Figure App.F.4.2 Mis4 binding on chromosome II, 11 out of 12

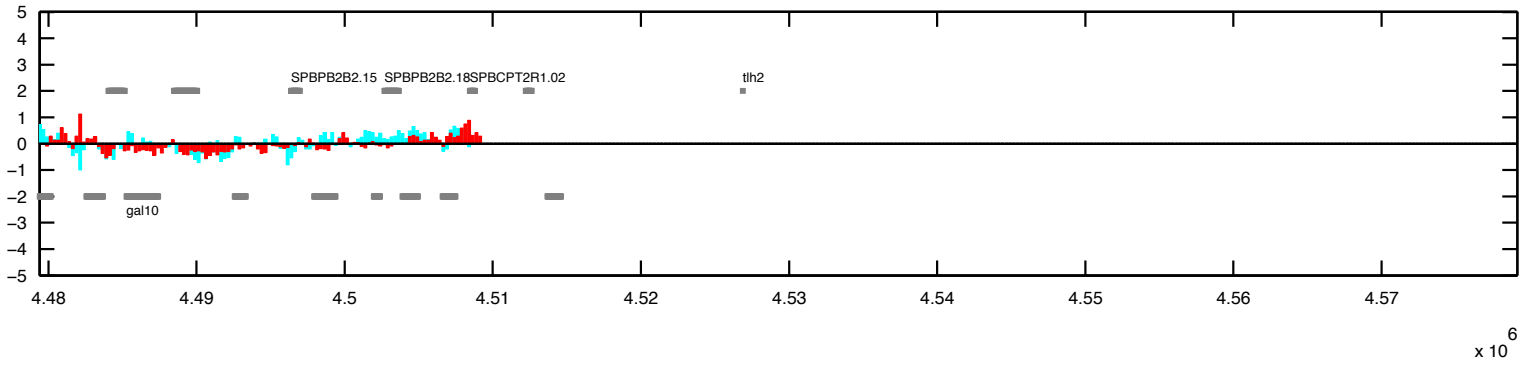

Figure App.F.4.2 Mis4 binding on chromosome II, 12 out of 12

F.5 Called Cohesin Peaks on chromosome III

Rad21 Peaks, chromosome III

Peak Number	Midpoint	Width
1	54000	7000
2	87500	5500
3	98750	7000
4	141125	15250
5	181375	6250
6	192125	2250
7	196750	4000
8	217125	3750
9	220000	1500
10	247000	1500
11	267000	4500
12	282125	1250
13	293625	6250
14	313625	3750
15	330375	4250
16	336875	4250
17	357750	7000
18	408125	6250
19	456875	1250
20	459000	2000
21	466750	7500
22	492500	5500
23	540250	1000

Peak Number	Midpoint	Width
24	546000	7500
25	554750	3500
26	558875	1750
27	561125	1250
28	585750	4500
29	611500	3000
30	624250	7000
31	637250	4000
32	659125	5250
33	662625	1250
34	688250	6000
35	692250	1500
36	704625	4250
37	707750	1000
38	727750	8000
39	756875	7750
40	768750	2000
41	773875	1250
42	776375	3250
43	803750	1500
44	817250	1000
45	832000	7500
46	849000	7000

Table App.F.5.1 - 1 out of 2: Called Rad21 peaks by my peak-calling algorithm on chromosome III.

Peak Number	Midpoint	Width
47	878375	3750
48	881125	1250
49	885625	7250
50	916375	2250
51	918500	1500
52	924625	1250
53	926625	2250
54	932500	4500
55	943875	9750
56	949875	1750
57	971000	1000
58	973875	3750
59	1014750	6500
60	1027000	7000
61	1033000	1500
62	1051375	9250

Peak Number	Midpoint	Width
63	1057125	1750
64	1065875	12750
65	1091100	1000
66	1146225	5750
67	1223975	7750
68	1233225	5750
69	1252850	3500
70	1264100	5000
71	1318600	10000
72	1327975	2750
73	1335475	4250
74	1347600	1000
75	1387350	7000
76	1393850	5500
77	1417975	2250

Table App.F.5.1 - 2 out of 2: Called Rad21 peaks by our peak-calling algorithm on chromosome III.

Mis4 Peaks, chromosome III

Peak Number	Midpoint	Width
1	43750	2500
2	49500	1000
3	63875	1750
4	89875	4250
5	109250	1000
6	114625	2750
7	118500	3000
8	136875	3750
9	140375	2750
10	187250	1500
11	331500	2500
12	352500	2500
13	358750	9000
14	401375	1750
15	409250	1000
16	470500	3500
17	474500	3500
18	482000	2500
19	489375	1750
20	499875	1750
21	506750	2500
22	523000	1000
23	538250	5500

Peak Number	Midpoint	Width
24	553625	2250
25	571250	3500
26	583750	2500
27	614875	1250
28	627125	2750
29	657625	2750
30	690500	1500
31	692000	1000
32	777375	1250
33	817875	2750
34	820250	1000
35	853750	1000
36	867000	1000
37	883125	1250
38	913375	1250
39	918375	2250
40	926625	2250
41	939875	3750
42	948375	6750
43	975000	3500
44	990000	6500
45	1027375	1250
46	1029500	2500

Table App.F.5.2 - 1 out of 2: Called Mis4 peaks by our peak-calling algorithm on chromosome III.

Peak Number	Midpoint	Width
47	1032625	2250
48	1041500	1000
49	1049375	4250
50	1068250	8000
51	1091100	1000
52	1096350	8000
53	1104350	1500
54	1145600	4500

Peak Number	Midpoint	Width
55	1256100	1500
56	1265725	3250
57	1295475	2250
58	1313725	5250
59	1362475	4750
60	1375850	1000
61	1403850	1500
62	1414975	2250

Table App.F.5.2 - 2 out of 2: Called Mis4 peaks by our peak-calling algorithm on chromosome III.

F.6 Cohesin distribution profiles on chromosome III

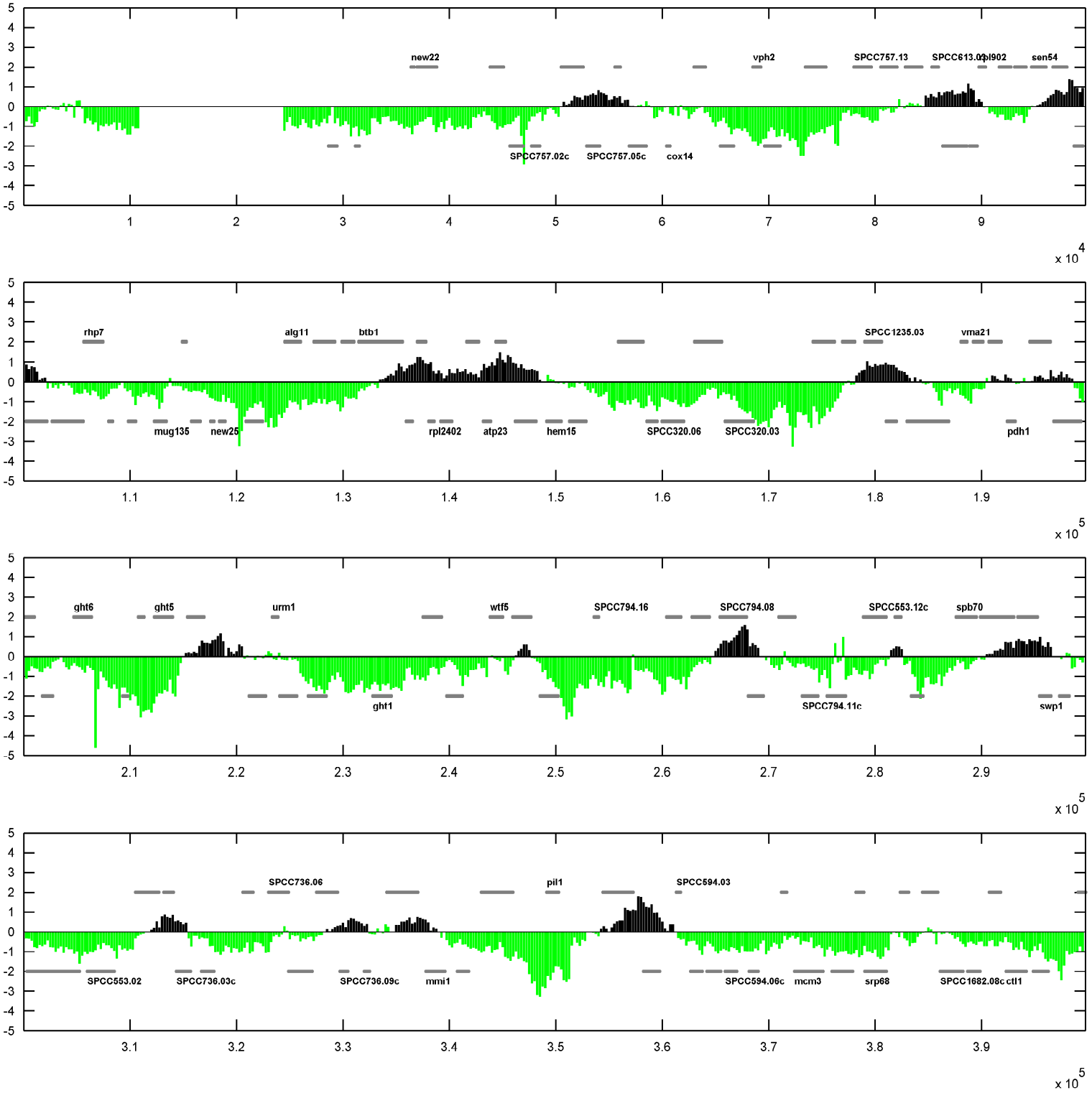

Figure App.F.6.1 Rad21 binding on chromosome III, 1 out of 4. Peaks are based on epitope-tagged Rad21-Pk9 data (Schmidt et al., 2009). Y-axis is in $\log_2$ scale. Called peaks are represented in black, while all other data is plotted in green. Peaks extend from their midpoint to either side until $\log_2$ reaches 0.

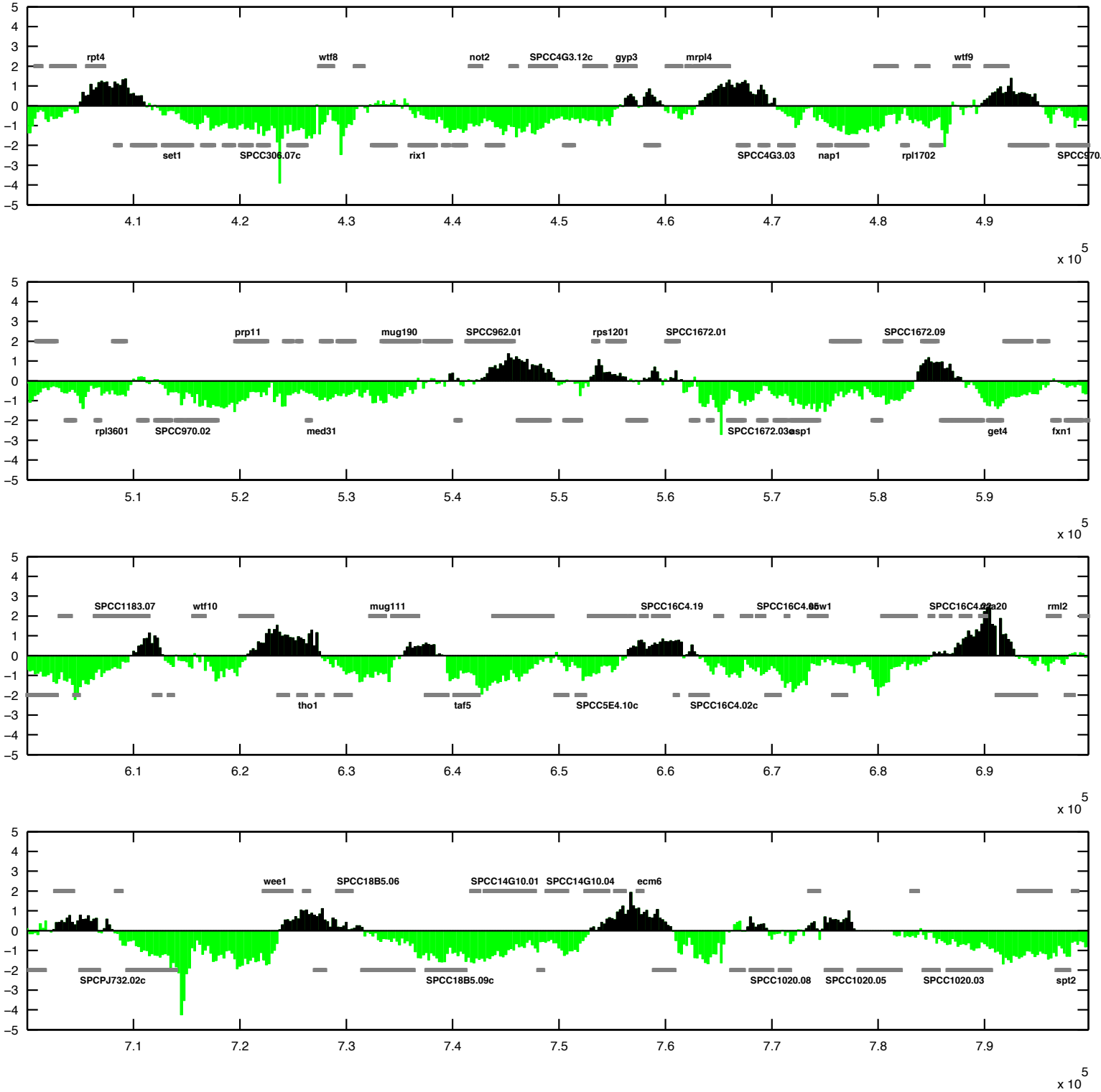

Figure App.F.6.1 Rad21 binding on chromosome III, 2 out of 4.

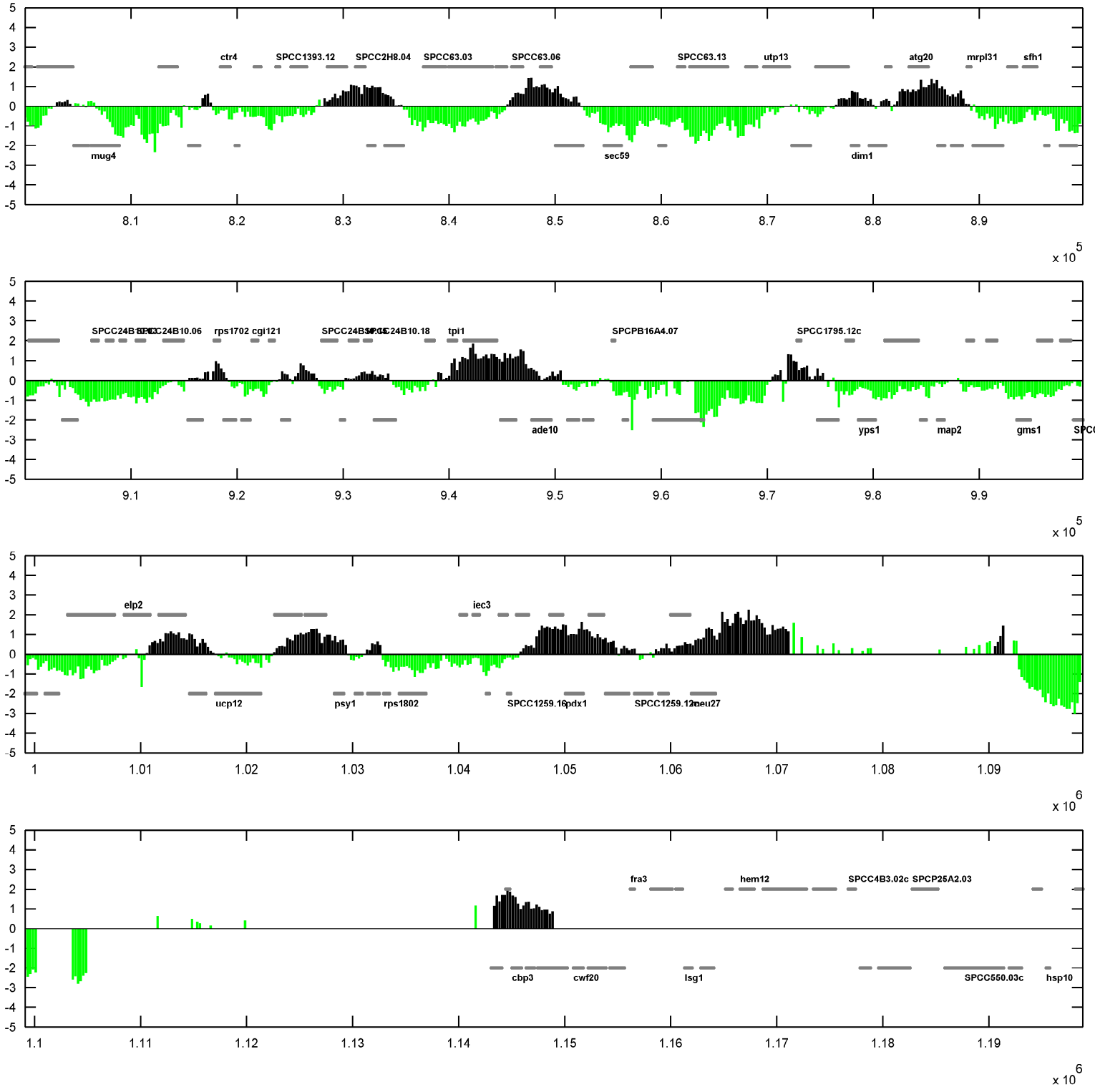

Figure App.F.6.1 Rad21 binding on chromosome III, 3 out of 12.

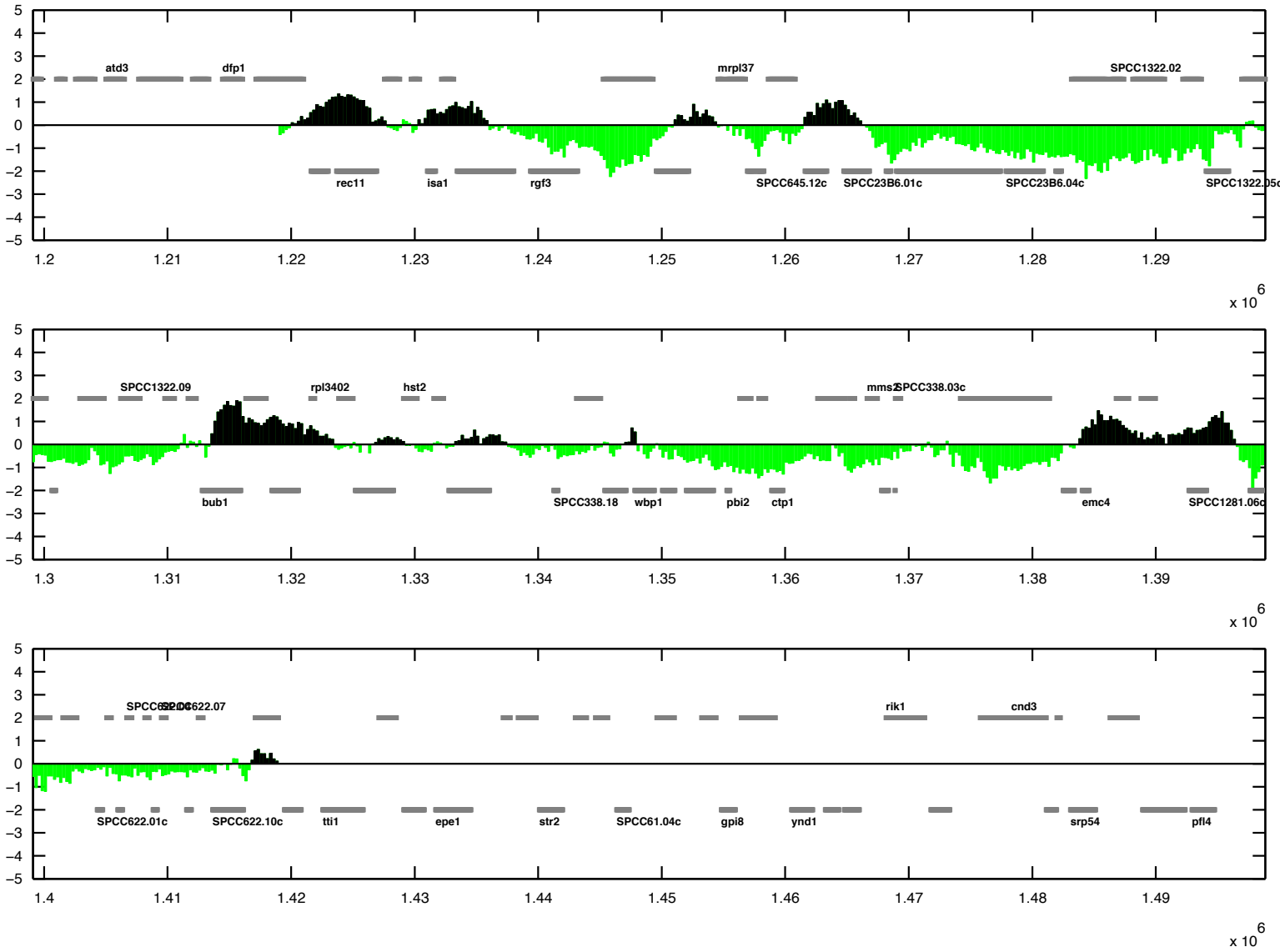

Figure App.F.6.1 Rad21 binding on chromosome III, 4 out of 12.

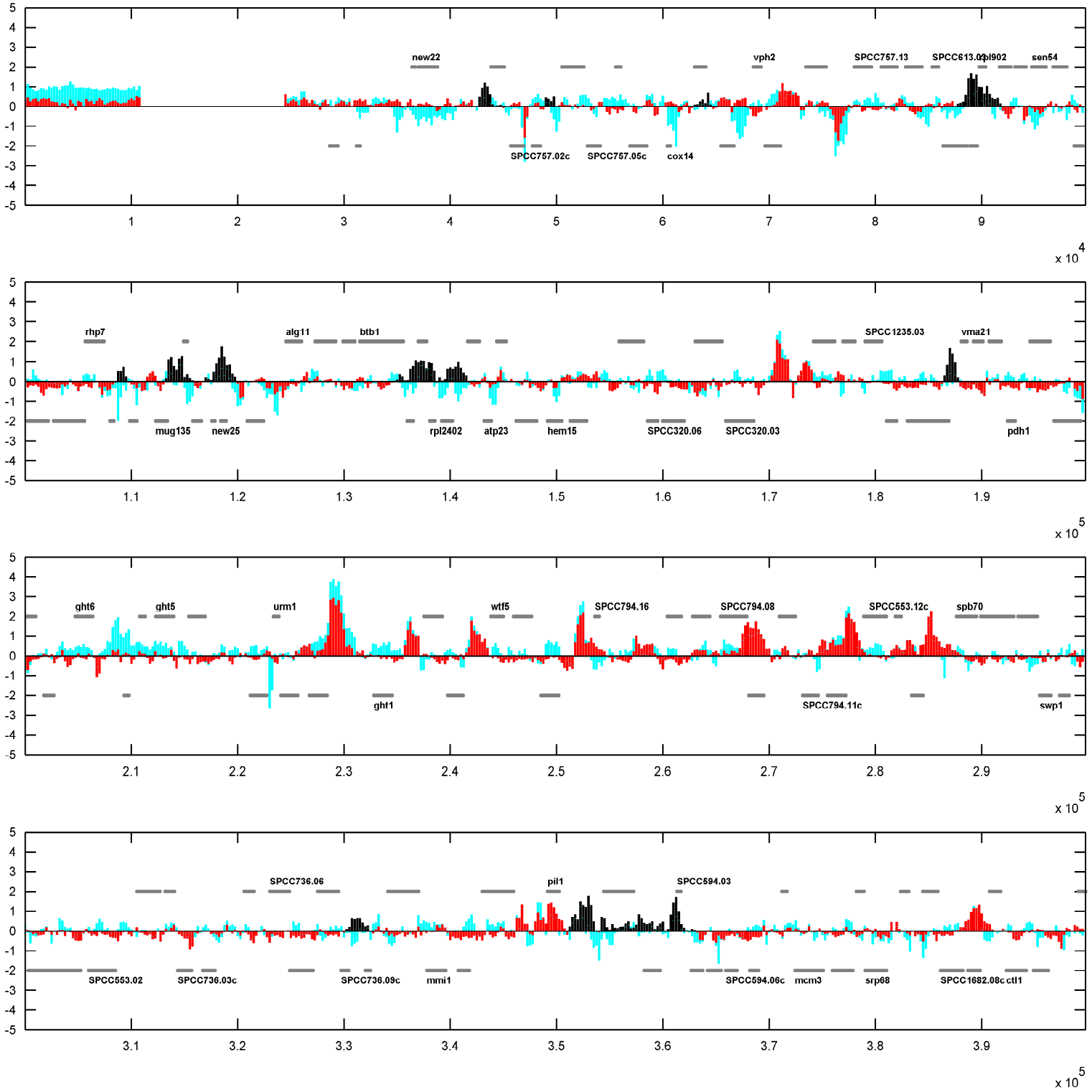

Figure App.F.6.2 Mis4 binding on chromosome III, 1 out of 4. Peaks were called based on comparison of epitope tagged Mis4-Pk9 data to epitope-untagged data Pk9. Y-axis is in $\log_2$ scale. Called peaks are represented in black, while all other data is plotted in red. Note that some regions may appear as peaks, but are not called. This is due to high data of the epitope-untagged data (light blue). Peaks extend from their midpoint to either side until $\log_2$ reaches 0.

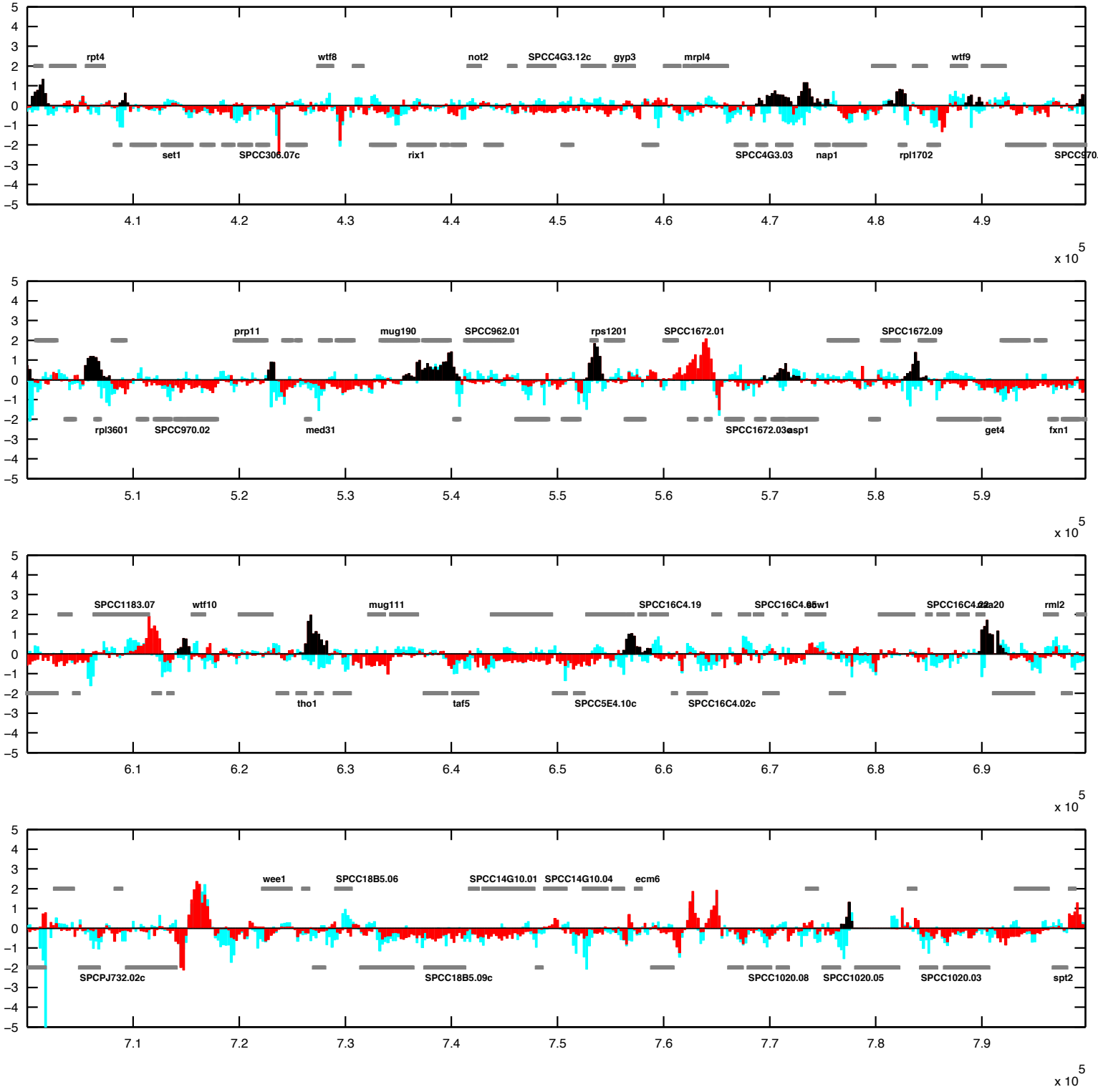

Figure App.F.6.2 Mis4 binding on chromosome III, 2 out of 4

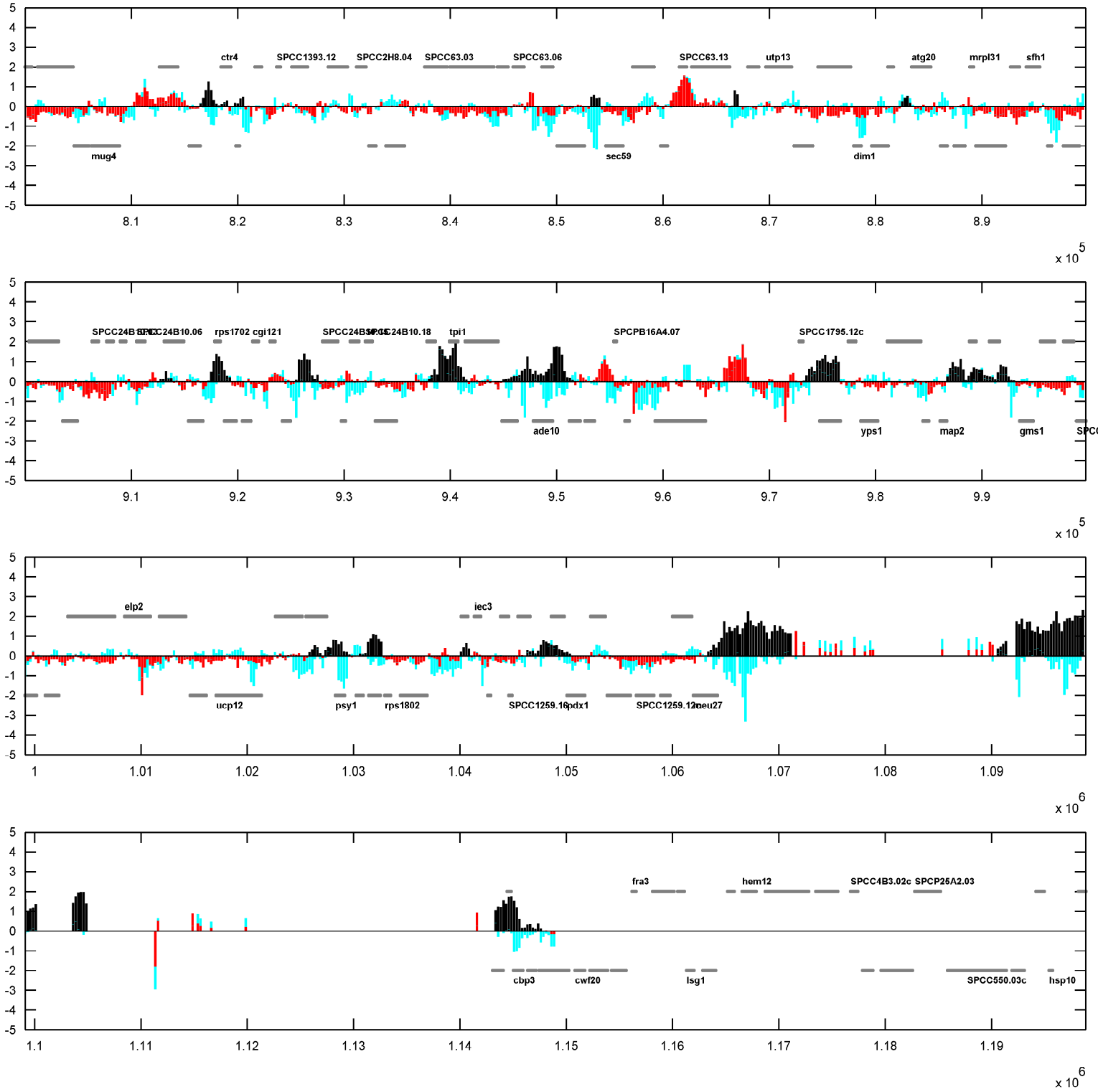

Figure App.F.6.2 Mis4 binding on chromosome III, 3 out of 4

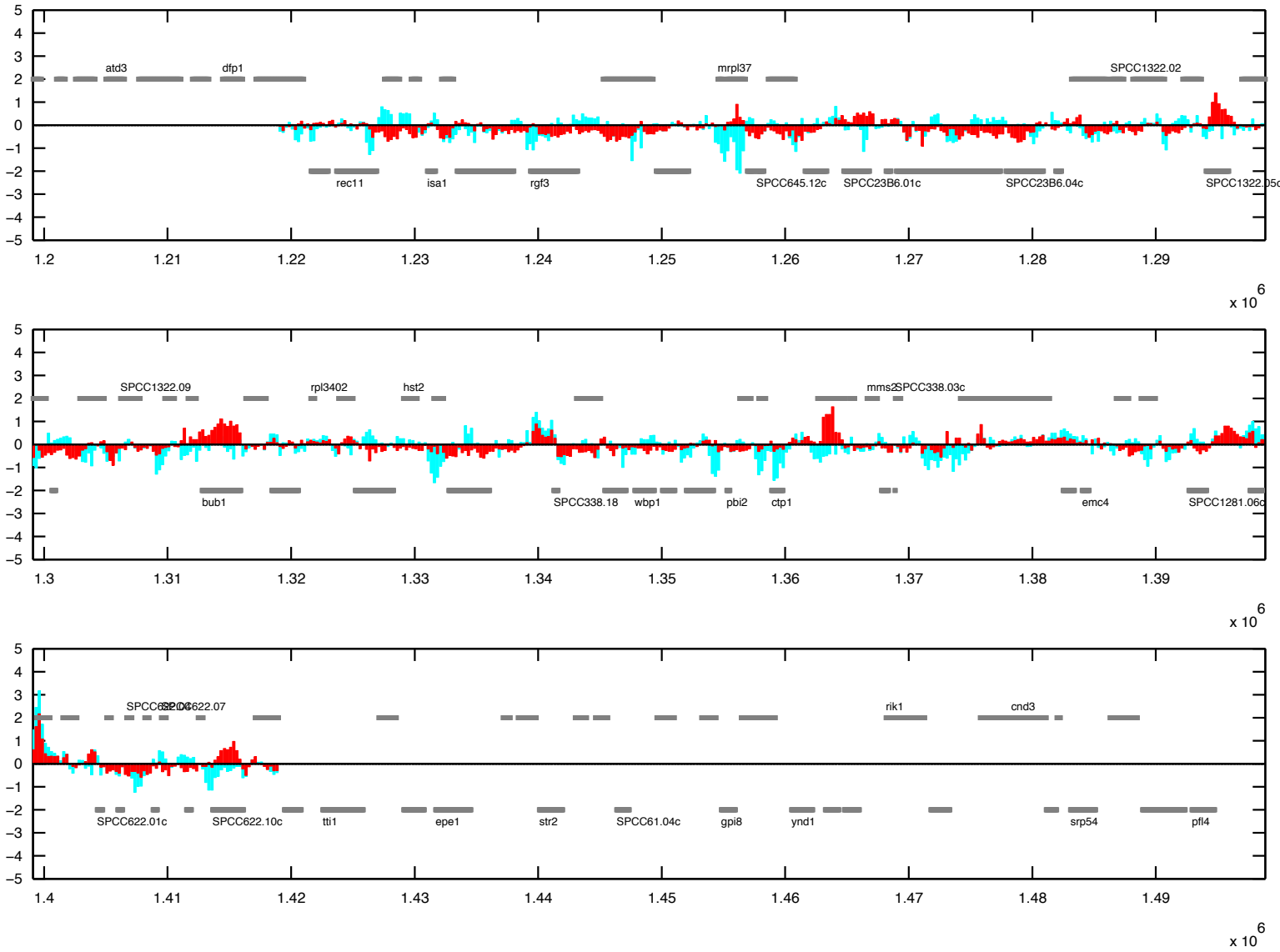

Figure App.F.6.2 Mis4 binding on chromosome III, 4 out of 4