THE UNIVERSITY
of EDINBURGH

COMPUTATIONAL METHODS FOR RNA INTEGRATIVE BIOLOGY

ALINA SELEGA

Doctor of Philosophy
School of Informatics
University of Edinburgh

2017

Alina Selega:
Computational methods for RNA integrative biology
Doctor of Philosophy, 2017

SUPERVISORS:
Guido Sanguinetti
Sander Granneman

ABSTRACT

Ribonucleic acid (RNA) is an essential molecule, which carries out a wide variety of functions within the cell, from its crucial involvement in protein synthesis to catalysing biochemical reactions and regulating gene expression. Such diverse functional repertoire is indebted to complex structures that RNA can adopt and its flexibility as an interacting molecule.

It has become possible to experimentally measure these two crucial aspects of RNA regulatory role with such technological advancements as next-generation sequencing (NGS). NGS methods can rapidly obtain the nucleotide sequence of many molecules in parallel. Designing experiments, where only the desired parts of the molecule (or specific parts of the transcriptome) are sequenced, allows to study various aspects of RNA biology. Analysis of NGS data is insurmountable without computational methods.

One such experimental method is RNA structure probing, which aims to infer RNA structure from sequencing chemically altered transcripts. RNA structure probing data is inherently noisy, affected both by technological biases and the stochasticity of the underlying process. Most existing methods do not adequately address the issue of noise, resorting to heuristics and limiting the informativeness of their output. In this thesis, a statistical pipeline was developed for modelling RNA structure probing data, which explicitly captures biological variability, provides automated bias-correcting strategies, and generates a probabilistic output based on experimental measurements. The output of our method agrees with known RNA structures, can be used to constrain structure prediction algorithms, and remains robust to reduced sequence coverage, thereby increasing sensitivity of the technology.

Another recent experimental innovation maps RNA-protein interactions at very high temporal resolution, making it possible to study rapid binding events happening on a minute time scale. In this thesis, a non-parametric algorithm was developed for identifying significant changes in RNA-protein binding time-series between different conditions. The method was applied to novel yeast RNA-protein binding time-course data to study the role of RNA degradation in stress response. It revealed pervasive changes in the binding to the transcriptome of the yeast transcription termination factor Nab3 and the cytoplasmic exoribonuclease Xrn1 under nutrient stress. This challenged the common assumption of viewing transcriptional changes as the major driver of changes in RNA expression during stress and highlighted the importance of

degradation. These findings inspired a dynamical model for RNA expression, where transcription and degradation rates are modelled using RNA-protein binding time-series data.

ACKNOWLEDGEMENTS

I would like to thank my supervisors Guido Sanguinetti and Sander Granneman, without whom this PhD project would not have been as successful as it was. Guido introduced me to the field of computational biology and taught me a great deal about both the theory and practical applications. He always had time for me when I (perhaps too often) needed advice, encouragement, or even simply a chat. Throughout the whole of my PhD, Sander provided me with a rich resource of data, which is invaluable for computational projects. Sander was also infinitely patient with me as I continued to ask undoubtedly repeating questions about various biological intricacies. I feel privileged to have worked with my supervisors who are as excellent experts as they are friendly and caring people.

I would also like to thank Ian Simpson for assisting with my annual reviews and giving me his support when I needed it. Many members of Informatics staff have spent their time talking to me about the future and sharing their experience, which was truly appreciated. Jane Hillston, Iain Murray, and Alison Edie deserve a special mention for their kindness and support. A special thanks goes to all students and staff of the DTC; it was a great pleasure to meet all of them and an incredible feeling to be a part of this community, which seemed so large back when I joined in 2013 and is almost at its end today.

I want to thank the research groups in CIBIO and FBK at the University of Trento, with whom I spent almost three months during my PhD. Toma Tebaldi, Gabriella Viero, Primoz Knap, and others made me feel welcome and made my time in Italy unforgettable. This too, goes back to Guido who put me in contact and organised the collaboration.

Our research group has provided me with much more than just colleagues. I will miss our daily morning coffee, our Christmas skittles with the customary pilgrimage to Arthur's seat, and our invariably outdoorsy summer socials (which, the Greeks were convinced, were intended to kill them). I am indebted to David, Gabriele, Botond, Edward, Giulio, Tom, Anastasis, Yuanhua, Andreas, Michalis, Dimitris, Michael, Emily, Ylenia, Fabio, Daniel, Van-Anh, Veronica, and, of course, Guido for creating such a supportive and fun atmosphere within the group over the years.

It would not be possible for me to get to the day when I am writing these lines if it wasn't for the endless and unconditional support of my friends and family. Matt, Paolo, Martino, Nathalie and others were there to experience the whole thing along-

side me. Sam, Svet, and Dragos were a part of our DTC13 team. Stefaniia, Liz, Artem, Dasha, Ivan and the *What? Where? When?* squad never failed to cheer me up. Thousands of kilometers didn't stop Viv, Natasha, and June from making me feel their warmth. John and Irene in their kindness gave me a beautiful and relaxing place to stay in the final stretch. My parents and Marina always knew which words to say (or not to say) and would drop everything to speak with me when things were not going well.

Finally, I dedicate this thesis to Gavin Gray, who did everything to support me every step of the way. I would not be here without him.

DECLARATION

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Edinburgh, 2017

Alina Selega, November 29, 2017

CONTENTS

ABBREVIATIONS

e.g. for example

i.e. that is to say

DNA deoxyribonucleic acid

RNA ribonucleic acid

mRNA messenger RNA

rRNA ribosomal RNA

tRNA transfer RNA

miRNA micro RNA

snRNA small nuclear RNA

snoRNA small nucleolar RNA

ncRNA non-coding RNA

CUT cryptic unstable transcript

SUT stable uncharacterised transcript

XUT Xrn1-sensitive unstable transcript

RNAP RNA polymerase

Pol II polymerase II

NNS Nrd1-Nab3-Sen1

GO gene ontology

NGS next-generation sequencing

RNA-Seq RNA sequencing

PCR polymerase chain reaction

cDNA complementary DNA

RT reverse transcriptase

RPKM reads per kilobase of transcript per million mapped reads

FPKM fragments per kilobase of transcript per million mapped reads

TPM transcripts per million

pH potential of hydrogen

PARS parallel analysis of RNA structures

FragSeq fragmentation sequencing

SHAPE selective 2′-hydroxyl acylation analysed by primer extension

SHAPE-MaP selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling

DMS dimethyl sulfate

NAI 2-methylnicotinic acid imidazolide

1M7 1-methyl-7-nitroisatoic anhydride

TCP^{EM} two-channel Poisson expectation maximization

BUM-HMM beta-uniform mixture hidden Markov model

RBP RNA-binding protein

UV ultraviolet

CLIP cross-linking immunoprecipitation

CRAC cross-linking and analysis of cDNA

PAR-CLIP photoactivatable-ribonucleoside-enhanced cross-linking and immunoprecipitation

iCLIP individual-nucleotide resolution CLIP

HMM hidden Markov model

EM expectation-maximisation

GP Gaussian process

ODE ordinary differential equation

PDE partial differential equation

LDR log-ratio between drop-off rates

AUC area under the curve

PR precision-recall

ROC receiver operating characteristic

ALS amyotrophic lateral sclerosis

1

INTRODUCTION

All life as we know it, from the first primitive cells to the spectacular diversity of the organisms that have inhabited our planet throughout its history, shares the basic principles of its biological organisation. All life forms are composed of cells and all cells use the same fundamental types of molecules to represent and manipulate their *genetic* information: a complete set of rules required for growth, development, functioning, and reproduction. Understanding the mechanisms of a great many intricate biological processes that take place in the cell every second can have a major impact on medical research, molecular and evolutionary biology, and many other research fields of great significance.

The main biological actors in the cell are the DNA, RNA, and proteins. All three types of molecules are necessary for correct cellular functioning and are tightly interconnected by the underlying processes. RNA (ribonucleic acid) fulfills an important role of transforming the instructions contained in the DNA code into real biological events, implemented by the proteins. It is now known that the functions of RNA extend much further, regulating various aspects of cell biology. The development of sequencing methods that can assay a wide range of phenomena happening on a genome-wide scale, together with speed gains and cost reductions which made these methods accessible for a wider research community, has generated a vast multitude of datasets elucidating all stages of gene expression. The analysis of *next-generation sequencing* data, collected in various biological conditions and at increasingly high spatial and temporal resolution, is insurmountable without advanced computational techniques.

In this thesis, I discuss computational methods that I developed for the analysis of next-generation sequencing data with the aim to investigate the two main aspects of RNA biology. Namely, I focus on two experimental technologies: RNA structure probing methods, which assay the molecular structure of RNA, and cross-linking methods, which map interactions between RNA and proteins. Both are experimentally challenging protocols studying inherently complex and noisy biological processes. The methods, discussed in this thesis, address these problems using statistical and machine learning techniques and provide the tools for the modelling and analysis of these experimental signals.

1.1 CONTRIBUTIONS

The contributions of this thesis can be summarised as follows:

1. The development of a computational analysis pipeline for modelling of RNA structure probing data, which explicitly captures biological variability of the data, provides automated strategies for correcting biases of the technology, and generates an interpretable probabilistic output. The software implementing the method was released as a peer-reviewed *Bioconductor* package (Selega et al., 2016) and the paper presenting the method was published in *Nature Methods* (Selega et al., 2017).

2. The development of a non-parametric method for modelling RNA-protein interactions time-series data of high temporal resolution, obtained with a novel cross-linking protocol χCRAC, which enables testing individual transcripts for differential binding with the protein of interest between biological conditions. The software implementing the method is freely available and the paper presenting χCRAC and using the method for parts of its analysis was published in *Nature Communications* (van Nues et al., 2017).

3. The conception and theoretical formulation of a dynamical model for RNA expression under nutrient stress, which uses RNA-protein interactions time-series data to infer the underlying transcription and degradation rates determining the total RNA abundance. The implementation of the proposed model is presently ongoing in collaboration with David Schnoerr, Edward Wallace, Sander Granneman, and Guido Sanguinetti.

1.1.1 *Collaboration with Alessandro Quattrone*

Additionally, during the duration of this PhD project, I collaborated with Alessandro Quattrone's lab at the Centre for Integrative Biology, University of Trento, Italy. The collaboration was a part of the "Axonomix" research project, further linked to the University of Edinburgh through a partnership with Thomas Gillingwater at the Centre for Integrative Physiology. The project aimed to investigate the effects of translational impairment in amyotrophic lateral sclerosis (ALS).

I developed exploratory deconvolution analysis for next-generation sequencing data of total and polysomal mRNA, isolated from the spinal cords of pre- and post-symptomatic SOD1 mice (a common model for ALS). The deconvolution computational approach I applied (Kuhn et al., 2011) aimed to identify post-transcriptional

disease-specific changes, while taking into account degenerative alterations in histological tissue composition, associated with the disease phenotype. As the method relies on reference genes expressed only by certain cell types, I performed the stability analysis, recovering cell type-specific differentially regulated genes for different combinations of reference genes. The software implementing these analyses was passed on to the computational biologist of the group Toma Tebaldi. This work did not fit in with the overall narrative of this PhD project and for this reason was not included in the thesis.

The first paper reporting the results obtained within the "Axonomix" research project was recently published in *Cell Reports* (Bernabo et al., 2017). The findings of this paper are outside of my involvement in the "Axonomix" project.

1.1.2 *Additional publication*

In the final year of my PhD, I presented my work on the modelling of RNA structure probing data at the Wellcome Trust Conference on Computational RNA Biology and co-authored an invited meeting report with my supervisor Guido Sanguinetti. The report reviewed the state of the art of this interdisciplinary field and was published in *Genome Biology* (Selega and Sanguinetti, 2016).

1.2 STRUCTURE OF THE THESIS

This thesis is structured as follows. Chapter 2 provides the necessary background to RNA biology, focussing on the experimental technologies relevant to this thesis. Chapter 3 gives a brief introduction to the probabilistic models, mathematical frameworks, and machine learning techniques used by the methods developed for this thesis. Chapter 4 discusses the BUM-HMM method for modelling RNA structure probing data and provides the associated journal paper, evaluating the method on real and synthetic data in various contexts. Chapter 5 presents the algorithm for identifying differential RNA-protein binding between conditions and applies it to study the role of different degradation pathways in the context of stress response, proposing a dynamical model for RNA expression motivated by the findings of the discussed analyses. Finally, Chapter 6 provides an overview of the methods presented in this thesis and discusses possible avenues for future research.

2

INTRODUCTION TO RNA BIOLOGY

It is estimated that there are more than 10 million species currently living on our planet (Alberts et al., 2014). All of them, dramatically ranging in their diversity, are composed of cells that share the same mechanisms for most of their basic processes. It is those universal features common to all known life that enable the notion of *heredity* — the ability of each species to yield progeny, faithfully reproducing itself.

Whether it is a single cell organism or a multicellular giant such as the human body, its entirety has been generated through cell division starting from the very first cell. Thus, a cell is the smallest unit equipped both with hereditary information and the necessary machinery required for constructing a new cell in its own image.

All cells store their hereditary information in the form of a linear chemical code provided by the double-stranded molecules of DNA (deoxyribonucleic acid). Given that living cells were estimated to have been evolving for more than 3.5 billions years (Alberts et al., 2014), such permanency of information representation is astonishing and seems to be fundamental to the definition of life.

2.1 DNA STRUCTURE AND REPLICATION

Each strand of the DNA is composed from a long chain of *nucleotides*, drawn from a 4-letter alphabet: A, T, G, and C. A nucleotide has two parts: a sugar (deoxyribose) with an attached phosphate group and a *base*, which can be adenine (A), guanine (G), thymine (T), or cytosine (C). Sugars are linked together with their phosphate groups, forming a chain with protruding bases. An isolated strand can be composed in any order. However, in living cells DNA is not synthesised in isolation, but from an existing template strand. The bases in the template strand bind to the bases in the synthesised strand according to the strict rule of base complementarity: A binds to T and G binds to C. The base-pairing holds the synthesised molecule in place and selects the monomer that should be added next, enlarging the chain. This process creates a double-stranded DNA molecule with two exactly complementary strands, which twist around each other forming the well-known double helix.

The bonds between the base pairs are weaker than the sugar-phosphate links, which makes it possible to separate the strands without breaking their sugar-phosphate

backbones. The separated strands can then become template strands, creating more copies of their hereditary information; a process called DNA *replication*.

2.2 RNA TRANSCRIPTION

Hereditary information is expressed with the help of another dedicated molecule, closely related to DNA. RNA (ribonucleic acid) copies a part of the DNA code and uses it to guide synthesis of proteins, another major class of macromolecules that put the cell's genetic information into action.

In its structure, RNA is very similar to DNA, with a few differences. The RNA backbone is comprised of *ribose*, a sugar with a hydroxyl group at the 2' position, which is replaced with a hydrogen atom in the related deoxyribose. Further, one of the RNA bases is different (uracil (U) instead of thymine (T)). All other bases are the same and the base complementarity rule still holds (however, other base pair interactions are possible (Reece et al., 2014)). During *transcription*, the RNA molecule is synthesised from the template DNA strand by the enzyme RNA polymerase, much like the DNA monomers are assembled during DNA replication. The resulting *transcript* thus encodes a part of the genetic information encoded by the DNA, even if it is in a slightly different alphabet.

2.3 GENE EXPRESSION

The main difference between RNA and DNA comes with their usage within the flow of genetic information within a cell. Many transcripts can be repeatedly synthesised from the same DNA segment. Thus, DNA represents the fixed "archive" of genetic information, whereas RNA molecules copy its certain parts and are mass-produced. The transcripts arising from the protein-coding segments of the genetic code (or *genes*) are called the *messenger RNA* (mRNA) and each transcript is *translated* into a molecule of the protein species it encodes.

These two steps — *transcription*, which transforms protein-coding genes into mRNA, and *translation*, which generates proteins those genes express from the mRNA — together constitute the process of *gene expression*, one of the most important functions implemented by the cell.

2.4 IMPORTANCE OF RNA IN CELLULAR FUNCTION

Another big difference between the two nucleic acids stems from the structure of RNA. Unlike DNA, it is predominantly single-stranded and thus has a flexible backbone, which allows the weak base pair bonds to form between the complementary parts of the same nucleotide chain. This base pairing leads to a variety of possible shapes the RNA molecule can fold into. Different shapes enable selective recognition and binding of other molecules and can even serve as a catalyst for chemical reactions.

These crucial abilities of RNA to store and copy genetic information, as well as catalyse biochemical reactions gave rise to the "RNA world" hypothesis, which suggests that self-replicating RNAs proliferated before the DNA and proteins evolved (Cech, 2012). It is thought that DNA could have taken on the information storage function due to its increased stability compared to the more fragile RNA, while proteins became the main biocatalysts as their abundance and the diversity of *amino acids*, their building blocks, makes them more versatile.

The wide-ranging capabilities of RNA make it a strong candidate for important regulatory roles in cellular function. These roles have indeed been confirmed as many novel experimental technologies have been recently generating vast amounts of data, which dramatically enriched our understanding of RNA biology. It is now clear that the importance of RNA extends much further than just an intermediate step in gene expression.

The majority of RNAs do not code for a protein (non-coding RNAs, ncRNAs) but instead have crucial regulatory functions (Mattick and Makunin, 2006). Some very abundant examples are ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs), both prominently involved in translation. Very short transcripts called microRNAs (miRNAs) can implement gene silencing by binding to mRNA and blocking its translation (Matzke and Matzke, 2004). Ribozymes (RNA enzymes) can catalyse biochemical reactions such as cleaving and ligating other molecules (Fedor and Williamson, 2005). RNA-protein complexes called spliceosomes perform RNA processing by splicing introns out of pre-mRNA (Berg et al., 2002). Some RNAs can even modify their own activity: a riboswitch is a regulatory segment of mRNA that can change the resulting protein production by binding a small molecule (Nudler and Mironov, 2004).

2.5 NEXT-GENERATION SEQUENCING TECHNOLOGIES

Our understanding of the transcriptome diversity and the richness of its functional repertoire was greatly aided by *next-generation sequencing* (NGS) (Schuster, 2007).

NGS is a technological advancement of Sanger sequencing, which was automated in the first generation of DNA sequencers. NGS provides significant improvements in speed, cost, accuracy and requires a smaller sample size. The main contribution to the improvement in sequencing speed came with the development of parallel analysis, giving the technology its second name "*high-throughput* sequencing".

Different high-throughput sequencing technologies exist, making use of various signals they detect, e.g. change in pH used in ion torrent sequencing. A common high-throughput sequencing method called *Illumina sequencing* relies on optical signals. A typical experiment involves cleaving the sample into shorter reads and amplifying them with the polymerase chain reaction (PCR). The amplified reads, separated into single strands, are attached to a slide which is flooded with a mixture of DNA polymerase and nucleotides, fluorescently labelled by base (Fig. 2.1a). These labelled nucleotides also have a terminator which stops the chain from growing once a nucleotide is added. An image is taken of the slide, detecting fluorescent signals corresponding to the bases added to the end of each read (Fig. 2.1b). For the next cycle, the terminator and the fluorescent label are removed from the added nucleotides, allowing the next base to be added. The process is repeated, imaging after adding each nucleotide, and thus, sequencing many reads at the same time (Fig. 2.1c). NGS is so massively parallel that 300 billion bases of DNA can be processed in a single run on a single chip. Other high-throughput sequencing methods differ in technical protocols but broadly follow similar logic.

As reads are amplified prior to sequencing, NGS relies on many short overlapping reads, sequencing each part of DNA multiple times. The more each section is repeatedly sequenced, the greater the *coverage*, leading to a more reliable sequence.

NGS methods have been extensively modified outside of their original scope of genome sequencing. Existing high-throughput technologies can map interactions between nucleic acids and other molecules (Johnson et al., 2007; Ule et al., 2005) and quantify transcript abundance genome-wide (Wang et al., 2009), at a greater speed and more affordable price than before.

2.5.1 *High-throughput RNA sequencing*

RNA-seq is a method for RNA sequencing, which generates a library of complementary DNA (cDNA) using *reverse transcription* (implemented by a reverse transcriptase enzyme) and sequences that library with NGS (Wang et al., 2009). The resulting reads are aligned to a reference genome or assembled *de novo* with bioinformatic analyses. RNA-seq can determine not only the sequence of RNA present in a sample, but also

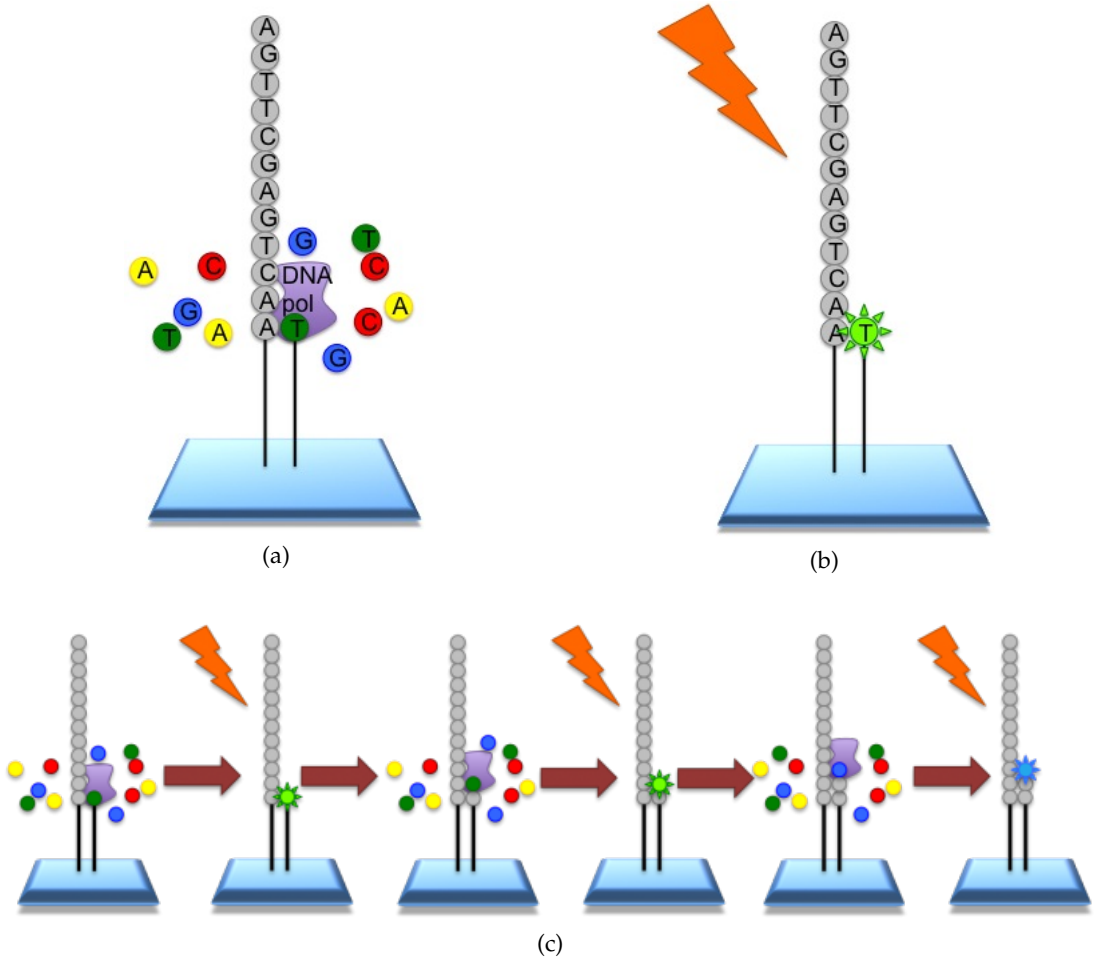


Figure 2.1: The basic process of an Illumina sequencing experiment. **a.** The reads are attached to a slide, flooded with DNA polymerase and fluorescently labelled nucleotides with terminators, which ensure that once a nucleotide is attached, the chain stops growing. **b.** The slide is imaged, detecting the fluorescent signal of the attached nucleotide. **c.** Its terminator is removed and the process is repeated, imaging one nucleotide at a time. Images reproduced from (EMBL-EBI, 2017).

the abundance of individual transcripts. However, to ensure accurate transcript quantification, various data normalisation strategies are required.

A number of biases can arise during the cDNA library preparation. As longer RNAs have to be cleaved into shorter fragments, different protocols achieve this with either RNA or cDNA fragmentation, each of which is known to be associated with a certain bias in the sequencing outcome (Wang et al., 2009; Marguerat and Bähler, 2010). Additionally, if amplification step is used by the protocol, corrections have to be made in order to distinguish the reads reflecting the genuine transcript abundance from the PCR artifacts.

Fragmentation steps result in the dependence of the number of reads mapped to a gene on its length. In order to be able to rank the abundances of different transcripts in the same sample, it must be noted that more reads are expected to arise from longer genes rather than shorter genes. Additionally, the effects of the library size on the number of mapped reads must be taken into account. These considerations are commonly addressed by normalising the raw counts of mapped reads by transforming them into measures such as *reads per kilobase of transcript per million mapped reads* (RPKM) (Mortazavi et al., 2008). RPKM is computed by dividing the sum of all reads in the sample by 10^6 and scaling all counts by this factor. Then the scaled counts are divided by their gene lengths in kilobases.

Other related normalisation transformations are *fragments per kilobase of transcript per million mapped reads* (FPKM) and *transcripts per million* (TPM). FPKM is similar to RPKM but is used for *paired-end sequencing*, where a fragment is sequenced from both ends, generating high-quality alignable sequence data. In this case, two reads can arise from a single fragment and the FPKM measure takes this into account by not counting this fragment twice. For TPM, the normalisation by gene length comes first in the transformation algorithm, making all normalised read counts sum to the same value in each sample. This can make it easier to compare the proportion of reads mapped to a gene in each sample. In contrast, the sum of the normalised counts may be different between samples when using FPKM normalisation; however, FPKM can be transformed into TPM using a single formula (Pachter, 2011).

Scaling counts to the library size intuitively reflects the expectation that sequencing a sample to half the coverage depth should yield half the number of reads mapped to each gene. However, this normalisation is unable to adequately capture differences in transcript abundance if the composition of the RNA population significantly changes between the compared samples (Robinson and Oshlack, 2010). Thus, FPKM and RPKM can be used for within-sample normalisation, taking care of the gene length and library size effects. TPM is more suited for between-sample normalisation; however, if a large number (but not the majority) of genes is unique or highly

expressed in one sample but not the other, it is advised (Conesa et al., 2016) to use an empirical normalisation method based on trimming the mean of gene-wise fold-changes (Robinson and Oshlack, 2010). It is worth noting that this method assumes that the majority of genes are not differentially expressed between samples.

2.5.2 *Biological relevance of NGS data*

The size and complexity of the NGS datasets unsurprisingly turned out to be insurmountable without computational methods to process them. However, even beyond the technical aspects of handling the data, it became increasingly clear that advanced statistical methodologies must be developed to interpret the data. This need gave rise to many analytical and modelling efforts focused on high-throughput sequencing data, a niche that is now firmly embedded into the interdisciplinary fields of bioinformatics and computational biology.

Many novel classes of transcripts were discovered as a result of these efforts. Additionally, they exposed a great variety of protein-coding transcripts that arises from different isoforms and synonymous variants.

A single gene can give rise to different protein isoforms via the process of alternative splicing, whereby some exons of the gene are included within or excluded from the resulting mRNA (Black, 2003). Alternative splicing is well regarded to be a major factor driving transcriptomic and proteomic complexity, as a high percentage of genes is thought to have alternative splice forms (Modrek and Lee, 2002). A novel class of highly conserved neuron-specific micro-exons was recently discovered and its misregulated alternative exclusion was shown to be linked to autistic phenotypes (Irimia et al., 2014). Further, many transcripts have multiple variants that differ only in their non-coding regions (Sandberg et al., 2008). Remarkably, while the proteins they generate have identical amino acid sequence, their functions can be drastically different in such important contexts as cell migration and cell survival (Berkovits and Mayr, 2015).

These important discoveries undoubtedly extended our understanding of RNA biology and its pervasive control at all steps of gene expression, with complex wide-ranging functions and implications in disease phenotypes. They were only made possible through collaborative efforts, which combined advanced experimental technologies and sophisticated computational strategies. High-throughput sequencing data, in all of its tremendous variety, provides an invaluable resource, yet is only informative when dealt with computationally. Thus, machine learning and computational methods for NGS data can shed light on two main aspects of RNA regulation —

its **structure** and **interactome**, both recently demonstrated to be the most influential research areas in RNA biology (Selega and Sanguinetti, 2016).

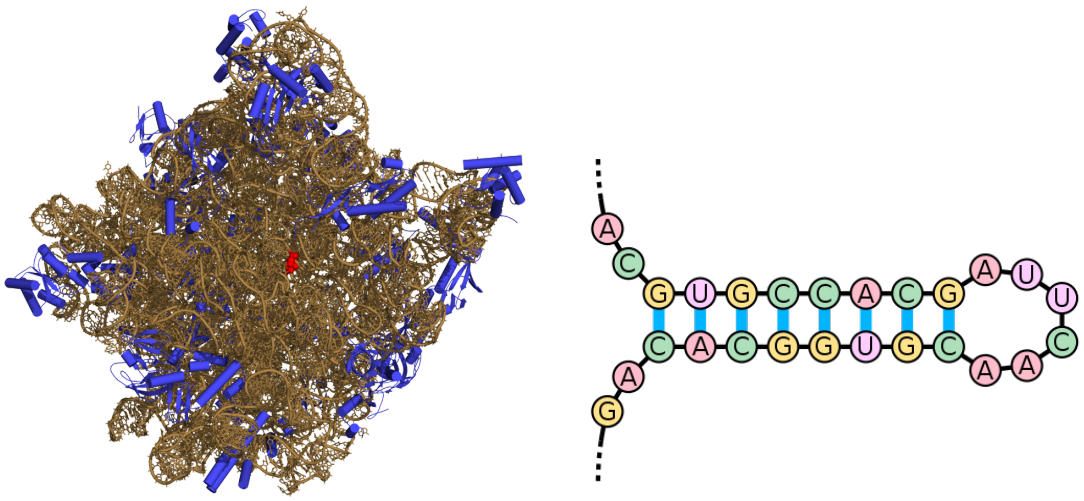
2.6 RNA STRUCTURE

An extra hydroxyl group in the ribose sugar enables RNA to form hydrogen bonds more easily. Thus, base pairs often form within the RNA molecule, creating folds and other structural elements. This gives rise to a large repertoire of highly complex structures that RNA can adopt (Adams, 2012). Fig. 2.2a shows a three-dimensional representation of the large 50S subunit of the prokaryotic ribosome, demonstrating the structural complexity of the RNA molecules.

As structure can facilitate (or prohibit) interactions with binding partners, it can thereby determine the biological function of the RNA. The structure of the target mRNA can affect its recognition by miRNAs, which increase its degradation or inhibit translation (Long et al., 2007). Further, most RNA processing reactions occurring post-transcriptionally are mediated by binding events with RNA-binding proteins and trans-acting RNAs (Glisovic et al., 2008). Thus, RNA structure prediction remains an important field which receives a lot of attention in the computational biology community.

RNA structure has been probed with methods such as X-ray crystallography (Speir et al., 1995) and nuclear magnetic resonance spectroscopy (Puglisi et al., 1992). While largely successful at solving three-dimensional RNA structures, these methods require complicated preparation procedures. Firstly, both methods require large quantities of purified RNA in order to generate useful structural information (Cheong et al., 2004). Secondly, RNA crystallisation is a complex multi-parametric process, for which the right solvent conditions are hard to determine *ab initio* (Kondo et al., 2014). Finally, understanding the RNA behaviour in the directly relevant physiological environment, its dynamic changes, and their impacts on the RNA binding and function is a task impossible to achieve with such experimental methodologies.

As RNA three-dimensional structures can be highly complex and understandably pose interpretation challenges, a useful representation is commonly used given by the RNA *secondary structure*: a list of base-pairing interaction patterns within the molecule. RNA secondary structure has many known functionally-relevant motifs; most common ones include hairpin loops (Svoboda and Cara, 2006). Fig. 2.2b shows the secondary structure representation of a hairpin loop, indicating the nucleotide sequence and the base-pairing interactions within it.



(a) (b)

Figure 2.2: **a.** 3D representation of the large ribosomal subunit (50S) of the prokaryotic ribosome (Yikrazuul, 2010). The ribosomal RNA are shown in ochre, the proteins are shown in blue. **b.** An example of a hairpin loop RNA secondary structure (Sakurambo, 2006). Blue bars indicate base-pairing interactions.

2.7 RNA STRUCTURE PREDICTION AND DETERMINATION

The problem of RNA secondary structure prediction has been extensively addressed with computational strategies. Many approaches tackle the complex problem of predicting the structure from a single nucleotide sequence.

2.7.1 *Single sequence structure prediction*

As the RNA structure is mainly determined by base pairing and base stacking interactions, the total sum of the free energy for these interactions should give an indication of the overall structure stability. This rationale motivates most single sequence structure prediction algorithms, which try to predict the folding free energy of a given secondary structure that could arise from the sequence and then pick the most stable structure by selecting that with the lowest free energy.

A common method for predicting the free energy for a structure is an empirical nearest-neighbour model (Mathews, 2006), where the change in free energy for each motif depends on its sequence and its neighbouring base pairs. The model is parameterised by estimates derived from calorimetric experiments measuring microscopic physical properties of the molecule (Mathews et al., 2004; Xia et al., 1998).

While in principle, finding the most stable structure is trivially achieved if all possible structures can be generated, this approach is completely infeasible in practice. The number of possible structures increases exponentially with the length of the RNA, e.g.

a transcript of a modest length of only 100 nucleotides has more than 10^{25} possible secondary structures (Mathews, 2006).

This problem was addressed with dynamic programming algorithms (Nussinov and Jacobson, 1980; Zuker and Stiegler, 1981), which remain a common solution in the field of secondary structure prediction. Dynamic programming allows to check all possible structures without explicitly generating them. The algorithm first determines the lowest possible free energy for each possible sequence fragment, starting with the shortest ones. The lowest free energy for longer fragments is determined recursively, using previously computed results for shorter sequences, which considerably speeds up the process. Once the lowest energy for the whole sequence is computed, the exact structure of the molecule is generated (Mathews, 2006).

Structure prediction methods that rely on free energy minimisation are potentially limited by their assumption that the most stable structure is the most probable. Not all known RNAs necessarily fold into structures which conform with the thermodynamic minimum (Freyhult et al., 2005). Further, some RNA sequences can have more than one structural conformation that is actively adopted *in vivo*, e.g. riboswitches (Nudler and Mironov, 2004). For these reasons, some structure prediction methods generate a list of "suboptimal" structures, which all have similarly low free energies (Zuker, 2003). Another approach is to generate statistically representative samples of possible structures from a Boltzmann ensemble (Ding and Lawrence, 2003).

Both dynamic programming and statistical sampling approaches frequently encounter a problem with prediction of complex structural elements such as pseudoknots (Andronescu et al., 2010). A pseudoknot contains two or more stem-loop structures, where a half of one stem is inserted between the two halves of another stem (Staple and Butcher, 2005). Many important RNA molecules rely on this complex three-dimensional structure, e.g. a ribozyme RNase P, which cleaves off extra RNA sequences on tRNA molecules (Guerrier-Takada et al., 1983), has a highly conserved pseudoknot region in its structure (Lee et al., 1996). Dynamic programming algorithms are well suited to detect "well-nested" structures, where formed base pairs do not overlap each other in sequence position. Pseudoknots are formed by interactions between distant nucleotides and so can only be detected by dynamical programming algorithms that have been substantially modified (Rivas and Eddy, 1999). However, these algorithms tend to be very computationally expensive. In fact, the general problem of pseudoknot prediction has been theoretically shown not to have a fast known solution (Lyngsø and Pedersen, 2000). Thus, most existing methods will not predict pseudoknots present in the query sequence, allowing the methods to perform at competitive speed (Knudsen and Hein, 2003; Zuker, 2003).

Another approach for three-dimensional structure prediction includes direct simulations of molecular dynamics (Sharma et al., 2008). These methods are already restricted to shorter sequences but even for those, full-resolution simulations can be computationally very costly. For this reason, coarse-grained models can be an attractive solution (Poblete et al., 2015).

2.7.2 *Comparative structure prediction*

In comparative structure prediction, the method relies on additional data (Gardner and Giegerich, 2004). Sequence covariation information is often used following the motivation that covariation of two distant nucleotides suggests the presence of a structural link between them. Using multiple sequences to predict consensus structure has a lot in common with sequence alignment.

A common strategy is to first align multiple sequences and then fold the determined consensus sequence, similarly as in single sequence structure prediction (Knudsen and Hein, 2003; Hofacker et al., 2002). In this case, the accuracy of structure prediction will depend on the quality of the alignment. The alignment step can be also addressed with a model-based approach, e.g. sequence covariation information was successfully used in maximum entropy global probability models to aid three-dimensional RNA structure inference (Weinreb et al., 2016).

Another method attempts to fold and align at the same time, merging sequence alignment and folding dynamic programming algorithms (Sankoff, 1985; Mathews and Turner, 2002; Holmes, 2005). As the general algorithm is very computationally expensive in both running time and required storage space, most practically used implementations apply restrictions to the maximal sequence length or variants of possible consensus structures.

2.7.3 *RNA structure probing experiments*

Another type of data that is now commonly incorporated into structure prediction algorithms (Reuter and Mathews, 2010) is obtained with *chemical* or *enzymatic RNA structure probing*, which became available with the advent of structure probing experiments. The experimentally derived scores are included as constraints to the model when computing free energy.

Structure probing experimental protocols utilise diagnostic chemical reagents, which can differentially modify parts of the RNA in a structure-dependent way. Typically, the modification exerted by the probing reagent is either cleaving the RNA (Kertesz et al., 2010) or adding a chemical group to its conformationally flexible nucleotides

(Wilkinson et al., 2006). Constrained nucleotides, either by complementary base pairing, protein interaction or local folding, should be protected from such modification. Thus, structure probing methods elucidate RNA secondary structure by characterising individual nucleotide accessibility.

This indirect readout of nucleotide accessibility is obtained using the fact that chemical modification causes the enzyme (reverse transcriptase, RT), synthesising the complementary DNA (cDNA) of the probed transcript, to terminate during the synthesis. The resulting cDNA is shorter than the original RNA and its length indicates the position of the modified nucleotide. Thus, if a read is observed terminating at a particular position, then the upstream neighbouring nucleotide could have been modified. However, further complications are introduced by the ability of the RT to randomly terminate in the absence of any chemical reagent. It is therefore necessary to correct the signal for background levels of RT termination (or *drop-off* events) that can arise simply by chance. This is achieved by pairing the structure probing experiment, in which the sample was treated with a chemical agent, with a *control* experiment performed in the same way but with no agent added. Structure probing experiments will be introduced again in mode detail in Chapter 4 (Section 4.1).

2.7.4 *Existing structure probing approaches*

Earlier structure probing techniques, such as parallel analysis of RNA structures (PARS) (Kertesz et al., 2010), made use of two enzymes RNase V1 and nuclease S1. V1 preferentially cleaves phosphodiester bonds 3′ of double-stranded regions while S1 cleaves 3′ of single-stranded regions. The quantitative measure of modification at each nucleotide position was defined as the log-ratio between the numbers of reads obtained in the two enzyme experiments. A similar method called FragSeq (Underwood et al., 2010) used a single nuclease specific to single-stranded regions and complemented it with a control nuclease-free experiment. However, both of these *in vitro* methods suffered from low resolution of measured structural information due to characteristic limitations of nuclease probing (Mauger and Weeks, 2010).

A recent, widely used probing reagent was first introduced by the influential method called SHAPE (2′-hydroxyl acylation analysed by primer extension) (Wilkinson et al., 2006), which gave rise to many related techniques. The SHAPE chemical acylates the 2′-hydroxyl group on the ribose of nucleotides in single-stranded and flexible regions, without discriminating by base (Mortimer and Weeks, 2007). When performed on intact crystals of bacterial ribosomes, SHAPE demonstrated strong agreement between its reactivities and nucleotide flexibility, independent of solvent or molecular accessibility (McGinnis et al., 2012).

Another common chemical agent dimethyl sulfate (DMS) selectively methylates nitrogen of unpaired adenine and cytosine residues (Zemora and Waldsich, 2010). The theoretical selectivity of DMS action can aid downstream validation analyses and the assessment of specificity and sensitivity of the experimental method.

Both agents, SHAPE and DMS, can be used *in vivo*, making them a popular choice in structure probing studies. Many factors inherent to the intracellular environment can influence RNA structure, e.g. transcription rates or presence of small molecules and RBPs (Zemora and Waldsich, 2010). Thus, the focus in the field has quickly shifted to structure probing *in vivo*. Combining SHAPE and DMS with a selective amplification strategy established a highly-sensitive method DMS/SHAPE-LMPCR, which allowed measuring structural information of lowly abundant RNAs *in vivo* in plants for the first time (Kwok et al., 2013).

Next, chemical structure probing *in vivo* was combined with high-throughput sequencing. DMS-Seq provided a parallel readout of a randomly fragmented pool of DMS-treated RNAs, selecting only prematurely terminated fragments (Rouskin et al., 2014). With DMS-seq, it was found that there were vastly fewer structured mRNA regions in dividing cells than *in vitro*, and even thermostable structures were often denatured in living cells. This finding once again highlighted the influence of cellular processes on RNA structure regulation and challenged the Anfinsen's dogma that the structure formed is the most thermodynamically favourable (Anfinsen, 1973).

Almost at the same time, structure-seq, another DMS-based high-throughput profiling method was globally applied *in vivo* in plants (Ding et al., 2014). It produced the first genome-wide nucleotide-resolution structure maps for any organism, quantitatively characterising more than 10,000 transcripts. The method computed per-nucleotide reactivity to DMS as a difference between the normalised *drop-off counts* of RT in the DMS (+) library and DMS (-) library (control). Raw drop-off counts were log-transformed and normalised by the average count of RT stops across each given transcript.

Notably, two years before, an *in vitro* method SHAPE-Seq combined SHAPE with paired-end high-throughput sequencing (Lucks et al., 2011). Paired-end sequencing enabled measurement of the sequence coverage information for each nucleotide. Using coverage and RT drop-off count, a *drop-off rate* was computed for each nucleotide, indicating its SHAPE reactivity in control (-) and treated (+) conditions. Introducing a reactivity measure instead of a raw drop-off count was an important development, as sequence coverage can vary dramatically along the transcriptome.

Moreover, SHAPE-Seq was extended by an automated mathematical framework, which implemented a maximum likelihood estimation strategy to infer relative reactivities from the observed fragment distribution (Aviran et al., 2011). This was the

first computational effort aiming to probabilistically model the underlying structural properties given the experimental measurements, drawing inspiration from models used in RNA sequencing analysis (Trapnell et al., 2010). Following the estimation of reactivity probabilities within the model, they are transformed into SHAPE reactivity scores (Lucks et al., 2011) using an empirically-derived normalisation strategy (Low and Weeks, 2010), which then aid the classification of nucleotide structural properties.

RNA-seq, which provides an integral part of RNA structure probing, is well known to be hampered by technological biases such as sequence-dependent bias (Shiroguchi et al., 2012) and coverage-dependent bias (Wang et al., 2009). The former has been routinely addressed by using barcodes with random nucleotides (Shiroguchi et al., 2012; Hector et al., 2014). Further, coverage-dependent bias can arise from various library preparation protocols, e.g. cDNA fragmentation is usually strongly biased towards the 3'end of the transcript (Wang et al., 2009).

Some of these consideration were addressed by SHAPE-MaP, a method combining SHAPE, high-throughput sequencing, and mutational profiling (Siegfried et al., 2014). It exploited conditions which caused RT to misread modified nucleotides and add a non-complementary base to the synthesised cDNA. This permanently stored positions and frequencies of SHAPE-formed 2'-O-adducts as mutations in the cDNA sequence. The method generated such mutational profiles for RNAs treated with a SHAPE reagent, a control solvent, and under denaturing conditions to control for sequence bias.

In SHAPE-MaP, nucleotide reactivities were computed by subtracting the control sample data from the treatment sample data and normalising it by the data from denaturing conditions. Its advantage stemmed from its insensitivity to biases arising from multi-step library-construction schemes. Moreover, single-stranded breaks, background degradation or signal decay did not affect it, unlike the structure probing methods depending on RT stops. However, SHAPE-MaP relied on a high read depth, with a recommendation of at least 5,000 reads required for accurate nucleotide-resolution structure mapping (Siegfried et al., 2014). Such sequencing depth cannot be guaranteed for more than a handful of transcripts in most transcriptome-wide experiments.

A related method ChemModSeq combined SHAPE, paired-end high-throughput sequencing, and statistical modelling (Hector et al., 2014). ChemModSeq utilised TCP^{EM}, an expectation maximisation algorithm with a Poisson model, to compute the likelihood of being modified for each nucleotide. A characteristic difference of the experimental design of ChemModSeq was the use of RT with oligonucleotides which randomly hybridised to the RNA template during cDNA synthesis and thus, circumvented problems with coverage-dependent bias. ChemModSeq was applied to

the intermediates of the yeast small ribosomal subunit 40S during its synthesis, providing insights into how ribosome assembly factors regulate the formation of 40S in eukaryotes (Hector et al., 2014).

2.7.5 *BUM-HMM: a computational analysis pipeline for structure probing data*

A recently launched curated repository (Norris et al., 2017) demonstrates the increasing pace at which high-throughput RNA structure probing studies are being performed. This explosion of available data once again recapitulates the need for computational strategies to analyse them.

The rich variety of genome-wide structure probing methods, of which the previous section only provides an incomplete description, commonly shares their form of output, given by some derivative of SHAPE reactivity scores. A handful of methods incorporate probabilistic models estimating the quantities of interest from the observed measurements of RT stops (Hector et al., 2014; Aviran et al., 2011). While the model-based approach is undoubtedly well-suited for accurately modelling the experimentally measured signal, some models' assumptions may not be applicable to specific properties of other experimental protocols. In addition, the overwhelming majority of existing methods makes use of semi-arbitrary thresholds when assessing structural properties of individual nucleotides (Siegfried et al., 2014; Kertesz et al., 2010; Lucks et al., 2011; Ding et al., 2014).

Further, most existing algorithms support a single pair of treatment-control experimental replicates (Siegfried et al., 2014; Rouskin et al., 2014; Kertesz et al., 2010; Underwood et al., 2010; Lucks et al., 2011; Kwok et al., 2013) and are thus unable to comprehensively quantify biological variability. Many algorithms do not or only superficially consider intrinsic biases of the technology. Finally, methods using log-ratio as a reactivity measure can assume different interpretations of the same score values, e.g. by setting negative log-ratio values to zero (Ding et al., 2014).

This PhD project aimed to address the above issues by developing a computational pipeline for modelling high-throughput RNA structure probing data while explicitly capturing biological variability (Selega et al., 2017). The pipeline performs data-driven bias-correcting steps and generates posterior probabilities of chemical modification for each nucleotide in the sequence. The description of the BUM-HMM (Beta-Uniform Mixture Hidden Markov Model) computational analysis pipeline is provided in Chapter 4.

2.8 INTERACTIONS BETWEEN RNA AND PROTEINS

RNA is exceptionally flexible as an interacting molecule, acting on DNA, other RNAs, and RNA-binding proteins (RBPs). Characterisation of these interactions both experimentally and computationally is a major research direction in RNA biology.

RNA is associated with proteins during all stages of its life cycle: transcription, splicing, nuclear export, and localisation to the cytoplasm where it is bound by ribosomes. RBPs, referred to as post-transcriptional regulators, interfere with many of these underlying processes. For instance, SR proteins recruit RNA-protein complexes that form the spliceosome, which itself consists of a large number of protein components (Long and Caceres, 2009). CPSF protein complex and poly(A)-binding protein ensure that mRNAs receive a 3′ poly(A) tail (Glisovic et al., 2008). ZBP1 transports mRNAs into the cytoplasm and is able to repress translation by preventing the ribosome from binding (Glisovic et al., 2008).

2.8.1 *Cross-linking methods for mapping RNA-protein interactions*

The primary technology to identify interactions between RBPs and RNA is given by cross-linking methods. Ultraviolet (UV) light-based methods in particular avoid problems associated with chemical cross-linkers, such as difficulties with entering the cores of large complexes. A revolutionary method which paved the path for many further technological modifications is CLIP (cross-linking immunoprecipitation) (Ule et al., 2005). Its workflow begins with exposing the living cells to UV light, which creates covalent bonds between interacting RNA and proteins (Fig. 2.3). The cells are then lysed and the protein of interest is isolated by immunoprecipitation. The resulting RNA-protein complexes are separated from free RNA and the protein and reverse transcribed into cDNA, allowing to identify the interacting RNA and its quantity.

A few years after its original application to study interactions of neuron-specific splicing factors (Ule et al., 2005), CLIP was combined with high-throughput sequencing, giving rise to HITS-CLIP (Licatalosi et al., 2008). Since then, HITS-CLIP (also called CLIP-seq (Yeo et al., 2009)) was used to generate genome-wide interaction maps for many RBPs, uncovering their roles in RNA regulation and connections with disease (Darnell, 2010). However, disadvantages of CLIP-like methodologies included DNA damage from the UV exposure, lower cross-linking efficiency compared to chemical cross-linkers, and low resolution of the generated interaction maps.

These problems were addressed by a number of related technologies. A method called CRAC (cross-linking and analysis of cDNA) was modified to have a denaturing affinity-purification step on nickel beads to improve specificity (Granneman

et al., 2009). CRAC was designed to be easier to apply in yeast and was used to study the architecture of preribosomes and their maturation. A further innovation based on HITS-CLIP was given by PAR-CLIP (photoactivatable-ribonucleoside-enhanced cross-linking and immunoprecipitation) (Hafner et al., 2010). PAR-CLIP incorporated photoreactive ribonucleoside analogs into RNA *in vivo* which left mutations at the interaction sites in the resulting cDNA, enabling single-nucleotide resolution of binding events. iCLIP (individual-nucleotide resolution CLIP) also generated highly-resolved RNA-protein interaction maps, but instead employed an enzyme which digests the cross-linked RNA, stopping at the interaction site (König et al., 2010). Unlike PAR-CLIP, iCLIP is not restricted to only those experimental systems that are amenable to RNA alteration.

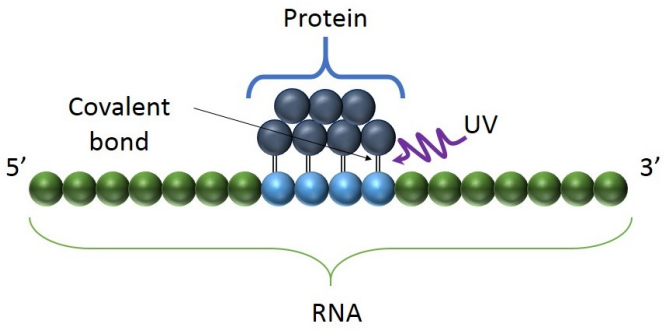


Figure 2.3: An illustration of the basic principle of the CLIP experimental protocol (Bluewhale22, 2014). Covalent bonds are created between interacting RNA and proteins upon exposure to UV light.

2.8.2 *Mapping rapid interactions with reduced irradiation times*

However, despite the improvements in the resolution of cross-linking site detection, the UV irradiation time, which could last up to tens of minutes (Beckmann, 2017), remained a rate-limiting step of the technology. This limitation posed a number of problems. Firstly, prolonged UV irradiation exposes cells to major additional stresses and could bias results towards the irradiation-specific conditions. Secondly, fast interactions are more difficult to capture with methods requiring a UV step longer than the duration of those interactions.

The longevity of the irradiation step was addressed by a recent method extended from CRAC (Granneman et al., 2009) called χCRAC (kinetic CRAC). χCRAC allows to quantitatively measure RNA-protein binding dynamics *in vivo* on a minute timescale (van Nues et al., 2017).

The required cross-linking time was dramatically reduced with the UV-irradiation device Vari-X-linker, developed as part of the technology. The fastest available device for cross-linking proteins to RNA in actively growing cells (Megatron (Granneman et al., 2011)) requires at least 100 seconds to achieve good cross-linking yields. In contrast, the Vari-X-linker has a number of features that enhance the effectiveness of UV cross-linking and allow 1-minute time resolution. The sample is presented in a specifically constructed UV transparent bag and flanked by two beds of lamps. The device uses a shutter system, enabling stable and repeatable UV exposure, and a fan cooling system, minimising thermal shock to the sample, while a vacuum pump rapidly extracts cells from the UV chamber (van Nues et al., 2017). This technological advancement underlying χCRAC made previously elusive rapid interactions between RNA and RBPs accessible for scientific scrutiny.

The paper presenting the χCRAC technology (van Nues et al., 2017), given in Chapter 5, among other results, describes the analyses of RNA-protein interactions time-course data that were performed with a non-parametric algorithm developed by myself as part of this PhD project. The following sections of this chapter motivate the context of the study presented in the paper and introduce the relevant computational methods developed for this thesis.

2.9 GENE EXPRESSION REGULATION IN STRESS RESPONSE

Survival under stress increasingly depends on the ability to rapidly change the gene expression programme, e.g. up-regulating some genes to deal with the hostile conditions and down-regulating others for energy considerations. Differential gene expression is affected by modulating cellular RNA levels, which are determined by the coordinated processes of transcription, RNA processing, and degradation.

2.9.1 *Roles of transcription and decay in gene expression regulation*

When explaining the dynamic changes to RNA levels underlying the adaptation response, a lot of emphasis has been placed on transcriptional control (Nachman et al., 2004; Ernst et al., 2007; Kim et al., 2009), while degradation rate was assumed to be constant (Barenco et al., 2009). However, this simplifying assumption was challenged by demonstrating that degradation can significantly affect the state of the transcriptome (Elkon et al., 2010; Shalem et al., 2008). Modelling of the data from gene expression time courses and the atlas of mRNA stability revealed that down-regulation of many transcripts during stress could not be explained by even a full shut-off of transcription (Elkon et al., 2010).

Further complications surrounding the study of changes in RNA expression were caused by the fact that transcription and degradation rates were determined via indirect methods. Degradation rate was routinely estimated by inhibiting transcription (Pelechano and Pérez-Ortín, 2008), which is ill-suited for dynamic settings and severely affects cell growth and survival, while the use of drugs can also lead to various side effects. Further, measuring degradation in isolation from transcription could be problematic as regulation of these two processes is likely to be tightly interconnected. On the other hand, metabolic labelling of RNA with 4-thiouridine provided a good technology to distinguish recently transcribed RNA with minimal adverse effects on cells (Friedel and Dölken, 2009). However, microarray platforms were often used (Cleary et al., 2005), which required large quantities of RNA and thus, longer labelling times, preventing a resolved dynamic analysis.

These issues were addressed by Rabani et al., who employed metabolic labelling coupled with massively parallel sequencing. They used a dynamic model to decompose RNA levels into the separate contributions of production and degradation and to estimate changes in degradation rates. While their model allowed a time-variant degradation rate, their results identified changes in transcription rates as a major determinant of temporal changes in RNA levels (Rabani et al., 2011).

Another study developed a 'switch' model, allowing for an additional instantaneous mRNA stabilization (or destabilization) event to happen at some time during the stress response, in contrast to the model with constant degradation rate (Marguerat et al., 2014). Their results showed that, while expression of most transcripts was best explained by the 'constant' model, for many mRNAs regulation of their turnover provided an additional control mechanism during the first minutes of the stress response, along with transcriptional changes. Marguerat et al. estimated transcription rates using the occupancy data of polymerase II (Pol II), a major RNA polymerase responsible for the mRNA synthesis. They specifically focused on the exonic regions located near the 3' ends of genes to control for Pol II stalling.

Pol II occupancy has been shown to correlate with metabolic labelling data (Miller et al., 2011) and was further validated by showing that its dynamic range is able to detect variations across the entire range of transcription rates (Marguerat et al., 2014). While confounding factors such as Pol II stalling have to be adequately taken into account, this non-invasive experimental method is still beneficial in terms of its relevance *in vivo*.

The results of the majority of studies identifying transcriptional control as the primary driver of changes in stress response could be underscored by the unavailability of direct measurements of degradation rates, especially at the early stages, previously suggested to be important (Marguerat et al., 2014).

2.10 DIRECTLY MEASURING DEGRADATION RATES

Exploiting the significantly shorter irradiation times, χCRAC was applied to the RBP mediating co-transcriptional degradation, Nab3, (van Nues et al., 2017) and the nuclease involved in cytoplasmic degradation, Xrn1 (currently unpublished data). These experiments generated the first direct and early measurements of degradation rates, examining their role in the context of stress response.

2.10.1 *Co-transcriptional degradation pathway*

The main nuclear degradation pathway identified in the yeast species *Saccharomyces cerevisiae* is regulated by the protein complex Nrd1-Nab3-Sen1 (NNS), which terminates transcription by interacting with the phosphorylated Pol II C-terminal domain (Vasiljeva et al., 2008) and with specific sequences in the nascent transcript (Carroll et al., 2007). These interactions then direct Pol II termination and processing of the nascent transcript (Arigo et al., 2006).

Widespread control of transcriptional regulation by NNS was elucidated by high-resolution transcriptome-wide maps of Nrd1 and Nab3 interactions (Webb et al., 2014; Creamer et al., 2011), measured *in vivo* with CRAC and PAR-CLIP (Granneman et al., 2009; Hafner et al., 2010). In addition to terminating transcription of non-coding RNAs, Nab3 and Nrd1 were shown to have many protein-coding targets (Webb et al., 2014). The function of both Nrd1 and Nab3 was also shown to be tightly integrated with the nutrient response pathway (Webb et al., 2014).

2.10.2 *Cytoplasmic degradation pathway*

The major cytoplasmic degradation pathway is governed by the activity of decapping enzymes and the 5′ to 3′ exoribonuclease Xrn1, which degrades the decapped transcripts completely (Berretta and Morillon, 2009). While Xrn1 is not essential for cell survival, its disruption was shown to markedly affect cell growth (Larimer and Stevens, 1990). In yeast, a novel class of Xrn1-sensitive unstable transcripts (XUT) has been identified, which implement various regulatory functions and are degraded in the cytoplasm (Van Dijk et al., 2011).

A summary of transcription and degradation mechanisms in yeast, governed by different pathways, is provided in Chapter 5 (Section 5.1).

2.10.3 *Identifying differential binding to degradation mediators under stress*

χCRAC was applied to Nab3, Pol II, and Xrn1 in glucose-deprived and glucose-rich *S. cerevisiae* cells at various time points as early as 1 minute after the nutrient shift. In this PhD project, I developed a model-based approach for the analysis and modelling of RNA-protein binding time-series data. This computational method, aimed at identifying differentially bound transcripts, was applied to the Nab3, Pol II, and Xrn1 χCRAC datasets. The analysis revealed pervasive rapid changes in Nab3 cross-linking to transcripts shortly after the stress induction, which were largely independent from changes in transcription (van Nues et al., 2017). The method can use different observation models, facilitating its application to datasets using raw or normalised cross-linking counts. This is demonstrated with the analysis of the Xrn1 dataset, which identified many of its interacting partners differentially bound under stress. The details of the developed method, the paper, presenting the Nab3 study, and the analysis of the currently unpublished Xrn1 χCRAC data are provided in Chapter 5.

2.10.4 *Modelling RNA expression kinetics*

In general, model-based studies of RNA expression kinetics assume that RNA decay can be summarized by a single mRNA half-life for each transcript, corresponding to a simple exponential decay process (Honkela et al., 2015; Rabani et al., 2011). This is despite the many identified RNA degradation pathways and their complexity (Deneke et al., 2013). This "constant degradation rate" assumption was recently further challenged by the dynamic behaviour of the transcriptional termination factor Nab3, revealed by the novel experimental technology χCRAC (van Nues et al., 2017).

Building on the insights obtained in the analyses of RNA-protein binding time-series, a dynamical model for RNA expression under stress was proposed in this PhD project. The proposed model aims to use the χCRAC dynamic binding data of the transcription catalyser Pol II, termination factor Nab3, and degradation factor Xrn1 to explain the changes in transcript abundance, associated with nutrient stress. The detailed description of the model is given in Section 5.9 of Chapter 5.

3

INTRODUCTION TO MACHINE LEARNING METHODS

This chapter outlines the necessary background on machine learning algorithms and models used in this thesis. It briefly introduces hidden Markov models, the EM algorithm, Gaussian processes, and dynamical models, mentions their applications in the field of computational biology, and specifies their usage in this thesis.

3.1 HIDDEN MARKOV MODELS

Hidden Markov models (HMMs) are a class of probabilistic models, in which the modelled system is assumed to be a *Markov process* with unknown (or *latent*) states (Baum and Petrie, 1966).

Definition 3.1.1. A discrete-time *Markov process* (or *Markov chain*) (Markov, 1906) is a sequence of random variables $X_1, X_2, X_3 ...$ which satisfy the *Markov property*, namely the condition that the probability of observing the value of the next variable depends *only* on the value of the present variable and *not* on the values of previous variables:

$$p(X_{t+1} = s | X_t = s_t) = p(X_{t+1} = s | X_1 = s_1, X_2 = s_2, ..., X_t = s_t), \text{ for } t \geqslant 1 \quad (3.1)$$

All possible values of each random variable X_t form a countable set $S = \{s_1, s_2, ...\}$ called the *state space* of the chain.

Let the random variables $X_1, X_2, X_3, ..., X_t$ describe the *state* of a stochastic process at times $1, 2, 3..., t$. Then the Markov property corresponds to the notion of *memorylessness*, whereby one could predict the state of the process at the next time $t + 1$, using only its state at time t, just as well as if one knew the full previous history of the process.

In an HMM, the states of the process are not directly visible but their *outputs*, which depend on the state at the corresponding time, are observed. Each hidden state has a probability distribution over its possible outputs and thus, the observed sequence of outputs can provide information about the sequence of hidden states that generated it.

Example 3.1.1. Suppose you start gambling with 5 Galleons[1] and bet 1 Galleon on a fair coin toss. You keep betting indefinitely or until you run out of gold. If X_t is the number of Galleons you have after t-th coin toss, then the sequence $X = \{X_t : t \in \mathbb{N}\}$ with $X_0 = 5$ is a Markov process. If an impartial observer knew that you currently had 7 Galleons, they they would expect that after the next toss, you would have either 8 Galleons or 6 Galleons, with equal odds. Knowing the history of your previous wins or losses does not improve the observer's prediction.

3.1.1 *HMM architecture*

Let us denote the hidden state of the process at time t with a random variable $h(t) \equiv h_t$. A hidden state h_t can adopt any value from the hidden state space H. The visible output at time t will be denoted with v_t and similarly, it takes values from the observed state space V. Typically, H is modelled as discrete, while V can be either discrete or continuous. In the graphical representation, the random variables are conventionally enclosed in circles, with arrows between the variables indicating conditional dependencies (Fig. 3.1).

According to the conditional independence relationships between variables implied by the graphical representation, given the values of h at all times, h_t depends only on the value of h_{t-1}. Similarly, given all values of h and v, v_t depends only on the value of its hidden state h_t. Using the notation $\{h_1, h_2, ..., h_t\} \equiv h_{1:t}$ and letting T to be the time of the last state, the above conditions are given in Eqs. 3.2, 3.3.

$$p(h_t|h_{1:T}) = p(h_t|h_{t-1}) \tag{3.2}$$

$$p(v_t|h_{1:T}, v_{1:T}) = p(v_t|h_t) \tag{3.3}$$

The joint distribution of the HMM is thus given by:

$$p(h_{1:T}, v_{1:T}) = p(h_1)p(v_1|h_1)\prod_{t=2}^{T} p(h_t|h_{t-1})p(v_t|h_t) \tag{3.4}$$

1 A Galleon is the most valued coin of the wizarding currency.

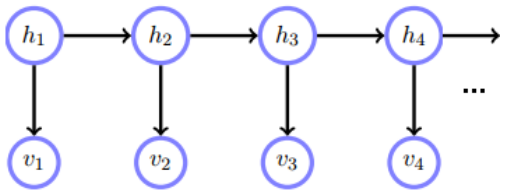


Figure 3.1: Graphical representation of the Hidden Markov Model with a hidden state h_t and output v_t.

3.1.2 *Transition and emission models*

Let us consider a discrete hidden state space H, which consists of N possible outcomes. *Transition probabilities* control the way the new state h_t is chosen given the value of the previous state at the time $t-1$. These probabilities can be summarised in a $N \times N$ transition matrix, where the entry at the position (i,j) is given by the conditional probability of moving to the state i from the state j. Transitions from any given state h_t must sum up to 1:

$$\sum_i p(h_t = i | h_{t-1} = j) = 1 \tag{3.5}$$

The observed state space can be modelled as discrete, with v_t taking one of M possible values according to a categorical distribution, or as a continuous M-dimensional multivariate distribution of choice. *Emission probabilities* describe the way the observed value v_t is chosen given the hidden state h_t at time t. Transition and emission probabilities are the *parameters* of the HMM.

3.1.3 *Inference in HMM*

One of the most useful inference questions for HMMs is to compute probabilities of some latent variables conditioned on the model parameters and observations, thereby recovering the true hidden state of the process at certain times, while using the information provided by the observed output.

The task of computing the distribution of any hidden variable given the observations is called *smoothing*:

$$p(h_k | v_{1:t}), \text{ for } k < t \tag{3.6}$$

Smoothed values (or *posterior marginals*) of the hidden variables can be efficiently computed with the *forward-backward algorithm*, which is a belief propagation approach.

The forward-backward algorithm computes the required values in two passes. The first, forward message pass computes the probabilities of being in any particular state h_k given the first k observations $\forall k \in \{1, ..., T\}$, using recurrently defined forward messages α_k:

$$\alpha_1(h_1) = p(h_1)p(v_1|h_1) \tag{3.7}$$

$$\alpha_k(h_k) = p(h_k, v_{1:k}) \propto p(h_k|v_{1:k}) \tag{3.8}$$

$$\alpha_k(h_k) = \sum_{h_{k-1}} p(v_k|h_k)p(h_k|h_{k-1})\alpha_{k-1}(h_{k-1}) \tag{3.9}$$

The second, backward message pass computes the probabilities of seeing the remaining observations given any starting time k, making use of the backward messages β_k:

$$\beta_k(h_k) = p(v_{k+1:T}|h_k) = \sum_{h_{k+1}} p(v_{k+1}|h_{k+1})p(h_{k+1}|h_k)\beta_{k+1}(h_{k+1}) \tag{3.10}$$

$$\beta_T(h_T) = 1 \tag{3.11}$$

Together, the forward and backward set of probabilities combine to yield the (unnormalised) probability distribution over any hidden state h_k given all observations (Eq. 3.12). This result is obtained using Bayes' rule and the conditional independence between the observations $v_{1:k}$ and $v_{k+1:T}$ given h_k.

$$\begin{aligned} p(h_k|v_{1:T}) = p(h_k|v_{1:k}, v_{k+1:T}) \propto p(v_{k+1:T}|h_k, v_{1:k})p(h_k, v_{1:k}) = \\ p(v_{k+1:T}|h_k)p(h_k, v_{1:k}) = \beta_k(h_k)\alpha_k(h_k) \end{aligned} \tag{3.12}$$

Another useful inference task is finding the *most likely explanation* of the observed data. This problem requires finding the maximum joint probability of the entire sequence of hidden states that generated the observations and can be solved with *Viterbi algorithm* (Forney, 1973).

3.1.4 *Expectation maximisation*

In order to perform inference in HMMs, the model parameters must be known. The task of finding the best set of transition and emission probabilities given the sequence

of observations is called parameter learning. The local maximum likelihood estimates of the parameters given the observations can be efficiently derived with the *expectation-maximisation* (EM) algorithm (Dempster et al., 1977).

Let us denote the HMM parameters with a vector θ and define a likelihood function of θ as:

$$L(\theta; h_{1:T}, v_{1:T}) = p(h_{1:T}, v_{1:T}|\theta) \tag{3.13}$$

Then the maximum likelihood estimate of the unknown parameter values is determined by the marginal likelihood of the observed data $p(v_{1:T}|\theta)$, which is often intractable as the number of all possible values for the sequence of hidden variables grows exponentially with the length of the sequence.

The EM algorithm aims to find the maximum likelihood estimate of the marginal likelihood by iterating between two steps. The *E-step* (or the expectation step) calculates the expected value of the log likelihood function with respect to the conditional distribution of $h_{1:T}$ given $v_{1:T}$, evaluated with the current parameter estimates $\theta^{(t)}$:

$$\mathbf{E}_{h_{1:T}|v_{1:T},\theta^{(t)}}[\log L(\theta; h_{1:T}, v_{1:T})] = Q(\theta|\theta^{(t)}), \tag{3.14}$$

The *M-step* (or the maximisation step) finds the new parameter values $\theta^{(t+1)}$ that maximise this quantity:

$$\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta|\theta^{(t)}) \tag{3.15}$$

The motivation for the algorithm uses the fact that if the model parameters are known, then the most likely sequence of hidden states can be found. Conversely, if the hidden states are known, then the parameters can be estimated by grouping together observations according to their latent state.

When both parameters θ and the hidden states $h_{1:T}$ are unknown, the algorithm proceeds as follows:

1. θ is initialised with some values.
2. The probabilities are computed for each possible value of the sequence $h_{1:T}$ using the current θ values.
3. The values of $h_{1:T}$ are used to find better estimates for the parameter values.
4. Steps 2 and 3 are iterated until convergence.

At each iteration of the EM algorithm, the parameter values are chosen to increase the value of $Q(\theta|\theta^{(t)})$. It can be shown that such selections result in the value of the log marginal likelihood $\log p(v_{1:T}|\theta)$ to be non-decreasing at all iterations (Little and Rubin, 2014). Thus, as the algorithm increases the marginal likelihood of the observed data, it monotonically approaches at least its local maximum. No guarantees exist for the EM algorithm to converge to a global maximum, which is why many heuristic approaches are used that aim to escape local maxima (e.g. starting with several random initialisations of the parameter values). The EM algorithm is an attractive method for finding maximum likelihood estimates as it doesn't require evaluating derivatives and often has closed-form update expressions for each step if the likelihood function belongs to an exponential family.

3.1.5 *Applications of HMMs in computational biology*

Even though HMMs are traditionally defined as models for a time-evolving process with a hidden state, they have become a common tool for modelling sequencing data. In these applications, each random variable h_t corresponds to the hidden state of the t-th nucleotide in the sequence (instead of the state of a process at time t), while v_t encodes some experimental measurement that provides information about the hidden state of the nucleotide.

HMMs have been widely used in such applications as gene prediction (Burge and Karlin, 1997; Korf, 2004), multiple sequence alignment (Durbin et al., 1998), and protein structure prediction (Bystroff et al., 2000). HMMs were also applied to RNA-seq data for discovery of splicing junctions (Dimon et al., 2010), non-coding RNA annotation (Weinberg and Ruzzo, 2005), and RNA structural alignment (Yoon and Vaidyanathan, 2008).

In this thesis, I developed a computational analysis pipeline for modelling RNA structure probing data (Selega et al., 2017), which used an HMM with a mixture model governing emission probabilities. The detailed description of the BUM-HMM pipeline is presented in Chapter 4.

3.2 GAUSSIAN PROCESSES

Gaussian Processes (GP) are Bayesian non-parametric models with a continuous input space (often representing time), where every point in space is associated with a Gaussian distributed random variable and every finite collection of those random variables is also jointly Gaussian. The distribution of a GP is thus the joint distribution

of infinitely many Gaussian random variables, which can be viewed as a distribution over *functions* with a continuous domain.

GP models are often used to solve problems concerned with *supervised learning*. Supervised learning problems, where the aim is to learn the mapping from input to output using training data, are divided into *classification* or *regression* problems. The usage of GPs in this thesis is concerned with *regression* problems, where the output is a continuous quantity.

3.2.1 *Motivation*

Let's denote the input variable as $\mathbf{x}$ and the output variable as $\mathbf{y}$. Then the training dataset with N observations can be written as a set of input-output pairs for each training item i, $\mathbf{D} = \{(\mathbf{x}_i, \mathbf{y}_i) | i = 1, ..., N\}$. Given $\mathbf{D}$, we would like to make predictions about the previously unseen inputs $\mathbf{x}_*$. Thus, the problem is to find a function f that maps every possible input to an output value.

One must make assumptions about the behaviour of f, to restrict the choice from the set of all possible functions consistent with the training data. One approach is to only consider a special class of functions, e.g. linear functions of the input. However, this can affect the accuracy of prediction if the expressiveness of this class is not powerful enough to adequately capture the data. In contrast, increasing the function's complexity can lead to problems with *overfitting*, when the function models the training data so well that it is unable to generalise to new examples.

Another approach is to attach some *prior* probability to every possible function, with higher probabilities given to functions that are deemed more likely. This is implemented by the Gaussian Process model (Rasmussen and Williams, 2006), which is a generalisation of the Gaussian probability distribution to *functions*. A useful way to think about a function in this context is in the form of a very long – infinite – vector, whose each value specifies $f(x)$ at an input x. Despite dealing with infinite-dimensional objects, GPs present a computationally tractable framework, where inference performed on a finite number of points would give results consistent with the case where infinitely many points were taken into account.

3.2.2 *GP specification*

Definition 3.2.1. *A* Gaussian Process is a collection of random variables any subset of which is jointly Gaussian distributed.

A Gaussian Process $f(\mathbf{x})$ is fully specified with *mean* and *covariance* functions defined on the inputs:

$$m(\mathbf{x}) = \mathbf{E}[f(\mathbf{x})] \tag{3.16}$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))] \tag{3.17}$$

$$f(\mathbf{x}) \sim GP(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{3.18}$$

In a GP, random variables represent the value of $f(\mathbf{x})$ at input $\mathbf{x}$. For notational convenience, f_i is often used to denote $f(\mathbf{x}_i)$. In this introduction, the mean function is taken as equal to 0 for simpler notation.

The choice of the covariance function specifies the distribution over functions. A common choice is a *squared exponential* covariance function (sometimes called *radial basis function*):

$$\mathrm{cov}(f(\mathbf{x}_p), f(\mathbf{x}_q)) = k(\mathbf{x}_p, \mathbf{x}_q) = \alpha \exp(-\frac{(\mathbf{x}_p - \mathbf{x}_q)^2}{2l}) \tag{3.19}$$

For this covariance function, the covariance is high for variables whose inputs are close to each other and decreases with the distance between the corresponding inputs. The *hyperparameter* α controls the variance of the random function $f(\mathbf{x})$ and the *length-scale* hyperparameter l indicates the distance between the inputs, after which the function value can significantly change.

Many different choices of covariance functions are used in order to encode the process' *stationarity* (invariance to translations in the input space), *isotropy* (invariance to rotations), *smoothness*, and *periodicity*. Specifying these properties reflects our beliefs about what a suitable function should look like.

Having specified the GP, we can draw samples from it by selecting input points, computing the corresponding covariance matrix for all pairs of points, and generating a Gaussian vector with this covariance matrix. Each generated vector is a sample function from the *prior* distribution.

3.2.3 *Inference in GPs*

In general, we'd like to incorporate the training data into the inference process when making predictions for new points. If we're modelling the training data in dataset $\mathbf{D}$ and assuming noise-free observations, then the output $\mathbf{y}_i$ gives the value of the

random variable f_i. Let the set of training inputs be X and the set of test inputs X_* (for which we wish to generate predictions).

According to the prior distribution, the joint distribution of training outputs $\mathbf{f}$ and test outputs $\mathbf{f}_*$ is given by:

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \sim N\left(0, \begin{bmatrix} K(X,X) & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right) \tag{3.20}$$

In the above, $K(X,X_*)$ denotes the covariances between all training and test inputs, and similarly for other block covariances. In order to compute the *posterior* distribution, we should select only those functions from the prior distribution that agree with the training data. This is achieved by *conditioning* the joint distribution on the observations:

$$\mathbf{f}_*|\mathbf{f},X_*,X \sim N(\bar{\mathbf{f}_*}, cov(\mathbf{f}_*)), \tag{3.21}$$

$$\bar{\mathbf{f}_*} = K(X_*,X)K(X,X)^{-1}\mathbf{f} \tag{3.22}$$

$$cov(\mathbf{f}_*) = K(X_*,X_*) - K(X_*,X)K(X,X)^{-1}K(X,X_*)) \tag{3.23}$$

In the case of noisy observations, the above inference process holds with a few modifications. Assuming additive Gaussian noise ϵ, a diagonal matrix is added to the covariance matrix between observations:

$$y = f(\mathbf{x}) + \epsilon \tag{3.24}$$

$$\epsilon \sim N(0, {\sigma_n}^2) \tag{3.25}$$

$$cov(\mathbf{y}) = K(X,X) + {\sigma_n}^2 I \tag{3.26}$$

The noise term then also appears in the covariance function of the joint distribution of observations and function values at test inputs:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim N\left(0, \begin{bmatrix} K(X,X) + {\sigma_n}^2 I & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right) \tag{3.27}$$

If we then introduce simplified notation for $K = K(X,X)$ and $K_* = K(X,X_*)$, the conditional predictive distribution for GP regression has the following mean and covariance functions:

$$\mathbf{f}_*|\mathbf{y}, X_*, X \sim N(\bar{\mathbf{f}}_*, \mathrm{cov}(\mathbf{f}_*)), \tag{3.28}$$

$$\bar{\mathbf{f}}_* = K_*[K + {\sigma_n}^2 I]^{-1}\mathbf{y} \tag{3.29}$$

$$\mathrm{cov}(\mathbf{f}_*) = K(X_*, X_*) - K_*[K + {\sigma_n}^2 I]^{-1}K_* \tag{3.30}$$

The mean prediction is a linear combination of observations $\mathbf{y}$, while the covariance depends only on the inputs as we are predicting the values of the random process $\mathbf{f}_*$. The variance is the difference between the prior covariance between test points and a positive term that represents the information about the function supplied by observations. If we wanted to predict the noisy observations $\mathbf{y}_*$ at test locations, it would suffice to add the diagonal noise matrix ${\sigma_n}^2 I$ to $\mathrm{cov}(\mathbf{f}_*)$.

3.2.4 *Marginal likelihood*

Definition 3.2.2. *Marginal likelihood* (or *evidence*) $p(\mathbf{y}|X)$ is the likelihood of observations given the input points.

$$p(\mathbf{y}|X) = \int p(\mathbf{y}|\mathbf{f}, X)p(\mathbf{f}|X)\,d\mathbf{f} \tag{3.31}$$

When computing marginal likelihood of the training data, all function values $\mathbf{f}$ are integrated out. This makes it possible to use marginal likelihood maximisation for optimising GP regression model parameters. If the observation noise is Gaussian then all terms under the integral are Gaussian and the log marginal likelihood has the following analytic form:

$$\log p(\mathbf{y}|X) = -\frac{1}{2}\mathbf{y}^\mathsf{T}(K + {\sigma_n}^2 I)^{-1}\mathbf{y} - \frac{1}{2}\log|K + {\sigma_n}^2 I| - \frac{n}{2}\log 2\pi \tag{3.32}$$

If the observation model is not Gaussian then the likelihood $p(\mathbf{y}|\mathbf{f}, X)$ can be approximated, e.g. with Laplace's method (MacKay, 2003).

3.2.5 *Applications of GPs in computational biology*

As GPs provide a rich framework for modelling time-series data, they have recently become a popular tool for analysis of gene expression time courses. One of the first such applications included modelling the dynamics of transcriptional regulation using expression levels of known target genes (Lawrence et al., 2007). Other applications

of the GP framework aimed to identify temporal intervals of differential gene expression (Stegle et al., 2010) and to detect quiet/active genes in microarray data (Kalaitzis and Lawrence, 2011).

A GP regression model was applied to RNA-seq time-series data, specifically accounting for experiment-specific biases in gene expression dynamics (Äijö et al., 2014). Notably, it was shown that modelling complete time-series instead of performing timepoint-wise analysis, as well as applying normalisation strategies for replicated measurements, allowed the discovery of differentially expressed genes that would otherwise have been missed. Further, a GP-based method was recently applied to single-cell RNA-seq data in order to detect branching dynamics for individual genes in populations of cells undergoing differentiation (Boukouvalas et al., 2017).

In this thesis, I developed an algorithm testing for differential binding, which used GP regression models for capturing the dynamics of RNA-protein interactions in different conditions (van Nues et al., 2017). The description of the model and its usage for hypothesis testing is presented in Chapter 5.

3.3 DYNAMICAL MODELS

Dynamical models describe the time-dependent changes in the state of a system. They are typically formalised with *differential equations*.

Definition 3.3.1. *A differential equation* relates a function with its derivatives. In applications, the function often represents some physical quantity and the derivatives describe its rates of change.

3.3.1 *Types of differential equations*

Differential equations can be divided into different types. The type of the equation provides information about the approach to solving it.

A commonly occurring type is *ordinary differential equations* (ODEs), which describe the relation between the function of a single independent variable and its derivatives:

$$\frac{dy}{dx} = F(x, y) \tag{3.33}$$

In the equation above, x is an independent variable and y is the dependent variable with $y = f(x)$ for some function f. An ODE can describe an n-th order derivative $d^n y/dx^n$ and the function F on the right hand side can be a function of x and all derivatives of order up to $n-1$.

In a *linear* ODE, F can be written as a linear combination of the derivatives of y. For linear ODEs, there exists a closed-form solution, whereas non-linear ODEs often have to be solved numerically as their exact solution is in series or integral form.

One approach is Euler's method (Butcher, 2016), which gives a first-order approximation of the unknown curve that satisfies the ODE and starts at a given point A. The method takes small steps of size h along the curve's tangent line starting from A, with each new point computed from the previous one. Each y_n is the approximation of the ODE solution at time x_n.

$$y_{n+1} = y_n + hF(x_n, y_n) \tag{3.34}$$

Differential equations that deal with multivariable functions and their partial derivatives are called *partial differential equations* (PDE). Both ODEs and PDEs can be generalised to their *stochastic* versions, where the unknown quantity is a stochastic process and the expression on the right involves a noise term. The solution for a *stochastic differential equation* is also a stochastic process and there are various methods for finding it numerically (Kloeden et al., 2012).

3.3.2 *Applications of dynamical models in computational biology*

Along with their applications in developmental biology and epidemiology, dynamical models have been widely used in computational systems biology. The following sources provide a few examples summarising the advances in this area. Alon (2006) gives an extensive overview of using dynamical systems for studying biological systems, particularly focussing on the design and dynamics of gene regulation networks. Lawrence et al. (2010) outlines inference approaches for biological networks, uncovering the structure and parameterisation of the underlying models of genetic regulation. Wilkinson (2011) provides an introduction to stochastic kinetic modelling of biological networks in the context of systems biology.

When studying the dynamics of RNA expression, the changes in RNA abundance are traditionally modelled as a combination of a time-variant transcription rate, which is seen as a main driver of change, and a constant degradation rate. However, many variants of this scenario have been considered. Transcription rate was modelled non-parametrically and included a processing delay (Honkela et al., 2015) and there have been attempts to allow a more flexible degradation rate schedule (Rabani et al., 2011; Marguerat et al., 2014).

In this thesis, a dynamical model for RNA expression under stress was proposed, based on the insights obtained with the differential binding analyses of RNA-binding proteins involved in degradation. The model aims to explain the changes in transcript abundance with the dynamical changes in the binding of proteins associated with transcription and degradation pathways. Further description of the theoretical formulation of the model and of the ongoing work on its development are given in Section 5.9 of Chapter 5.

4

BUM-HMM: MODELLING HIGH-THROUGHPUT RNA STRUCTURE PROBING DATA

This chapter introduces the use of RNA structure probing experiments for informing structure prediction algorithms, identifies existing problems surrounding the analysis of the resulting data, and proposes a probabilistic modelling pipeline to analyse high-throughput RNA structure probing data.

Section 4.1 provides an illustration of the experimental methodology and states the main question that the BUM-HMM pipeline aims to answer. Then, Section 4.2 gives an intuitive overview of the proposed method. This introductory material is intended to aid the understanding of the paper which presents the methodology and evaluates its performance on real data (Selega et al., 2017).

In accordance with the University of Edinburgh regulations, the paper is included in its published form in Section 4.3. The supplementary figures that are relevant for a detailed description of the pipeline's features are presented throughout this chapter. All Supplementary Figures for the paper are included in Section A.2 of Appendix A.

The proposed modelling pipeline was developed by myself and all computational analyses evaluating the method were performed by myself unless stated otherwise. The other authors contributed in the following manner. Christel Sirocchi, Ira Iosub, and Sander Granneman carried out the experiments. Christel Sirocchi, Sander Granneman, and Guido Sanguinetti performed exploratory computational analyses, informing the early pipeline development. Sander Granneman processed raw sequencing data and performed computational analyses informing the bound on coverage required for effective structure probing and investigating the link between structural flexibility and ribosome occupancy. Guido Sanguinetti and Sander Granneman jointly supervised my work on the BUM-HMM pipeline, with Guido Sanguinetti providing supervision on the modelling side. The manuscript was written by myself, Guido Sanguinetti, and Sander Granneman.

The chapter proceeds by providing a detailed description of the pipeline's automated bias-correcting strategies in Section 4.4. The optional parameter optimisation strategy is referenced in Section 4.5. The chapter concludes by discussing recent developments in the field and summarising the key contributions of the proposed methodology in Section 4.6.

4.1 ILLUSTRATION OF AN RNA STRUCTURE PROBING EXPERIMENT

RNA structure is known to be a key regulator of many important cellular mechanisms. RNA structural regulatory elements are interrogated with chemical and enzymatic structure probing (Kubota et al., 2015). In these experiments, a chemical agent reacts with the RNA molecule in a structure-dependent way, cleaving or otherwise modifying its structurally flexible parts. These modified positions can then be detected, providing valuable structural information that can improve structure prediction (Wu et al., 2015).

Specifically, chemical modification terminates the reverse transcription reaction, resulting in the reverse transcriptase (RT) dropping off just before the modified position. Thus, the modified position is 1 nucleotide upstream in the sense direction of the position of the RT drop-off. The drop-off positions can be mapped back to the reference sequence, identifying structurally flexible parts of the transcript. However, the challenge lies in the stochasticity of this process as the RT can also drop off randomly. To address this, a complementary control experiment is routinely performed to monitor random RT drop-offs when no reagent is used.

Let us consider a toy example of data obtained in a paired-end sequencing RNA structure probing experiment, when fragments are sequenced from both ends (Fig. 4.1). We'll focus on a particular nucleotide **G** in the sequence and consider the data from a control experiment (with no reagent added) and a treatment experiment (with RNAs modified by the reagent). In control conditions, we mapped 5 fragments overlapping with the nucleotide **G**, one of which also terminated at that position. Thus, this nucleotide had a coverage of 5 and a drop-off count of 1 (the number of times the RT dropped off at this position), giving it a *drop-off rate* of $\frac{1}{5}$. In treatment conditions, more fragments terminated at this position and we measured a drop-off rate of $\frac{4}{5}$. This seems to suggest that the next nucleotide **T** has been modified by the reagent and perhaps corresponds to a flexible site within the transcript molecule.

However, would our conclusion remain the same had we observed a higher drop-off rate in control conditions to start with? In fact, how high would this control drop-off rate have to be for us to dismiss the drop-off rate of $\frac{4}{5}$ as a noisy measurement of random drop-off rather than an indication of real modification by the chemical reagent?

This question reinforces the need for deciding statistically whether the drop-off rate in treatment conditions is significantly higher than the drop-off rate in control. To do this, we must understand how much noise can be expected in control conditions. If the treatment drop-off rate is outside of this range of drop-off rate variability, then

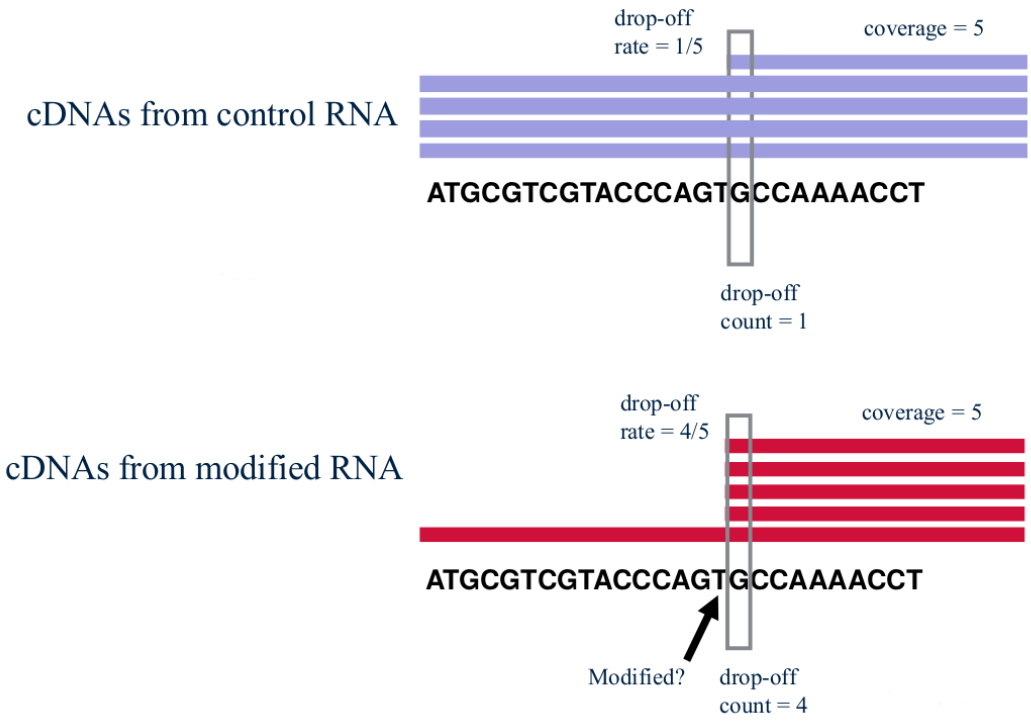


Figure 4.1: A toy example illustrating the data obtained in a paired-end RNA structure probing experiment.

we could deem it as being significantly higher and conclude that a real chemical modification signal is present.

4.2 DATA REPRESENTATION AND THE MODEL

The pipeline uses two types of experimental information for each nucleotide position: the *coverage*, defined as the total number of reads that cover a nucleotide, and the *drop-off count*, defined as the total number of reads stopping at a nucleotide. It combines them in a measure of a *drop-off rate*, defined for each nucleotide as the ratio between its drop-off count k and coverage n:

$$r = \frac{k}{n} \tag{4.1}$$

The drop-off rate represents a nucleotide's reactivity to the chemical probing reagent and ranges between 0 and 1. Its usage also provides a normalisation strategy for different read depths.

4.2.1 *Measuring drop-off rate variation*

The pipeline uses information collected in replicate experiments in control and treatment conditions. Any number of experimental replicates is supported, with no constraint on them being paired as treatment-control.

The goal is to identify the variability in drop-off rates that arise in the absence of a chemical agent in order to be able to deem a signal measured in treatment conditions as real. The biological variability is quantified with the log-ratio between the drop-off rates (LDR) at the same nucleotide position in a pair of control replicates i and j:

$$\text{LDR} = \log \frac{r_i}{r_j} = \log(r_i) - \log(r_j) \tag{4.2}$$

If the drop-off rate is stable across control replicates, LDR at a nucleotide will be close to 0, indicating little to no variability. In contrast, big changes in drop-off rates between replicates will result in a large absolute value of the measure. Due to a log transform, the nucleotide positions with the drop-off rates $r = 0$ in either replicate of the pair are discarded from the analysis.

LDRs corresponding to all pair-wise comparisons between control replicates collectively describe the variability in drop-off rates at all nucleotide positions that could be observed in the absence of the probing reagent. A collection of these define the *null distribution* of LDRs.

4.2.2 *Computing empirical* p*-values*

For each combination of treatment-control replicates, LDRs are computed for all nucleotide positions, quantifying the difference between the drop-off rate observed in a treatment experiment with respect to a control replicate. The goal is to decide whether this difference is larger than what we expect at random. This is achieved by comparing the treatment-control LDRs to the quantiles of the null distribution:

$$\text{p-value} = 1 - q, \text{where } q \text{ is the closest quantile} \tag{4.3}$$

The resulting p-value for each treatment-control LDR represents the probability of it being insignificantly different from what could be observed by chance.

For instance, if a particular LDR is closest to the 99^{th} quantile, then the difference between the drop-off rates at that nucleotide in the compared treatment and control

experiments is larger than most differences that could be observed in the absence of a reagent. This LDR would be assigned an empirical p-value of 0.01 or a 1% probability of it being insignificantly different from the null distribution. This result would provide evidence towards chemical modification of the nucleotide position in question.

Conversely, if an LDR is closest to the 10^{th} quantile, it would receive a p-value of 0.9, suggesting that this difference between the drop-off rates is not unusual. Thus, a *low* empirical p-value is indicative of a significant difference between the drop-off rates in a pair of treatment and control experiments.

4.2.3 *Computing posterior probabilities of modification with HMM*

The empirical p-values for each nucleotide position and each treatment-control comparison are used as observations in a hidden Markov model, introduced in Section 3.1 (Chapter 3). The hidden state h_t ($t = 1...T$ for T nucleotides) corresponds to the true binary state of the t-th nucleotide (chemically modified or unmodified) and the observed variable v_t is given by the empirical p-value attached to that position.

Modelling p-values directly enabled the definition of the HMM emission distribution as a Beta-Uniform mixture model. This design exploits the result that p-values are uniformly distributed under the *null hypothesis* (e.g. Murdoch et al. (2012)). When the null hypothesis is true, the drop-off rates in a given treatment-control pair *do not* differ significantly, which would lead us to believe that the corresponding nucleotide position t was unmodified in the treatment experiment. This scenario is represented by the **u**nmodified binary state of the nucleotide $h_t = U$ and thus the distribution of p-values given this state is modelled as Uniform.

For the *alternative* hypothesis, a significant difference between the drop-off rates would suggest that the nucleotide was chemically modified. In this case of the **m**odified state $h_t = M$, we expect to see large LDRs and small associated p-values, which are modelled with a Beta distribution favouring small values. The p-value distribution computed for the transcriptome-wide dataset strongly agrees with this model (Panels **a**, **b** in Fig. 4.2 show the null distributions computed from the transcriptome-wide data for two strands). P-values at the same nucleotide position that correspond to different comparisons of treatment-control experimental replicates $\{v_t^n : n = 1..N\}$ for N different comparisons are assumed to be independent measurements.

$$p(v_t|h_t = U) \sim U(0,1) \tag{4.4}$$

$$p(v_t|h_t = M) \sim Beta(\alpha,\beta), \text{with } \alpha = 1, \beta = 10 \tag{4.5}$$

$$p(v_t|h_t) = \prod_{n=1}^{N} p(v_t^n|h_t), v_t = v_t^1, ..., v_t^N \tag{4.6}$$

Transition probabilities of the HMM were defined using empirically derived lengths of single- and double-stranded regions of nucleotides. The model assumes expected uninterrupted stretches of 20 double-stranded (or constrained) nucleotides and 5 single-stranded (or flexible) nucleotides. The exact expressions are given in Section A.1 of Appendix A.

Posterior probabilities of being in the modified state $p(h_t = M|v_{1:T})$ are computed per-nucleotide with the forward-backward algorithm, generating a statistically interpretable output which obviates the need for heuristically chosen thresholds.

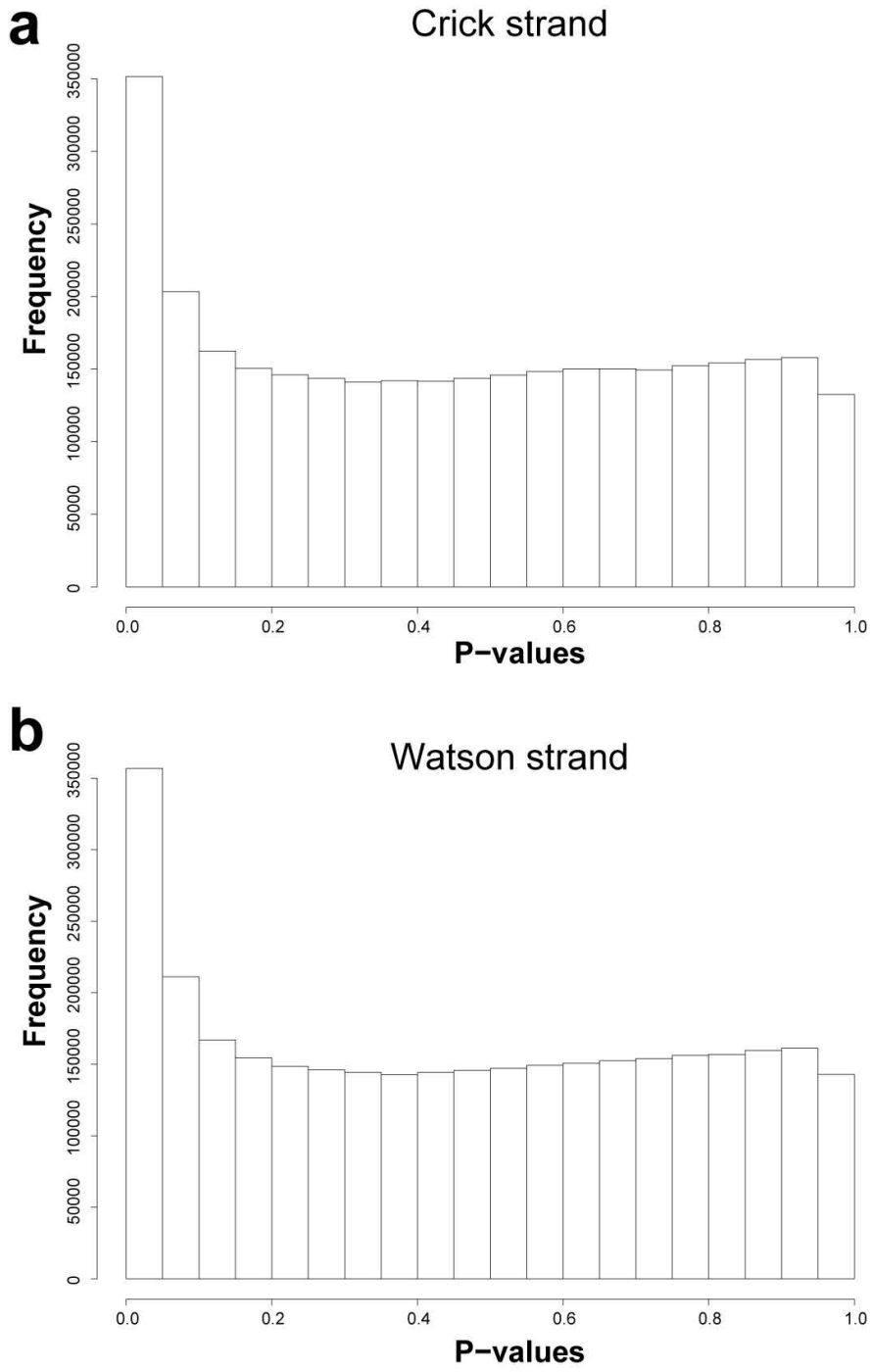


Figure 4.2: **a, b.** The histograms show the distributions of empirical p-values associated with LDRs between all combinations of treatment and control samples on the transcriptome data set for both strands. Reproduced from Supplementary Figures of Selega et al. (2017).

Robust statistical modeling improves sensitivity of high-throughput RNA structure probing experiments

Alina Selega[1], Christel Sirocchi[2], Ira Iosub[2], Sander Granneman[2] & Guido Sanguinetti[1,2]

Structure probing coupled with high-throughput sequencing could revolutionize our understanding of the role of RNA structure in regulation of gene expression. Despite recent technological advances, intrinsic noise and high sequence coverage requirements greatly limit the applicability of these techniques. Here we describe a probabilistic modeling pipeline that accounts for biological variability and biases in the data, yielding statistically interpretable scores for the probability of nucleotide modification transcriptome wide. Using two yeast data sets, we demonstrate that our method has increased sensitivity, and thus our pipeline identifies modified regions on many more transcripts than do existing pipelines. Our method also provides confident predictions at much lower sequence coverage levels than those recommended for reliable structural probing. Our results show that statistical modeling extends the scope and potential of transcriptome-wide structure probing experiments.

RNA structure plays a key role in regulating RNA stability, transcription, and mRNA translation rates. In order to identify novel RNA structural regulatory elements, chemical and enzymatic structure probing is routinely used to interrogate RNA structure both *in vivo* and *in vitro*[1]. Current *in silico* RNA structure prediction programs rely on thermodynamic estimates to generate the most likely secondary structure models. By incorporating data from structure probing experiments, the accuracy of secondary and tertiary RNA structure prediction can be significantly improved[2,3].

Most chemical RNA structure probing methods rely on the formation of adducts or cleavage of the RNA backbone, using as probes dimethylsulfate (DMS) and SHAPE reagents such as 1-methyl-7-nitroisatoic anhydride (1M7) and 2-methylnicotinic acid imidazolide (NAI)[4,5]. In all of these methods, the reagents terminate reverse transcription (RT), enabling detection of the sites of cleavage or modification by primer extension analyses, followed by mapping the RT drop-off position back to the reference sequence. These methods can be combined with next-generation sequencing (NGS) to simultaneously probe thousands of RNA molecules, as well as very long RNAs, in a single RT reaction. Insights obtained by these techniques include the largely unstructured state of stress-responsive transcripts in yeast and plants[6,7]. Recently, we developed the ChemModSeq structure probing pipeline to gain deeper understanding of RNA structural changes in long ribosomal RNA precursors during ribosome assembly[8].

NGS is certainly revolutionizing the RNA structure probing field; however, several data analysis issues need to be addressed. First, NGS is often plagued by sequencing representation and coverage biases introduced during library preparation[9]. Identifying and correcting such biases is essential for avoiding erroneous interpretations; however, to our knowledge, current methods do not address these issues. Second, statistical assessments must be informed by an analysis of inter-replicate variability in both control and treatment samples. Except for Mod-seq[10], current methods do not exploit replicate information; as a result, their output scores are not readily statistically interpretable, and interpretation of these scores often requires setting arbitrary thresholds and other postprocessing. Finally, a major question in the field concerns the coverage per nucleotide necessary to get reliable chemical reactivity values. Partly as a result of unresolved statistical issues in handling variability, current recommendations indicate that very high coverage levels are required[10,11], and this requirement is normally only met for a handful of transcripts in transcriptome-wide experiments.

To tackle these important issues, we developed beta-uniform mixture hidden Markov model (BUM-HMM), a statistical machine-learning pipeline for modeling NGS RNA structure probing data. BUM-HMM uses inter-replicate variability to identify transcript regions that are significantly more modified compared with control conditions, incorporating coverage and sequence bias information within the model. The output of BUM-HMM is probabilistic, giving a transparent statistical interpretation which obviates the need for arbitrary thresholds and postprocessing. We demonstrate that, compared with existing bioinformatic pipelines, BUM-HMM is highly sensitive and remarkably robust even at low coverage.

RESULTS

To demonstrate the strength of the BUM-HMM method, we reanalyzed high-throughput DMS and 1M7 RNA structure probing experiments performed on yeast 40S ribosomes[8]. This study generated biological triplicates of each chemical probing experiment with high sequence coverage, both in treatment and control samples (**Supplementary Table 1**). As secondary structure

[1]School of Informatics, University of Edinburgh, Edinburgh, UK. [2]Centre for Synthetic and Systems Biology, University of Edinburgh, Edinburgh, UK. Correspondence should be addressed to S.G. (sgrannem@staffmail.ed.ac.uk) or G.S. (gsanguin@inf.ed.ac.uk).

RECEIVED 25 MARCH; ACCEPTED 3 OCTOBER; PUBLISHED ONLINE 7 NOVEMBER 2016; DOI:10.1038/NMETH.4068

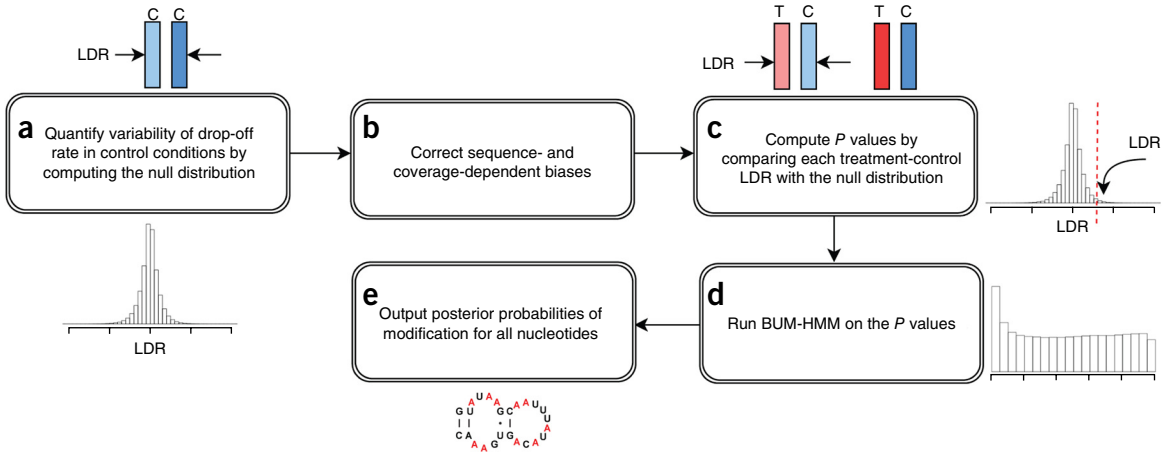


Figure 1 | Overview of the BUM-HMM computational analysis pipeline. (**a**) Null distribution of LDRs at each nucleotide position is computed for all pairs of control replicate samples (bars labeled "C" each represent per-nucleotide drop-off rates in a given sample), quantifying variability in drop-off rate observed by chance. (**b**) Coverage-dependent bias is corrected by applying a variance stabilization transformation. For transcriptome-wide data sets, different null distributions are computed for different nucleobase patterns to address sequence-dependent bias. (**c**) Per-nucleotide empirical *P* values are computed for all pairs of treatment (bars labeled "T") and control replicate samples by comparing the corresponding LDRs to the null distribution (in the figure, the LDR is larger than most items in the null distribution). (**d**) BUM-HMM is run on *P* values as observations, leaving out any nucleotides with missing data. (**e**) The output is a posterior probability of modification, ranging from 0 to 1, carrying structural information for each nucleotide included in the analysis.

models for rRNAs and crystal structures of yeast ribosomes are now readily available[12,13], these data allowed us to investigate the sensitivity and specificity of BUM-HMM compared with existing methods. In addition, we also generated two *in vivo* yeast mRNA transcriptome data sets using NAI as the chemical probe (see Online Methods), which enabled us to test the performance of BUM-HMM in the context of a transcriptome-wide mRNA structure probing experiment. For these analyses, between 36 and 55 million paired cDNA sequences were analyzed per sample (see **Supplementary Table 1** and Online Methods).

Data preparation and model

All cDNA libraries were generated by random priming[6,8,11,14] and paired-end sequenced (see Online Methods and **Supplementary Fig. 1**). Paired-end sequencing allows normalization for different read depths through calculating drop-off rates, which we define as the total number of reads stopping at a nucleotide divided by the total number of reads that cover that nucleotide[8,14]. The full procedure is described in detail in the Online Methods and schematically illustrated in **Figure 1**.

Briefly, we quantified biological variability using the log ratio between the drop-off rates at the same nucleotide in a pair of control replicates (log-dor ratio, LDR), for all possible pairs. We assembled all control LDRs in a null distribution (step A) and corrected sequence and coverage biases (step B) to control for confounders (see Online Methods and **Supplementary Fig. 2**). We then evaluated empirical *P* values for all treatment-control LDRs at each nucleotide (step C) and modeled these *P* values using a BUM-HMM (step D) with hidden states corresponding to presence or absence of modification (see Online Methods and **Supplementary Fig. 3** for a theoretical justification of the beta-uniform choice). We used BUM-HMM to compute posterior probabilities of chemical modification for all nucleotides (step E), providing a robust and statistically interpretable readout.

It is important to note that, while single molecules are either modified or unmodified at a particular locus, interpreting structure probing data as binary may appear overly simplistic. Transcripts *in vivo* exhibit dynamic secondary structures and may be bound by different proteins, so that different molecules of the same transcript may be accessible to chemical reagents at different positions. Furthermore, not all accessible nucleotides will be modified at low reagent concentrations, such as those nucleotides typically used in structure probing experiments. The correct interpretation of the probabilistic output of BUM-HMM is therefore not that all transcript molecules with high posterior probability at a locus are in a specific state of accessibility, but that the proportion of modified molecules is sufficiently large to lead to an LDR value which cannot be explained by random variability alone.

Performance comparisons

Interpreting and evaluating the outcome of structure probing experiments is a notoriously difficult task because of a lack of 'ground truth' examples to validate model predictions (see also "Discussion"). In this respect, yeast 18S ribosomal RNA represents a case of a high-abundance transcript with a well-defined and stable secondary structure. Therefore, we first evaluated BUM-HMM's performance in terms of recovering the 18S structure from a recently published chemical probing data set[8]. These data sets have extremely high coverage (with a mean coverage per nucleotide close to 1 million for some samples; **Supplementary Table 1**), which clearly cannot be achieved on many transcripts in transcriptome-wide studies. We thus later examine the performance of BUM-HMM on a transcriptome data set that reflects a more realistic coverage scenario. We demonstrate through a number of case studies how BUM-HMM can aid the use of structure prediction algorithms and recover structural features in conserved areas of transcripts and we examine the robustness of BUM-HMM toward reductions in coverage.

BUM-HMM recovers the structure of 18S with readily interpretable output

Guided by the available 80S and 40S structures[12,13], we determined which nucleotides were accessible and single stranded

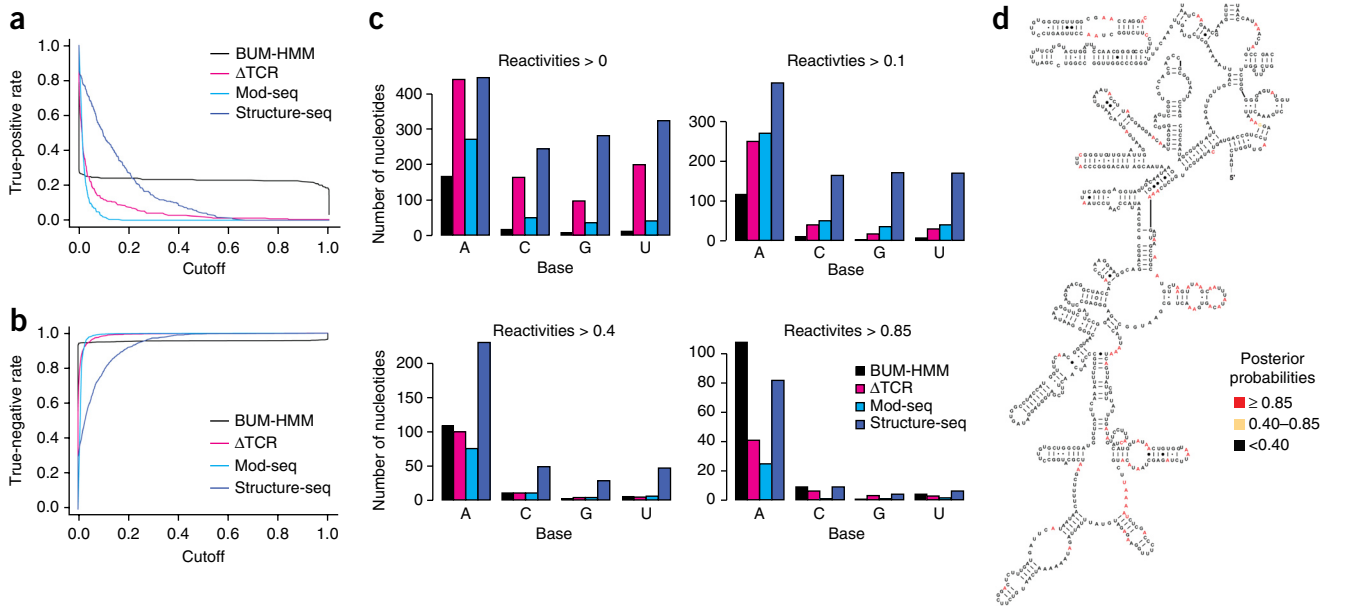


Figure 2 | BUM-HMM identifies many modified nucleotides of 18S ribosomal RNA with high accuracy and specificity. (**a**,**b**) True-positive rate and true-negative rate of BUM-HMM as compared with those of ΔTCR, Mod-seq, and structure-seq for reconstructing secondary structure of 18S rRNA, as evaluated against the known crystal structure. (**c**) Base composition of called nucleotides for all methods, when considering reactivity scores greater than: a value close to zero (10^{-6}), a low-reactivity threshold (0.1), a medium-reactivity threshold (0.4), and a high-reactivity threshold (0.85). (**d**) A fragment of the 18S secondary structure with bases colored according to the BUM-HMM posterior probability at the corresponding nucleotide position.

and should therefore, in theory, be modified by 1M7 or DMS. Notice that this crystallographic structure is different from the phylogenetic (predicted) structure used in other studies[15]. As DMS preferentially reacts with A and C nucleotides, we were able to examine the sensitivity and specificity of BUM-HMM. From many existing bioinformatic approaches[6–8,10,14,16], we chose the following methods to compare our model to: structure-seq[6]; ΔTCR[14] (which was the strongest performer in a recent review[16]); and Mod-seq[10], which to our knowledge is the only method supporting multiple biological replicates. We evaluated all methods using the receiver operating characteristic (ROC), which plots the false-positive rate against the true-positive rate for different discrimination thresholds. A random predictor would have the area under the ROC curve (the AUC statistic) equal to 0.5, and the higher the AUC value, the better the predictor performs. When evaluated against the known crystal structure, BUM-HMM and ΔTCR were clearly the best performers (with AUCs of 0.73 and 0.74, respectively), outperforming structure-seq and Mod-seq (AUCs of 0.68 and 0.64, respectively). The 1M7 data set demonstrated similar performance between methods (**Supplementary Table 2**).

However, the dynamic output ranges of the methods vary dramatically; to enable comparisons with BUM-HMM while taking into account these differences, we separately examined the true-positive and true-negative rate for different discrimination thresholds (scaling the scores to range between 0 and 1). BUM-HMM demonstrated a 20% increase of the true-positive rate throughout most of the dynamic range compared with the other methods, and it demonstrated only a small decrease of the true-negative rate (**Fig. 2a**,**b**).

Figure 2c shows the proportions of nucleobases called as modified by all methods when discriminating the scores at low, medium, and high thresholds or considering all scores greater than zero. BUM-HMM has excellent specificity to A and C throughout its dynamic range. On the contrary, structure-seq and ΔTCR do not discriminate as well between C, G, and U when considering all scores, demonstrating these methods' reliance on arbitrary thresholds as the means to remove noise. BUM-HMM identifies over a hundred modified nucleotides with high posterior probabilities, many more nucleotides than the other methods do when considering high reactivity thresholds. It is interesting to observe that on the 18S DMS data, BUM-HMM generates an almost binary output, with few values between 0 and 1. This reflects the stability of the 18S transcript clearly evident from the data, rather than a property of the model; BUM-HMM generates many more intermediate values on the transcriptome data set.

Figure 2d shows a fragment of the 18S secondary structure as predicted by BUM-HMM, with many single-stranded As and Cs correctly identified. The results for all methods are shown on the 18S secondary structure models in **Supplementary Figure 4**.

BUM-HMM output aids computational prediction of secondary structures

As explained earlier, the output posterior probabilities of BUM-HMM should not be directly interpreted as secondary structure readouts in general. These probabilities can, however, provide valuable constraints to energy-based structure prediction software such as RNAstructure[17], ViennaRNA[18], and others. Such software predicts secondary structures of transcripts by minimizing the free energy associated with a particular 'sequence–structure' configuration. For all but the shortest transcripts, this is a difficult combinatorial optimization problem, resulting in many nearly equivalent optima corresponding to different structures. Transcripts *in vivo* are highly dynamic and can therefore exist in many different such configurations. However, under physiological constraints, it can be expected that only a subset of all possible

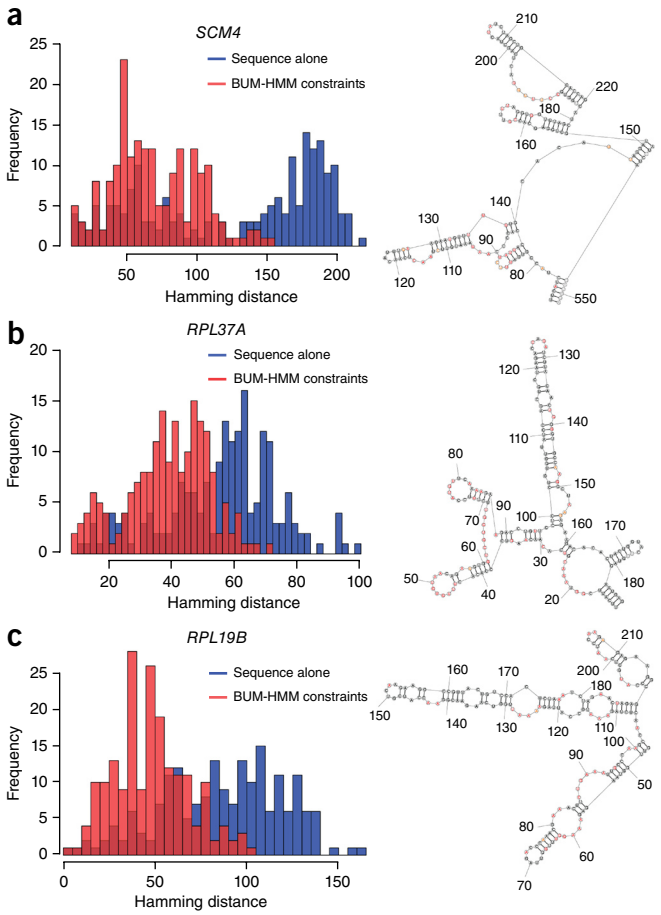


Figure 3 | Using BUM-HMM output results in more consistent secondary structure prediction. (**a**) Left, distribution of Hamming distances between all pairs of secondary structures (n = 20) predicted for *SCM4* by Fold when using only sequence (blue) and adding the BUM-HMM output as constraints (red). Right, a fragment of the lowest free energy structure. (**b**,**c**) Same as in **a**, for *RPL37A* (**b**) and *RPL19B* (**c**).

structures (from a free energy point of view) will be present. We therefore used the BUM-HMM output as constraints for structure prediction with the RNAstructure web server[17].

To quantify the improvement provided by the BUM-HMM constraints, we selected as representative examples the coding sequences of *SCM4*, which encodes a mitochondrial outer membrane protein, and of *RPL37A* and *RPL19B*, which encode ribosomal 60S subunit proteins. These genes all have good coverage levels (mean coverage per nucleotide: 799; 38,711; and 15,798, respectively), thus problems with missing information are avoided; they are also relatively long transcripts (564, 260, and 568 nucleotides long, respectively), and hence challenging for structure prediction algorithms. We used the Fold[17] method in RNAstructure, with and without the BUM-HMM constraints, to predict the secondary structure of these genes. Fold generally returns an ensemble of around 20 low-free-energy structures, and we quantified the distance between two structures by using the binary Hamming distance. Constraining the algorithm with the BUM-HMM output considerably narrowed down the search space for free energy minimization, as demonstrated by smaller Hamming distances between the resulting structures (**Fig. 3a–c**). Furthermore, these structures were more similar to the output of the alternative method MaxExpect[17] compared with only using sequence (**Supplementary Fig. 5**). We conclude that using posterior probabilities generated by BUM-HMM as algorithm constraints can improve secondary structure prediction for relatively long transcripts.

BUM-HMM correctly predicts structure of conserved regions in U3 small nucleolar RNA

While transcripts may coexist in several different structural configurations, it is likely that some of their sections present increased structural stability for correct cellular functioning (e.g., in order to be bound by proteins). It is reasonable to expect highly conserved regions of a transcript to represent its more stable parts. To validate our model in a more realistic transcriptome-wide coverage scenario, we turned to the small nucleolar RNA (snoRNA) U3. U3 is a model for evolutionary fitness studies[19] and has an accepted secondary structure in yeast[20], making it a good candidate for validation.

Even though the coverage on U3 was uneven and did not allow structural predictions on the whole molecule, BUM-HMM achieved an AUC of 0.76 when evaluated on the highly conserved regions located in boxes A, A′, B, C, C′, and D. Furthermore, when considering the longest conserved region with 16 nucleotides (box A and one highly conserved upstream nucleotide), BUM-HMM demonstrated excellent prediction accuracy of 0.88.

BUM-HMM increased informativeness on transcriptome-wide analysis of RNA structure probing data

To evaluate the applicability of the methods in the transcriptome-wide scenario, we generated synthetic data sets by randomly selecting subsets of reads from the 18S DMS data set and evaluated the consistency of the methods at lower coverage (see Online Methods). BUM-HMM showed excellent consistency, as the mean coverage along the transcript was progressively reduced (**Fig. 4**), retaining accuracy significantly above that of random performance even at a reduction of almost 2,000 times (**Supplementary Fig. 6**). This performance challenges recent recommendations for the minimum coverage level for chemical probing experiments[11], indicating that BUM-HMM can obtain reliable predictions on a large fraction of transcripts in a standard transcriptomic experiment. Mod-seq and structure-seq exhibited considerably lower levels of consistency (**Fig. 4c**,**d**) and behaved as random predictors at the lowest coverage level. Highly consistent reactivity scores generated by ΔTCR (**Fig. 4b**) were largely due to its extreme conservatism at the chosen threshold of 50% of the dynamic range, at which it called no more than 20 nucleotides at all coverage levels. Notably, all methods identified fewer modified nucleotides than BUM-HMM both on the full data set and at all coverage levels; this difference was particularly striking with ΔTCR and Mod-seq (**Fig. 4b**,**c**).

While performance analysis is hampered by a lack of a ground truth for most transcripts, a more general assessment of the informativeness of the methods' outputs is possible and instructive. We therefore quantified how many transcripts had at least 5% of their length called as modified by BUM-HMM and ΔTCR. We considered those nucleotides which obtained a score above 50% of the dynamic range of the model (having removed outliers for ΔTCR) to be 'called as modified'. With this procedure, BUM-HMM identified 2,219 transcripts; while ΔTCR only retrieved 285. The low number of transcripts identified by ΔTCR is at odds with

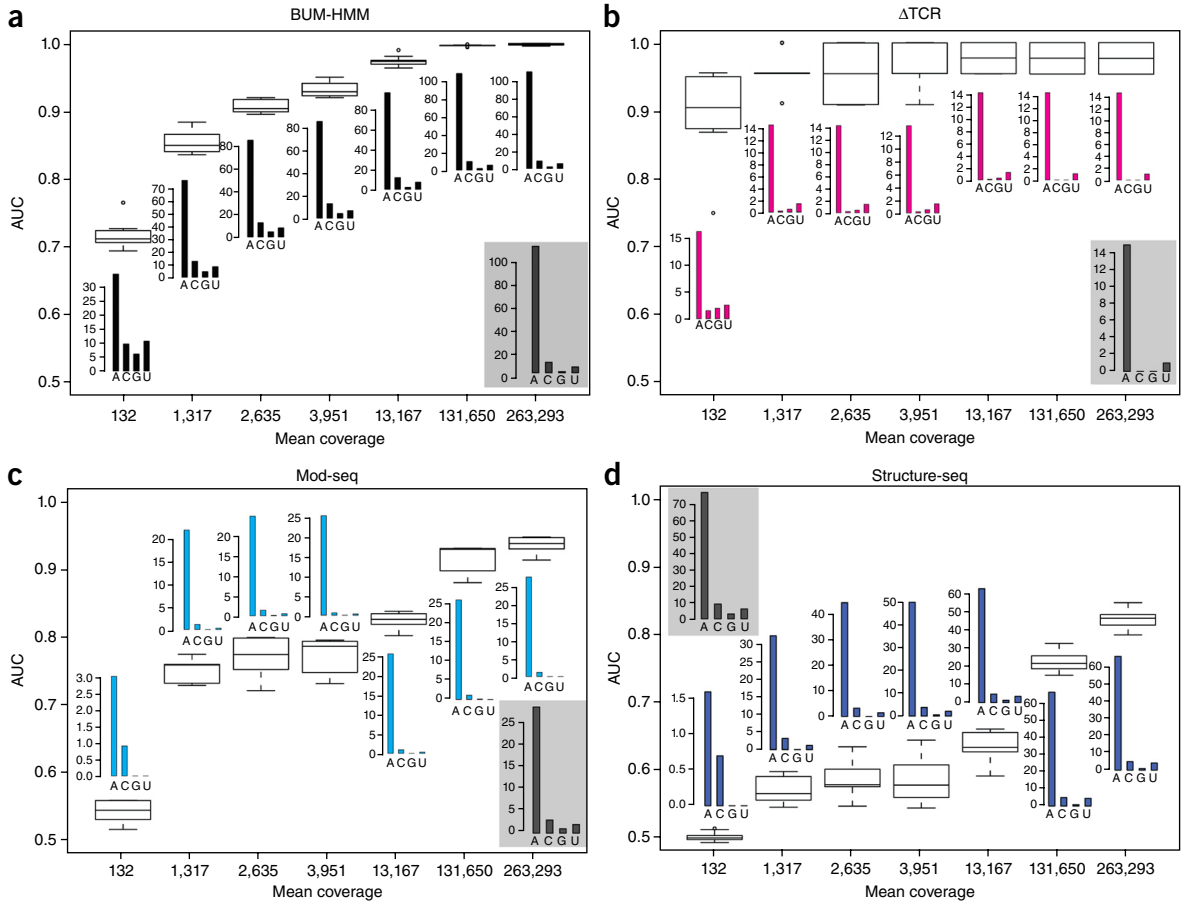


Figure 4 | BUM-HMM is highly consistent at low coverage and calls more nucleotides modified at all coverage levels. (**a**) Consistency of posterior probabilities generated by BUM-HMM on data sets with progressively lower mean coverage (shown on the *x*-axis), measured with the AUC statistic between them and the predictions made for the full 18S DMS data set. Error bars quantify variability across random selections of each reads subset, synthesized from the full data set (see Online Methods). For each coverage level, base composition of nucleotides called as modified is shown in a corresponding barplot, averaged across the selections of each subset. The barplot in a shaded rectangle corresponds to the base composition of called nucleotides on the full data set. (**b**–**d**) Consistency of reactivity scores generated by ΔTCR (**b**), Mod-seq (**c**), and structure-seq (**d**) on the same synthetic data sets, with prior outlier removal.

previous studies[6,7], suggesting that many RNAs are largely accessible and unstructured *in vivo*; this conservativeness may be due to the normalization procedures of ΔTCR[14] (see **Supplementary Fig. 7** for illustration of associated problems).

We next analyzed the distribution of posterior probabilities across those mRNA transcripts which had a nonzero score attached to more than 75% of their length, transcripts which we call effectively probed. BUM-HMM selected 363 mRNA genes (**Fig. 5a**), a striking contrast with ΔTCR's 43 selected transcripts. When relaxing this criterion to (still highly informative) effective probing of more than 50% of the length, the number of mRNAs selected by BUM-HMM increased dramatically to 1,764. Analyses of the 363 selected genes revealed that many appeared to have long segments of almost completely unstructured regions (such as *TDH3*, **Fig. 5b**) and many had significant structure in the coding sequence (such as *YOR365W*, **Fig. 5b**). We next calculated the average fragments per kilobase of transcript per million mapped reads (FPKMs) for these genes using the read counts from the control and treated sequencing data. This revealed a broad distribution with a median 191 (**Fig. 5b**) and the lowest FPKM of 60 (*YOR385W*, **Fig. 5b**,**c**). The *YOR385W* gene had an average coverage of 335 reads per nucleotide, which we propose can be an indicative guideline of the lower bound on coverage required for high-throughput RNA structure probing experiments to effectively probe long transcripts.

Metabolic transcripts are generally flexible around the translation start site

Structure in untranslated regions (UTR) and around the translation start site (AUG) can reduce translation efficiency[21,22]. Recent high-throughput RNA structure probing also revealed a weak but significant negative correlation between RNA structure at AUG *in vitro* and ribosome occupancy[23]. To test whether RNA structure measured *in vivo* also correlates with ribosome occupancy, we plotted the distribution of posterior probabilities around the translation start sites and performed a *k*-means clustering to identify patterns in the data. This revealed five clusters with different reactivity profiles (**Fig. 5d**). For the majority of transcripts, the region around the AUG had high posterior probabilities and therefore appeared to be largely unstructured (genes in clusters 0, 2, 3, and 4). Interestingly, KEGG pathway analyses revealed that these clusters were highly enriched for transcripts encoding for ribosomal and metabolic proteins, in particular proteins involved in glycolysis or gluconeogenesis and amino

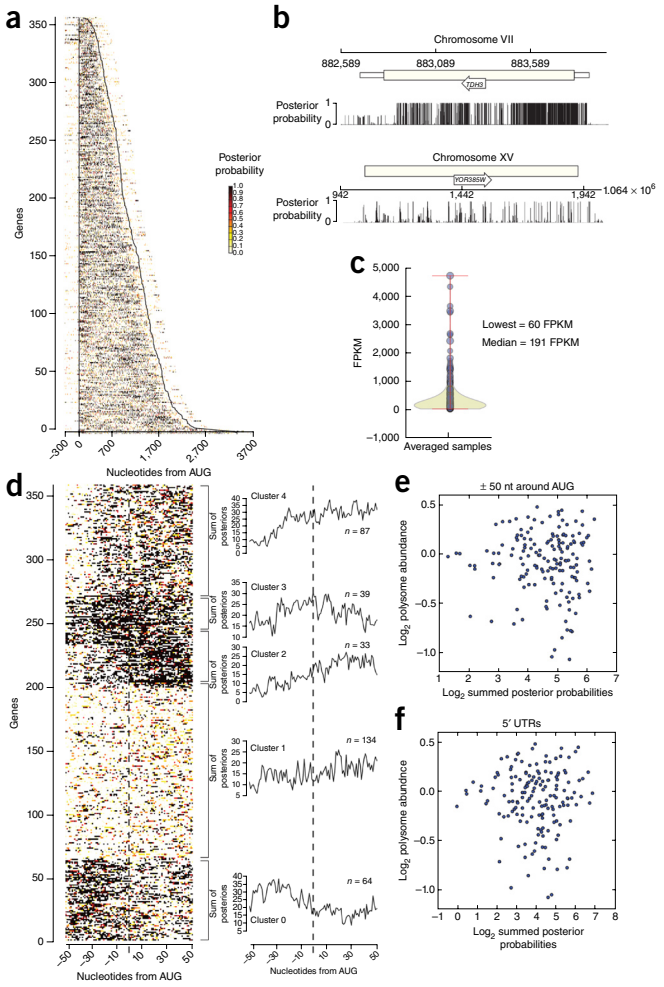


Figure 5 | Flexibility of 5′ UTR and ribosome occupancy do not show a significant positive correlation *in vivo*. (**a**) Distribution of posterior probabilities over 363 protein-coding transcripts. The heatmap displays posterior probabilities for 363 mRNA sorted by length (from short to long) and extended at each end by 300 nucleotides. The two black lines indicate the position of the start codon and stop codon, respectively. (**b**) Genome browser examples showing posterior probabilities of a highly expressed gene (*TDH3*; average FPKM = 3,491) and a weakly expressed gene (*YOR385W*; average FPKM = 60). (**c**) Violin plot showing the distribution of average FPKMs, calculated using the sequence reads from the control and NAI data sets. (**d**) The plot shows the distribution of posterior probabilities 50 nucleotides around the translation start site (AUG), clustered in five groups. On the right side of the heatmap, cumulative plots and number of genes (*n*) in each cluster are shown. (**e**) For each gene, we calculated $\log_2$ of the sum of posterior probabilities from the heatmap data shown in **d** and plotted it against the $\log_2$ of the reported enrichment of the transcript in polysomes[25]. Nt, nucleotide. (**f**) Same as in **e** but with the entire 5′ UTR.

acid biosynthesis (**Supplementary Table 3**). Remarkably, the more structured transcripts in cluster 1 were mostly enriched for transcripts encoding proteins involved in mitochondrial translation (**Supplementary Table 3**).

One possible explanation for why the metabolic transcripts appear largely unstructured *in vivo* could be because they were occupied by ribosomes, which have an intrinsic RNA helicase activity to unfold structured regions within mRNAs[24]. We therefore asked whether there was a significant correlation between RNA flexibility within that region and ribosome occupancy on the transcripts. To test this, we calculated $\log_2$ of the sum of posterior probabilities within 50 nucleotides around AUG and compared it with the translational efficiency obtained from recently published polysome microarray data[25] (**Fig. 5e**). This revealed that flexibility around the AUG did not positively correlate with polysome occupancy (Pearson correlation: −0.196; *P* value = 0.0014). Similar results were obtained when using the entire 5′ UTR region (**Fig. 5f**). Taken together, these results suggest that high ribosome occupancy alone is not sufficient to explain why certain transcripts were highly flexible in our *in vivo* NAI chemical probing data.

DISCUSSION

Our statistical pipeline addresses a number of important problems in the analysis of high-throughput RNA secondary structure probing data. First, it explicitly models the biological variability of the data, providing a statistical basis for determining the significance of the observed signal. As such, it removes the need to set arbitrary thresholds and perform extensive postprocessing of the analysis results, yielding a clean and statistically interpretable pipeline. This is a direct consequence of the probabilistic formulation of BUM-HMM. In this respect, it is indebted to earlier probabilistic models of SHAPE-Seq data[26]; notably, however, recent developments in the experimental technology—and in particular, the shift to random-primed experimental designs—force a major change in model architecture and motivate the nonparametric approach we take.

Our analysis identified important biases in the technology, especially prominent in transcriptome-wide experiments, which can have severe downstream consequences in any analysis. While random-priming designs effectively resolve the 3′ biases of earlier SHAPE technologies, significant sequence and coverage biases remain. Our method provides automated empirical strategies for correcting these biases, potentially extending the applicability of the technology.

Finally, the BUM-HMM model generates accurate and more informative results compared with the results of other methods. Crucially, its predictions remain consistent even with reduced sequence coverage, demonstrating that the choice of an appropriate modeling framework can greatly increase the robustness of the technology. This is borne out by the effectiveness of BUM-HMM on a transcriptome data set with relatively low coverage; while current state-of-the-art methods can only provide information over a handful of transcripts, BUM-HMM selected more than 360 transcripts, some of which had a per-nucleotide coverage as low as 335, heralding the advent of truly transcriptome-wide structure probing experiments.

However, it is important to stress that significant issues remain unresolved with the interpretation of RNA structure probing data. Many factors may affect accessibility (protein binding being a prime example), and in general transcripts *in vivo* may coexist in multiple configurations, cautioning against simplistic interpretations in terms of secondary structure. How structure probing data may be used to inform model-based structure prediction is an important and active research field[27,28]. Our results show that BUM-HMM constraints, when incorporated in structure prediction algorithms, lead to more consistent structure models for many transcripts, demonstrating the importance of statistically sound data analysis strategies for downstream analyses.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. The transcriptome-wide chemical probing sequencing data are available in the Gene Expression Omnibus under accession number GSE78208.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank all the members of the Granneman and Sanguinetti labs for critically reading the manuscript. This work was supported by grants from the Wellcome Trust to S.G. (091549) and I.I. (102334), a European Research Council grant to G.S. (MLC306999) and the Wellcome Trust Centre for Cell Biology core grant (092076). A.S. is supported in part by grants from the UK Engineering and Physical Sciences Research Council, Biological Sciences Research Council, and the UK Medical Research Council (EP/F500385/1 and BB/F529254/1 to the University of Edinburgh Doctoral Training Centre in Neuroinformatics and Computational Neuroscience). Next generation sequencing was carried out by Edinburgh Genomics, The University of Edinburgh. Edinburgh Genomics is partly supported through core grants from NERC (R8/H10/56), MRC (MR/K001744/1) and BBSRC (BB/J004243/1).

AUTHOR CONTRIBUTIONS

All authors contributed to planning the experiments and computational procedures. C.S., I.I., and S.G. carried out the experiments. G.S. and A.S. developed the computational analysis pipeline. A.S., C.S., S.G., and G.S. performed the bioinformatics and computational analyses of the sequencing data. All authors contributed to writing the manuscript and approved the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Kubota, M., Tran, C. & Spitale, R.C. Progress and challenges for chemical probing of RNA structure inside living cells. *Nat. Chem. Biol.* **11**, 933–941 (2015).
2. Wu, Y. *et al.* Improved prediction of RNA secondary structure by integrating the free energy model with restraints derived from experimental probing data. *Nucleic Acids Res.* **43**, 7247–7259 (2015).
3. Ouyang, Z., Snyder, M.P. & Chang, H.Y. SeqFold: genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data. *Genome Res.* **23**, 377–387 (2013).
4. Mortimer, S.A. & Weeks, K.M. A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *J. Am. Chem. Soc.* **129**, 4144–4145 (2007).
5. Spitale, R.C. *et al.* RNA SHAPE analysis in living cells. *Nat. Chem. Biol.* **9**, 18–20 (2013).
6. Ding, Y. *et al.* *In vivo* genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature* **505**, 696–700 (2014).
7. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J.S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures *in vivo*. *Nature* **505**, 701–705 (2014).
8. Hector, R.D. *et al.* Snapshots of pre-rRNA structural flexibility reveal eukaryotic 40S assembly dynamics at nucleotide resolution. *Nucleic Acids Res.* **42**, 12138–12154 (2014).
9. van Dijk, E.L., Jaszczyszyn, Y. & Thermes, C. Library preparation methods for next-generation sequencing: tone down the bias. *Exp. Cell Res.* **322**, 12–20 (2014).
10. Talkish, J., May, G., Lin, Y., Woolford, J.L. Jr. & McManus, C.J. Mod-seq: high-throughput sequencing for chemical probing of RNA structure. *RNA* **20**, 713–720 (2014).
11. Siegfried, N.A., Busan, S., Rice, G.M., Nelson, J.A.E. & Weeks, K.M. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nat. Methods* **11**, 959–965 (2014).
12. Ben-Shem, A. *et al.* The structure of the eukaryotic ribosome at 3.0 Å resolution. *Science* **334**, 1524–1529 (2011).
13. Aylett, C.H.S., Boehringer, D., Erzberger, J.P., Schaefer, T. & Ban, N. Structure of a yeast 40S-eIF1-eIF1A-eIF3-eIF3j initiation complex. *Nat. Struct. Mol. Biol.* **22**, 269–271 (2015).
14. Kielpinski, L.J. & Vinther, J. Massive parallel-sequencing-based hydroxyl radical probing of RNA accessibility. *Nucleic Acids Res.* **42**, e70 (2014).
15. Tang, Y. *et al.* StructureFold: genome-wide RNA secondary structure mapping and reconstruction *in vivo*. *Bioinformatics* **31**, 2668–2675 (2015).
16. Kielpinski, L.J., Sidiropoulos, N. & Vinther, J. Reproducible analysis of sequencing-based RNA structure probing data with user-friendly tools. *Methods Enzymol.* **558**, 153–180 (2015).
17. Reuter, J.S. & Mathews, D.H. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* **11**, 129 (2010).
18. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
19. Puchta, O. *et al.* Network of epistatic interactions within a yeast snoRNA. *Science* **352**, 840–844 (2016).
20. Méreau, A. *et al.* An *in vivo* and *in vitro* structure-function analysis of the *Saccharomyces cerevisiae* U3A snoRNP: protein-RNA contacts and base-pair interaction with the pre-ribosomal RNA. *J. Mol. Biol.* **273**, 552–571 (1997).
21. Kudla, G., Murray, A.W., Tollervey, D. & Plotkin, J.B. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* **324**, 255–258 (2009).
22. Tuller, T., Waldman, Y.Y., Kupiec, M. & Ruppin, E. Translation efficiency is determined by both codon bias and folding energy. *Proc. Natl. Acad. Sci. USA* **107**, 3645–3650 (2010).
23. Kertesz, M. *et al.* Genome-wide measurement of RNA secondary structure in yeast. *Nature* **467**, 103–107 (2010).
24. Takyar, S., Hickerson, R.P. & Noller, H.F. mRNA helicase activity of the ribosome. *Cell* **120**, 49–58 (2005).
25. Arribere, J.A., Doudna, J.A. & Gilbert, W.V. Reconsidering movement of eukaryotic mRNAs between polysomes and P bodies. *Mol. Cell* **44**, 745–758 (2011).
26. Aviran, S. *et al.* Modeling and automation of sequencing-based characterization of RNA structure. *Proc. Natl. Acad. Sci. USA* **108**, 11069–11074 (2011).
27. Deng, F., Ledda, M., Vaziri, S. & Aviran, S. Data-directed RNA secondary structure prediction using probabilistic modeling. *RNA* **22**, 1109–1119 (2016).
28. Eddy, S.R. Computational analysis of conserved RNA secondary structure in transcriptomes and genomes. *Annu. Rev. Biophys.* **43**, 433–456 (2014).

ONLINE METHODS

ChemModSeq library preparation. The 18S DMS and 1M7 data sets were previously described[8] and can be accessed under the accession code GSE52878 at the Gene Expression Omnibus repository. To generate the NAI transcriptome-wide data set, yeast cells (BY4741 strain) were grown to exponential phase and harvested by centrifugation. Cells were subsequently resuspended in 1 volume of phosphate buffer saline (PBS). NAI (dissolved in DMSO) was added to the suspension in a final concentration of 100 mM (5% DMSO final) and incubated for 10 min at room temperature. Cells were harvested by centrifugation, washed with ice-cold PBS, and snap frozen in liquid nitrogen. Total RNA was extracted as previously described[29]. The mRNAs were isolated using the PolyATtract mRNA isolation kit according to manufacturer's procedures (Promega). Two biological replicates were generated for the transcriptome-wide analyses. The ChemModSeq libraries were generated as previously described[8]. Briefly, cDNA was generated by random priming using a random hexamer oligo[8]. Subsequently, a DNA adaptor was ligated to the 3′ end of cDNAs using CircLigase. These adaptors contained a random nucleotide at the 5′ end to minimize the sequence representation biases introduced during the linker ligation reaction. Following PCR, libraries were resolved on 2% Metaphor gels, and fragments between 200–700 were gel purified. Samples were sequenced on Illumina HiSeq2500 systems.

Sequence data processing and raw data analysis. To process the fastq files the pyCRAC package was used[30]. To demultiplex the raw sequencing data we used pyBarcodeFilter.py, after which the remaining random nucleotide was removed from the 5′ end of the forward reads. The data were subsequently collapsed using pyFastqDuplicateRemover.py that utilizes the random barcode information present in the 5′ adaptors to remove potential PCR duplicates. The resulting fasta file was mapped to the *Saccharomyces cerevisiae* genome (version R64, ENSEMBL) using novoalign 2.05, and only uniquely mapped reads were considered. PyReadCounters.py was subsequently used to generate read counts and FPKMs for all annotated features. The resulting GTF output files were converted to tab-delimited files containing three columns: chromosome, genomic position, and coverage or drop-off counts using pyGTF2sgr.py. These files were then fed to the BUM-HMM model to generate posterior probabilities.

Software. The software implementing the BUM-HMM pipeline can be accessed in the following repository: https://github.com/alinaselega/BUMHMM.

Data characterization. Using the final output files (see sequence data processing and raw data analysis), the drop-off rate was computed for all nucleotide positions in each replicate as a measure of nucleotide's reactivity to the probing reagent in a given experiment. By definition, the drop-off rate ranges between 0 and 1. All drop-off rates were normalized to a common median across replicate samples.

$$r = k/n$$

where r is the drop-off rate, k is the drop-off count, and n is the coverage.

A measure of inter-replicate variability at each nucleotide position is defined as the log-ratio of drop-off rates (LDR) in a pair of replicate samples i and j:

$$\log\frac{r_i}{r_j} = \log(r_i) - \log(r_j)$$

If the drop-off rates are similar in both samples, the LDR will be close to 0, indicating little variability. In contrast, different drop-off rates would result in an LDR large in absolute value. LDRs in control conditions collectively describe the variability in drop-off rates that could be observed in the absence of the probing reagent. The set of these define the null distribution of LDRs.

LDRs are then computed for each combination of treatment-control replicates, quantifying the difference between the drop-off rate observed in a treatment experiment with respect to a control replicate. These are compared with the null distribution giving rise to empirical P values. For efficiency, LDRs are compared with the precomputed quantiles of the null distribution. The P value of an LDR represents the probability of it being insignificantly different from what could be observed by chance.

$$P\text{ value} = 1 - q, \text{where } q \text{ is the closest quantile}$$

Preprocessing. In order to use the log transform, it is necessary to ensure that no nucleotides have zero drop-off rates. Therefore, only those nucleotides with nonzero drop-off counts for a corresponding pair of replicate samples are used. The pipeline also features a user-defined parameter describing the minimum level of coverage that nucleotides should have to be included in the analysis (set to 1 in our analyses).

Model. Empirical P values, computed for each nucleotide position and each treatment-control comparison (of which there are nm for n treatment and m control experimental replicates) are passed onto a hidden Markov model. The model has a hidden state h_t ($t = 1 \ldots T$ for T nucleotides) representing the true binary state of the t-th nucleotide (modified, 1; or unmodified, 0) and the observed variable v_t, corresponding to the empirical P value at that position. P values corresponding to different pairs of treatment-control replicates are assumed to be independent measurements. Notice that, since P values are used as features and not for decision making, no issues of multiple hypothesis testing arise.

Transition probabilities are defined through empirically derived lengths of single- and double-stranded stretches of nucleotides. The model assumes expected uninterrupted stretches of 20 double-stranded, or constrained, nucleotides and 5 single-stranded, or flexible, nucleotides.

Emission probabilities come from a beta-uniform mixture (BUM) model. This design exploits the result that P values are uniformly distributed under the null hypothesis[31]. P values corresponding to accessible nucleotides are modeled with a Beta distribution, which favors small values, accommodating the fact that accessible nucleotides would have LDRs greater than most values in the null distribution. The P value distribution computed for the transcriptome-wide data set strongly agrees with this model (**Supplementary Fig. 3**). The HMM is run separately on continuous stretches of nucleotides with a user-specified minimum coverage threshold and a nonzero drop-off rate in at least one treatment sample.

$$P(v_t \mid h_t = 0) \sim U(0,1)$$

$$P(v_t \mid h_t = 1) \sim \text{Beta}(\alpha, \beta), \text{ with } \alpha = 1; \beta = 10$$

doi:10.1038/nmeth.4068

The default values for the beta parameters were chosen heuristically as to attach approximately equal likelihood under both hypotheses to nucleotides with LDR in the top quintile of the empirical distribution.

Statistics. Quantification of *P* values associated with each nucleotide in treated data sets is done by comparing log-dor ratio (LDR) values to the quantiles of the empirical LDR distribution in control data sets.

Optimization of parameters. We provide a strategy to optimize parameters of the beta distribution with respect to the data. This strategy uses the expectation-maximization (EM) algorithm[32] and Newton's optimization method.

The iterative EM algorithm starts with the initial values of $\alpha = 1$ and $\beta = 10$, with which the posterior probabilities are computed. It then computes new estimates for α and β using Newton's optimization method. Newton's method finds the maximum of the expected complete data log likelihood or, more precisely, its relevant terms. The shape parameters α and β only appear in the emission term and, within that, only in the component corresponding to the modified state of the latent variable h_t.

The expected complete data log likelihood is given by the following expression (all expectations are with respect to corresponding distributions):

$$\langle \log p(v_{1:T}, h_{1:T} \mid \alpha, \beta) \rangle = \langle \log p(h_1) \rangle + \left\langle \sum_{t=1}^{T} \sum_{n=1}^{N} \log p(v_t^n \mid h_t) \right\rangle + \left\langle \sum_{t=1}^{T-1} \log p(h_{t+1} \mid h_t) \right\rangle,$$

for $t = 1 \ldots T$ nucleotides and $n = 1 \ldots N$ number of treatment-control comparisons.

The relevant term corresponds to emission probabilities (second term in the previous expression):

$$\left\langle \sum_{t=1}^{T} \sum_{n=1}^{N} \log p(v_t^n \mid h_t) \right\rangle = \sum_{t=1}^{T} \sum_{n=1}^{N} \log p(v_t^n \mid h_t = 0) p(h_t = 0 \mid v_{1:T}^n) + \sum_{t=1}^{T} \sum_{n=1}^{N} \log p(v_t^n \mid h_t = 1) p(h_t = 1 \mid v_{1:T}^n)$$

Within that expression, the relevant term corresponds to the modified state of the hidden variable (second term in the previous expression):

$$\sum_{t=1}^{T} \sum_{n=1}^{N} \log p(v_t^n \mid h_t = 1) p(h_t = 1 \mid v_{1:T}^n) = \sum_{t=1}^{T} \sum_{n=1}^{N} \gamma_t \log \frac{(v_t^n)^{\alpha-1}(1-v_t^n)^{\alpha-1}}{B(\alpha,\beta)} = F.$$

Where $\gamma_t = p(h_t = M \mid v_{1:T}^n)$ is the responsibility.

The first order derivatives of *F* are:

$$\frac{\delta F}{\delta \alpha} = \sum_{t=1}^{T} \sum_{n=1}^{N} \gamma_t \log v_t^n - \gamma_t (\psi_0(\alpha) - \psi_0(\alpha + \beta))$$

$$\frac{\delta F}{\delta \beta} = \sum_{t=1}^{T} \sum_{n=1}^{N} \gamma_t \log (1 - v_t^n) - \gamma_t (\psi_0(\beta) - \psi_0(\alpha + \beta))$$

The second order derivatives of *F* are:

$$\frac{\delta^2 F}{\delta \alpha^2} = \sum_{t=1}^{T} \gamma_t N (\psi_1(\alpha + \beta) - \psi_1(\alpha)$$

$$\frac{\delta^2 F}{\delta \alpha \delta \beta} = \sum_{t=1}^{T} \gamma_t N \psi_1(\alpha + \beta)$$

$$\frac{\delta^2 F}{\delta \beta^2} = \sum_{t=1}^{T} \gamma_t N (\psi_1(\alpha + \beta) - \psi_1(\beta)),$$

here ψ is the polygamma function. Log transform is applied at the beginning of the algorithm to ensure that the estimated α and β are positive. Posterior probabilities are recomputed with the new estimates of α and β, and the process is repeated a maximum number of ten times or until the parameter values stop changing within the small predefined tolerance range. We remark that, in our experiments, the EM optimization appeared severely vulnerable to local minima, and we therefore opted to keep the beta parameters fixed.

Bias correction. We used the transcriptome-wide data set to identify potential confounding factors which influence the LDRs in the absence of a reagent. The aim is to transform all LDRs accordingly and eliminate the revealed biases.

Coverage bias. The coverage bias was identified by plotting the control LDRs as a function of the inter-replicate mean coverage at the corresponding nucleotide position (**Supplementary Fig. 2a**,**b**).

This bias is corrected by learning the functional dependency between these variables and transforming the data to reduce the variance of LDRs. We model drop-off count as a binomially distributed variable, which thus has the following s.d.:

$$\sigma[k] = \sqrt{np(1-p)}$$

with probability of drop-off p for a nucleotide covered n times and a drop-off count of k.

Consequently, LDR has a s.d. of:

$$\sigma[LDR] \propto \frac{\sigma[k]}{n} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

Therefore, the functional relationship between log-ratios and coverage can be modeled as $k\frac{1}{\sqrt{n}} + b$, with some unknown parameters k and b, which are learned from the data using a nonlinear least-squares technique. Then, all LDRs are rescaled by this model with fitted parameters. For efficient runtime on

transcriptome-wide data sets, the LDRs are split in bins of equal coverage ranges and the 95th quantile of LDRs and mean coverage are computed for each bin. These are used for parameter fitting. **Supplementary Figure 2c**,**d** shows that the transformed LDRs have reduced dependency on coverage.

Sequence bias. We compared the resulting LDR null distributions when separately considering nucleobase patterns of length 3 (AAA, AAT, AAG,…). For each of the 64 combinations of nucleobases, the transcriptome sequence was searched for all places of its occurrence. The LDRs of the middle nucleotide at these occurrences defined the null distribution specific to this nucleobase combination. **Supplementary Figure 2e**,**f** demonstrates significant differences between these null distributions.

To correct for this sequence-dependent bias, we store the quantiles of each of the 64 different null distributions and compute empirical *P* values by keeping track of which nucleobase triplet corresponds to the current nucleotide position and looking up values from the corresponding null distribution.

On account of the short length of the 18S ribosomal RNA molecule, the sequence-bias correcting step was omitted from the analysis when handling the corresponding data sets.

Handling of missing data and outliers. The methods used in the evaluation[6,10,14] not only generate scores with drastically differing dynamic ranges, but they also assume different interpretations of the same score values. For instance, ΔTCR makes no distinction between the equal drop-off rates in control and treatment conditions and no coverage, assigning a score of 0 in both cases. Structure-seq marks missing data with a dummy value, whereas Mod-seq clamps the scenarios of no coverage and no significant modification to the same score of 0. Further, the outputs of these methods have clear outliers, with a handful of values being much larger than the 99th quantile of the output distribution. Therefore, simply choosing the midpoint of the dynamic range for binarizing the resulting classifications would result in as few as a single true positive for some methods.

Thus, when performing evaluation, we set the missing data (for those methods that use it) and the outliers (computed as the values greater than the 99.5th quantile of the output distribution) to 0. Considering other strategies, such as removing outliers or only evaluating on the nonmissing data, resulted in grossly limited outputs generated by some methods for the simulated low-coverage levels. Our choice, while circumventing these problems and enabling comparisons, follows the commonly used assumption that the reactivity of zero does not carry significant structural information.

When computing true-positive and true-negative rates, the output scores of all methods were normalized to the range of BUM-HMM. AUCs and true-positive and true-negative rates were computed with the ROCR package[33]. When characterizing the methods' sensitivities using the DMS data set specific to As and Cs, the outputs of ΔTCR and Mod-seq were normalized with the 2–8% normalization rule[34] to enable comparisons at the same (previously used) low-, medium-, and high-reactivity thresholds[34,35].

Secondary structure prediction. When generating secondary structures informed by BUM-HMM, posterior probabilities were uploaded to the RNAstructure web server[17] as a SHAPE constraints file with default parameter values used. For *RPL37A* and *RPL19B*, the structure was predicted for the longest CDS region.

Performance evaluation of BUM-HMM on the conserved regions of U3 snoRNA. Conservation scores associated with the human U3 snoRNA were taken from Rfam[36]. Highly conserved parts of the box regions, matching in sequence between the human[37] and yeast transcripts[20], were selected, with three weakly conserved nucleotides allowed in the middle of the regions (a total of 40 nucleotides). Evaluation was performed on those nucleotides with an attached posterior probability $P > 0$ (28 of those nucleotides).

Lower coverage simulation analysis. To evaluate the output consistency of the methods at lower coverage levels, we generated synthetic data sets by randomly selecting subsets of 2 million; 1 million; 100,000; 30,000; 20,000; 10,000; and 1,000 reads from the 18S DMS data set. For each subset, ten such selections were made. Files with coverage and drop-off counts were generated for each selection and passed to BUM-HMM. Consistency was evaluated with the AUC statistic between the output scores generated by each method for a given synthetic subset selection and the whole data set. For all methods, outliers were handled as described above and calling of modified nucleotides (used for the barplots of base composition) was performed at the threshold of 50% of the dynamic range of each method after having dealt with the outliers.

Code Availability. All of the code used in this study can be accessed in the following repository: https://github.com/alinaselega/BUMHMM.

29. Tollervey, D. A yeast small nuclear RNA is required for normal processing of pre-ribosomal RNA. *EMBO J.* **6**, 4169–4175 (1987).
30. Webb, S., Hector, R.D., Kudla, G. & Granneman, S. PAR-CLIP data indicate that Nrd1-Nab3-dependent transcription termination regulates expression of hundreds of protein coding genes in yeast. *Genome Biol.* **15**, R8 (2014).
31. Murdoch, D.J., Tsai, Y.-L. & Adcock, J. *P*-values are random variables. *The American Statistician* **62**, 242–245 (2008).
32. Dempster, A.P., Laird, N.M. & Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B* **39**, 1–38 (1977).
33. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **21**, 3940–3941 (2005).
34. Low, J.T. & Weeks, K.M. SHAPE-directed RNA secondary structure prediction. *Methods* **52**, 150–158 (2010).
35. Lucks, J.B. *et al.* Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc. Natl. Acad. Sci. USA* **108**, 11063–11068 (2011).
36. Nawrocki, E.P. *et al.* Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.* **43**, D130–D137 (2015).
37. Granneman, S. *et al.* Role of pre-rRNA base pairing and 80S complex formation in subnucleolar localization of the U3 snoRNP. *Mol. Cell. Biol.* **24**, 8600–8610 (2004).

doi:10.1038/nmeth.4068

4.4 BIAS CORRECTION

This section aims to provide a more detailed description of the automated empirical bias-correcting strategies employed by the BUM-HMM computational pipeline, complementing the brief corresponding sections in the Online Methods of the paper.

The aim of the bias investigation, performed during the development of the pipeline, was to identify the dependency of control LDRs, per-nucleotide variability measures of drop-off rates in the absence of a reagent, on some (if any) confounding factors. These factors would correspond to the *intrinsic* biases of the technology and would not be representative of the processes underlying chemical modification. The aim of the bias-correcting steps was to then transform the treatment-control LDRs (as well as control LDRs) to eliminate the revealed dependencies, in theory leaving only the 'real' signal.

The bias investigation was performed on the transcriptome-wide dataset, as it should be able to provide a rich (and in the ideal case of a perfect dataset, almost extensive) repertoire of scenarios necessary for detecting biases. The analysis revealed the sequence coverage level and the nucleotide sequence to be the confounding factors influencing the control LDRs.

4.4.1 *Coverage bias*

The transcriptome-wide dataset had two experimental replicates in control conditions and two replicates treated with NAI as the chemical probe (details provided in the paper). Two control datasets resulted in a single control-control comparison, and thus the null distribution contained one LDR for each nucleotide position (provided it had a positive coverage level and positive drop-off count in both control replicates).

The coverage bias was identified by plotting the control LDRs as a function of the mean coverage between the replicates at the corresponding nucleotide position (Panels **a**, **b** in Fig. 4.3 show the corresponding transcriptome-wide data for two strands). The plot demonstrated that the majority of nucleotides had a mean coverage well under 1000 in control conditions. It also showed that the variability of the drop-off rate at nucleotides was roughly inversely proportional to the mean coverage. This can be expected as the faithfulness of the signal should improve with increased read sequencing depth. Consequently, the plot revealed that most nucleotides had a large LDR (in absolute value), which means that their measured drop-off rates did not agree well between control experimental replicates due to lower coverage levels.

The aim of the coverage bias-correcting strategy was to transform the data in order to reduce the variance of control LDRs as a function of coverage. This would quan-

tify the drop-off rate variabilities at nucleotide positions across replicates in control conditions independently of the coverage levels achieved at these positions.

To make progress, let us consider an idealised model where the drop-off count k of a nucleotide covered n times can be modelled as a binomially distributed random variable with some unknown success probability p. This simplified model assumes that the event of random RT drop-off is equiprobable across the transcriptome. A more sophisticated model would require extensive biological insight into the machinery of reverse transcription and its spontaneous termination; such a model is not required for our purposes here. The standard deviation of the random variable k is thus given in Eq. 4.8.

$$k \sim B(n, p) \tag{4.7}$$

$$\sigma(k) = \sqrt{np(1-p)} \tag{4.8}$$

Correspondingly, the random variable representing the drop-off rate r at that nucleotide would have a standard deviation as given in Eq. 4.9.

$$\sigma(r) = \frac{\sigma(k)}{n} = \frac{\sqrt{p(1-p)}}{\sqrt{n}} \tag{4.9}$$

In order to determine the variance of the transformed random variable $\log(r)$, we use the following result. The moments of a function f of a random variable X can be approximated with Taylor expansion, provided that f is sufficiently differentiable and the moments of X are finite (Hogg and Craig, 1995). Specifically, the first-order approximation of the second moment of $f(X)$ can be derived as shown in Eq. 4.10. Let's denote $E[f(X)] = \mu_f$ and $E[X] = \mu_x$.

$$\begin{aligned} \operatorname{var}[f(X)] = E[(f(X) - \mu_f)^2] \approx \\ E[(f(\mu_x) + f'(\mu_x)(X - \mu_x) - f(\mu_x))^2] = \\ E[(f'(\mu_x)(X - \mu_x))^2] = \\ (f'(\mu_x))^2 \operatorname{var}[X] \end{aligned} \tag{4.10}$$

Similarly, one can approximate the covariance between functions of random variables X and Y with the expression in Eq. 4.11, denoting $E[Y] = \mu_y$.

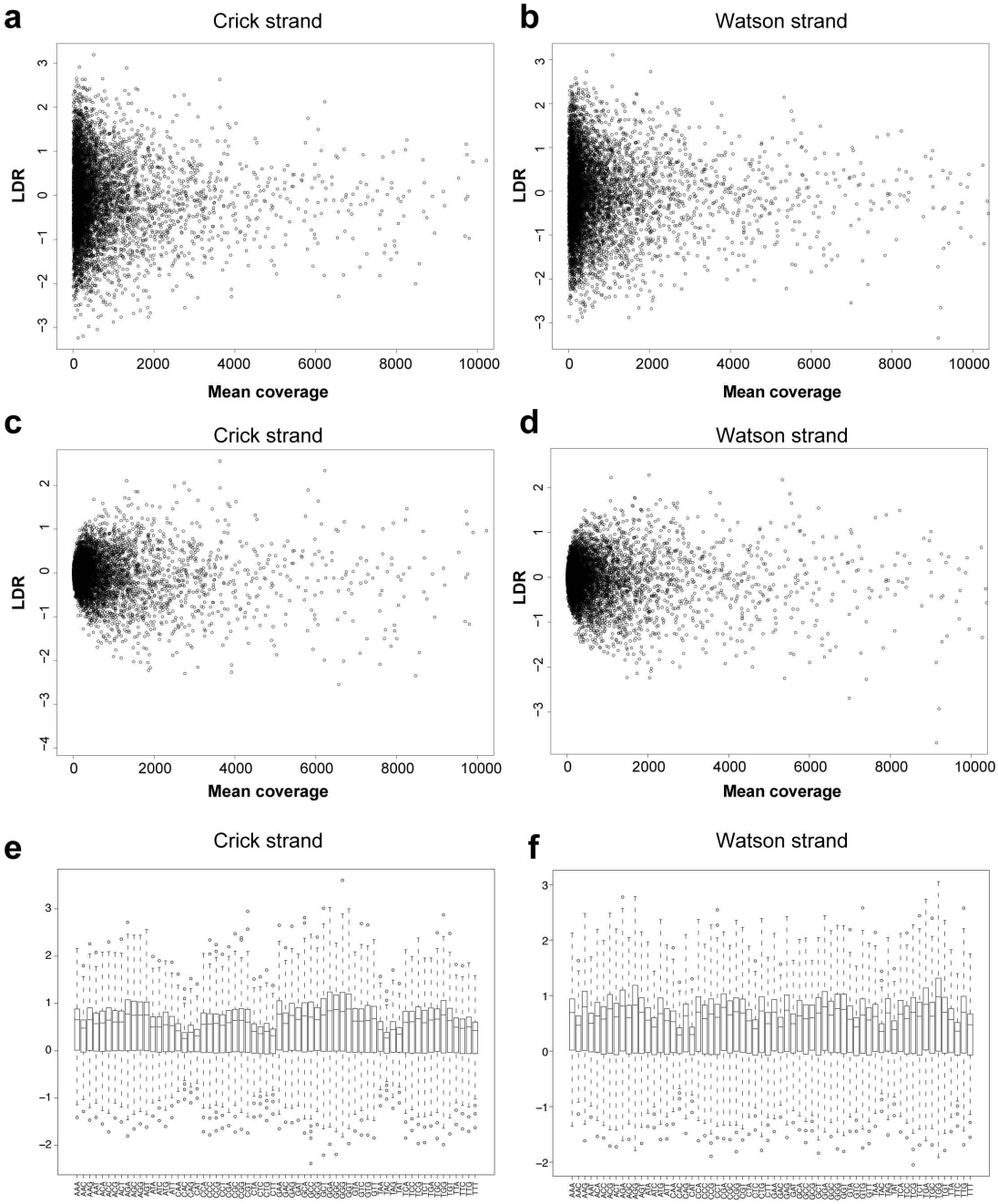


Figure 4.3: **a, b.** Presence of a coverage-dependent bias, reflected by the dependency between the LDR and the mean coverage at each nucleotide position in a pair of control replicate samples, for all such pairs, computed from the yeast transcriptome-wide data set on both strands. **c, d.** Same dependency plotted as in **a**, **b** after applying a bias-correcting strategy to the LDRs. **e, f.** Presence of a sequence-dependent bias, reflected by differing null distributions of LDRs. Each boxplot represents the null distribution (y-axis shows LDR) computed only for the nucleotide positions corresponding to a given trinucleotide pattern (indicated on the x-axis). Reproduced from Supplementary Figures of Selega et al. (2017).

$$\begin{aligned} \operatorname{cov}[f(X), f(Y)] = E[(f(X) - E[f(X)])(f(Y) - E[f(Y)])] \approx \\ f'(\mu_x) f'(\mu_y) \operatorname{cov}(X, Y) \end{aligned} \tag{4.11}$$

As the logarithm function is differentiable and the moments of r are assumed to be finite, the variance of the log-transformed drop-off rate can be approximated as given in Eq. 4.13. As the expected value of r is independent of n under these modelling assumptions (Eq. 4.12), the logarithm transformation only adds a constant factor $\frac{1}{p^2}$ to the variance of r in the expression for the approximated variance of log-transformed drop-off rate as a function of coverage n.

$$E[r] = \frac{E[k]}{n} = \frac{np}{n} = p \tag{4.12}$$

$$\operatorname{var}[\log(r)] \approx \left(\frac{1}{E[r]}\right)^2 \operatorname{var}[r] \tag{4.13}$$

Now let's examine the variance of the LDR measure at some nucleotide position for control replicates i and j:

$$\begin{aligned} \operatorname{var}[\mathrm{LDR}] = \operatorname{var}[\log(r_i) - \log(r_j)] = \\ \operatorname{var}[\log(r_i)] + \operatorname{var}[\log(r_j)] - 2\operatorname{cov}(\log(r_i), \log(r_j)) \end{aligned} \tag{4.14}$$

We can approximate the covariance between $\log(r_i)$ and $\log(r_j)$ using Eq. 4.11, yielding:

$$\operatorname{cov}(\log(r_i), \log(r_j)) \approx \frac{1}{E[r_i]E[r_j]} \operatorname{cov}(r_i, r_j) \tag{4.15}$$

The last thing to do is to consider the covariance between the drop-off rates in two experimental replicates, r_i and r_j. Even though we expect these random variables to measure a similar (ideally, identical) signal, which reflects structural properties of the transcript at that position, the *noise* associated with their measurements in *independent* experimental replicates is independent of each other. We can then factorise the covariance, showing that r_i and r_j are *conditionally* independent given the experimental procedure:

$$\begin{aligned} \operatorname{cov}(r_i, r_j) &= E[(r_i - p)(r_j - p)] = \\ &E[r_i - p]E[r_j - p] = 0 \end{aligned} \tag{4.16}$$

Putting together our results and approximating the coverage levels in control replicates with their mean n, the binomial model implies that the variance of a control LDR has an approximately inverse dependency on coverage n:

$$\begin{aligned} \operatorname{var}[LDR] &= \operatorname{var}[\log(r_i)] + \operatorname{var}[\log(r_j)] \approx \\ &\left(\frac{1}{E[r_i]}\right)^2 \operatorname{var}[r_i] + \left(\frac{1}{E[r_j]}\right)^2 \operatorname{var}[r_j] = \\ &\frac{2(1-p)}{pn} \end{aligned} \tag{4.17}$$

We would like to transform the data in a way that rids the variance of LDR of this dependency on n. We model the standard deviation of a control LDR with an unknown parameter k as shown in Eq. 4.18 and allow an intercept b to account for the variance we still observe in control LDRs for large coverage levels.

$$\sigma[LDR] = \frac{k}{\sqrt{n}} + b \tag{4.18}$$

In order to learn k and b from the data, all control LDRs were split in bins of equal coverage ranges. For each bin, the mean coverage n and the 95^{th} quantile f were computed. As the 95^{th} quantile approximately corresponds to 2 standard deviations from the mean, the bin-wise values of f with subtracted mean and mean coverage levels n were fitted to the model in Eq. 4.18 with a non-linear least squares technique, thus determining the parameters k and b.

Then, all LDR values were divided by $\frac{k}{\sqrt{n}} + b$, with their corresponding mean coverage n and the determined values of k and b. The standard deviation of these rescaled LDR values should thus be independent of n.

$$\sigma\left(\frac{LDR}{\frac{k}{\sqrt{n}} + b}\right) = \frac{\sigma(LDR)}{\frac{k}{\sqrt{n}} + b} \approx \frac{\sigma(LDR)}{2\sigma(LDR)} = \frac{1}{2} \tag{4.19}$$

Plotting the rescaled control LDRs against the corresponding mean coverage demonstrated that their variance as a function of coverage was reduced (Panels **c**, **d** in Fig. 4.3). The coverage bias-correcting strategy is aimed at reducing the influence of cov-

erage levels on the measure of drop-off rate variability employed in the BUM-HMM pipeline and demonstrates that sequence coverage is a factor that should be taken into account when modelling RNA structure probing data (and other sequencing data).

4.4.2 *Sequence bias*

To investigate whether sequence has an impact on the distribution of LDRs in control conditions, three-nucleotide patterns were separately considered. Combinations of three were chosen in order to efficiently search the transcriptome sequence for pattern occurrences and following the interpretation of looking at a nucleotide and both of its immediate neighbours. However, the pipeline software implementation features a user-defined parameter for the number of nucleotides in the pattern.

For each of the $4 \cdot 4 \cdot 4 = 64$ combinations of three nucleotides, the transcriptome sequence was searched for all places of its occurrences. Then, control LDRs were extracted from the null distribution, transformed with the coverage bias strategy, for all positions of the middle nucleotide at these occurrences, generating a null distribution specific to each pattern. Resulting null distributions demonstrated significant differences (Panels **e**, **f** in Fig. 4.3 show the corresponding transcriptome-wide data for two strands).

The sequence bias-correcting step of the pipeline stores the quantiles of each of the 64 different null distributions and computes empirical p-values for each nucleotide according to the three-nucleotide pattern it belongs to, comparing its treatment-control LDR to the corresponding null distribution.

The sequence bias correction is designed for transcriptome-wide studies and should be omitted from the analysis for short molecules as the limited number of pattern occurrences in the sequence will be unable to sufficiently inform the null distributions. The sequence bias-correcting strategy is aimed at correctly performing LDR comparisons in scenarios when the variability of the drop-off rate is affected by the local sequence, revealing it to be a factor that needs to be accounted for when handling RNA structure probing data.

4.5 PARAMETER OPTIMISATION

The BUM-HMM pipeline provides an optional automated strategy to optimise shape parameters of the Beta distribution component of the mixture model, which defines the emission model of the HMM. The strategy combines the EM algorithm, introduced in Section 3.1.4 (Chapter 3), with Newton's optimisation method.

Briefly, the algorithm iterates between computing new estimates for the shape parameters α, β (initialised at the default values $\alpha = 1, \beta = 10$) with Newton's method based on the expected complete data log-likelihood and using these estimates to generate posterior probabilities. Further details and derivations are given in the Online Methods section of the paper.

4.6 CONCLUSIONS AND OUTLOOK

The arsenal of experimental technologies for rapid and high precision characterisation of structural features within complex RNA populations has been steadily growing since the last decade, gradually moving from *in vitro* to *in vivo* and from single-transcript studies to transcriptome-wide. Chemical probing replaced nuclease cleaving, experimental designs utilising random priming provided a strategy to resolve technology bias associated with the 3′ end, generated structural maps became single-nucleotide resolution, and high-throughput sequencing allowed simultaneous probing of thousands of transcripts. However, the technological progress underlying these important innovations remained unmatched by justified statistical algorithms to analyse and interpret the mass of resulting data, which were advancing at a slower pace. Specifically, methods, existing at the beginning of this PhD project, differed significantly both in measures used in the analysis and the pre- and post-processing strategies applied to these measures.

4.6.1 *Breadth of existing analysis strategies for RNA structure probing data*

Some methods defined the measure of chemical modification using raw counts of RT stops, transformed with various normalisation strategies (Kertesz et al., 2010; Ding et al., 2014). Perhaps a legacy of experimental designs employing single-paired sequencing, the sole usage of drop-off counts is potentially vulnerable to differences in coverage, which can vary dramatically along the transcriptome. A more commonly used (Rouskin et al., 2014; Kwok et al., 2013; Siegfried et al., 2014) definition relied on a notion of *drop-off rate*, obtained by using the measured number of reads at each position and representing the *reactivity* of a nucleotide.

Normalisation strategies for either raw drop-off counts or calculated drop-off rates in general shared a common goal of defining a scale for the measurements in order to enable comparisons and interpretation. However, in practice, the strategies ranged from normalising values to the highest reactivity (Rouskin et al., 2014), scaling lowest values in treatment experiments to the values at corresponding positions in control experiments (Low and Weeks, 2010), and dividing values by the total sum of mea-

surements on the whole transcript, followed by normalisation by the length (Ding et al., 2014). Many methods employed the log-transform (Kertesz et al., 2010; Ding et al., 2014; Underwood et al., 2010), but differed in which steps of the normalisation strategy they applied it to.

Most methods applied additional post-processing to their transformed measures of reactivity. A commonly used outlier detection strategy was the 2-8% rule, which removed the highest 2% of reactivities and scaled the rest by the average reactivity of the next highest 8% (Low and Weeks, 2010). This heuristic strategy was empirically derived for SHAPE-directed structure prediction experiments and has since been used by many other methods using SHAPE and DMS (Lucks et al., 2011; Kwok et al., 2013; Ding et al., 2014). Caution should be applied when using heuristic approaches as their derived thresholds might be specific to conditions of a given experimental protocol or a lab-specific setup. Further, this and other outlier removing methods varied in their application: some methods removed outliers from raw data before normalising it (Rouskin et al., 2014), some performed it after scaling their measures of interest (Ding et al., 2014; Lucks et al., 2011; Kwok et al., 2013).

Once the reactivity measures have been transformed with the chosen normalisation strategies, they needed to be compared with their counterparts derived in control conditions in order to account for background RT stops. Some methods performed their analysis directly on treatment measurements (Rouskin et al., 2014), but most methods tried to correct for random drop-offs, even though using a simplistic log-ratio criterion (Ding et al., 2014; Kertesz et al., 2010; Underwood et al., 2010; Kwok et al., 2013). A direct subtraction of control log-reactivities from the treatment values meant that methods often had to define the difference to be 0 if more drop-off was observed in control than in treatment. Further, this scenario became impossible to tease apart from the situation when the same drop-off rate was observed in both conditions.

Finally and perhaps, most importantly, analyses relying on this statistic retained no information about the measurement variability and were thus unable to accommodate the fact that the measurements are noisy observations of a stochastic process, influenced by many factors that were both intrinsic to the technology and represented the underlying biological process.

4.6.2 *Existing model-based approaches*

One model-based approach which aimed to probabilistically decide whether drop-off in treatment conditions is significantly higher than in control is TCP^{EM} (two-channel Poisson expectation maximization) (Hector et al., 2014). The algorithm does

not model random and chemically-induced RT stops separately; instead, it models two settings for the drop-off rate at each nucleotide λ_1, λ_2 (high and low). Assuming these high and low rates are constant along the transcript, each nucleotide is assigned to one of three classes: high drop-off in both conditions (corresponding to random RT stops), low drop-off in both conditions (assigning the unmodified state), and low drop-off in control and high drop-off in treatment experiments (assigning the modified state). The position-wise class probabilities and λ_1, λ_2 are estimated with the EM algorithm and upon convergence, each nucleotide is classified based on its class probabilities.

This approach benefits from the incorporated noise model for the measurements and probabilistic classification of the nucleotide's structural state. However, its assumption of constant high and low drop-off rates λ_1, λ_2 enforces constant propensities to reacting with the probe for all structurally flexible nucleotides. Further, restricting attention to only high and low drop-off rates disregards nucleotides exhibiting intermediate number of drop-offs and implies a somewhat binary view on the structural state of nucleotides. In reality, many factors can influence the structural properties of a nucleotide, as well as multiple structural conformations can exist for a transcript *in vivo*.

Another probabilistic approach separately modelled a set of natural drop-off propensities $\{\Gamma_i\}_{i=1}^n$ and a set of relative reactivities to the probe $\{\Theta_i\}_{i=1}^n$ for all n nucleotide positions in a transcript (Aviran et al., 2011). The sets $\{\Gamma_n\}$, $\{\Theta_n\}$, and the expected number of modifications per molecule c, modelled with a Poisson distribution, were estimated with maximum likelihood optimisation, yielding probabilistic reactivities for all nucleotides in a transcript that best described the observed counts under this model.

An important assumption of this approach was that all fragments started at the 3′ end and a fragment of length k was generated in a scenario when the k-th position was the *first* modified position encountered by RT, regardless of the number of formed adducts upstream from that. This assumption rendered the model not applicable for modelling data collected in experiments employing random priming, whereby RT can randomly hybridise to the RNA. This illustrates how assumptions enforced by a parametric model can limit the generalisation abilities of a method without suitable extensions. However, this approach fulfilled a very important purpose by being the first fully probabilistic stochastic model for RNA structure probing data, treating the experimental counts as noisy measurements of underlying processes and using them to infer the quantities of interest.

4.6.3 *Contributions to the field*

At the time when this project was beginning, the existing computational methods for analysis of RNA structure probing data differed wildly in normalisation techniques and various post-processing. The employed transformations were not always statistically justified and therefore hard to compare and evaluate for correctness. Importantly, most methods deduced the structural state of a nucleotide based on a single comparison of intrinsically very noisy measurements. Even those methods that aimed to treat them as noisy measurements still only utilised information from a single experimental replicate in each condition.

The main contribution of the BUM-HMM computational analysis pipeline is its formulation in terms of multiple experimental replicate datasets, whereby the performance of the pipeline can only be improved with more biological replicates. The goal of BUM-HMM is to determine whether the drop-off in treatment conditions is significantly above the drop-off in control while using statistically justified methods. This is achieved by quantifying the variability of random drop-off and thus, making a statistical assessment for the drop-off at every position in the presence of a chemical reagent.

BUM-HMM assumes no parametric form for drop-off events, except when empirically correcting the coverage-dependent bias. The only parametric distributions used in the model are chosen according to justified statistical results describing p-value distributions, which was the main motivation to use empirical p-values as observation in an HMM instead of some measure of difference between reactivities in two conditions. The non-parametric approach makes BUM-HMM flexible to experimental variations (such as random priming) as it would work on any setup that characterises reactivity at a nucleotide position.

As part of the BUM-HMM development, bias investigative analysis was performed in order to identify confounding factors influencing the measure of choice, the log-ratio of drop-off rates between conditions (LDR). As many existing methods use the same or a very similar statistic, the results of this investigation are not only useful for improving the BUM-HMM performance, but also are of general interest to the field. Coverage levels and sequence have been shown to influence the resulting distributions of LDRs in control conditions, suggesting that this dependency should be corrected for in order to expose the signal pertinent to chemical modification rather than background processes. The BUM-HMM pipeline implements data-driven strategies to remedy these biases and perform correct comparisons.

As BUM-HMM is probabilistically formulated, its output is given by *posterior* probabilities of chemical modification in the treatment experiment for each nucleotide po-

sition given the observed data. This output is directly and statistically interpretable and circumvents the need for defining empirical scoring thresholds that might not translate between different datasets.

Even though validation of algorithms analysing RNA structure probing data is notoriously difficult due to lack of "ground truth" for most transcripts, a number of experiments demonstrated the performance of BUM-HMM. The pipeline generates a more informative output, assessing the likelihood of chemical modification for more nucleotides than other methods, while agreeing well with known structures. The generated output can be used as constraints for free energy-based algorithms to produce tighter ensembles of possible secondary structures of transcripts, possibly explained by escaping local minima. Simulated data analysis showed that the method is robust to variations in coverage, with results remaining consistent as the coverage of the dataset decreases to much lower coverage levels that have been previously recommended for effective structure probing. This result is likely to be significant for experimental planning and cost-effectiveness.

To summarise, the main features of the BUM-HMM pipeline that addressed gaps in the field of RNA structure probing data analysis are:

- explicit modelling of biological variability,
- automated data-driven strategies to address intrinsic biases,
- probabilistic and directly interpretable output,
- the choice of statistical model confirmed by data and greatly extending the sensitivity of the technology when coverage levels are lower than recommended.

Finally, the software implementing the BUM-HMM statistical modelling framework has been released on *Bioconductor*, a peer-reviewed platform for open source software for bioinformatics, in a package called `BUMHMM` (Selega et al., 2016). The implementation features user-defined parameters setting the minimum threshold for nucleotide-wise coverage and the length of a nucleotide pattern for correcting the sequence-dependent bias. The transition of the R software implementing the BUM-HMM pipeline into the form of an accepted *Bioconductor* package, including unit testing and integration with existing classes, was performed by myself. All R software for developing, testing, and evaluating the pipeline was developed by myself.

4.6.4 *Recent developments since publication*

Perhaps the most prominent recent development in the field tackling similar problems to BUM-HMM came with *PROBer*, a general statistical analysis pipeline for

sequencing-based transcriptase drop-off assays (Li et al., 2017). Published a few months after our paper, it once again recapitulated the need for well-justified statistical methods for analysing vast amounts of available data collected with many similar experimental protocols. PROBer is solving a bigger problem, proposing a general framework for modelling data from high-throughput sequencing *toeprinting* assays, which measure a signal of interest (such as the structural state of a nucleotide) via RT drop-offs and recover them by mapping the resulting cDNA *toeprints*. Such assays undoubtedly include RNA structure probing, but also assays mapping RNA-protein interactions (König et al., 2010) and detecting RNA modification such as pseudouridylation (Carlile et al., 2014) and 2'-O-methylation (Incarnato et al., 2017).

The motivation of the method is based on the fact that inference of quantities of interest heavily depends on accurately estimating the RT drop-off profiles from data, which are simultaneously influenced by RT noise, variable transcript abundance, and ambiguous read mapping. PROBer thus jointly infers transcript abundances and modification probabilities, combining together models for RNA-seq (Trapnell et al., 2010) and RNA structure probing data (Aviran et al., 2011).

The method builds upon on the previously discussed model-based approach for structure probing data (Aviran et al., 2011), whereby at each nucleotide position of a cDNA fragment encountered by the RT, there is a probability of premature synthesis termination due to modification or random drop-off, or in the case of reaching the fragment's end. However, it additionally models the generation process of a cDNA by selecting a transcript from the transcriptome based on its lengths and abundance, randomly priming or fragmenting it, and primer extending it one position at a time.

Note that commonly used experimental designs employing random priming are now supported by the PROBer model, in contrast to Aviran et al. (2011). However, this and other modelling choices lead to a very large number of model parameters, which are estimated with expectation maximisation algorithm as before (Aviran et al., 2011). In order to lighten the computational burden, PROBer assumes that transcript abundances do not change between treatment and control conditions and further imposes a parametric form on the chemical modification and RT noise profiles. This could, perhaps, again argue in favour of a non-parametric approach, such as the one we take with the BUM-HMM method, which would remain generalisable even in the case of major changes to technical procedures of experimental protocols.

PROBer's evaluation in the context of simulated data demonstrated that it could generate structural estimates of equal or better accuracy compared with other methods while requiring up to 90% less reads. This result echoes our own finding (Selega et al., 2017), whereby BUM-HMM would generate informative results even as the

coverage was vastly reduced. This once again supports our statement that rigorous statistical modelling can greatly extend the scope of experimental technologies.

As the simulated data was partly produced using PROBer itself, the authors carried out additional evaluation analyses on real data. The performances of PROBER and other methods against the "ground truths" (e.g. known secondary structures and RBP binding motifs) were evaluated in terms of a variety of different metrics (area under the curve of precision-recall (PR) and receiver operating characteristic (ROC) curves, sensitivity, positive predictive value). While PROBer mostly scored the highest on the shown metrics, the difference between its performance and performance of other methods was often very small. Also in comparison with BUM-HMM on the 18S and 25S rRNAs, the difference between both methods in sensitivity and positive predictive value was in the range of few percent.

This once again draws attention to the notorious problems with validation in the field of RNA structure probing and prediction. The aforementioned issues stemming from the lack of known structures are further complemented by the incomplete understanding of what properties of a molecule are measured in a structure probing experiment and how they relate to its structure. Even for the highly abundant transcripts with a well-defined and stable secondary structure, such as 18S, the best achieved performance only achieves 52% PR classification accuracy and 84% ROC classification accuracy on its crystallographically informed secondary structure, which is far from a perfect reconstruction. (This is keeping in mind that PR area under the curve is more suited for class imbalance problems with many more negative examples than positive examples so 52% might be a closer estimate of accuracy.) This suggests that new predicted structures for less abundant transcripts might have an even lower accuracy. Further, crystallographically obtained structures are likely to differ from the structural conformations found *in vivo* and some transcripts might even exist in multiple structural variants (e.g. caused by ribosnitches (Wan et al., 2014)). These factors provide severe limitation for the development and evaluation of computational methods for RNA structure inference.

Additionally, in PROBER's evaluation, the differences between the performances of compared methods were not equally spaced when measured with different metrics: some metrics yielded much closer results than others. It would be enlightening to understand what caused this behaviour, thereby providing a systematic comparative analysis between different methods. Otherwise, it is not immediately obvious which metric should be chosen in order to select the best performing method from many existing ones.

This goes to show that evaluation remains a very hard problem not only for modelling structure probing data but also data from other high-throughput sequencing

assays based on RT drop-offs. However, PROBer is an important milestone that both aims to provide a unified framework for many experimental protocols sharing key steps and reminds the field that accounting for noise and other confounding factors is a necessary requirement for accurate inference.

The breadth of computational analysis methods for RNA structure profiling data, available both at the beginning of this project (briefly summarised in Section 4.6.1) and still to this date, has been comprehensively outlined in a recent review (Choudhary et al., 2017a). The review pointed out common conceptual frameworks employed by most methods in estimating reactivity to chemical modification, yet also noted vast differences in carried out methodologies, confirming the motivation behind the development of BUM-HMM.

Further, the authors paid special attention to the importance of informing the analysis with datasets from multiple biological replicates. They note that information derived from replicate experiments can help to identify significant biological variation, the idea which lies at the very cornerstone of the BUM-HMM architecture. In line with this reasoning, an interactive tool for quality control of RNA structure probing data was recently proposed (Choudhary et al., 2017b). *SEQualyzer* allows one to gauge the agreement between replicates and perform exploratory analysis of high-throughput data, identifying regions with poor quality data or conversely, screening for transcripts with available high-quality information. The tool implements commonly used strategies for optimising and normalising reactivity scores (Aviran et al., 2011; Tang et al., 2015). In its aim to provide a standardised tool for data quality assessment, it mirrors the motivation behind BUM-HMM, which proposes a justified modelling strategy that could unite disparate computational methods.

Another recent effort has directly demonstrated the utility of RNA structure probing data for the problem of structure prediction. Building on the theoretically formulated probabilistic framework of Eddy (2014), which derived the likelihood-based expressions for pseudo-energy terms from a statistical model for structure probing data, Deng et al. (2016) implemented its extension to multiple structural contexts within the RNAstructure software for RNA structure prediction (Reuter and Mathews, 2010). The study further carried out theoretical investigations, assessing the information content of various reactivity values and showing that high reactivities are the major drivers of structure prediction. This result pointed out that increasing the information content of moderate reactivity values may hold the key to further improvements in structure prediction.

The authors additionally noticed the importance of upstream modelling efforts for RNA structure probing, the examples of which are Aviran et al. (2011) or the BUM-HMM (Selega et al., 2017). The reactivities estimated within a justified statisti-

cal model may better represent the underlying signals in the data and thus be informative when incorporated into the algorithms optimising free energy. The authors further point out that multiple replicates are an essential component for adequately capturing the data from a statistical analysis perspective.

In conclusion, the last two years have witnessed important work in the field of computational analysis for RNA structure probing data. These studies drew attention to such vital aspects of modelling as carefully accounting for noise and confounding factors, aspiring to propose rigorous, well-justified statistical frameworks, and thus moving towards more global and unified computational approaches. In many ways, they echoed the problems addressed by the BUM-HMM pipeline, showing that statistical analysis of high-throughput data can be of great assistance to informing experimental design and generating highly informative estimates without necessarily increasing the sequencing depth and experimental cost.

5

MODELLING THE DYNAMICS OF RNA-PROTEIN INTERACTIONS

This chapter presents a non-parametric approach for the analysis of RNA-protein interactions time-course data. Specifically, the algorithm identifies transcripts which are differentially bound by the protein of interest between conditions. The algorithm was applied to longitudinal RNA-protein binding data of high temporal resolution acquired with a novel experimental UV cross-linking methodology called χCRAC. χCRAC was applied to the mediators of transcription and degradation in the context of stress response.

The chapter begins by introducing the machinery underlying transcription and degradation in yeast in Section 5.1. It then formulates the main question that the proposed algorithm aims to answer and explains the utilised experimental design in Section 5.2. Section 5.3 provides an illustrative overview of the proposed computational method for time-series analysis. These introductory materials are intended to aid understanding of the paper which uses the proposed method to study the role of the yeast transcription termination factor Nab3 in regulating gene expression during stress (van Nues et al., 2017).

In accordance with the University of Edinburgh regulations, the paper is included in its published form in Section 5.4. Supplementary Information is included in Section B.4 of Appendix B.

The application of the proposed method for the differential cross-linking test led to one of the main findings of the paper. Pervasive changes in Nab3 binding to the transcriptome during the early stages of stress response revealed that transcription termination can provide an important control mechanism of gene expression. The analyses underlying this result are provided in the following Results sections of the paper:

> *Monitoring in vivo dynamics of protein-RNA interactions*
>
> *χCRAC provides insights into transcription kinetics*
>
> *Nab3-RNA interaction dynamics during glucose starvation.*

The introductory section provides motivation for the study and summarises its key findings and contributions, while *Discussion* gives an overview of the results and suggests directions for future work.

The remaining sections of the paper present other important results characterising the χCRAC experimental methodology and the underlying function of Nab3 during stress. Namely, the first section provides evidence for the cross-linking mapping efficiency of χCRAC and its ability to generate more biologically relevant results that are not affected by biases associated with prolonged UV irradiation. The Results section that follows the sections stated above characterises the changes in Nab3 binding site distribution under stress. The following sections made use of the anchor-away experimental system that depletes nuclear Nab3 to identify known and novel target genes regulated by Nab3 and demonstrate its effect on expression regulation of retrotransposon genes. While providing important insights into the function of the transcription termination factor, these sections are not directly relevant to this PhD thesis.

The method for differential binding analysis was developed by myself. The χCRAC time-series datasets were normalised and the analysis generating protein-specific transcript targets was performed by myself. The other authors contributed in the following manner. Sander Granneman conceived the χCRAC method. Sander Granneman, Rob van Nues, and Peter Wadsworth conceived the filtration unit used for cell harvesting. Sander Granneman and Rob van Nues designed the experiments and themselves and Erica de Leau performed the experiments. Sander Granneman performed the comparisons between the Vari-X-linker and Megatron, statistical analyses of data replicability, GO-term and clustering analyses of differentially bound transcripts, comparisons between changes in Nab3 and Pol II binding, and analyses of anchor-away data. Gabriele Schweikert performed the analyses for Nab3 transcriptomic redistribution. Guido Sanguinetti provided supervision on data modelling and the development of the differential cross-linking testing method. Sander Granneman, Guido Sanguinetti, and myself wrote the relevant Results sections specified above. Guido Sanguinetti and myself wrote the relevant Methods sections: *Data normalization* and *Testing for differential dynamic response*.

Following the paper, this chapter proceeds by providing the formal definition of the model and the associated derivations in Section 5.5. Section 5.6 provides a correction for the presented paper. The chapter then presents the analysis of the cross-linking time-series of the cytoplasmic degradation factor Xrn1 in Section 5.7, illustrating its role in gene expression regulation. This analysis was performed with the modified method for differential binding analysis that used an observation model more suitable for the data. Section 5.8 compares the results of the two method modifications by applying them to the previously discussed cross-linking datasets. The chapter concludes with Section 5.9 by summarising the main results gained from the cross-linking time-series analyses and identifying the future research directions motivated

by these findings. A dynamical model for RNA expression in stress conditions is proposed, which aims to explain transcript abundance through relative contributions of transcription and two degradation pathways (nuclear and cytoplasmic).

5.1 TRANSCRIPTION AND DEGRADATION IN YEAST

Transcription in eukaryotic cells is catalysed by the enzyme RNA polymerase (RNAP). There are multiple types of nuclear RNAP, each of which is responsible for the synthesis of distinct functional types of RNA. RNA polymerase II (Pol II) is the most studied type and it synthesises precursors of mRNA, small nuclear RNA (snRNA), and microRNA. In yeast (and human), Pol II has 12 subunits. The largest subunit is called RPB1 and it forms a part of the DNA-binding domain, necessary for catalysing transcription.

In yeast, the process of transcription can be modulated by the activity of the Nrd1-Nab3-Sen1 (NNS) protein complex. Assisted by another co-factor TRAMP (Houseley and Tollervey, 2009), the complex gives rise to the nuclear co-transcriptional degradation pathway. Proteins Nrd1 and Nab3 bind specific sequences in transcribed RNA and interact with the Pol II C-terminal domain. As a result of these interactions, transcription is terminated and the nascent transcript is degraded by the NNS-recruited exosome. Nrd1 and Nab3 have also been shown to participate in the nutrient response pathway (Webb et al., 2014).

Mature transcripts can be degraded in the cytoplasm by the cooperative effort of decapping enzymes, which remove the 5'cap, and the exoribonuclease Xrn1, which then degrades the transcript completely.

In summary, RNA expression in yeast arises from the combination of transcript production (via transcription) and decay (Fig. 5.1). The latter is implemented by various degradation pathways, two of which are mediated by the NNS complex in the nucleus and the Xrn1 nuclease in the cytoplasm.

5.2 EXPERIMENTAL DESIGN

The study, within which the proposed testing algorithm was applied, used the novel experimental method χCRAC to characterise the role of degradation in gene expression regulation. χCRAC is a UV cross-linking method for mapping RNA-protein interactions, which reduces the irradiation time to seconds. This cross-linking efficiency enabled quantitative measuring of short-lived RNA-protein interactions, which would be impossible to capture with other methods that required longer irradiation times.

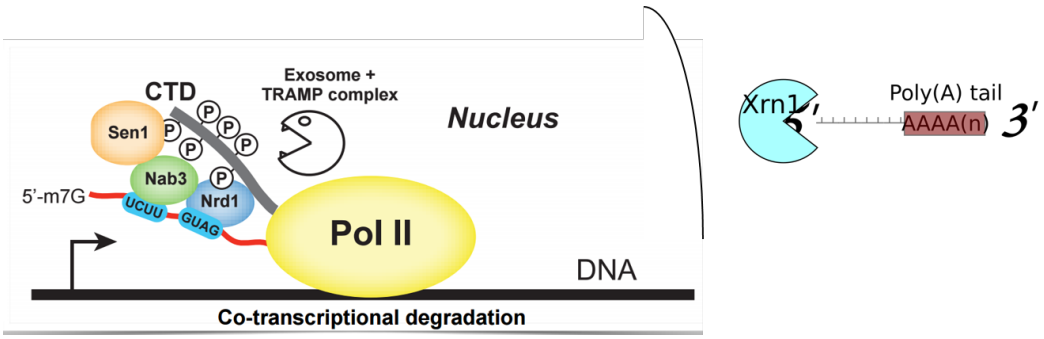


Figure 5.1: Transcription and degradation mechanisms in yeast. In the nucleus, Pol II synthesises nascent RNAs. By interacting with the C-terminal domain of the Pol II and binding sequences in the RNA, the NNS complex can terminate transcription. Outside of the nucleus, the exoribonuclease Xrn1 degrades mature transcripts. Image by S. Granneman.

The χCRAC experimental protocol was previously introduced in more detail in Chapter 2 (Section 2.8.2).

The study drew motivation from the combination of the following results about stress adaptation. Firstly, dynamical modelling of Marguerat et al. suggested that degradation might play a role in shaping gene expression at the early stages of stress response, which were previously inaccessible experimentally. Secondly, the proteins of the NNS complex involved in co-transcriptional degradation were known to be functionally relevant in nutrient stress response (Webb et al., 2014). Thirdly, glucose deprivation in yeast has been shown to lead to significant changes in mRNA levels and transcriptome-wide redistribution of the NNS complex components (Darby et al., 2012). Finally, the lack of direct experimental measurements of degradation and the general view of attributing the main importance in gene expression regulation to transcription, illustrated in Section 2.9 of Chapter 2, put together with the aforementioned results, led to the conceived experimental design.

The study aimed to investigate the role of transcription termination in adaptation response of yeast to nutrient stress. Exponentially growing *S. cerevisiae* cells were UV cross-linked to the protein of interest Nab3. Then, a portion of cells was rapidly harvested with a custom-made device (described in the Methods section of the paper) and placed in a glucose-lacking medium. These glucose-starved cells were then repeatedly UV cross-linked to Nab3 at various timepoints after stress induction (e.g. 1, 2, 4... minutes after cells were starved). The first cross-linking experiment before the transfer corresponds to the timepoint of 0 minutes. To control for cell transfer, a complementary series of control experiments was performed, which mirrored the "treatment" experiments except that the cells were transferred to the same glucose-rich medium. The experimental design is illustrated in Fig. 5.2.

Experiments following the same design were also performed for Pol II. Additionally, transcript abundance was quantified at various timepoints since cell transfer to

the glucose-starved medium. The paper featured other experiments, but they are not directly relevant to the computational method described in this section.

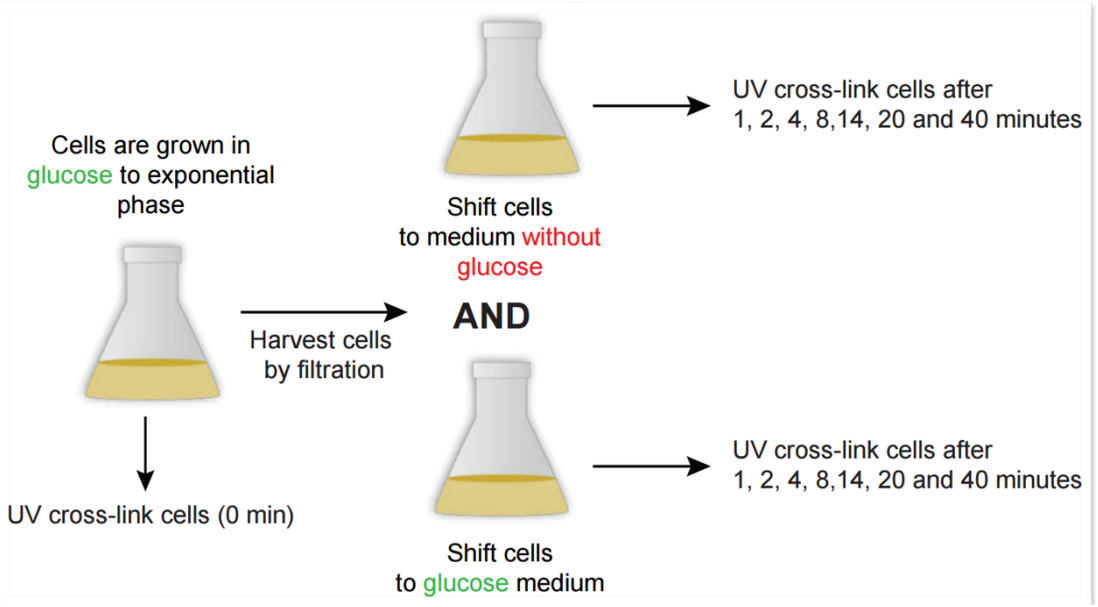


Figure 5.2: Experimental design of χCRAC experiments studying the dynamics of gene expression regulation under nutrient stress. Cells were UV cross-linked to the protein of interest at various times after transferring them to media with and without glucose. Reproduced from van Nues et al. (2017).

5.3 IDENTIFYING DIFFERENTIAL CROSS-LINKING BETWEEN CONDITIONS

The proposed algorithm aims to identify transcripts whose regulation by the protein of interest significantly changed in response to nutrient stress compared to control conditions. The algorithm is based on the Gaussian process (GP) model, introduced in Section 3.2 of Chapter 3. The GP model is a Bayesian non-parametric regression model and an infinite-dimensional generalisation of the multivariate Gaussian distribution, which specifies a prior distribution over the space of all possible functions used to model the data.

The algorithm follows a similar approach to that of Äijö et al., in that it fits GPs to time-series under different models and uses the *Bayes factor* to compute the evidence of one model over another. Specifically, two possible models are considered:

- *null hypothesis* or model M_0: all time-series, collected both in stress and control conditions, can be explained as noisy observations of a *single* underlying function that describes the dynamics of the system.

- *alternative hypothesis* or model M_1: time-series data collected in control and stress conditions arise from *two distinct* underlying functions, i.e. the dynamics of the system change between conditions.

The observed binding response $y_j(t)$ between the protein and a transcript (corresponding to the measured counts in a cross-linking experiment) is modelled as being generated by the underlying process $f_j(t)$ with the addition of Gaussian noise. Let j denote the condition (stress or control) and t denote time in minutes since the induction of stress.

$$y_j(t) = f_j(t) + \epsilon \tag{5.1}$$

$$\epsilon \sim N(0, \sigma^2) \tag{5.2}$$

Under the null hypothesis, both **c**ontrol and **s**tress time-series $\mathbf{y_c}$ and $\mathbf{y_s}$ can be explained by a single function $\mathbf{f}$ (given in vector notation in Eq. 5.3, 5.4). Under the alternative hypothesis, time-series in each condition is generated by its own function, $\mathbf{f_c}$ and $\mathbf{f_s}$, correspondingly (Eq. 5.5, 5.6). Graphically, this corresponds to fitting one function to all data, treating time-series from control and stress conditions as replicates, and fitting two different functions to control time-series and stress time-series separately (Fig. 5.3). The algorithm examines whether two separate functions explain the data better than a single one.

$$\mathbf{y} = [\mathbf{y_c}\mathbf{y_s}]^T \tag{5.3}$$

$$\mathbf{y} = \mathbf{f} + \epsilon \tag{5.4}$$

$$y_c(t) = f_c(t) + \epsilon \tag{5.5}$$

$$y_s(t) = f_s(t) + \epsilon \tag{5.6}$$

The marginal likelihood of the data is computed given each model. Under the null hypothesis model M_0, we compute the marginal likelihood jointly over the time-series in both conditions, $p(\mathbf{y_c}, \mathbf{y_s}|M_0)$. Under the alternative hypothesis model M_1, we compute the marginal likelihood of the data as the product of the marginal likelihood of the control time-series, $p(\mathbf{y_c}|M_1)$, and the marginal likelihood of the stress time-series, $p(\mathbf{y_s}|M_1)$. The Bayes factor (BF) is defined as the ratio of the marginal likelihoods of the the data given each model and evaluates the evidence of the al-

ternative model M_1 (Eq. 5.7). It is assumed that both models are equally likely. The treatment of hyperparameters is discussed in Section 5.5.4.

$$BF = \frac{p(\mathbf{y_c}|M_1)p(\mathbf{y_s}|M_1)}{p(\mathbf{y_c}, \mathbf{y_s}|M_0)} \tag{5.7}$$

We follow the suggestion based on Bayes factor interpretation (Jeffreys, 1998) that a Bayes factor greater or equal to 10 provides strong evidence in favour of the model M_1 over the model M_0. Thus, transcripts with Bayes factors exceeding this threshold are selected as targets, whose binding to Nab3 changes in response to stress.

The testing question is thereby reformulated as a model selection problem, where for selected targets, the null hypothesis, stating that the differences in the binding profiles between the RBP and each transcript can be explained solely by noise, is rejected. The algorithm is applied independently to the cross-linking data of each transcript.

The algorithm was used to identify transcripts that showed significant changes in Pol II or Nab3 cross-linking profiles after the shift to a medium lacking glucose. Results are presented in the earlier specified relevant sections of the paper, which is included below.

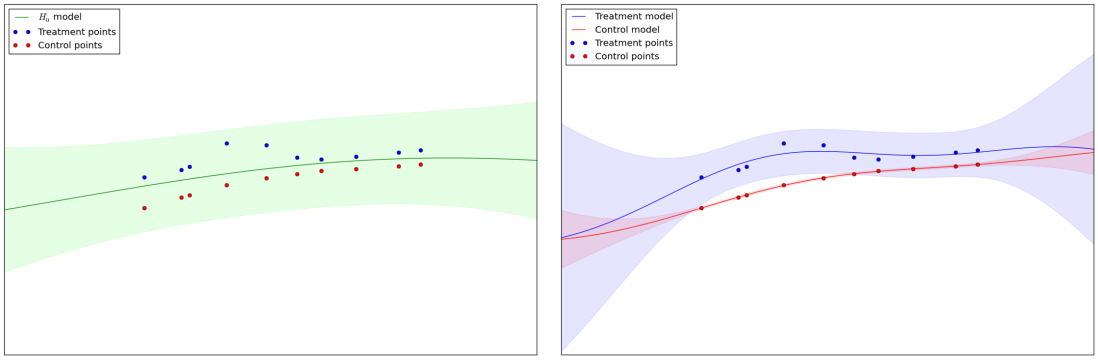


Figure 5.3: A graphical illustration of the GP-based algorithm testing for differential binding response. On the left, the toy cross-linking data corresponding to control conditions (red points) and stress conditions (blue points) are explained with a single hidden process (mean shown by the green curve, variance indicated by the green shaded region). On the right, the two time-series are explained by two separate hidden functions (indicated in red and blue, correspondingly).

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-017-00025-5 OPEN

Kinetic CRAC uncovers a role for Nab3 in determining gene expression profiles during stress

Rob van Nues[1,4], Gabriele Schweikert[2], Erica de Leau[1,5], Alina Selega[2], Andrew Langford[3], Ryan Franklin[3], Ira Iosub[1], Peter Wadsworth[3], Guido Sanguinetti[1,2] & Sander Granneman[1]

RNA-binding proteins play a key role in shaping gene expression profiles during stress, however, little is known about the dynamic nature of these interactions and how this influences the kinetics of gene expression. To address this, we developed kinetic cross-linking and analysis of cDNAs (χCRAC), an ultraviolet cross-linking method that enabled us to quantitatively measure the dynamics of protein–RNA interactions in vivo on a minute time-scale. Here, using χCRAC we measure the global RNA-binding dynamics of the yeast transcription termination factor Nab3 in response to glucose starvation. These measurements reveal rapid changes in protein–RNA interactions within 1 min following stress imposition. Changes in Nab3 binding are largely independent of alterations in transcription rate during the early stages of stress response, indicating orthogonal transcriptional control mechanisms. We also uncover a function for Nab3 in dampening expression of stress-responsive genes. χCRAC has the potential to greatly enhance our understanding of in vivo dynamics of protein–RNA interactions.

[1] Centre for Synthetic and Systems Biology (SynthSys), University of Edinburgh, Edinburgh EH9 3BF, UK. [2] School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, UK. [3] UVO3 Ltd, Unit 25 Stephenson Road, St Ives, Cambridgeshire PE27 3WJ, UK. [4] Present address: Institute of Cell Biology, University of Edinburgh, Edinburgh EH9 3FF, UK. [5] Present address: Institute for Molecular Plant Sciences, University of Edinburgh, Edinburgh EH9 3BF, UK. Gabriele Schweikert, Erica de Leau and Alina Selega contributed equally to this work. Correspondence and requests for materials should be addressed to S.G. (email: sgrannem@staffmail.ed.ac.uk)

RNA-binding proteins (RBPs) control almost all aspects of gene expression, including the stability of the RNA, its structure, the rate at which the RNA is processed, how efficiently it is translated and its subcellular localization. Not surprisingly, because of these important functions, RBPs are often found associated with many diverse genetic and somatic diseases, including muscular disorders, autoimmune diseases, and cancer[1]. RBPs also play a very important role in adapting to dynamic environments, such as those encountered by microbes when exposed to stress. Survival under stress is contingent on the ability to rapidly reprogram gene expression and, while this ability has been largely attributed to the activity of transcription factors, it is becoming increasingly clear that RBPs also play a primary role in shaping gene expression response profiles by modulating RNA processing and decay[2]. RBPs involved in RNA decay are believed to play an important role during the first few minutes of the adaptation response during which major transcriptional reprogramming events happen[3, 4]. However, direct measurement of protein–RNA interactions during these early stages has so far proved elusive. Consequently, little is known about the contribution of individual RNA decay factors during rapid rewiring of the gene expression program in response to environmental changes.

In recent years, ultraviolet (UV) cross-linking and immunoprecipitation (CLIP) followed by deep sequencing has emerged as the main technology to map protein–RNA interactions in vivo[5]. UV-irradiation is used to forge covalent bonds (cross-links) between proteins and directly bound RNAs. Proteins of interest are then purified under stringent conditions and high-throughput sequencing of the cross-linked RNA enables mapping of the interaction sites. A number of CLIP-related techniques have been developed over the years, such as CRAC (cross-linking and analysis of cDNAs), iCLIP, and PAR-CLIP (photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation)[6–8]. For CRAC the protein of interest is fused to a tandem affinity purification tag (HTP; His_6-TEV-ProtA) to enable purification of cross-linked RNAs under completely denaturing conditions[6]. Recent advancements have enhanced the efficiency of the library preparation, increased the data complexity and improved the resolution of RNA-binding site detection[7–11]. Despite such advances, current protocols are ill-suited to quantitatively measure dynamic changes in protein–RNA interactions. Using current commercially available UV-irradiation equipment, the cross-linking step can take up to 30 min to reach the desired dose (depending on organism and wavelength)[12–14]. This limitation rules out measurements of the early stress responses, which can happen on the minute time scale[3, 4]. In addition, due to prolonged UV-irradiation, cells are exposed to major additional stresses, such as DNA damage, which can confound the results and insert a bias toward RNA transcripts that are specific for the irradiation conditions.

To tackle these problems, we have improved the original CRAC protocol and developed a UV-irradiation device that cross-links proteins to RNA in vivo in seconds. These advancements enabled us to perform quantitative time-resolved in vivo measurements of direct protein–RNA interactions at 1-min time-point resolution. We refer to this method as kinetic CRAC (χCRAC).

We have applied χCRAC to glucose-deprived *Saccharomyces cerevisiae* to investigate the dynamic interactions of the RBP Nab3 during the adaptation process. Nab3 is a component of the Nrd1-Nab3-Sen1 (NNS) transcription termination complex that is involved in degradation of diverse classes of lnc-RNAs, such as cryptic unstable transcripts (CUTs), Nrd1 unterminated transcripts, and various messenger RNAs (mRNAs) and in the maturation of snoRNAs[15, 16]. Depriving yeast of glucose results in

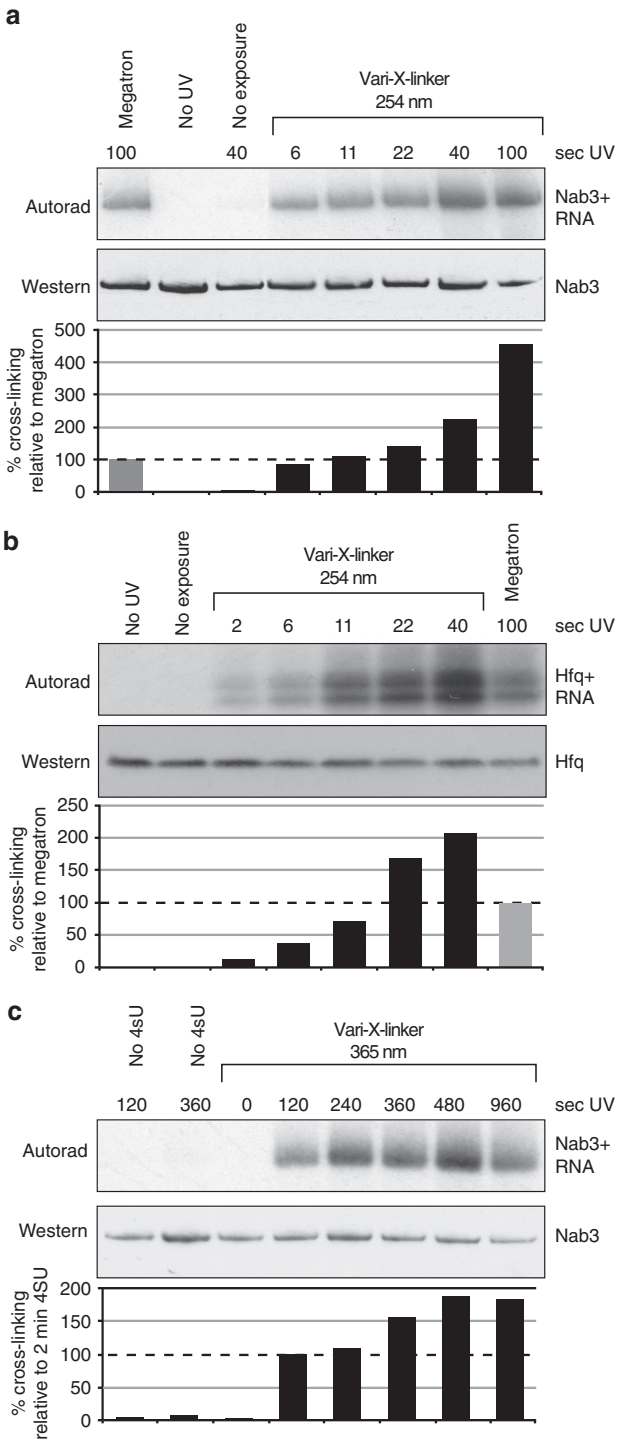


Fig. 1 TheVari-X-linker cross-links proteins to RNA in seconds. **a** The Vari-X-linker standard lamps are ~10× more efficient in cross-linking proteins to RNA in vivo compared to the Megatron unit. Cells were UV irradiated in the Megatron for 100 s. Cross-linking in the Vari-X-linker was performed at the indicated times (seconds). The western blot shows that comparable amounts of Nab3 protein was purified during the CRAC experiments. The autoradiogram shows the ^{32}P-labeled RNA cross-linked to Nab3 in each sample. These scans were used to quantify the level of cross-linking relative to the Megatron by normalizing the autoradiogram signal to the protein levels. **b** As in **a** but now monitoring the cross-linking of the *E. coli* Hfq protein. **c** Results of PAR-CLIP experiments performed using variable 365 nm UV-irradiation times, indicated in seconds. For experimental details, see the Methods section

NATURE COMMUNICATIONS | DOI: 10.1038/s41467-017-00025-5 ARTICLE

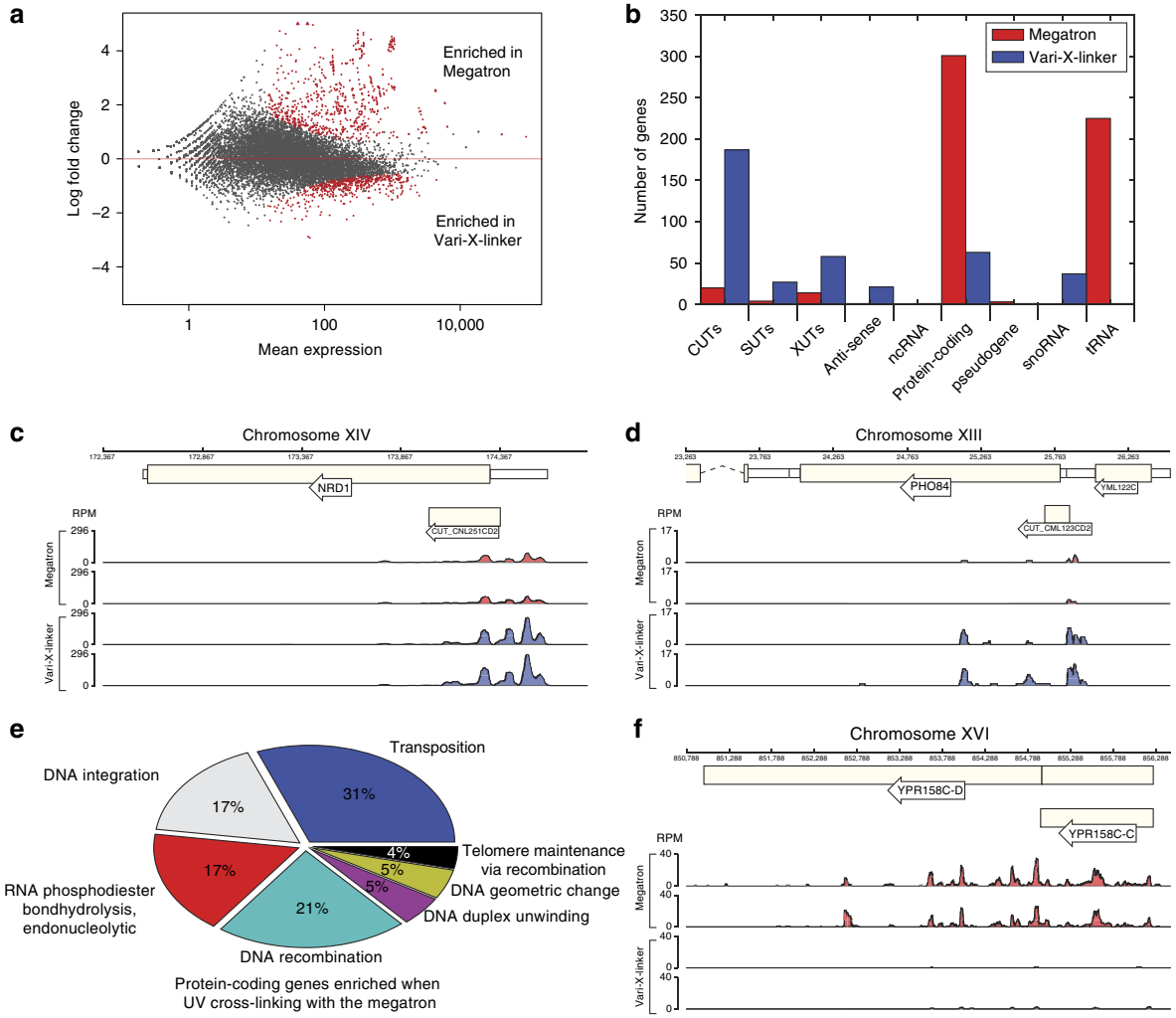


Fig. 2 TheVari-X-linker allows better detection of Nab3 binding to short-lived RNA species and reduces the induction of the DNA damage response. **a** DESeq2 differential expression analysis of Megatron (two replicates) and the Vari-X-linker Nab3 CRAC data (four replicates). The *red dots* in the plot indicate the transcripts that differentially cross-linked in data from the two different UV cross-linkers. Transcripts with positive log2-fold change values are enriched in the Megatron data, whereas transcripts with negative log2-fold change values are enriched in the Vari-X-linker data. **b** Feature analyses of differentially cross-linked transcripts. The bar plot shows the number of genes (*y*-axis) in each genomic feature (*x*-axis) that were found to be significantly enriched (adjusted *p*-value < = 0.05) in the Megatron (*red bars*) and the Vari-X-linker (*blue bars*) data. SUTs: Stable Uncharacterized Transcripts. XUTs: Xrn1 Unstable Transcripts. CUTs: Cryptic Unstable Transcripts. ncRNA: non-coding RNA. **c**,**d** Genome browser examples of CUTs that show higher Nab3 binding in the Vari-X-linker data. The *y*-axis shows reads per million (RPM). **e** Pie chart showing the significantly enriched GO-terms (FDR < 0.01) in the ~300 protein-coding transcripts enriched in the Megatron data. **f** Genome browser graph of Nab3 cross-linking to transcripts originating from retrotransposable elements *YPR158C-C* and *YPR158C-D*. The *y*-axis shows reads per million (RPM)

significant changes in mRNA levels and a transcriptome-wide redistribution of NNS components[17–20]. χCRAC accurately detected these widespread changes in Nab3 binding; importantly, the high temporal resolution enabled us to document transient changes in cross-linking of Nab3 to many transcripts, indicating a potentially pervasive importance of termination factors in the early stages of stress response. We also uncover a role for Nab3 in regulating the expression kinetics of stress-responsive genes and found that Nab3 is required for suppression of retrotransposon transcription during late stages of the glucose deprivation response. This suggests that Nab3 could play an important role in maintaining genome integrity during stress.

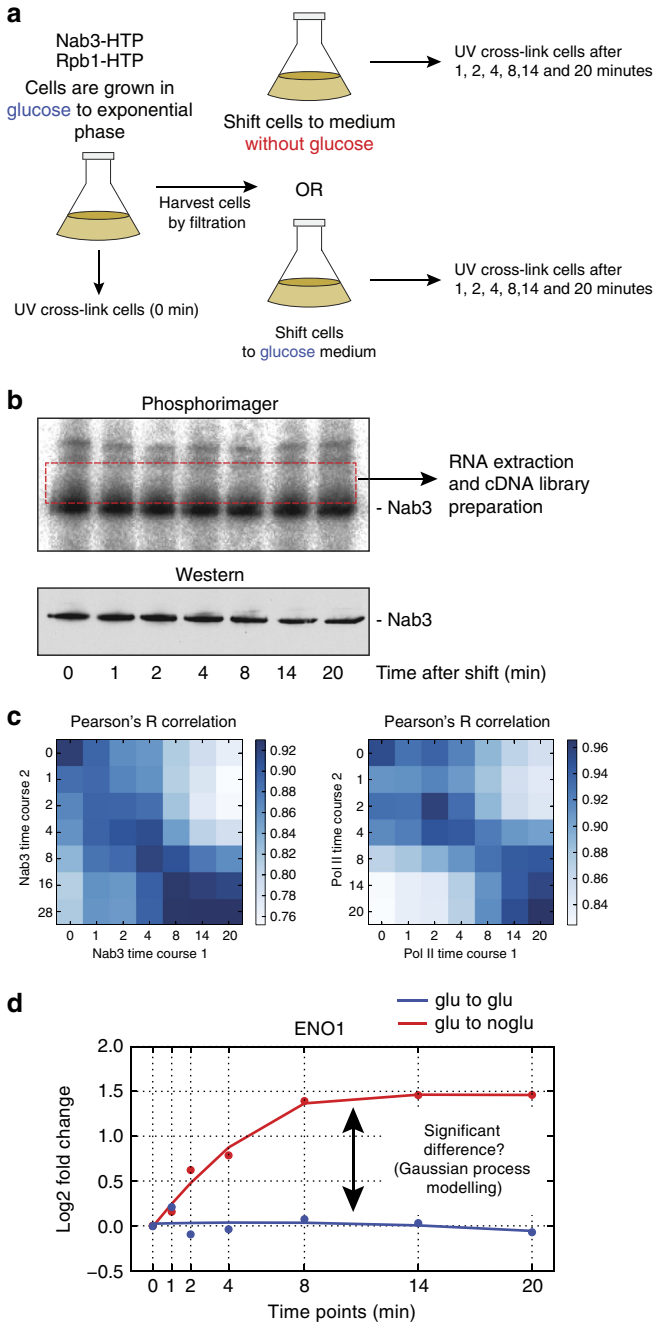
Results

Very fast protein–RNA cross-linking in vivo. To establish χCRAC, we developed a UV-irradiation apparatus (Vari-X-linker) to improve the in vivo protein–RNA cross-linking efficiency (Supplementary Fig. 1; see Methods for a more detailed description of the apparatus). To test the effectiveness of the Vari-X-linker, we performed CRAC experiments on yeast strains expressing HTP-tagged (His_6-TEV-ProtA) Nab3 that were UV-irradiated in the Vari-X-linker or in the Megatron, which (to the best of our knowledge) is currently the most efficient UV cross-linker on the market for cross-linking cell cultures[12]. Our tests with the Vari-X-linker's 254 nm lamps showed that it can cross-link yeast proteins to RNA in seconds, up to tenfold more efficiently than the Megatron (Fig. 1a). Combined with a cell filtration device that we developed, it is possible to cross-link cells and harvest 1 L of cells in ~1 min. We also tested the Vari-X-linker standard lamps on an *E. coli* strain expressing a His6-TEV-FLAG-tagged Hfq protein[21], which showed a sevenfold improvement in cross-linking time (Fig. 1b). The Vari-X-linker can also be used with 365 nm lamps (350 W) to

perform PAR-CLIP experiments, providing a high cross-linking efficiency after 2 min of UV-irradiation at considerably lower 4-thio-Uracil concentrations and shorter labeling times (see Methods for more details) (Fig. 1c).

Thus, we can perform time-resolved (PAR-)CLIP/CRAC experiments on very short time-scales, enabling the measurement of dynamic protein–RNA interactions in living cells with high temporal resolution.

Differential expression analysis of Nab3 CRAC data generated using the Megatron and the Vari-X-linker revealed significant differences between the two UV-irradiation conditions (Fig. 2a, DESeq2[22]; adjusted p-values $<= 0.05$). The Vari-X-linker data were more highly enriched for short-lived lncRNA species (Stable Uncharacterized Transcripts (SUTs), Xrn1 Unstable Transcripts (XUTs), Cryptic Unstable Transcripts (CUTs), and anti-sense transcripts) (Fig. 2b). This suggests that very short UV-irradiation times significantly improve the recovery of these unstable lncRNAs, as shown for two known CUTs that originate from the Nrd1 and Pho84 genes (Fig. 2c,d)[23, 24]. The ~300 protein-coding genes enriched in the Megatron data (Fig. 2b) were highly enriched for genes that are upregulated during DNA damage (Fig. 2e,f; FDR < 0.01). Although the steady state levels of retrotransposons did not significantly change during the 100 s UV-irradiation in the Megatron (Supplementary Fig. 2), the DESeq2 analyses revealed significantly higher cross-linking of Nab3 to these transcripts, suggesting that Nab3 actively targets these transcripts during long UV-irradiation times (Fig. 2e,f). Additionally, we also detected a significant enrichment of almost all transfer RNAs (tRNAs) in the Megatron data, which we believe reflects Nab3-dependent degradation of tRNAs that accumulate in the nucleus during the DNA damage response[25]. While these data suggest a role for the NNS complex in regulating DNA damage response and suppressing retrotransposon transcription (see below), it also illustrates that long UV-irradiation times increase the likelihood of detecting alterations in transcription that are the result of the activation of the DNA damage response.



Monitoring in vivo dynamics of protein–RNA interactions. Yeast cells deprived of glucose redistribute NNS components over the transcriptome[17]. Therefore, to test the feasibility of our χCRAC method, we measured changes in Nab3 cross-linking during glucose deprivation. Because Nab3 co-transcriptionally binds RNA, we also performed χCRAC on RNA polymerase II using a strain expressing an HTP-tagged Rpo21 subunit[26]. This enabled us to determine how well Nab3 binding correlates with changes in Pol II transcription. To measure changes in steady-state RNA levels, we performed RNA-Seq on ribosomal RNA-depleted total RNA. We devised a simple experimental set-up that would enable us to rapidly shift cells to a new medium (Fig. 3a). Cells were grown to exponential phase in glucose medium after which a fraction of the cells were harvested ($t = 0$ time-point). The rest was rapidly harvested by filtration and transferred to a flask with medium lacking glucose. After the shift,

Fig. 3 Time-resolved cross-linking analyses during glucose deprivation. **a** Outline of experimental set-up. Cells are grown in glucose medium to exponential phase. A fraction is cross-linked and harvested ($t = 0$ sample). The rest is rapidly harvested by filtration and transferred into medium lacking glucose or medium with glucose (control experiment). Subsequently, the cells were UV irradiated at the indicated times. **b** Nab3 cross-linking to RNA during glucose deprivation. Shown is a result of a typical χCRAC experiment. After resolving purified protein with RNAse digested radiolabeled cross-linked RNA on NuPAGE gels, the cross-linked RNA is detected by autoradiography. Western blotting was performed to ensure that comparable amounts of protein was recovered in each time-point. A cDNA library was subsequently prepared from RNA extracted from a single membrane slice containing RNA from all time-points. **c** Early time-points are highly correlated. The heat map shows a Pearson's R correlation analysis of each individual time-point from Pol II and Nab3 replicate χCRAC experiments. The darker the *blue color*, the higher the Pearson's correlation. Pearson correlations were calculated from log2 transformed FPKM (fragments per kilobase transcript per million reads) values. **d** A Gaussian process model was used to select genes that show significantly different cross-linking profiles between the control (glucose to glucose) and treated (glucose to no glucose) experiment. The example shows the *ENO1* Pol II cross-linking profiles from a control (*blue*) and treated (*red*) experiment. The *x*-axis shows the time-points (minutes) at which samples were taken during the time-course. All data were normalized to the 0 time-point. The *y*-axis shows the log2-fold changes in FPKMs

NATURE COMMUNICATIONS | DOI: 10.1038/s41467-017-00025-5 ARTICLE

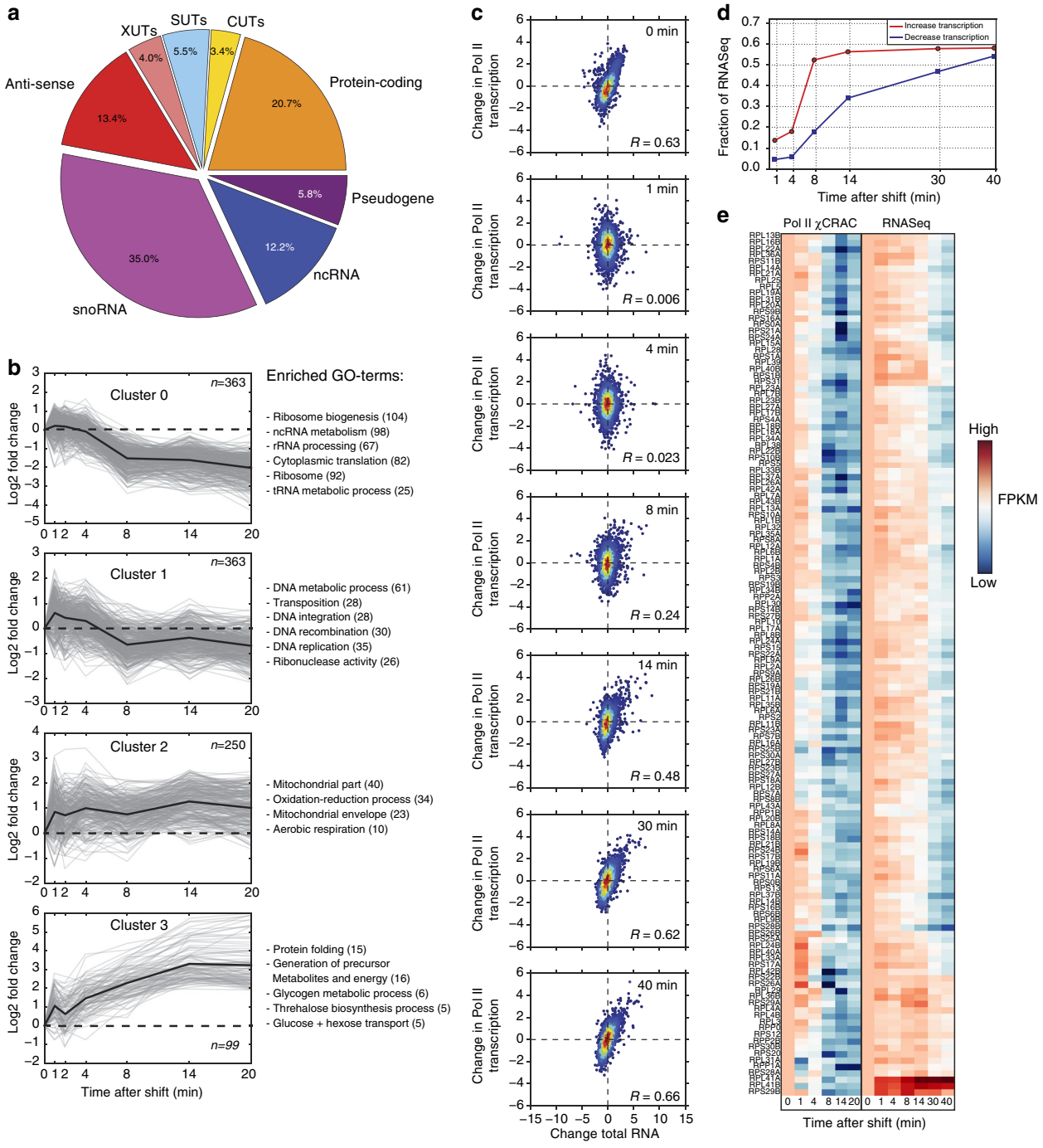


Fig. 4 RNA polymerase II χCRAC shows rapid changes in Pol II transcription during glucose deprivation. **a** The pie chart shows what percentage of each RNA class showed changes in Pol II transcription during glucose deprivation. **b** Pol II cross-linking profiles for protein-coding genes were generated by K-means clustering, performed using STEM[60]. Only mRNA profiles were selected that showed a maximum fold-change of at least 1.5 and had a mean pairwise correlation over two biological replicates of 0.7. The *gray* lines indicate profiles from individual genes. The *dark black lines* show the average profile for each cluster. Enriched gene ontology (GO) terms are indicated on the *right* side of each graph. The *y*-axis shows the log2-fold change of each time-point relative to time-point 0 (glucose sample). The *x*-axis shows the time-points that were analyzed (minutes) during the glucose starvation time-course. **c** Comparison of changes in Pol II transcription (*y*-axis) to changes in total RNA levels (*x*-axis) for several time-points. *Red* and *blue* colored dots indicate high and low data point density, respectively. To compare the data sets, we Z-normalized the fragments per kilobase transcript per million reads (FPKM) values. *R*-values indicate Pearson correlations. **d** RNA degradation is a rate limiting step during the glucose deprivation response. For each time-point, we calculated what fraction of genes that showed an increase or decrease in transcription (>=2-fold) also showed a similar change in the RNASeq data (*y*-axis). The *x*-axis shows the time-points (minutes) after induction of glucose starvation that were analyzed. The *red line* shows the results for the transcriptionally upregulated genes. The *blue line* shows the results for transcriptionally downregulated genes. **e** Transcription of most r-protein genes is shut down within 4–8 min, but total RNA levels only decrease many minutes later. The *x*-axis shows the time-points (minutes) after induction of glucose starvation that were analyzed. The heat map shows a side-by-side comparison of Pol II χCRAC and RNASeq r-protein data from a glucose starvation time-course. The higher the FPKM, the redder the color. The lower the FPKM the darker blue the color. Note that the RNASeq data has two longer time-points (30 and 40 min)

cells were cross-linked at various time-points. To control for changes in gene expression caused by the filtration process, we performed experiments where filtered cells were transferred back to glucose containing medium. To accurately quantify differences in cross-linking between time-points, we made several improvements to the original CRAC protocol[6] to reduce sequence representation biases (Supplementary Fig. 3) and to improve the preparation of complementary DNA (cDNA) libraries (see Methods). After resolving the purified protein–RNA complexes on SDS-polyacrylamide gel electrophoresis (PAGE) gels, they were transferred to nitrocellulose (Fig. 3b). Western blotting was performed to assess the efficiency of protein recovery after the nickel purification steps (Fig. 3b). RNA from each sample was ligated to 5′ adapters with unique barcodes (Supplementary Table 2). To reduce technical noise, cross-linked RNAs from all the time-points were pooled by extracting RNA from a single membrane slice containing the radioactive signal just above the main bands (Fig. 3b, red dashed rectangle) from which a single cDNA library was generated.

Statistical analyses of biological replicates revealed that χCRAC generates highly reproducible results (Fig. 3c, Supplementary Fig. 4). The early (1–4) min time-points were also highly correlated (Fig. 3c), followed by a sharp drop in correlation coefficients, suggesting that major changes in cross-linking profiles take place shortly after the first 4 min of glucose deprivation.

To identify transcripts that showed significant differential cross-linking profiles between the control (glucose to glucose) and treated (glucose to no glucose), we fitted a Gaussian process (GP) regression model to both time series[27, 28] to compute the likelihood that the control and treated originated from different profiles (see Methods; Fig. 3d).

Finally, to validate our findings we used the anchor-away system[29]. By tagging Nab3 with the FKBP12-rapamycin-binding (FRB) domain in the anchor-away strain, we were able to rapidly and effectively deplete Nab3 from the nucleus by adding rapamycin to the glucose medium 1 h before shifting the cells to medium lacking glucose (see Methods, Supplementary note 1, Supplementary Figs. 5 and 6).

χCRAC provides insights into transcription kinetics. GP analyses identified 2431 Pol II transcripts that showed significant changes in Pol II cross-linking profiles after the shift to medium lacking glucose (Bayes Factor > 10 supporting different response dynamics (see Methods)). The largest changes were observed in snoRNAs, protein-coding genes and anti-sense RNAs (Fig. 4a). To determine how well our data agrees with previous transcriptome-wide studies, we analyzed the Pol II cross-linking profiles for protein-coding genes in more detail. For the majority of protein-coding genes the changes in profiles could be summarized into 4 K-means clusters (Fig. 4b). Genes in all four clusters were enriched for distinct GO-terms. Cluster 0 is highly enriched for genes from the Ribiregulon (ribosomal proteins (r-proteins) and ribosome assembly factors) that are down-regulated during glucose deprivation. Cluster 1 is enriched for genes involved in DNA metabolism and transposition. Clusters 2 and 3 contain many stress responsive genes and genes involved in respiration that are known to be upregulated during glucose starvation[18, 20]. These results are in excellent agreement with previous studies[18, 20], demonstrating that Pol II χCRAC can accurately measure changes in gene expression at high temporal resolution.

Another reason for performing Pol II χCRAC studies, was to determine whether the data could potentially be used to develop statistical models for estimating RNA half-lives or to generate mechanistic models for RNA transcription and processing during stress. As a first step in this direction, we asked how well the Pol II mRNA χCRAC data (Fig. 4c, *y*-axis) correlated with changes in the levels of total mRNA as measured by RNASeq at each time point (Fig. 4c, *x*-axis). Only at the 0 (glucose) and late 30 and 40-min (no glucose) time-points a highly positive correlation between changes in total RNA levels and changes in Pol II transcription (Pearson's $R = 0.63$ to 0.68; p-values < 0.01) was observed. These results suggest that it takes about 30–40 min to adjust total RNA levels to mirror Pol II transcription levels. After about 14 min of glucose deprivation ~60% of the transcriptionally upregulated genes also showed a comparable increase in total RNA levels. This percentage only marginally increased at later time-points, (Fig. 4d, red line). This indicates that transcription regulation plays a dominant role during the first 8 min of the glucose deprivation response. In contrast, many genes with decreasing transcription levels only showed a similar decrease in total RNA levels during late stages of the adaptation response, suggesting that the adjustment of steady-state RNA levels for these genes is relatively slow (Fig. 4c and blue line in Fig. 4d). This especially was the case for r-protein coding genes: We observed that transcription of most r-proteins was reduced to basal level already 8 min after the shift, whereas total RNA levels of most r-protein transcripts decreased more slowly (Fig. 4e). The average mRNA half-life of r-protein coding transcripts during rapid glucose removal was estimated to be around 16 min[30], which is consistent with the slow decrease in total mRNA levels that we observed during the adaptation response. Interestingly, although both *RPL41A* and *RPL41B* were down-regulated on the transcriptional level, total mRNA level of these transcripts increased during glucose starvation (Fig. 4e).

Collectively, our data indicate that during glucose deprivation the bulk of the transcriptional changes take place within the first 8 min and that degradation of transcripts from down-regulated genes could be a rate-limiting step during the adaptation process.

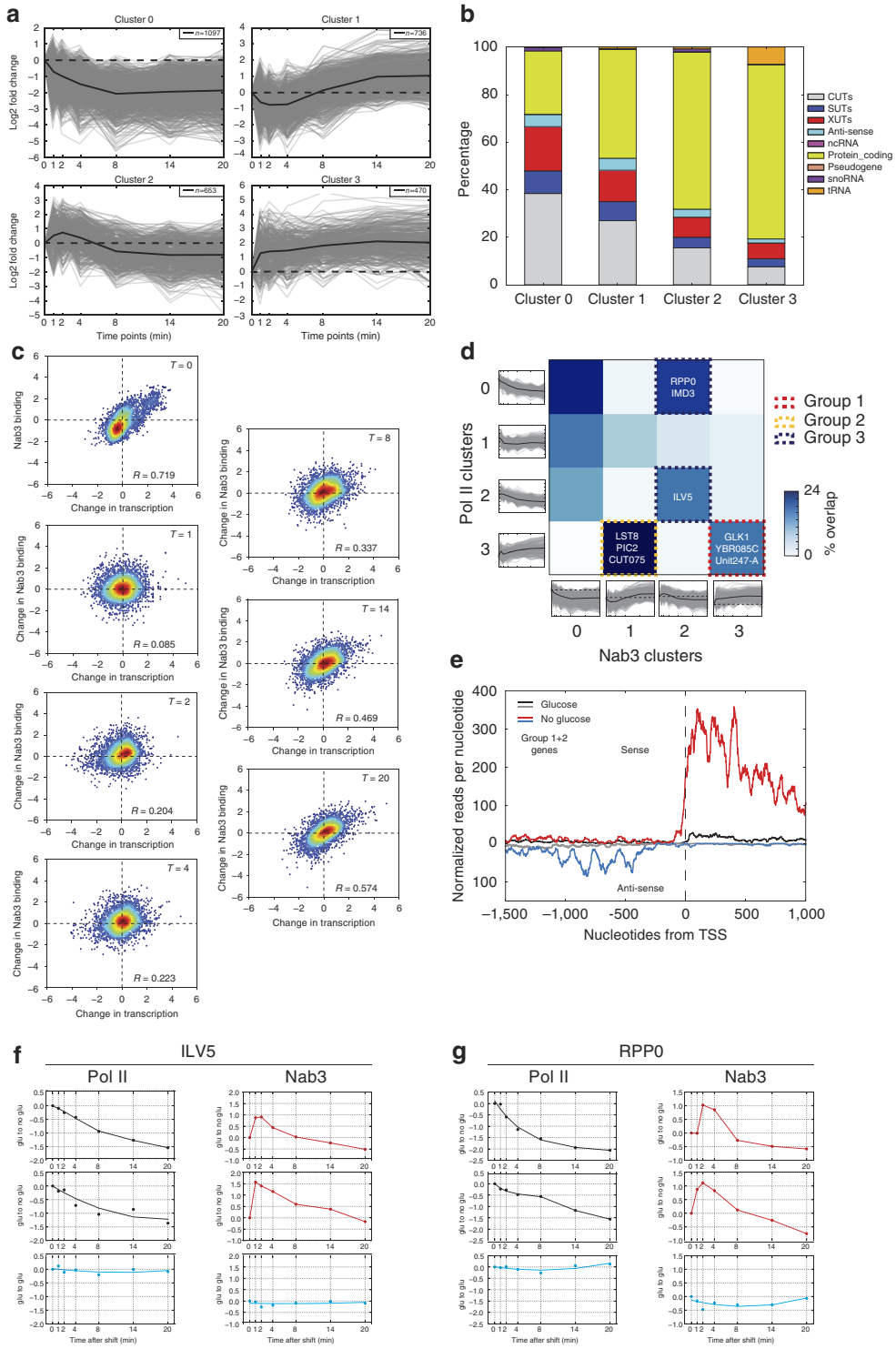
Nab3-RNA interaction dynamics during glucose starvation. We detected differential cross-linking of Nab3 to over 4100 transcripts (~37% of all features) during glucose deprivation. Using K-means clustering, we divided the Nab3 cross-linking profiles of the differentially bound transcripts into four clusters (Fig. 5a). Interestingly, both clusters 2 and 3 show a very rapid increase in Nab3 cross-linking during the first few minutes of the adaptation response. Clusters 1 and 2 also indicated transient changes in Nab3 cross-linking during the first 8 min of glucose deprivation. These data suggest that Nab3 binding very rapidly changes during glucose deprivation and is dynamic. Cluster 0 contains transcripts that generally show a decrease in Nab3 binding. About three quarters of the transcripts in this cluster are ncRNAs (XUTs, CUTs, SUTs, anti-sense RNAs, and snoRNAs), suggesting that Nab3 binding to this class of transcripts decreases during glucose starvation (Fig. 5b). Cluster 3 genes showed an increase in Nab3 binding during the time-course. These generally were underrepresented in ncRNAs, but contained the largest group of protein-coding genes and a number of tRNAs. The vast majority of the reads mapping to the 3′ end of these tRNAs contained CCA trinucleotides, suggesting that they are mature transcripts.

We next asked how well these changes in Nab3 cross-linking correlated with changes in Pol II transcription (Fig. 5c). The Nab3 and Pol II glucose χCRAC data ($t = 0$) were highly positively correlated, which is consistent with the co-transcriptional binding of Nab3 to the nascent transcript[31, 32]. However, 1 min after the shift to medium lacking glucose, the Nab3 and Pol II data

NATURE COMMUNICATIONS | DOI: 10.1038/s41467-017-00025-5 ARTICLE

decoupled, suggesting rapid changes in Nab3 binding that were independent of alterations in Pol II transcription. Thus, during early stages of glucose deprivation, binding of Nab3 to nascent transcripts might be regulated by additional factors, providing an additional level of control, which is largely orthogonal to transcriptional regulation. At later time-points, however, the correlation between the data sets improved, indicating that the cells have started to adjust to their new environment. Consistent with this, comparison of Nab3 and Pol II χCRAC profiles revealed that at later time-points Nab3-binding profiles generally followed Pol II transcription (Fig. 5d, groups 2–3). Notably, group 1 contained many genes involved in the heat-shock response, transmembrane sugar transporters and glycolysis/gluconeogenesis. Group 2 is enriched for genes involved in the oxidative stress response and starch/sucrose metabolism, hinting at a role for Nab3 in regulating the oxidative stress response.

As many yeast promoters are intrinsically bi-directional[24, 33], the induction of the group 1 and 2 genes during glucose starvation frequently resulted in the appearance of divergent anti-sense transcripts (Fig. 5e), which are also bound by Nab3. Two examples are *CUT075* (Fig. 5d, group 2) and *Unit247/CUT246* (Fig. 5d, group 1) that are readily detected upon the



induction of the heat-shock proteins *SSA2* and *HSP78* (both group 3 genes; Supplementary Fig. 7). *Unit247/CUT246* and *CUT075* are anti-sense to *POM33* and *YAP6*, respectively. Interestingly, our Nab3-depletion data suggests that Nab3 prematurely terminates these transcripts, preventing the polymerases from reaching the 5′ ends of *POM33* and *YAP6*, which could result in silencing of these genes[15].

For a number of the downregulated genes we observed a transient increase in Nab3 cross-linking and a reduction in Pol II transcription (Fig. 5d, group 3). Two examples are shown in Fig. 5f,g. Transcription of the *ILV5* and *RPP0* genes decreased almost linearly during glucose starvation (Fig. 5f,g, top two graphs), however, Nab3 cross-linking reproducibly increased about 2–3-fold during the first 4 to 8 min (Fig. 5f,g, top two graphs). These examples demonstrate that χCRAC can detect rapid changes in protein–RNA interactions at very high temporal resolution.

***In cis* changes in Nab3 binding during glucose deprivation**. Nab3-binding profiles within transcripts also changed for many genes during glucose starvation (Fig. 6). As anticipated, the majority of the Nab3 cross-linking peaks identified in the glucose data ($t = 0$) clustered near the 5′end of protein-coding genes where Nab3 is known to act[31, 32, 34] (Fig. 6a). However, 14 min after the shift to medium lacking glucose, the binding pattern of Nab3 appeared to spread more into the coding sequence (Fig. 6b). The Nab3 peak distribution plot in Fig. 6c confirmed that the Nab3-binding site distribution in the no-glucose data was significantly different from the glucose data (two sample Kolmogorov–Smirnov test; p-value $< 10^{-5}$). A striking example was the enolase (*ENO1*) gene, an enzyme involved in gluconeogenesis and glycolysis (Fig. 6d), which is strongly upregulated during glucose deprivation (Fig. 6d, top panel). In the glucose to glucose Nab3 control data (Fig. 6d, Nab3 (glu to glu)) we mainly observed three Nab3 peaks in the 5′UTR of *ENO1* that overlapped with two CUTs. In the no glucose data (Fig. 6d, Nab3 glu to noglu) the main Nab3 peaks were located further downstream in the coding sequence. This transition happens very quickly: already after the first few minutes of glucose starvation we see a change in the intensity of Nab3 binding at various sites in *ENO1* (Fig. 6d).

These results demonstrate that χCRAC is also capable of detecting rapid *in cis* changes in protein–RNA interactions.

We hypothesized that this redistribution of Nab3 in *ENO1* could be linked to the use of alternative TSSs when cells are grown in glucose. Such a mechanism is sometimes employed to regulate expression of genes encoding metabolic proteins, such as *IMD2*[35, 36], under specific conditions. To test whether expression of *ENO1* is controlled by a similar mechanism, we analyzed Cap-binding protein (Cbp1) CRAC[37], ChIP-Seq[38] and TIF-Seq[39] data to identify transcription start sites and transcript isoforms, respectively. All data sets show that transcription can initiate upstream of the *ENO1* TATA box (Supplementary Fig. 8a–c).The high ChIP signal near the CUT TSS indicates that CUT transcription is regulated by TFIID. All the available data indicate that CUT transcription is driven by a different promoter. Relative to the orthologous *ENO2*, formation of transcription initiation complexes at the *ENO1* promoter appears to be inefficient, and we speculate that this is partly the result of transcriptional interference from the upstream CUT (Supplementary Fig. 8c). Transcription of this CUT probably terminates between the *ENO1* TATA box and TSS as high levels of Nab3 cross-linking was detected in this region (Supplementary Fig. 8a). Indeed, analysis of reads containing non-encoded oligo-A tails, which are a hallmark for NNS-exosome degradation[40], revealed many degradation intermediates that overlapped with the Nab3 cross-linking sites (Fig. 6d panel Oligo-A tails), but mainly in cells grown in glucose. Nab3 depletion resulted in a ~8-fold increase in the CUT levels, confirming that Nab3 binding triggers the degradation of the CUT. However, overall the expression levels did not change during the time-course (Fig. 6e). Relative to *ACT1*, *ENO1* mRNA levels in glucose were low and Nab3-depleted cells showed a modest increase in *ENO1* quantitative reverse transcription-PCR (qRT-PCR) signal (Fig. 6f), possibly because CUT transcripts no longer terminate at the Nab3-binding sites, and less termination in the *ENO1* coding sequence (Fig. 6f, left plot). However, Nab3 depletion did not significantly affect gene expression levels of *ENO1* during glucose starvation (Fig. 6f).

We speculate that the upstream CUT helps to suppress transcription initiation at the *ENO1* transcription start site (TSS) when cells are grown in glucose (also see Discussion and Supplementary note 2).

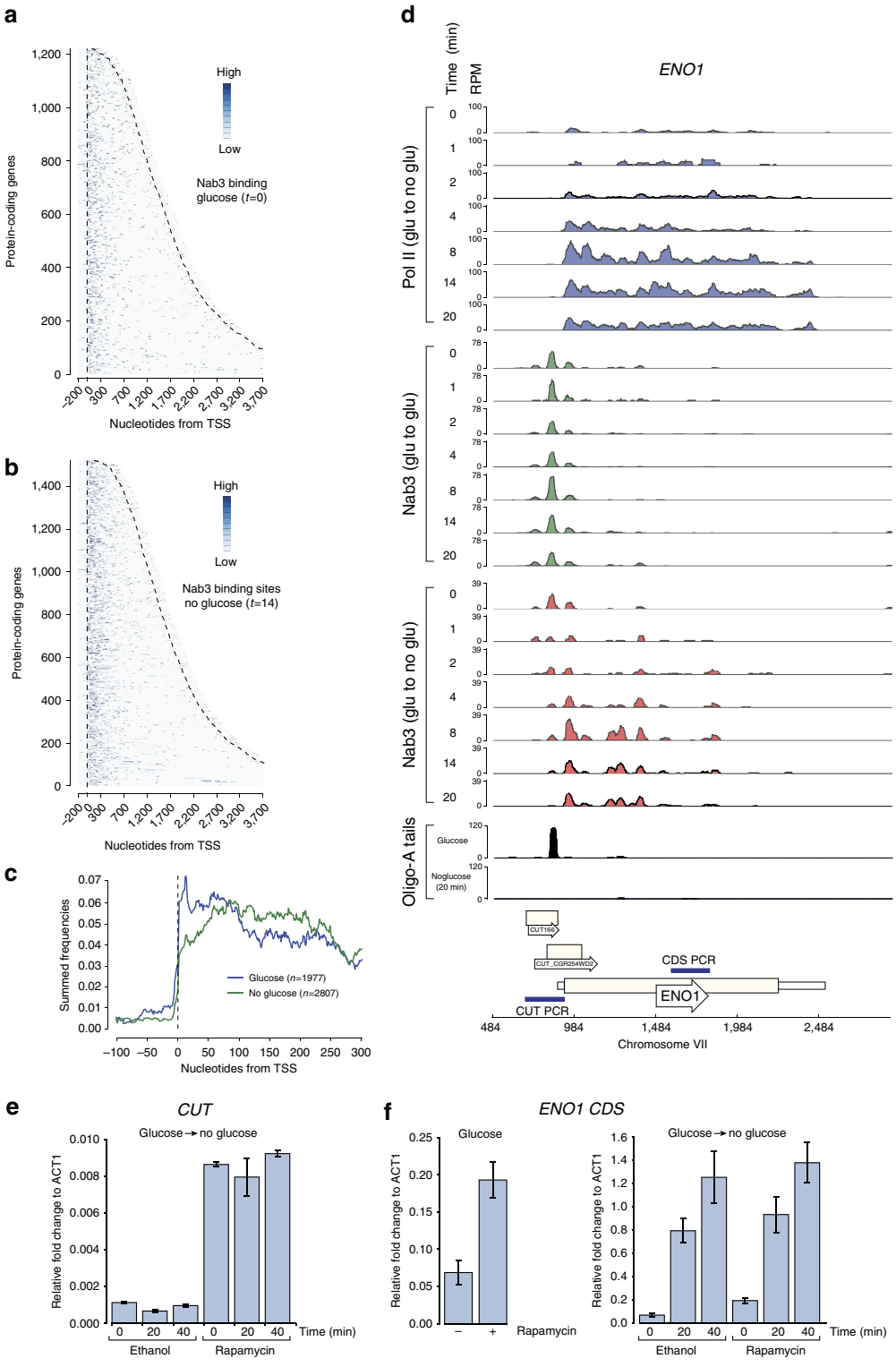
Nab3 dampens the expression of stress-responsive genes. Our data analyses (Fig. 5b,d) revealed that almost a quarter of the protein-coding transcripts are differentially bound by Nab3 during glucose starvation. We hypothesized that Nab3 could play a role in regulating the kinetics of these genes during stress. To test this model, we again employed the anchor-away method to deplete Nab3 from the nucleus and asked how this globally affected Pol II transcription. To identify Nab3-regulated protein-coding genes we calculated Pol II escape indices (EI[41]; Fig. 7a) that are a measure of changes in Pol II distribution over the gene upon rapamycin treatment. We assumed that genes that are tightly controlled by Nab3 would show high read densities near Nab3-binding sites in the promoter proximal region as a result of Pol II pausing. Upon Nab3 depletion, we expected that these “pileups” would largely dissolve, leading to an increased Pol II density over the body of the gene (see example in Supplementary Fig. 6b). Thus, genes with an EI > 1 are potentially regulated by

Fig. 5 Dynamic binding of Nab3 to many transcripts during glucose deprivation. **a** Clusters of all Nab3 cross-linking profiles generated by K-means clustering, performed using STEM[60]. Only profiles were selected that showed a maximum fold-change of at least 1.5 and had a mean pairwise correlation over two biological replicates of 0.7. The *y*-axis shows the log2-fold change of each time-point relative to time-point 0 (glucose sample). The *x*-axis shows the time-points (minutes) after the shift to medium lacking glucose that were analyzed. **b** Bar chart indicating the percentage of different RNA classes in each cluster. **c** Scatter plots comparing the Nab3 binding (*x*-axis) to Pol II transcription (*y*-axis) for the indicated time-points (minutes) after inducing glucose starvation. To compare the Nab3 and Pol II time-point 0 data we Z-normalized the FPKM values. For time-points 1 to 20 we divided the FPKM values at each time-point by the time-point zero data, which were then Z-normalized. **d** The heat map shows what fraction of the genes belonging in each Pol II K-means cluster (*y*-axis) were also found in each Nab3 cluster (*x*-axis). Dashed lines indicate groups of genes with specific Nab3 and Pol II cross-linking profiles. **e** Shown is the cumulative read density of genes belonging to groups 1 and 2 around the annotated TSS. The *black* and *gray* lines show the sense and anti-sense read densities, respectively, for the glucose data. The *red* and *blue* lines show the sense and anti-sense read densities, respectively, for the glucose-deprived cells (14 min after the shift). **f**,**g** Examples of genes (*ILV5* and *RPP0*) showing a decrease in Pol II transcription and transient cross-linking of Nab3. Biological replicates of the glucose to no glucose (*black* and *red* lines) are shown. The *blue lines* show data from glucose to glucose control experiments. *y*-axis shows fold-change relative to time-point 0

NATURE COMMUNICATIONS | DOI: 10.1038/s41467-017-00025-5 ARTICLE

Nab3. To reduce noise, we only considered highly expressed genes with an EI > = 2 that showed at least a 1.5 increase in Pol II transcription upon rapamycin treatment (see Methods for more details; Fig. 7b, top-right red quadrant).

Five of the 14 genes selected from the glucose data (Supplementary Table 4) were previously shown to be regulated by NNS, including *SER3*, *IMD3*, *NRD1,* and *URA8* and four others (*FLX1, HEM4, TRS31,* and *SEN2*) were picked up as a result of snoRNA read-through. We identified a number of new genes that are regulated by Nab3-dependent attenuation during glucose deprivation (Fig. 7b; middle and right plot). These include the aldolase *GRE3,* the mitochondrial copper transporter *PIC2* (Fig. 5 group 2), the maltose fermentation regulatory protein *MAL33* and the small GTPase *RHO5*, all of which were upregulated during glucose starvation. Comparison of the Pol II χCRAC profiles of the Nab3-depleted and control (ethanol treated) data for these genes showed a clear accumulation of Pol II around the Nab3 cross-linking sites in the ethanol-treated cells and little transcription downstream, indicative of Pol II pausing and termination (Supplementary Fig. 9b–e). In the rapamycin-



treated cells more Pol II could be detected in the body of these genes. Importantly, rapamycin treatment of the anchor-away strain expressing Nab3 without the FRB domain did not noticeably affect the Pol II transcription profiles of these genes (Supplementary Figs 6b,10), demonstrating that the observed changes in Pol II distribution is a direct result of Nab3-FRB depletion from the nucleus.

To substantiate those results, we performed qRT-PCRs on total RNA isolated from cells treated with rapamycin or the solvent (ethanol), focusing on *MAL33* and *PIC2* (Fig. 5d; group 2) (Fig. 7c). As positive controls we analyzed the levels of *IMD3* (Fig. 5d; group 3) and *NRD1*, two genes known to be regulated by Nab3-dependent attenuation[16, 23]. As negative controls we selected three genes (*LST8, GLK1,* and *YBR085C-A*; blue dots in Fig. 7b) that showed both a high increase in Nab3 binding and Pol II transcription in glucose-deprived cells (Fig. 5d groups 1 and 2), however, based on the calculated EIs were less likely to be affected by Nab3 binding. Although nuclear depletion of Nab3 resulted in only a modest increase in Pol II cross-linking of *IMD3, NRD1*, *PIC2.* and *MAL33* (generally less than twofold), total RNA levels increased quite dramatically (Fig. 7c). This suggests that in the absence of Nab3 a substantially higher number of polymerases reach the 3′ end of these genes and are terminated by the canonical cleavage and polyadenylation machinery. After 1 h of rapamycin-treatment total mRNA levels of *IMD3* and *NRD1* increased about fivefold in glucose, consistent with a role for Nab3 in terminating transcription of these genes[16, 23]. In Nab3-depleted cells, *IMD3* transcription and total RNA levels cells were generally higher throughout the time-course, suggesting that Nab3 is important for repressing *IMD3* expression in glucose and during glucose starvation (Supplementary Fig. 9b; Fig. 7c). In contrast, *PIC2* and *MAL33* mRNA levels only increased in glucose-deprived cells treated with rapamycin (Fig. 7c). Thus, *PIC2* and *MAL33* are clear examples of stress-specific Nab3 targets. Except for *LST8* (Welch's *t*-test; p-value < 0.01), nuclear depletion of Nab3 did not significantly alter total mRNA levels of the control group genes under normal or stress-conditions (Fig. 7d, Supplementary Fig. 9f).

The observation that Nab3-depletion did not affect *YBR085C-A* gene expression levels was surprising given that we detected a strong increase in Nab3 cross-linking near the 5′ end of the transcript (Fig. 8a,b) and identified hundreds of oligo-A-tailed reads in the sequencing data (Fig. 8a), strongly suggesting that the NNS terminates *YBR085C-A* transcription in glucose-deprived cells. We, therefore, engineered a strain in which the Nab3 and Nrd1 motifs in the 5′ region of *YBR085C-A* were mutated (without affecting the amino-acid sequence) (Fig. 8c). Quantitative RT-PCR analyses revealed that although the difference in mRNA levels between the mutant and the wild-type gene was always less than twofold during the 20-min time-course, the mutant was upregulated faster than the wild-type gene during glucose starvation, demonstrating a role for the NNS in regulating the kinetics of *YBR085C-A* expression (Fig. 8d). Thus, we predict that changes in transcription kinetics induced by NNS-dependent termination is more widespread than the Nab3 anchor away depletion data would suggest.

We conclude that Nab3 controls the induction kinetics as well as the maximum mRNA expression levels of stress-responsive genes during glucose deprivation.

Nab3 suppresses retrotransposon transcription during stress. We showed that prolonged UV-exposure substantially increased Nab3 cross-linking to Ty retrotransposon transcripts (Fig. 2b,e, and f). To investigate whether Nab3 regulates Ty retrotransposon expression, we measured their transcription and total RNA levels in the Nab3 anchor-away strains treated with ethanol or rapamycin. Yeast expresses five different classes of Ty retrotransposons (Ty1 to Ty5)[42]. In line with transposon-abundance, very few reads mapped to Ty5, which was, therefore, not further considered. Consistent with our initial results (Fig. 4b), in the ethanol-treated cells we observed a transient increase in Pol II cross-linking to the highly abundant Ty1 and Ty2 retrotransposons (Fig. 9a,b). Rapamycin treatment did not affect Ty1 transcription kinetics during the first 8 min, however, at later time-points Pol II transcription was significantly higher (Fig. 9a; Welch's *t*-test; p-value < 1.0×10^{-6}). This suggests that Nab3 activity is required to suppress transcription of Ty1 transposable elements primarily during late stages of the glucose adaptation response. Consistent with this idea, Nab3 cross-linking to Ty1 was highest at the late time-points (Fig. 9c). Remarkably, Nab3 appears to control transcription of Ty2 retrotransposons more tightly; In Nab3-depleted cells Ty2 transcription was significantly higher in glucose medium (Welch's *t*-test; p-value < 1.0×10^{-7}) and continued to increase during the glucose deprivation response (Fig. 9b). Nab3 cross-linking to Ty2 transcripts was dynamic, peaking at 14 min after the medium shift (Fig. 9d). These data demonstrate how χCRAC can be used to measure alterations in Pol II transcription kinetics during changes in the environment or in mutant strains.

For the less-abundant Ty3 and Ty4 retrotransposons the pattern was noisy, however, we could detect an increase in Ty3 Pol II transcription during the last three time-points, indicating

Fig. 6 Nab3 binds to different sites in protein-coding transcripts during glucose deprivation. **a** The heat map displays the distribution of Nab3-binding sites across protein-coding genes (*y*-axis) that were aligned by the TSS (*x*-axis) and sorted by length. The *dashed lines* indicate the TSS and 3′-end, respectively. Shown is the glucose data ($t = 0$). **b** Same as in **a** but now for the $t = 14$ no glucose time-point. **c** Distribution of Nab3-binding sites around the TSS. For each Nab3 protein-coding target, the distribution frequency of the binding sites was plotted around the TSS (*x*-axis). These frequencies were subsequently summed (*y*-axis) to generate this distribution plot. The *blue line* indicates the data from the glucose experiment ($t = 0$). The *green line* shows the data from the no glucose $t = 14$ time-point. **d** Genome browser images showing the results of the Pol II control χCRAC experiment (*top panel*; *blue*), Nab3 control χCRAC experiment (*green*), the Nab3 glucose to no glucose χCRAC experiment (*red*) and the total amount of oligo-A tailed reads for the *ENO1* gene. The time-points (minutes) at which samples were harvested after shifting the cells to medium lacking glucose is indicated on the *left* side of each track. **e**,**f** qRT-PCR analyses of *ENO1* and upstream *CUT* levels. Cells were grown in glucose, treated with rapamycin or ethanol for 1 h and subsequently rapidly shifted to medium lacking glucose. RNA was extracted from cells before (0) and 20, 40 min after the shift. The qRT-PCR data were normalized to the levels of *ACT1*, as both the mRNA levels and the Pol II cross-linking profiles for this gene did not significantly change during the time-course (Supplementary Fig. 9a). *ENO1* mRNA levels were quantified using RT-PCR oligonucleotides that amplify a region that is located downstream of the main Nab3 cross-linking sites (see **d**, *bottom track*). To detect the upstream CUT in **e**, we used oligonucleotides that amplify the CUT region, including the Nab3-binding sites upstream of the *ENO1* TSS. The *left bar* in **f** shows the effect of Nab3 depletion on *ENO1* mRNA levels in cells grown in glucose. The *right bar* plot in **f** shows the results for the whole time-course. Error bars indicate s.d. from three to four experimental replicates

NATURE COMMUNICATIONS | DOI: 10.1038/s41467-017-00025-5 ARTICLE

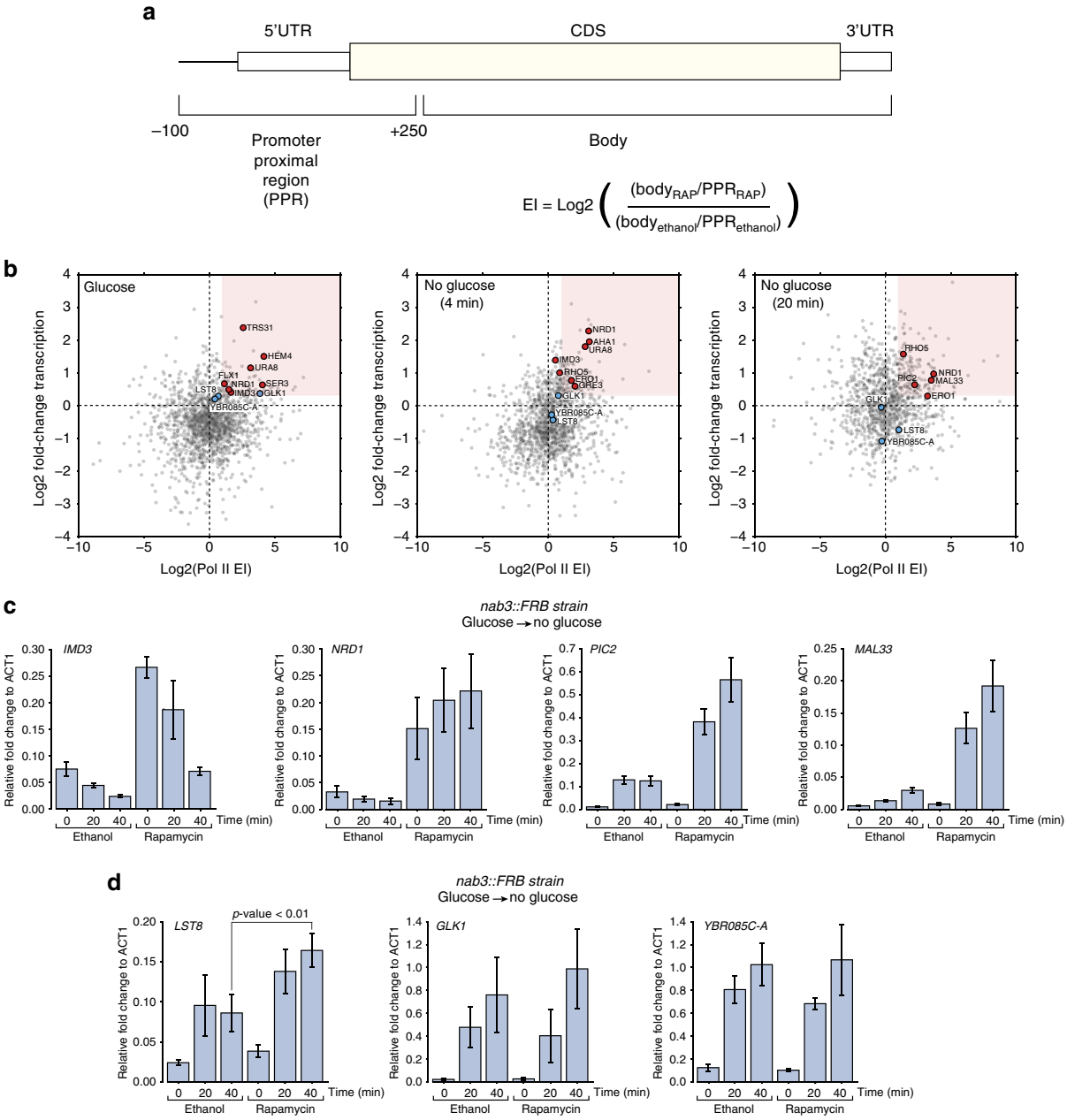


Fig. 7 Nab3 regulates the timing of expression of stress-responsive genes. **a** Schematic representation of how escape factors (EI) were calculated. For more details, see the Methods section. **b** Nab3 targets different transcripts during glucose deprivation. The scatter plot shows the comparison of escape indices (EIs) and changes in Pol II transcription for protein-coding genes before the shift (0) and 4 and 18 min after the shift to medium lacking glucose. The *red square* indicates genes that showed at least a 1.5-fold increase in transcription and an EI of at least 2. The *red dots* indicate genes that could potentially be attenuated by Nab3. The *blue dots* indicate genes that, based on the EI, are less likely to be regulated by Nab3. **c** Quantitative RT-PCR analyses of *IMD3, NRD1, PIC2,* and *MAL33* transcripts during a glucose starvation experiment. Cells were grown in glucose to exponential phase, treated with rapamycin or ethanol for 1 h and subsequently rapidly shifted to medium lacking glucose (but supplemented with rapamycin). RNA was extracted from cells before (0) and 20, 40 min after the shift to medium lacking glucose. **d** Same as in **c** but now for genes that based on the calculated EI are less likely to be regulated by Nab3. Error bars indicate s.d. from three to four experimental replicates. The *p*-value was calculated using an Welch's *t*-test on the data from the 40-min time-points

that Nab3 is also involved in regulating Ty3 expression (Supplementary Fig. 11). Quantitative RT-PCR analyses confirmed that Ty1 and Ty2 total RNA levels increased during the time-course (Fig. 9e,f), most significantly (Welch's *t*-test; p-values < 0.01) at late stages of the adaptation response.

To ascertain why Nab3 appears to more tightly regulate Ty2 transcription, we next compared the Nab3 and Pol II cross-linking profiles over Ty1 and Ty2 retrotransposon genes in the Nab3-FRB anchor-away strain grown in glucose ($t = 0$) or deprived of glucose for 20 min, with and without rapamycin treatment (Fig. 9g,h). To normalize for the differences in Ty transcript lengths, we divided the reads over an equal number of bins (Fig. 9g,h, *x*-axis). Strikingly, Nab3 cross-linked primarily to a single region in Ty2 retrotransposons (Fig. 9h, panel II), whereas Nab3 cross-linking over Ty1 transcripts was more diffuse (Fig. 9g, panel II). Although Nab3 cross-linking to Ty1 and Ty2 transcripts increased over time during glucose deprivation (Fig. 9c,d), the cross-linking pattern did not

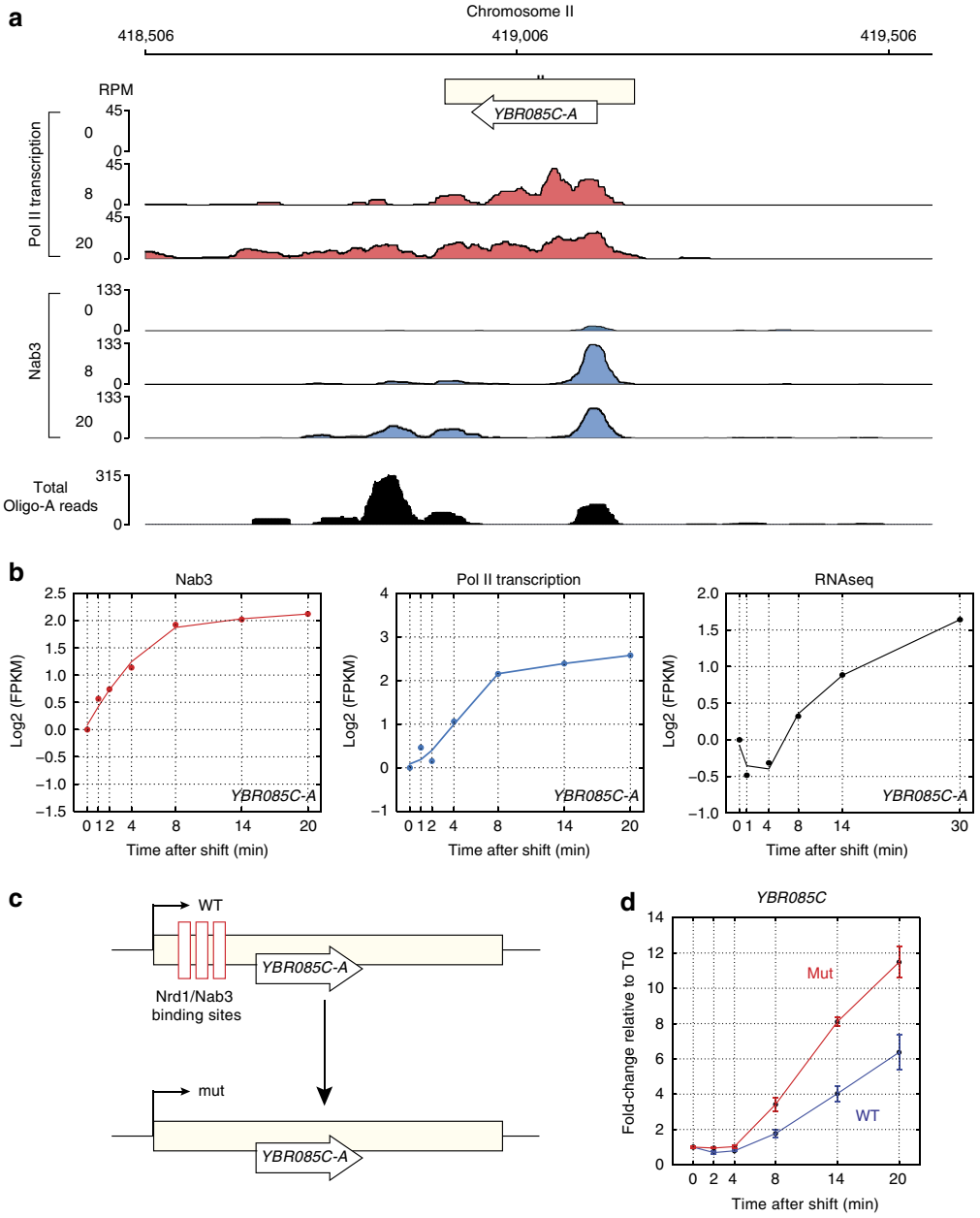


Fig. 8 Nab3 induces changes in *YBR085C-A* expression kinetics. **a** Genome browser image showing the Pol II (*red*) and Nab3 (*green*) χCRAC data for the *YBR085C-A* region from cells harvested before (0) or 8 and 20 min after the shift to medium lacking glucose. The *bottom panel* shows the total number of reads with short oligo-A tails mapped to this region. **b** The plots show the log2-transformed FPKMs (*y*-axis) for the *YBR085C-A* transcript from the Nab3 χCRAC, Pol II χCRAC, and RNASeq data. The *x*-axis indicates the time (in minutes) after the shift to medium lacking glucose. **c** Schematic representation of how the *YBR085C-A* Nrd1-Nab3 site mutant was generated. Nrd1 and Nab3 motifs that overlapped with the main Nab3 cross-linking sites in the 5′ end of *YBR085C-A* were mutated (without changing the amino-acid sequence). **d** Quantitative RT-PCR results on total RNA isolated from the wild-type (WT) and *YBR085C-A mutant* (mut) strain during a glucose deprivation time-course. The *y*-axis shows fold change in signal relative to the 0 (glucose) sample. Error bars indicate s.d. from three to four experimental replicates

dramatically change (Fig. 9g,h, compare panels II and III). Ty1 and Ty2 Pol II cross-linking profiles in cells grown in glucose were very similar, with the read densities roughly evenly distributed over the genes. Rapamycin treatment of cells in glucose only modestly increased the read density downstream of the main Nab3 peaks (Ty1 EI = 0.15; Ty2 EI = 0.37; Fig. 9g,h, compare panels IV and V). These data suggest that only a small fraction of Ty1 and Ty2 transcripts is terminated by Nab3 in glucose. However, 20 min after the shift to medium lacking glucose, Pol II cross-linking downstream of the Nab3 sites in both Ty1 and (in particular) Ty2 transcripts was substantially reduced (Fig. 9g,h, panel VI). Nab3 depletion by rapamycin treatment did not completely resolve the Pol II pileups near the Nab3 cross-linking sites (Fig. 9h, compare panels VI with VII), however, still a much higher fraction of reads was detected in downstream regions (Ty1 EI = 0.47, Ty2 EI = 1.0). These data support the notion that Nab3-dependent transcription termination is mostly active on retrotransposons during glucose starvation.

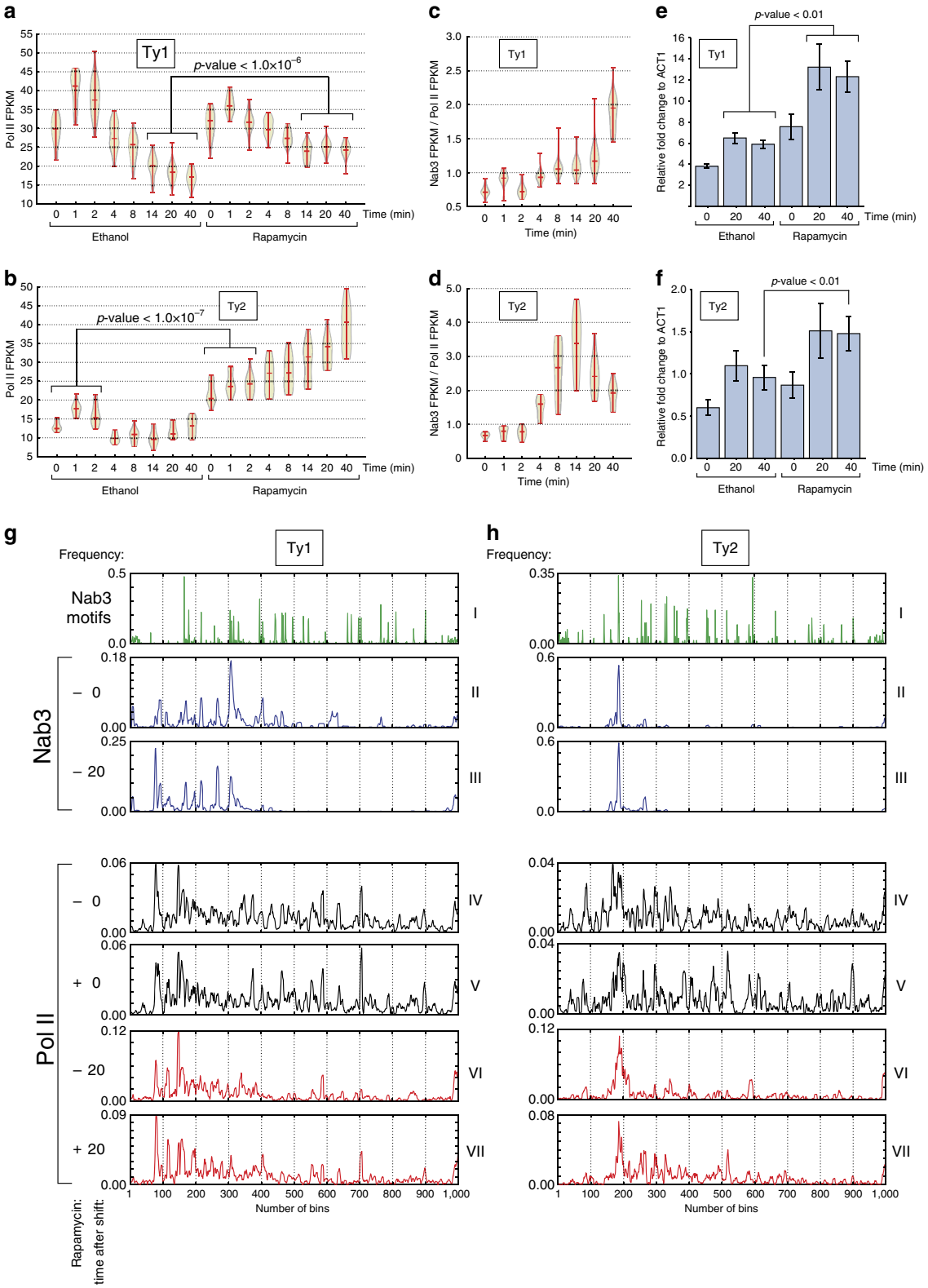
Collectively, these results demonstrate that Nab3 plays an important role in regulating the kinetics of retrotransposon gene expression during glucose deprivation.

NATURE COMMUNICATIONS | DOI: 10.1038/s41467-017-00025-5 ARTICLE

Discussion

The methodological advances underpinning the development of χCRAC enabled us to glimpse the highly dynamic reprogramming of RBP–RNA interactions in response to stress. Our data reproducibly show widespread relocation of the yeast transcriptional termination factor Nab3 within a minute of the imposition of stress. Given the central role of RBPs in all aspects of RNA life, it is plausible that other RBPs show similar dynamic behaviors. Because χCRAC is generally applicable we anticipate that its use will enable future studies to unearth novel mechanistic insights into the function of RBPs during stress conditions.

Our results provide evidence that Nab3 "dampens" the induction of several stress-responsive genes during stress. What could be the benefit of this? One possibility is that Nab3 functions as part of a negative feedback loop that reduces noise in gene expression by preventing transcription levels from overshooting



(Fig. 10a). Nab3 could also regulate the timing of expression of these stress-responsive genes, although both models are not mutually exclusive. Many of these stress-responsive genes are controlled by the same group of transcription factors. Although the benefits of expressing many stress-responsive genes simultaneously undoubtedly has advantages, in some cases it might be required that expression of certain genes is delayed (Fig. 10b) or only strongly induced when the stress signal has reached a certain amplitude. Such a dampening system would also reduce activation of gene expression due to false or noisy signals. In this respect, the role of Nab3 might be similar to that of the Set3c histone deacetylase, although the mechanism is different[43]. Deletion of Set3c induces the expression of certain genes much faster when cells are subjected to changes in carbon sources and, therefore, it was proposed that Set3c dictates the expression timing of these genes[43].

We also uncovered a role of Nab3 in regulating the expression kinetics of Ty retrotransposons, which are upregulated during a variety of stress conditions[44]. Their expression needs to be carefully controlled as recombination between Ty elements can lead to chromosomal rearrangements, which are detrimental for gene expression and genome stability[45]. In Nab3-depleted cells, transcription of Ty1 and Ty2, in particular, is significantly upregulated at later stages of the glucose deprivation response and we observed a transient increase of Nab3 binding to Ty2 retrotransposons at late stages of the adaptation response. We propose that Nab3 activity contributes to stress adaptation by rapidly shutting down transcription of retrotransposons (Fig. 10c).

In many fungi, as in mammals, retrotransposon transcription is regulated by the RNAi machinery[46, 47]. *S. cerevisiae*, however, does not have RNAi components, and it is tempting to speculate that, in view of its poor conservation, the NNS complex, together with other factors, may have taken up the function of controlling the expression of retrotransposons.

We show that χCRAC can also measure rapid *in cis* changes in protein–RNA interactions *in vivo*. We demonstrate that Nab3 binding to the *ENO1* transcript changes within the first few minutes of the glucose starvation response (Fig. 6). Follow-up analyses revealed that Nab3 binds a CUT that initiates upstream of the *ENO1* TATA box and we predict that transcription of this CUT upstream of the *ENO1* promoter helps to suppress *ENO1* expression when cells are grown in glucose. Deciphering the mechanism of regulation of *ENO1* expression is not trivial as this gene is controlled by many different transcription factors (See Supplementary note 2 and Supplementary Fig. 12). However, it is worth mentioning here that the upstream CUT initiates from an element referred to as the upstream repressor element (URS), which is a binding site for many transcription factors, such as Reb1 and Tye7 (Sgc1), and mediates ~20-fold repression of *ENO1* in glucose[48]. Interestingly, this URS has directionality as reverting this element relieves inhibition of *ENO1* expression when cells are grown on glucose[49] (Supplementary Fig. 12). This begs the question whether transcription of the CUT is also reversed in this mutant.

There are many biological scenarios where measurements of *in cis* changes, as observed in *ENO1*, could shed light on RBP–RNA interaction dynamics. A major area of interest is the assembly of large macromolecular RNP complexes, such as the ribosome and the spliceosome, which involves dynamic interactions between many proteins and RNAs. It is likely that some assembly factors contact different sites on their RNA substrates or may not occupy all of their binding sites simultaneously during the assembly process, as is the case for ribosomal proteins during ribosome assembly[50]. We envision that χCRAC could be used to perform high-resolution time-resolved analyses of dynamic changes of protein–RNA interactions in RNP particles during their assembly in vivo. Such studies would require the development of protocols to synchronize the cells in a way that the assembly could be monitored from start to finish.

Another major potential area of application is the study of the kinetics of RNA expression, which is a balance of RNA transcription and degradation. Most studies use indirect methods to estimate RNA decay rates, which can rely on Pol II mutants, metabolic labeling or drugs to inactivate transcription[51, 52]. Although these studies have provided a wealth of interesting results, the data generated by these indirect approaches are not always highly correlated[53]. In general, model-based studies assume that RNA decay can be summarized by a single mRNA half-life for each transcript, corresponding to a simple exponential decay process[52, 54], despite the complexity of RNA degradation pathways[55]. The highly dynamic behavior of the termination factor Nab3 indeed challenges this assumption. Since RNA degradation involves the activity of many nucleases, dissecting the dynamics of individual proteins is likely to be crucial for understanding how the rate of RNA decay is determined. We envision that χCRAC analyses on individual nucleases would enable us to directly measure such interactions, providing invaluable data to constrain and refine our understanding of the kinetics of gene expression.

Methods

The Vari-X-linker. The Vari-X-linker incorporates a number of new features that enhance the effectiveness of UV cross-linking. The sample is presented in a controlled 1 cm thick layer contained in a specially constructed UV transparent bag and flanked by two beds of powerful 254 nm (400–550 W) or 365 nm lamps (350 W) that were assembled on trays for easy exchange of the lamps. As far as we are aware, the Megatron is currently the fastest cross-linker available on the market for cross-linking proteins to RNA in actively growing cells[12]. Despite this, it still requires about 100 s to get good cross-linking yields with this machine (or more depending on the protein), which is not fast enough to do time-resolved analyses with minute time-point resolution. Another problem we faced with the Megatron system is that it was not trivial to cross-link cells when the lamps were at full output. As a consequence, cells would not always receive the same level of 254 nm

Fig. 9 Nab3 regulates the expression of retrotransposons during glucose deprivation. **a**,**b** Violin plot showing the Pol II FPKMs for Ty1 and Ty2 retrotransposons from the *nab3::frb rpo21-HTP* χCRAC data generated in the presence of solvent (ethanol) or rapamycin. Shown are the averaged FPKMs from two biological replicates. Time (min) indicates the number of minutes in medium lacking glucose (but supplemented with rapamycin). The *p*-values were generated using Welch's *t*-test. **c**,**d** Dynamic cross-linking of Nab3 to Ty1 and Ty2. The violin plot shows Ty1 and Ty2 FPKM distribution from a Nab3 χCRAC time-course experiment. The Nab3 χCRAC data were normalized to the average Pol II ethanol data shown in **a**,**b**. **e**,**f** Quantitative RT-PCR analysis of Ty1 and Ty2 retrotransposon transcript levels during a glucose starvation time-course. The *x*-axis shows the time (minutes) after the shift to medium lacking glucose at which samples were harvested. The *p*-values were generated using a Welch's *t*-test. **g**,**h** Plots showing the distribution of Nab3 motifs (CUUG and UCUU; panel I), Nab3 cross-linking and Pol II cross-linking to Ty1 and Ty2 transcripts. To normalize for transcript lengths, each gene was divided into 1000 bins (*x*-axis). Roman numerals indicate the results from individual experiments. The *black plots* show the Pol II profiles for cells grown in glucose in the presence or absence of rapamycin. The *red plots* show the Pol II profiles 20 min after the shift to medium lacking glucose, in the presence or absence of rapamycin. For each Ty transcript we calculated the fraction of reads that mapped to each bin for each individual transcript. These were subsequently summed (*y*-axis) to generate these profiles

NATURE COMMUNICATIONS | DOI: 10.1038/s41467-017-00025-5 ARTICLE

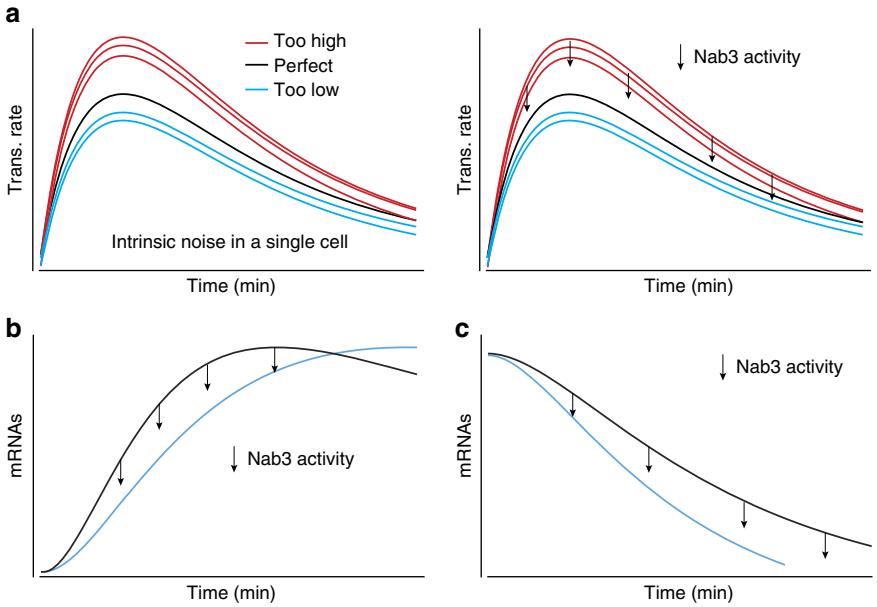


Fig. 10 Models for how Nab3 could contribute to regulating gene expression during stress. **a** Shown is a schematic representation of typical gene expression profiles observed during stress responses. The *black lines* in the plots indicate the ideal gene expression profile. The *red* and *cyan* lines indicate variability in gene expression (either too high or too low). Here, Nab3-dependent transcription termination may function to prevent transcription from over-shooting. **b**,**c** Nab3 activity could also contribute to stress adaptation by either dampening the expression of a gene **b**, which would increase the response time, or its termination activity could contribute to rapidly shutting down expression of genes that are downregulated during stress **c**

UV intensity during the 100 s of UV-irradiation, resulting in variation in cross-linking efficiencies between samples and noise in the data. It was also not possible to control the temperature inside of the unit. Although the Megatron works well for normal cross-linking studies, these issues made it very difficult to perform time-resolved analyses during short periods. To overcome this, we incorporated a shutter system to allow the lamps to be at full power with stable, repeatable output before exposing the cells. A fan cooling system was installed to minimize thermal shock to the sample. With the Vari-X-linker, the lamps can be left on throughout the experiment and cells will only be exposed when the shutters are opened (Supplementary Fig. 1; shutter release). Using a vacuum pump the cells can be quickly extracted from the UV chamber. The bag in the UV-chamber can be exchanged for a tray that allows for the cross-linking of small volumes or adherent cells in petri-dishes. Cross-linking of adherent cells could be improved by growing the cells on UV-transparent plastic. Using a vacuum pump the cells can be quickly extracted from the UV chamber. We also developed a new filtration device that enables harvesting of 1 L of cells in ~30 s. The Vari-X-linker and the filtration device can be purchased from UVO_3 (www.vari-x-link.com; sales@uvo3.co.uk).

Kinetic CRAC (χCRAC). For an eight time-points time-course, 8 L of cells were grown in synthetic medium with glucose (SD-TRP) to exponential phase (OD_{600}~0.5) at 30 °C. For time-point zero, 1 L of cells were cross-linked in the Vari-X-linker using the high-output 254 nm lamps for 12 s and then harvested by rapidly passing the cells through a 0.8 µm filter (Millipore) using a new vacuum filtration device (see above). The remaining 7 L of cells were harvested on filters and quickly resuspended in S-TRP (no glucose samples) or SD-TRP (glucose control samples) and maintained at 30 °C. For each time-point 1 L of cells were cross-linked in the Vari-X-linker and harvested by filtration as above. This yielded ~1 g of cells for each time-point.

For the PAR-CLIP experiments (Fig. 1c), for each condition 1 L of cells expressing Nab3-HTP were grown to exponential phase in SD-URA-TRP and incubated with 4-thio-Uracil for 5 min (final concentration = 20 µM). After labeling, the cells were rapidly harvested by filtration onto 0.8 µm membranes to remove the free 4-thio-Uracil and resuspended in SD-TRP before UV-irradiation at 365 nm. Removing the free 4-thio-Uracil greatly enhanced the cross-linking efficiency (data not shown).

Cells were lysed in 1 V/w of TN150 (50 mM Tris pH 8.0, 150 mM NaCl, 0.1% NP-40, 5 mM β-mercaptoethanol) and 3 V of Zirkonia beads (0.5 mm; Thistle Scientific) by vortexing the cells five times for 1 min, with a 1-min incubation on ice between each step. Three milliliter of lysis buffer was added and extracts were clarified by centrifugation (20 min at 4500×*g* and 20 min at 20,000×*g* at 4 °C). Extracts were incubated with 250 µl of equilibrated IgG Sepharose beads (GE Healthcare) for 2 h at 4 °C. Beads were washed three times 5 min with 10 ml of TN1000 (50 mM Tris pH 7.5, 0.1% Nonidet P-40, 5 mM β-mercaptoethanol, 1 M NaCl) and three times 5 min with TN150 (50 mM Tris pH 7.5, 0.1% Nonidet P-40, 5 mM β-mercaptoethanol, 150 mMNaCl).

For the Nab3 anchor-away nuclear depletion experiments[29], cells were incubated with 1 µg/ml rapamycin (Sigma) for 1 h before the cells were shifted to medium lacking glucose (but supplemented with 1 µg/µl rapamycin).

For the Pol II (Rpo21-HTP) χCRAC experiments, cells were lysed in 1 V/w TMn150 (50 mM Tris-HCl pH 8.0, 10 mM $MnCl_2$, 0.1% NP-40, 5 mM β-mercaptoethanol, 150 mM NaCl, Roche Midi protease inhibitors; 1 ml per gram of cells). Subsequently, 1 V/w of TMn150 was added containing 1U/ml of RQ1 RNase-free DNAse (Promega) and the suspension was incubated for half an hour on ice to degrade the chromatin.

For the Hfq cross-linking test (Fig. 1b), 0.5 L of bacteria were grown in Luria-Bertani medium (LB) to an OD_{600} of 0.4 and cross-linked in the Megatron and Vari-X-linker for the indicated times. To purify the *E. coli* Hfq-HTF, one gram of cells was lysed as described above. Extracts were incubated with 35 µl of anti-Flag magnetic beads (Sigma) for 2 h at 4 °C. Beads were washed three times 10 min with TN1000 and rinsed three times with TN150.

For the TEV cleavage step, beads (IgG or Flag) were resuspended in 600 µl of TN150 and incubated for 2 h with 10 µg of home-made GST-TEV protease at 18 °C. The TEV eluates were subsequently incubated with 0.1 unit of RNace-IT (Agilent) for 5 min at 37 °C after which 0.4 g of guanidine HCl (Sigma) was added to the TEV eluates to inactivate the RNAses. NaCl and Imidazole was added to a final concentration of 300 and 10 mM, respectively and the samples were incubated overnight with 50 µl of Nickel agarose beads (Qiagen) at 4 °C. Beads were transferred to a Snap Cap columns (Pierce), washed three times with 500 µl wash buffer I (50 mM Tris-pH 7.5, 6 M guanidium-HCl, 0.1% Nonidet P-40, 5 mM β-mercaptoethanol, 300 mM NaCl, 10 mM Imidazole) and three times with 1×PNK buffer (50 mM Tris pH 7.5, 0.1% Nonidet P-40, 5 mM β-mercaptoethanol, 10 mM $MgCl_2$). Beads were subsequently incubated with 80 µl of 1xPNK buffer containing eight units of TSAP alkaline phosphatase (Promega) and 80 units of recombinant RNasin (Promega) for 1 h at 37 °C. After one 500 µl wash with wash buffer I and three 500 µl washes with 1xPNK buffer, the beads were resuspended in 80 µl of 3′ linker ligation mix (1xPNK buffer, App-PE 3′ adapter (see Supplementary Table 2; 0.6 µM final concentration), 10% PEG8000, 30 units of T4 RNA ligase 2 truncated K227Q (NEB), 60 units RNAsin (Promega)). The samples were incubated at 25 °C for 4–6 h. Following one 500 µl wash buffer I wash and three 1xPNK buffer washes, the beads were incubated with 60 µl of 5′end labeling mix (1xPNK buffer, 30µCi ^{32}P-γATP (Perkin Elmer) and 30 units of T4 polynucleotide kinase (NEB)) for 40 min at 37 °C. ATP (Roche) was added to 1 mM final concentration, followed by another 20-min incubation at 37 °C. Beads were subsequently washed three times with 500 µl of wash buffer I and three times with 500 µl of 1xPNK buffer and incubated with 80 µl of 5′ linker ligation mix (1xPNK buffer, 10 mM ATP, 80 units RNAsin (Promega) 40 units of T4 RNA ligase 1 (NEB) and 5′ adapter (1.25 µM final concentration; see Supplementary Table 2) overnight at 16 °C. Beads were subsequently washed three times with 500 µl wash buffer I and three times with 500 µl of wash buffer II (50 mM Tris pH 7.5, 0.1% Nonidet P-40, 5 mM β-mercaptoethanol, 50 mM NaCl, 10 mM Imidazole). Proteins were eluted from the nickel beads using wash buffer II

containing 250 mM Imidazole, TCA precipitated (20% final concentration) and resolved on 1 mm thick 4–12% NuPAGE gels (Thermo Fisher Scientific), transferred to nitrocellulose membranes and visualized by autoradiography. Bands corresponding to the size of the protein of interest, including a region ~ 1 cm above the band, were cut from the nitrocellulose membrane and pooled in a single 2 ml tube. Radiolabeled RNA was extracted by incubating the membrane slices with 200 µg of proteinase K in 800 µl of wash buffer II containing 1% SDS and 5 mM EDTA. The solution was transferred to a new tube and the RNA was subsequently phenol-chloroform-extracted and ethanol-precipitated. Reverse transcription with SuperScript III was performed as per the manufacturer's procedures (Thermo Fisher Scientific) using the reverse transcription primer listed in Supplementary Table 2. The cDNAs were purified using the Zymo DNA Clean & Concentrator 5 kit and eluted into a final volume of 10 µl. Five microliter of cDNA was PCR amplified using Pfu polymerase (Promega) for 20–24 cycles (95 °C 30 s, 52 °C 30 s and 72 °C 1 min) using PCR oligonucleotides listed in Supplementary Table 2. PCR products were resolved on 2% Metaphor agarose gels (Lonza) and 160–300 bp fragments were gel purified using the miniElute kit (Qiagen) according to the manufacturer's procedures. Paired-end sequencing (50 bp) was performed by Edinburgh Genomics using the IlluminaHiSeq 2500 and 4000 platforms. This improved the detection of high-confidence cross-linking induced mutations[6]. Following sequencing, samples were demultiplexed using the 5′ adapter barcode sequences and collapsed reads were mapped to the yeast genome. To improve T4 RNA ligation efficiencies, we added random nucleotides to adapter termini that ligate to the RNA (Supplementary Table 2). We found that there is a significant preference for specific donor–acceptor nucleotide combinations for both 5′ and 3′ T4 RNA ligase reactions (Supplementary Fig. 3).

The uncropped images for Figs. 1 and 3b are provided in Supplementary Figs. 13 and 14.

Procedures used for western, northern, qRT-PCR and a description of the yeast strains and media can be found in the Supplementary Methods.

Processing of raw sequencing data. Sequencing was performed on IlluminaHiSeq 2500 and HiSeq 4000 machine by our Edinburgh Genomics facility. The complete pipeline for the processing of paired-end kinetic CRAC data is available on https://bitbucket.org/sgrann/kinetic_crac_pipeline. The entire pipeline can be run using a single script (CRAC_pipeline_PE.py) that divides the tasks over multiple processors. The pipeline performs the following steps: demultiplexing of raw fastq files by pyBarcodeFilter.py version 2.3.3 from the pyCRAC tool suite[56] (version 1.2.2.6). Flexbar[57] then trims the reads and removes 3′ adapters sequences (Supplementary Table 2). Reads are then collapsed using random barcode information provided in the in-read barcodes using the pyCRAC tool pyFastqDupicateRemover.py (see Supplementary Table 2 for adapter sequences). Reads are then aligned to the reference sequence (yeast genome R64 in our case) using novoalign (www.novocraft.com) version 2.0.7 and those that mapped to multiple genomic regions were randomly distributed over each possible location. PyReadCounters then makes read count and fragments per kilobase transcript per million reads (FPKM) tables for each annotated genomic features. Only genes for which cross-linking could be detected in all time-points were considered. Genomic feature files were obtained from ENSEMBL (version R64-1-1.75). Coordinates for anti-sense transcripts, CUTs, XUTs, SUTs, and retrotransposons[24, 33, 58, 59] were obtained from the Saccharomyces Genome Database (sgd; yeastgenome.org).

Identification of Nab3-binding sites and oligo-A reads. PyCalculateFDRs.py was used to find significantly enriched Nab3-binding peaks using default settings. Only peaks with at least five reads were considered and the minimum width of the peak interval was set to 20 nucleotides. Oligo-A reads were identified using blast and in-house perl and python scripts. PyBinCollector.py was used to generated the Nab3 and Pol II cross-linking distribution figures.

K-means clustering. K-means clustering of cross-linking profiles was performed using the STEM clustering program[60]. Only profiles were selected that showed a fold-change in FPKM of at least 1.5 and had a mean pairwise correlation over two biological replicates of 0.7.

Escape indices and selection of Nab3 attenuated genes. To calculate the Pol II transcription EI[41], we first selected protein-coding genes that had a minimum coverage of 10 FPKM at the indicated time-points (0, 4 and 18 min after the induction of the glucose deprivation response). The EI was calculated by summing the nucleotide density in the promoter proximal region (PPR; −100 to + 250 from the 5′UTR) and dividing this number by the nucleotide density of the body region (+251 to 3′ end). Subsequently, we divided the values for the 4 and 18-min time-point by the values for the 0 time-point to calculate the EI. These data were then compared to changes in Pol II transcription, which was calculated by dividing the total normalized nucleotide density of the whole gene from the 18-min sample by the total normalized nucleotide density of the 0-minute sample. We then only selected those genes that showed an increase in transcription of at least 1.5, a coverage of at least 10 FPKM and EI of at least 2. From the resulting list of genes, only genes were selected that (a) had Nab3-binding sites near the 5′ end of the gene and (b) showed reproducible profiles in a replicate experiment.

Data normalization. We scaled the FPKM values of all transcripts within each time point by a constant factor such that the sum of FPKMs for all time points and all experimental replicates is the same. This was done for all data sets that were analyzed simultaneously, e.g., Nab3 and Pol II data sets (see below). For all data sets, the same time points were used (on a few occasions, temporally close time points were deemed identical for experimental purposes). Finally, for all analyzed data sets, we divided each time series of each experimental replicate by its steady state value before the imposition of stress (at 0 min after the nutrient shift). This way, all time series start at the same normalized binding value of 1 a.u. before the nutrient shift and the other values for later time points are relative to the background binding signal. We only keep those transcripts for the analysis that have real values for all time points in all experimental replicates after all steps of the normalization procedure.

Differential gene expression analyses. For the differential expression analyses we used DESeq2[22] in which two Megatron data sets were compared to four Vari-X-linker Nab3 glucose data sets. Only differentially expressed genes were selected that had an adjusted p-value of 0.05 or lower.

Testing for differential dynamic response. To determine whether the imposition of stress results in differential dynamics of RBP binding, we used a Bayesian non-parametric regression approach. Let $f_j(t)$ represent the binding response in condition j (stress or control) at time t, relative to time 0. Our main assumption is that this response, averaged over a population of cells, can be well modeled as a smooth function of time. To capture this assumption, we formulate a probabilistic model for the response function in terms of GPs (see e.g.[27]). A GP is an infinite-dimensional generalization of the multivariate Gaussian distribution, which provides a suitable prior distribution over a space of functions. Here we enforce the smoothness assumption by modeling correlations between function values at times s and t using a squared-exponential covariance function:

$$\mathrm{cov}(f_j(s), f_j(t)) = a^2 \exp\left[-\frac{(s-t)^2}{2\lambda^2}\right]$$

This covariance depends on two hyperparameters α, λ that are fitted to the data as described below.

We assume that observations $y_j(t)$ of binding to a specific transcript in condition j at time t (i.e., FPKM from the CRAC experiment at time t in condition j) are obtained from the unobserved function $f_j(t)$ by addition of zero-mean Gaussian noise with standard deviation σ. These assumptions enable us to marginalize exactly the unobserved function values to obtain an estimate of the data evidence (or marginal likelihood).

$$p(y(0), \ldots, y(T)|\alpha, \lambda, \sigma)$$

We then reformulate the testing question as a model selection problem. We consider two competing models:

- H0, all the time series (control and stress) can be explained as noise corrupted observations of a single underlying function $f(t)$ describing the dynamics of the system (null hypothesis).
- H1, control and stress time series result from two distinct underlying dynamics, i.e., there are two functions $f_c(t)$ and $f_t(t)$, which generate the observations (alternative).

The ratio of the evidence under the two hypotheses (Bayes factor) quantifies the ratio of posterior probabilities of each model being correct, and hence provides a criterion for selecting one hypothesis over the other. We follow Kass and Raftery[28] in adopting a Bayes factor greater or equal to 10 as strong evidence of one hypothesis over the other. The Bayes factor computation is performed independently for every transcript; it should be noted that, as this is a Bayesian method, no issues of multiple testing arise, since sampling variability is already accounted for in the marginalization process.

It should be noted that the evidence calculation can only be performed exactly when the covariance hyperparameters α, λ as well as the observation noise with standard deviation σ are known. Such parameters can also be assigned a prior distribution and marginalized for a fully Bayesian treatment; however, this greatly complicates the computational task of computing the evidence as exact marginalization is not possible. To avoid these additional overheads, we adopted an empirical Bayesian strategy and fixed the hyperparameters in a data-driven fashion. The scaling hyperparameterα^2 was defined as 50% of the variance of all binding time series, in control and treatment experimental replicates, for a given transcript. The length scale hyperparameter λ, which determines the number of units the data can be extrapolated away from, was globally set as 1 min. The variance of the observation noise σ^2 was chosen to be the variance of the binding time series of a given transcript in control conditions, i.e., under the glucose-to-glucose shift. Notice that the GP models under both hypotheses were provided with the same hyperparameter values, to avoid over fitting to the data.

Code availability. The pyCRAC package[56] used for the data analyses is available on https://bitbucket.org/sgrann/pycrac. The complete data analysis pipeline is available on https://bitbucket.org/sgrann/kinetic_crac_pipeline/. Other python and perl scripts used for isolating oligo-A tailed reads are available upon request.

Data availability. Fastq and processed sequencing data that support the findings of this study have been deposited in the Gene Expression Omnibus (GEO) under the accession code GSE85545 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?&acc=GSE85545). The data that support the findings of this study are available from the corresponding author upon request.

Received: 22 September 2016 Accepted: 20 February 2017
Published online: 11 April 2017

References

1. Castello, A., Fischer, B., Hentze, M. W. & Preiss, T. RNA-binding proteins in Mendelian disease. *Trends Genet.* **29**, 318–327 (2013).
2. Pérez-Ortín, J. E., de Miguel-Jiménez, L. & Chávez, S. Genome-wide studies of mRNA synthesis and degradation in eukaryotes. *Biochim. Biophys. Acta* **1819**, 604–615 (2012).
3. Marguerat, S., Lawler, K., Brazma, A. & Bähler, J. Contributions of transcription and mRNA decay to gene expression dynamics of fission yeast in response to oxidative stress. *RNA Biol.* **11**, 702–714 (2014).
4. Kresnowati, M. T. A. P. *et al.* When transcriptome meets metabolome: fast cellular responses of yeast to sudden relief of glucose limitation. *Mol. Syst. Biol.* **2**, 49 (2006).
5. Ule, J. *et al.* CLIP identifies Nova-regulated RNA networks in the brain. *Science* **302**, 1212–1215 (2003).
6. Granneman, S., Kudla, G., Petfalski, E. & Tollervey, D. Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. *Proc. Natl. Acad. Sci. USA* **106**, 9613–9618 (2009).
7. Hafner, M. *et al.* Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141**, 129–141 (2010).
8. König, J. *et al.* iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.* **17**, 909–915 (2010).
9. Van Nostrand, E. L. *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods* **13**, 508–514 (2016).
10. Zarnegar, B. J. *et al.* irCLIP platform for efficient characterization of protein–RNA interactions. *Nat. Methods* **13**, 489–492 (2016).
11. Flynn, R. A. *et al.* Dissecting noncoding and pathogen RNA-protein interactomes. *RNA* **21**, 135–143 (2015).
12. Granneman, S., Petfalski, E. & Tollervey, D. A cluster of ribosome synthesis factors regulate pre-rRNA folding and 5.8S rRNA maturation by the Rat1 exonuclease. *EMBO. J.* **30**, 4006–4019 (2011).
13. Schaughency, P., Merran, J. & Corden, J. L. Genome-wide mapping of yeast RNA polymerase II termination. *PLoS Genet.* **10**, e1004632 (2014).
14. Beckmann, B. M. RNA interactome capture in yeast. *Methods* doi:10.1016/j.ymeth.2016.12.008 (in the press; 2016)
15. Porrua, O. & Libri, D. Transcription termination and the control of the transcriptome: why, where and how to stop. *Nat. Rev. Mol. Cell Biol.* **16**, 190–202 (2015).
16. Schulz, D. *et al.* Transcriptome surveillance by selective termination of noncoding RNA synthesis. *Cell* **155**, 1075–1087 (2013).
17. Darby, M. M., Serebreni, L., Pan, X., Boeke, J. D. & Corden, J. L. The S. cerevisiae Nrd1-Nab3 transcription termination pathway acts in opposition to ras signaling and mediates response to nutrient depletion. *Mol. Cell. Biol.* **32**, 1762–1775 (2012).
18. Gasch, A. P. *et al.* Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell* **11**, 4241–4257 (2000).
19. Jamonnak, N. *et al.* Yeast Nrd1, Nab3, and Sen1 transcriptome-wide binding maps suggest multiple roles in post-transcriptional RNA processing. *RNA* **17**, 2011–2025 (2011).
20. Wu, J., Zhang, N., Hayes, A., Panoutsopoulou, K. & Oliver, S. G. Global analysis of nutrient control of gene expression in Saccharomyces cerevisiae during growth and starvation. *Proc. Natl Acad. Sci. USA* **101**, 3148–3153 (2004).
21. Tree, J. J., Granneman, S., McAteer, S. P., Tollervey, D. & Gally, D. L. Identification of bacteriophage-encoded anti-sRNAs in pathogenic Escherichia coli. *Mol. Cell* **55**, 199–213 (2014).
22. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
23. Arigo, J. T., Carroll, K. L., Ames, J. M. & Corden, J. L. Regulation of yeast NRD1 expression by premature transcription termination. *Mol. Cell* **21**, 641–651 (2006).
24. Neil, H. *et al* Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature* **457**, 1038–1042 (2009).
25. Ghavidel, A. *et al.* Impaired tRNA nuclear export links DNA damage and cell-cycle checkpoint. *Cell* **131**, 915–926 (2007).
26. Milligan, L. *et al.* Strand-specific, high-resolution mapping of modified RNA polymerase II. *Mol. Syst. Biol.* **12**, 874 (2016).
27. Rasmussen, C. E. & Williams, C. K. I. in *Gaussian Processes for Machine Learning* (Mit Press, 2006).
28. Kass, R. E. & Raftery, A. E. Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795 (2012).
29. Haruki, H., Nishikawa, J. & Laemmli, U. K. The anchor-away technique: rapid, conditional establishment of yeast mutant phenotypes. *Mol. Cell* **31**, 925–932 (2008).
30. Munchel, S. E., Shultzaberger, R. K., Takizawa, N. & Weis, K. Dynamic profiling of mRNA turnover reveals gene-specific and system-wide regulation of mRNA decay. *Mol. Biol. Cell* **22**, 2787–2795 (2011).
31. Vasiljeva, L., Kim, M., Mutschler, H., Buratowski, S. & Meinhart, A. The Nrd1-Nab3-Sen1 termination complex interacts with the Ser5-phosphorylated RNA polymerase II C-terminal domain. *Nat. Struct. Mol. Biol.* **15**, 795–804 (2008).
32. Gudipati, R. K., Villa, T., Boulay, J. & Libri, D. Phosphorylation of the RNA polymerase II C-terminal domain dictates transcription termination choice. *Nat. Struct. Mol. Biol.* **15**, 786–794 (2008).
33. Xu, Z. *et al.* Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**, 1033–1037 (2009).
34. Kim, H. *et al.* Gene-specific RNA polymerase II phosphorylation and the CTD code. *Nat. Struct. Mol. Biol.* **17**, 1279–1286 (2010).
35. Kopcewicz, K. A., O'Rourke, T. W. & Reines, D. Metabolic regulation of IMD2 transcription and an unusual DNA element that generates short transcripts. *Mol. Cell. Biol.* **27**, 2821–2829 (2007).
36. Kuehner, J. N. & Brow, D. A. Regulation of a eukaryotic gene by GTP-dependent start site selection and transcription attenuation. *Mol. Cell* **31**, 201–211 (2008).
37. Tuck, A. C. & Tollervey, D. A transcriptome-wide atlas of RNP composition reveals diverse classes of mRNAs and lncRNAs. *Cell* **154**, 996–1009 (2013).
38. Rhee, H. S. & Pugh, B. F. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature* **483**, 295–301 (2012).
39. Pelechano, V., Wei, W. & Steinmetz, L. M. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* **497**, 127–131 (2013).
40. Wlotzka, W., Kudla, G., Granneman, S. & Tollervey, D. The nuclear RNA polymerase II surveillance system targets polymerase III transcripts. *EMBO. J.* **30**, 1790–1803 (2011).
41. Brannan, K. *et al.* mRNA decapping factors and the exonuclease Xrn2 function in widespread premature termination of RNA polymerase II transcription. *Mol. Cell* **46**, 311–324 (2012).
42. Kim, J. M., Vanguri, S., Boeke, J. D., Gabriel, A. & Voytas, D. F. Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete Saccharomyces cerevisiae genome sequence. *Genome Res.* **8**, 464–478 (1998).
43. Kim, T., Xu, Z., Clauder-Münster, S., Steinmetz, L. M. & Buratowski, S. Set3 HDAC mediates effects of overlapping noncoding transcription on gene induction kinetics. *Cell* **150**, 1158–1169 (2012).
44. Oliver, K. R. & Greene, W. K. Transposable elements: powerful facilitators of evolution. *Bioessays* **31**, 703–714 (2009).
45. Lesage, P. & Todeschini, A. L. Happy together: the life and times of Ty retrotransposons and their hosts. *Cytogenet. Genome Res.* **110**, 70–90 (2005).
46. Billmyre, R. B., Calo, S., Feretzaki, M., Wang, X. & Heitman, J. RNAi function, diversity, and loss in the fungal kingdom. *Chromosome Res.* **21**, 561–572 (2013).
47. Heras, S. R., Macias, S., Cáceres, J. F. & Garcia-Perez, J. L. Control of mammalian retrotransposons by cellular RNA processing activities. *Mob. Genet. Elements* **4**, e28439 (2014).
48. Carmen, A. A. & Holland, M. J. The upstream repression sequence from the yeast enolase gene ENO1 is a complex regulatory element that binds multiple trans-acting factors including REB1. *J. Biol. Chem.* **269**, 9790–9797 (1994).
49. Carmen, A. A., Brindle, P. K., Park, C. S. & Holland, M. J. Transcriptional regulation by an upstream repression sequence from the yeast enolase gene ENO1. *Yeast* **11**, 1031–1043 (1995).
50. dela Cruz, J., Karbstein, K. & Woolford, J. L. Functions of ribosomal proteins in assembly of eukaryotic ribosomes in vivo. *Annu. Rev. Biochem.* **84**, 93–129 (2015).
51. Coller, J. Methods to determine mRNA half-life in Saccharomyces cerevisiae. *Methods Enzymol.* **448**, 267–284 (2008).

52. Rabani, M. *et al.* Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nat. Biotechnol.* **29**, 436–442 (2011).
53. Sun, M. *et al.* Comparative dynamic transcriptome analysis (cDTA) reveals mutual feedback between mRNA synthesis and degradation. *Genome Res.* **22**, 1350–1359 (2012).
54. Honkela, A. *et al.* Genome-wide modeling of transcription kinetics reveals patterns of RNA production delays. *Proc. Natl Acad. Sci. USA* **112**, 13115–13120 (2015).
55. Deneke, C., Lipowsky, R. & Valleriani, A. Complex degradation processes lead to non-exponential decay patterns and age-dependent decay rates of messenger RNA. *PLoS ONE* **8**, e55442 (2013).
56. Webb, S., Hector, R. D., Kudla, G. & Granneman, S. PAR-CLIP data indicate that Nrd1-Nab3-dependent transcription termination regulates expression of hundreds of protein coding genes in yeast. *Genome Biol.* **15**, R8 (2014).
57. Dodt, M., Roehr, J. T., Ahmed, R. & Dieterich, C. FLEXBAR-Flexible barcode and adapter processing for next-generation sequencing platforms. *Biology (Basel)* **1**, 895–905 (2012).
58. van Dijk, E. L. *et al.* XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. *Nature* **475**, 114–117 (2011).
59. Yassour, M. *et al.* Strand-specific RNA sequencing reveals extensive regulated long antisense transcripts that are conserved across yeast species. *Genome Biol.* **11**, 1 (2010).
60. Ernst, J. & Bar-Joseph, Z. STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics* **7**, 191 (2006).

Acknowledgements

We would like to thank Lidia Vasiljeva and members of the Granneman lab for critically reading the manuscript. We thank Frank Pugh for helpful suggestions. We are grateful to Ola Helwak for testing the Vari-X-linker with mammalian cells and Francesca Storici for providing the Delitto perfetto plasmids. This work was supported by grants from the Wellcome Trust (to S.G.; 091549) and a BBSRC Sparking award (to S.G.; S.I.2013.0202). G.S. and G.Sch. acknowledge support from the European Research Council under grant MLCS306999. A.S. is supported by the EPSRC. Next generation sequencing was carried out by Edinburgh Genomics, The University of Edinburgh. Edinburgh Genomics is partly supported through core grants from NERC (R8/H10/56), MRC (MR/K001744/1) and BBSRC (BB/J004243/1).

Author contributions

S.G. conceived the χCRAC method. S.G. and R.v.N. designed the experiments. R.v.N., E.d.L., and S.G. performed the experiments. P.W., S.G., and R.v.N. conceived the filtration unit. A.S., G.Sch., S.G., and G.San. designed and performed the computational analyses. All authors contributed to writing the paper and read and approved the manuscript.

Additional information

Supplementary Information accompanies this paper at doi:10.1038/s41467-017-00025-5.

Competing interests: P.W. was the director and owner of UVO_3 that sells equipment for water sterilization. A.L. and R.F. are employees of UVO_3. P.W., A.L., and R.F. have been involved in the development of the Vari-X-linker and the filtration unit. All remaining authors declare no competing financial interests.

5.5 FORMAL DEFINITION OF THE MODEL

This section formally defines the models used by the algorithm and provides the corresponding derivations.

As before, denote the binding of the protein to a transcript in condition j (stress or control) at time t as $f_j(t)$. The binding response at time t is relative to that at the timepoint 0, which corresponds to the cross-linking experiment before the stress induction. Thus, the time-series starts at the value of 1 a.u. and the following values represent the fold-change in binding compared to control conditions. For the time being, let's omit the index j specifying the condition as the derivations presented below apply to both conditions.

We assume that the RNA-protein cross-linking dynamics, averaged over a population of cells, can be modelled as a smooth function of time. We reflect this assumption by using a squared-exponential covariance function (which depends on two hyperparameters α and λ) to model dependencies between the function values at times t_a and t_b:

$$K_{ab} = \mathrm{cov}(f(t_a), f(t_b)) = \alpha^2 \exp\left(\frac{-(t_a - t_b)^2}{2\lambda^2}\right) \tag{5.8}$$

This makes the function values that are close to each other in time have larger covariance. The hyperparameter α controls the amplitude of the process and the lengthscale λ specifies the distance between timepoints, after which function values can change significantly.

5.5.1 *Gaussian observation model*

As before, we denote the observation of the binding response of a transcript at time t as $y(t)$ (temporarily omitting index j for simplicity). Let us further use the notation such that $y(t) \equiv y_t$ to declutter the following derivations.

We assume that observations differ from the hidden function values by additive noise and we further assume for this noise to come from an independent identically distributed Gaussian. This yields the likelihood of one data point y_i given the value of the hidden process f_i to be Gaussian distributed (Eq. 5.9). Following the independence assumption, the likelihood of the whole time-series $\mathbf{y} = \{y_i\}_{i=1}^N$ given the function $\mathbf{f} = \{f_i\}_{i=1}^N$ for $i = 1...N$ timepoints factorises as shown in Eq. 5.10. Below, I_N is a diagonal $N \times N$ matrix.

$$p(y_i|f_i) = N(0, \sigma^2) \tag{5.9}$$

$$p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{N} p(y_i|f_i) = \frac{1}{(\sqrt{2\pi}\sigma)^N} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{f})^T I_N (\mathbf{y}-\mathbf{f})\right) \tag{5.10}$$

As the study included multiple replicates of Pol II and Nab3 cross-linking experiments in treatment conditions, we further include replicate information in the model, assuming independent replicates. Let $\mathbf{y}^k$ correspond to the k-th replicate of a time-series for $k = 1...M$ experimental replicates. The likelihood expression including all replicates is then defined in Eq. 5.11.

$$p(\mathbf{y}|\mathbf{f}) = \frac{1}{(\sqrt{2\pi}\sigma)^{NM}} \exp\left(\sum_{k=1}^{M} -\frac{1}{2\sigma^2}(\mathbf{y}^k-\mathbf{f})^T I_N (\mathbf{y}^k-\mathbf{f})\right) \tag{5.11}$$

5.5.2 *Computing marginal likelihood*

We are interested in computing the marginal likelihood of the data, marginalising out the unobserved function values $\mathbf{f}$. The marginal likelihood can be computed by integrating the joint distribution $p(\mathbf{y}, \mathbf{f})$ over all values of $\mathbf{f}$ (Eq. 5.13), which has an analytical solution if $p(\mathbf{y}|\mathbf{f})$ is a Gaussian distribution.

The prior of $\mathbf{f}$ is defined to have the mean of 0 and the covariance K, specifying the dependencies between the function values at various times (Eq. 5.12). The mean of zero was chosen both for simplifying the following derivations and as a suitable choice for representing the cross-linking values in the control time-series, which are expected to be around 1 a.u. The same prior is used for the processes under both models to facilitate fair comparison.

$$\mathbf{f} \sim N(\mathbf{0}, K) \tag{5.12}$$

$$p(\mathbf{y}) = \int p(\mathbf{y}, \mathbf{f}) d\mathbf{f} = \int p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}) d\mathbf{f} \tag{5.13}$$

Writing out the expression for the marginal likelihood, we get:

$$p(\mathbf{y}) = \frac{1}{(\sqrt{2\pi}\sigma)^{NM}} \cdot \frac{1}{(\sqrt{(2\pi)^N |K|})} \cdot \int \exp\left(\sum_{k=1}^{M}\left(-\frac{1}{2\sigma^2}(\mathbf{y}^k - \mathbf{f})^T I_N (\mathbf{y}^k - \mathbf{f})\right) - \frac{1}{2}\mathbf{f}^T K^{-1}\mathbf{f}\right) d\mathbf{f} \tag{5.14}$$

The integral over $\mathbf{f}$ can be evaluated exactly if the expression under the exponent is written in the quadratic form with respect to $\mathbf{f}$ for some mean $\boldsymbol{\mu}$ and covariance C:

$$\int \exp\left(-\frac{1}{2}(\mathbf{f}-\boldsymbol{\mu})^T C^{-1}(\mathbf{f}-\boldsymbol{\mu})\right) d\mathbf{f} = \sqrt{(2\pi)^N\ |C|} \tag{5.15}$$

The expression for the joint distribution $p(\mathbf{y}, \mathbf{f})$ can be manipulated into the quadratic form given in Eq. 5.15 by completing the square. Full derivations are given in Section B.1 of Appendix B.

After taking the terms not containing $\mathbf{f}$ outside the integral and evaluating it using Eq. 5.15, the expression for the marginal likelihood of the data $\mathbf{y}$ is given by:

$$p(\mathbf{y}) = \frac{1}{(\sqrt{2\pi}\sigma)^{NM}} \cdot \frac{1}{(\sqrt{(2\pi)^N |K|})} \cdot \exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{M}\left((\mathbf{y}^k)^T I_N \mathbf{y}^k\right) + \frac{1}{2\sigma^4}\left(\sum_{k=1}^{M}\mathbf{y}^k\right)^T C\left(\sum_{k=1}^{M}\mathbf{y}^k\right)\right) \cdot \sqrt{(2\pi)^N\ |C|} \tag{5.16}$$

In Eq. 5.16 above, the covariance matrix is defined as $C = \left(K^{-1} + \frac{M}{\sigma^2} I_N\right)^{-1}$, where K is prior covariance, M is the number of replicates, and σ^2 is the observation noise.

5.5.3 *Model selection*

For the null hypothesis M_0 model, the marginal likelihood is computed for all time-series replicates, in both control and stress conditions. For the alternative hypothesis M_1 model, the marginal likelihood is separately computed for the control time-series and for all time-series replicates in treatment conditions, and their product is taken. For numerical stability, log-marginal likelihoods were computed under both models and the log-Bayes factor was compared with $\log(10)$ for selecting target transcripts with significant changes in cross-linking profiles.

$$\log \mathrm{BF} = \log p(\mathbf{y_c}|M_1) + \log p(\mathbf{y_s}|M_1) - \log p(\mathbf{y_c}, \mathbf{y_s}|M_0) \tag{5.17}$$

It is worth noting that the more complex model M_1, allowing two different underlying functions, does not necessarily always provide a better explanation for the data than the simpler M_0 model, which assumes a single hidden function $\mathbf{f}$. This corresponds to the notion of *Occam's razor*, which encourages simplicity in explanations. We can see this from the expression of the marginal likelihood of the data. The expression in Eq. 5.16 simplifies as follows:

$$\begin{aligned} p(\mathbf{y}) = \frac{1}{(\sqrt{2\pi}\sigma)^{NM}} \cdot \sqrt{\frac{|C|}{|K|}} \cdot \\ \exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{M}\left((\mathbf{y}^k)^T I_N \mathbf{y}^k\right) + \frac{1}{2\sigma^4}\left(\sum_{k=1}^{M}\mathbf{y}^k\right)^T C \left(\sum_{k=1}^{M}\mathbf{y}^k\right)\right) \end{aligned} \tag{5.18}$$

Using the fact that $\frac{|C|}{|K|} = |CK^{-1}|$, we can write down the expression for this determinant as shown in Eq. 5.19. Full derivations are given in Section B.1.3 of Appendix B.

$$|CK^{-1}| = \left|\frac{\sigma^2}{M}\left(\frac{\sigma^2}{M}\cdot I_N + K\right)^{-1}\right| = \left(\frac{\sigma^2}{M}\right)^N \left|\frac{\sigma^2}{M}\cdot I_N + K\right|^{-1} \tag{5.19}$$

In the expression for the log-marginal likelihood of the data, the term under the exponent measures how close the hidden function is to the data, while the term from Eq. 5.19 penalises the marginal likelihood as $-\frac{1}{2}\log\left|\frac{\sigma^2}{M}\cdot I_N + K\right|$. Under the null hypothesis M_0 model, the marginal likelihood is penalised once by this term. In contrast, under the alternative model M_1, the expression for the marginal likelihood has two such terms, arising from both hidden functions: $-\frac{1}{2}\log\left|\frac{\sigma^2}{N_c}\cdot I_N + K\right|$ and $-\frac{1}{2}\log\left|\frac{\sigma^2}{N_t}\cdot I_N + K\right|$, where N_c and N_t sum to M and are the numbers of replicates in control and stress conditions, correspondingly. Thus, under the alternative model, the fit of the hidden functions to the data has to overcome the sum of the penalising terms in order to provide a better explanation over the M_0 model.

5.5.4 *Defining hyperparameters*

The marginal likelihoods of the data can be computed exactly only when the hyperparameters are known. The fully Bayesian treatment of hyperparameters involves placing priors on them and marginalising them out, which is impossible to do exactly. Thus, in order to do exact marginalisation and avoid complicating the computational task, an empirical Bayesian strategy was adopted and the hyperparameters were fixed in a data-driven fashion for each transcript.

The scaling hyperparameter α^2 was defined to be 50% of the variance of all time-series in both conditions. The variance of the observation noise σ^2 was chosen to be the variance of the time-series in control conditions, representing our beliefs about the noise associated with experimental measurements. The lengthscale hyperparameter λ was set to 1 minute, corresponding to the temporal resolution of the data and thus, representing the detectable unit of time after which the binding can change significantly.

5.6 CORRECTION FOR THE PAPER

Since the publication of the paper, a numerical error was found in the implementation of the proposed algorithm, which was used to produce lists of transcripts that demonstrated significantly different binding to the protein of interest (Pol II or Nab3) under stress compared to control conditions. The error was found in the function computing the constant which the integral for the marginal likelihood evaluates to, as shown in Eq. 5.15. Specifically, the implementation incorrectly used the inverse of the constant when computing the marginal likelihoods of the data under both models.

The constant was erroneously inverted in expressions corresponding to all marginal likelihoods of the data, $p(\mathbf{y_c}|M_1)$, $p(\mathbf{y_s}|M_1)$, and $p(\mathbf{y_c}, \mathbf{y_s}|M_0)$. The constant depends on the matrix C, which is defined in terms of the prior covariance K, the observation noise σ^2, and the number of replicates M used under each model (Eq. B.4). While K and σ^2 are the same in both models, the number of replicates changes for each marginal likelihood: M equals to the number of control replicates, the number of treatment replicates or their sum when fitting the control time-series, the treatment time-series or all time-series, correspondingly.

Thus, the differential binding analysis was performed again for the Pol II and Nab3 χCRAC datasets using the corrected implementation. During processing of raw sequencing data used for the analyses presented in the paper, the untranslated regions (UTR) at each end of the gene were manually set to 200 nucleotides. Reads were

mapped and FPKM (fragments per kilobase per million) values were computed as described in the Methods section "*Processing of raw sequencing data*". Prior to the differential binding analysis, the FPKM values were transformed with the normalisation procedure described in the Methods section "*Data normalization*".

The results for Nab3 remained largely identical, with 93% of newly selected targets overlapping with the original selection used in the paper (4245 targets were originally selected, 4032 were selected with the corrected implementation, 3955 transcripts were present in both selections). Therefore, all presented analyses involving transcripts differentially bound by Nab3 under stress remain virtually unchanged. This result means that for the majority of transcripts, the error didn't affect the magnitude of the marginal likelihoods ratio in comparison to the chosen threshold of 10. This could be attributed to the effect of other terms in the expression for the marginal likelihood (Eq. 5.16) on the Bayes factor. As the exponential term depends on the data, this could suggest a confident model selection, whereby the M_1 model consistently produced better fits to the control and treatment cross-linking data than the M_0 model did to all time-series.

For the Pol II dataset, the results were more different. The number of transcripts quoted in the paper as showing significant changes in Pol II cross-linking during stress (Results section "χ*CRAC provides insights into transcription kinetics*") corresponded to the number of transcripts for which cross-linking was detected at all timepoints. The total number of unfiltered Pol II targets selected by the algorithm was 4372. The correct implementation selected 2695 transcripts, of which 2621 transcripts were also present in the original selection. As the correctly identified Pol II targets constituted 60% of the originally selected targets, the affected analyses presented in the paper were repeated on the corrected selection.

Specifically, the figures corresponding to Fig. 4a and 4b of the paper were generated again for the new transcript selection. The selection was filtered for the time-series with more than 0 reads mapped in any timepoint, as before. Panel **a** in Fig. 5.4 shows what percentage of each RNA class showed changes in Pol II binding during glucose deprivation. The RNA class split among the selected transcripts remained similar to the results presented in the paper. The largest changes were observed among small nucleolar RNAs (snoRNAs), protein-coding RNAs, and non-coding RNAs (ncRNAs). Changes in Pol II binding of pseudogenes appear slightly larger than before, while changes in the anti-sense RNA class seemed to have been overestimated. Smaller changes in binding are still observed among the cryptic unstable transcripts (CUTs), stable uncharacterised transcripts (SUTs), and Xrn1-sensitive unstable transcripts (XUTs).

The clusters of Pol II binding patterns found among the filtered selection largely resembled the results shown in Fig. 4b in the paper. The originally identified cluster 0 seemed to correspond to clusters 0 and 1 shown in Panel **b** (Fig. 5.4), obtained with clustering the new transcripts selection. The profiles shown in cluster 2 (Fig. 5.4**b**) seemed to include the binding patterns that were originally split into clusters 2 and 3 (Fig. 4b in the paper). The newly identified clusters 0, 1, and 2 also agreed with the enriched GO (gene ontology) terms indicated for their counterparts in the original clusters shown in the paper. The binding profile shown in cluster 3 (Fig. 5.4**b**) resembled the profiles of the cluster 1 shown in Fig. 4b in the paper, albeit consisted of fewer elements. The new cluster 3 had fewer enriched GO terms associated with it. It is possible that some transcripts that were previously allocated to cluster 1 (shown in the paper) now appear in the newly identified cluster 1, as it recovers enrichment in transcripts involved in DNA integration. The clustering analysis on the new selected targets was performed as described in the Methods section "*K-means clustering*".

The correction for the presented paper (van Nues et al., 2017) is in preparation by the senior authors Guido Sanguinetti and Sander Granneman. It will draw attention to the corrected algorithm implementation, include changes to the affected analyses summarised above, and add clarifications regarding the quoted numbers of selected transcripts (filtered and unfiltered) and data processing involving manually edited gene coordinates. Notably, all discussed changes are minor and do not affect the conclusions of the paper. Specifically, the χCRAC experimental protocol and the algorithm for detecting differential cross-linking presented in this chapter provide a powerful tool for studying RNA-protein interactions at high temporal resolution. Our results revealed pervasive changes in Nab3 binding to the large fraction of the transcriptome in response to nutrient stress. These differential Nab3 binding patterns were not simply co-occurring with transcriptional changes and thus, suggest an additional control mechanism for gene expression regulation.

In the light of finding an implementation error after the publication, it is necessary to stress the importance of research reproducibility and careful data handling. All R software developed by myself for implementing the proposed algorithm and applying it to the data is kept in a distributed version control system online. This made it possible for me to speedily reproduce all results presented in the paper and evaluate the arising differences. Sander Granneman recommended the use of the interactive computational environment IPython Notebook, which was used to create figures for the paper and made it easy to repeat analyses collaboratively. The collaboration with Sander Granneman was further made more transparent by storing data in the shared server repository, which enabled myself to easily interact with the data and track changes over time. These and other more advanced practices for software

development and collaborative data handling are of principal importance in the research process and, while requiring considerable amounts of time to configure for useful operation, must not be overlooked.

5.7 TIME-SERIES ANALYSIS OF XRN1 CROSS-LINKING

Continuing to examine the role of degradation in stress response, further χCRAC experiments, which followed the design described in Section 5.2, were performed on the exoribonuclease Xrn1, a major nuclease involved in cytoplasmic degradation. Experiments were performed by Sander Granneman and Rob van Nues. One time-series was collected in control conditions (shift to a glucose-rich medium) and three time-series were collected in stress conditions (shift to a glucose-deprived medium). The analysis presented in this section additionally used the RNA-seq dataset on ribosomal RNA-depleted total RNA collected in stress conditions (described in the Results section "*Monitoring in vivo dynamics of protein-RNA interactions*" of the paper), which quantified transcript abundance at various times since the nutrient shift.

5.7.1 *Dataset quality control*

Instead of working with normalised FPKM values, as was previously done for the analyses presented in the paper, we directly examined the raw counts of reads mapped to transcripts in each χCRAC experiment, in order to assess the quality of the Xrn1 dataset. Reads were mapped and raw counts were generated as described in the Methods section "*Processing of raw sequencing data*", using the yeast genome annotation by Nagalakshmi et al. As the differential binding analysis is concerned with independent modelling of each transcript's time-series, we directly monitored the average number of reads mapped per timepoint in different replicates for each transcript.

The transcripts' χCRAC time-series were selected for testing using the same normalisation procedure as described in the Methods section "*Data normalization*", with the addition of linearly interpolating the cross-linking counts between different time points. This was done due to the fact that the Xrn1 χCRAC experiments were performed at different times following the shift of cells to a different medium. Specifically, the control time-series was collected at 0, 1, 2, 4, 8, 16, 28, and 40 minutes following the shift to the glucose-rich medium. The first experimental replicate in stress conditions was collected at the same timepoints, except for the last experiment, which was performed at a later time of 45 minutes since the nutrient shift. For the second replicate, the later timepoints varied to be 11, 18, 29, and 46 minutes since the nutrient shift. Finally, the third replicate in stress conditions had only 4 timepoints

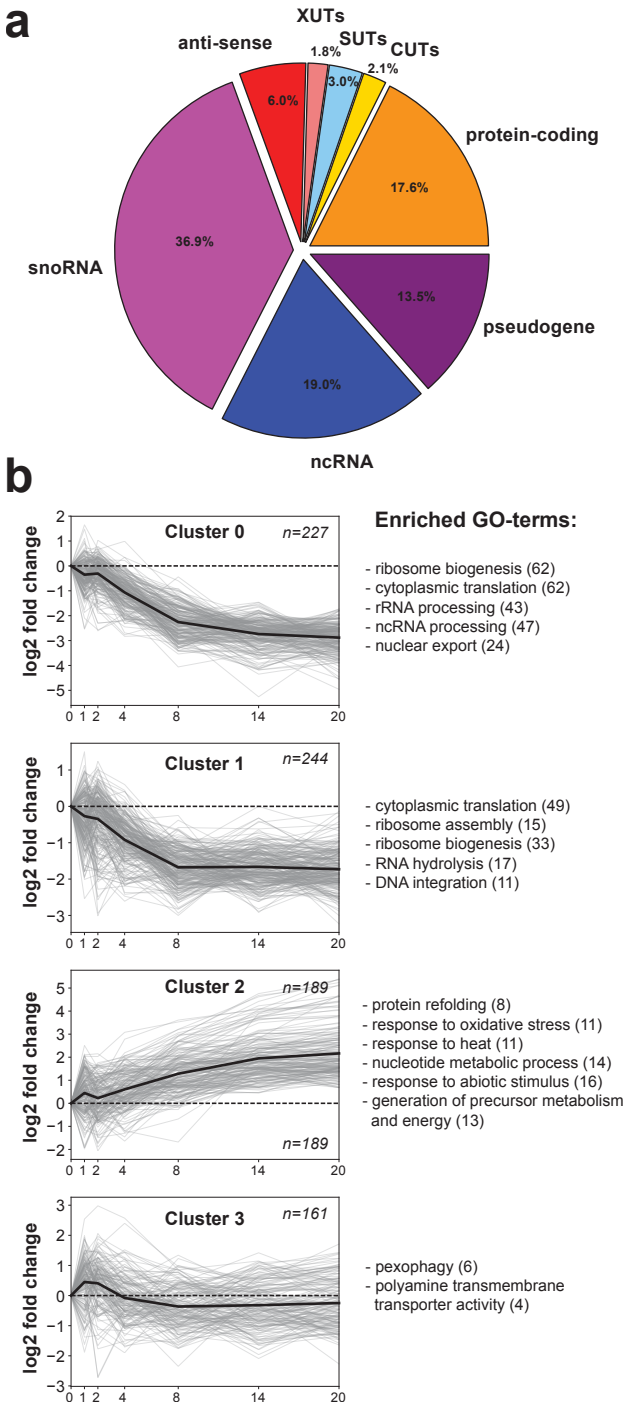


Figure 5.4: Repeated analyses for the selected targets differentially bound by Pol II under nutrient stress. **a.** Percentages of each RNA class with detected changes in Pol II transcription. **b.** Cross-linking profiles for all selected targets. Both panels generated as described for Fig. 4a, 4b in the paper.

mapped in total: 0, 4, 16, and 48 minutes following the shift to the glucose-deprived medium. Xrn1 datasets were processed together with the RNA-seq dataset, in order to test for significant changes specific to Xrn1 binding and exclude cases where they can be explained by changes in total abundance of mature transcript. To control for this, time-series for stress replicates were normalised by the RNA-seq time-series (interpolated to the same timepoints) and the control time-series were normalised by the RNA-seq values at 0 minutes. All time-courses in each condition were normalised to their value at 0 minutes, measured before the stress induction and corresponding to control conditions. Transcripts with no reads mapped at more than half of all considered timepoints were removed from the analysis.

For the transcripts selected for testing, the medians were computed across their unprocessed raw count time-series in each condition (Fig. 5.5). The median was used instead of the mean as many time-series commonly had one or two high counts mapped at some timepoints alongside low (or zero) counts mapped in the rest of the time-course. Examination of median counts per time-series confirmed the lower quality of the control and stress replicate 2 datasets (the median values of distributions across all transcripts were 2 and 4.5 reads, correspondingly), which was first detected by comparing the total number of mapped reads at each timepoint. For this reason, it was decided not to use these datasets.

As the experimental control time-course was not being used due to its limited coverage, a synthetic control time-course was generated from the values mapped at 0 minutes in the two remaining Xrn1 time-series in stress conditions. This modelling choice followed the assumption that there should be no significant time-specific changes in Xrn1 binding to the transcriptome in control conditions. Following the Gaussian observation model (Eq. 5.9), the control time-series was generated from a Gaussian distribution with the mean and variance estimated from the two available measurements at 0 minutes in stress conditions. Negative numbers in generated control time-courses were set to 0.

Once the datasets with low coverage were removed from the analysis, more transcripts were selected for testing after the normalisation procedure (5241 instead of 1899). This was due to the fact that only those transcripts were selected that had non-zero values in all timepoints and all experimental replicates after normalisation. The medians of mapped reads at each timepoint for the selected transcripts were 6.5 reads in stress replicate 1 and 12 reads in stress replicate 3, correspondingly. This indicated that 6 reads or less were on average mapped at each timepoint for half of all selected transcripts in the dataset of replicate 1. This suggested that working with FPKM values can misrepresent the true coverage levels of datasets and lead to modelling of transcript binding time-series with only a few reads mapped by χCRAC

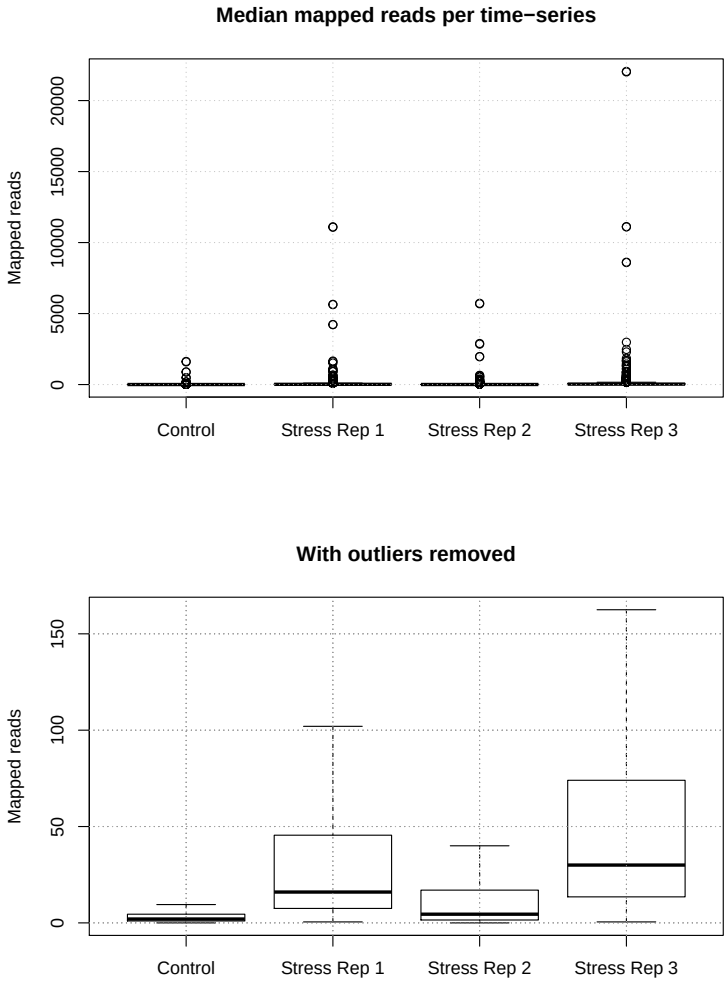


Figure 5.5: Boxplots of median values of mapped reads per Xrn1 time-series of each transcript. Each boxplot corresponds to the specified experimental replicate (1 replicate in control conditions, 3 replicates in stress conditions). The bottom panel shows boxplots with removed outliers.

in each timepoint. It was thus decided to work directly with raw counts, preserving information about the coverage of time-series for each transcript.

5.7.2 *Poisson observation model*

The decision to work with raw integer counts instead of a continuous measure such as FPKM resulted in a change to the definition of the observation model used by the proposed algorithm. The Gaussian observation model could no longer be used and the Poisson distribution was chosen instead as a standard choice for modelling raw sequencing counts. Indeed, the Poisson distribution expresses a probability of observing a given number of cross-linking events between Xrn1 and a transcript in a fixed interval of time given that they occur with a constant rate, which we model with the hidden function $\mathbf{f}$. The observation model is thus defined as below for N timepoints.

$$p(y_i|f_i) = \text{Pois}(f_i) \tag{5.20}$$

$$p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{N} \frac{{f_i}^{y_i}}{y_i!} \exp(-f_i) \tag{5.21}$$

Including the replicate information and following the notation y_i^k for the k-th replicate out of M replicates, the full likelihood of the data given the hidden process is given by:

$$\begin{aligned} p(\mathbf{y}|\mathbf{f}) = \prod_{k=1}^{M} \prod_{i=1}^{N} \frac{(f_i)^{y_i^k}}{(y_i^k)!} \exp(-f_i) = \\ \exp\left(M \sum_{i=1}^{N} (-f_i)\right) \prod_{k=1}^{M} \prod_{i=1}^{N} \frac{(f_i)^{y_i^k}}{(y_i^k)!} \end{aligned} \tag{5.22}$$

The difference between the Gaussian and Poisson observation models lies with their handling of noise. Under the Gaussian model, the zero-mean noise with the same variance was added to all observations, regardless of their value. In the Poisson distribution, the variance is equal to the mean, so the noise associated with each measurement depends on its value, with more noise applied to higher observation values.

Under the new Poisson observation model, the control time-series of Xrn1 binding was generated from the Poisson distribution with the mean estimated from the two cross-linking measurements mapped at 0 minutes in stress replicates 1 and 3. This

procedure is preferred to generating a synthetic time-course under the Gaussian noise model as it generates natural numbers, rendering any post-filtering unnecessary, and avoids estimating variance from only two measurements. Further, the normalisation procedure applied to all time-series was modified to scale them such that they start from the same number. This achieved the same as dividing each time-series by their first value at 0 minutes (as was done previously), but also preserved the original scale. All time-series were then rounded to the nearest integer. As rounding can introduce many zero values per time-series, the selected transcripts were filtered by the number of zero values in all time-courses, removing any transcripts with more than half of zero entries. Finally, we removed transcripts with more than 100 counts mapped in any timepoint due to taking factorial in the expression for $p(\mathbf{y}|\mathbf{f})$ (Eq. 5.22). Those transcripts with high number of counts should be processed by the original algorithm using the Gaussian observation model.

Now that we define the function values $\{f_i\}_{i=1}^N$ to be the rates of Poisson distributions (which must be positive), the prior of the hidden function $\mathbf{f}$ should be adjusted to have a non-zero mean (Eq. 5.23). We will further use an exponential transform to restrict the values of $\mathbf{f}$ to positive values.

$$p(\mathbf{f}) = N(\boldsymbol{\mu}, K) \tag{5.23}$$

5.7.3 *Computing marginal likelihood*

Using a Poisson likelihood $p(\mathbf{y}|\mathbf{f})$ makes it impossible to exactly compute the marginal likelihood. This ability for exact computation of marginal likelihood and our choice to use (continuous) FPKM values to represent binding response served as motivation for originally using the Gaussian observation model as described in Section 5.5.

The marginal likelihood is now given by the following integral:

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f})d\mathbf{f} =$$
$$= \int \exp\left(M \sum_{i=1}^{N} (-f_i)\right) \prod_{k=1}^{M} \prod_{i=1}^{N} \frac{(f_i)^{y_i^k}}{(y_i^k)!} \cdot \frac{1}{(\sqrt{(2\pi)^N |K|})} \exp\left(-\frac{1}{2}(\mathbf{f}-\boldsymbol{\mu})^T K^{-1}(\mathbf{f}-\boldsymbol{\mu})\right) d\mathbf{f} \tag{5.24}$$

Taking the terms constant in $\mathbf{f}$ outside of the integral yields the following expression:

$$p(\mathbf{y}) = \frac{1}{(\sqrt{(2\pi)^N |K|})} \cdot \prod_{k=1}^{M} \prod_{i=1}^{N} \frac{1}{(y_i^k)!} \cdot \int \exp\left(M \sum_{i=1}^{N} (-f_i) - \frac{1}{2}(\mathbf{f}-\boldsymbol{\mu})^T K^{-1} (\mathbf{f}-\boldsymbol{\mu}) \right) \prod_{k=1}^{M} \prod_{i=1}^{N} (f_i)^{y_i^k} d\mathbf{f} \tag{5.25}$$

The integral can be approximated as a Gaussian distribution using the Laplace's method (MacKay, 2003). This method approximates any single-mode distribution with a Gaussian distribution with the mean set as its mode and the variance computed using the curvature at the mode. The full derivations are given in Section B.2 of Appendix B.

As mentioned before, to restrict the values of $\mathbf{f}$ to positive values, which is a requirement for the Poisson observation model, we use the following reparameterisation:

$$\mathbf{f} = \exp(\mathbf{h}) \tag{5.26}$$

Under this reparameterisation, the resulting integral in the general form becomes:

$$p(\mathbf{y}) = \int p(\mathbf{y}|\exp(\mathbf{h}))p(\exp(\mathbf{h}))\exp(\mathbf{h})d\mathbf{h} \tag{5.27}$$

And thus, the integral to be approximated is given by:

$$\int \exp\left(M \sum_{i=1}^{N} (-\exp(h_i) - \frac{1}{2}(\exp(\mathbf{h}) - \boldsymbol{\mu})^T K^{-1} (\exp(\mathbf{h}) - \boldsymbol{\mu}) \right) \cdot \prod_{k=1}^{M} \prod_{i=1}^{N} (\exp(h_i))^{y_i^k} \prod_{i=1}^{N} \exp(h_i) d\mathbf{h} \tag{5.28}$$

5.7.4 *Model selection*

The expression in Eq. 5.28 was optimised for each transcript, setting the prior mean μ to the mean of the time-series in control conditions. This recovered the mean x_0 and the covariance C of the approximating Gaussian. The amplitude and lengthscale hyperparameters for the prior covariance were empirically set for each transcript as before.

The final expression for the marginal likelihood $p(\mathbf{y})$ under each model was computed as follows, denoting the integrand evaluated at the optimised mean as $P(x_0)$:

$$p(\mathbf{y}) \approx \frac{\sqrt{|C|}}{\sqrt{|K|}} \cdot \prod_{k=1}^{M} \prod_{i=1}^{N} \frac{1}{(y_i^k)!} \cdot P(x_0) \tag{5.29}$$

For numerical stability, log-marginal likelihoods were computed as before and transcripts with Bayes factors greater or equal to 10 were taken as exhibiting significant changes in Xrn1 binding between control and stress conditions.

5.7.5 *Differential binding analysis*

The Xrn1 differential binding analysis was performed using two Xrn1 χCRAC replicate datasets and the RNA-seq dataset in stress conditions. The normalisation procedure applied to all time-courses is summarised below. The raw counts for all transcripts were scaled within each timepoint by a constant factor such that the sum of all counts at all timepoints in all experimental replicates was the same. The counts were linearly interpolated across the set of timepoints common to all time-series. Xrn1 binding time-series were normalised by the RNA-seq time-series in order to exclude the effect of changes in transcript abundance on changes in Xrn1 binding. The Xrn1 binding time-series in control conditions was generated using the counts mapped in the χCRAC experiments before the stress induction (at 0 minutes). The control time-series was generated from a Poisson distribution with the mean given by the average of the specified measurements for each transcript. All time-series were rescaled to start at the same value (equal to the first value in the control time-series). Finally, all time-series were rounded to the nearest integer number. The set of remaining transcripts was filtered by the number of zero values in all time-series. Those with more than half of zero values were excluded from the analysis. Additionally, transcripts with a count > 100 in any timepoint were removed from the analysis due to computing factorials of observed values under the Poisson observation model. These transcripts (148) were tested with the original algorithm using Gaussian regression.

The final set of transcripts to be tested for differential binding under stress contained 2872 transcripts. 627 were selected as significantly changing their Xrn1 binding patterns in stress response compared to the synthetic control time-series. Out of the 148 transcripts with at least one high count in all time-series, 97 were selected for the Gaussian regression analysis (as described in Section 5.7.1) and 29 were selected as demonstrating significant changes in binding. Thus, a total of 656 transcripts (or

22% of all tested transcripts) were selected as showing significant changes in Xrn1 binding in response to nutrient stress.

These selected transcripts belonged to 10 different RNA classes. The pie chart in Fig. 5.6 shows what percentage of each RNA class showed changes in Xrn1 binding during stress, similar to the analysis for Pol II presented in the paper (Fig. 4a). The percentages of differentially bound transcripts in each class (shown in brackets for each class) were computed as ratios between the numbers of selected annotated transcripts in each RNA class and the total number of annotated transcripts in that class. As the pie chart only shows percentages, the actual numbers of transcripts are shown in Fig. 5.8.

Additionally, the percentage of differentially bound transcripts was computed out of the number of *tested* annotated transcripts in each RNA class rather than out of *all* annotated transcripts (numbers are shown in Fig. 5.8). This was important to examine as only a subset of transcripts, which passed various filtering during the normalisation procedure, was selected for testing with the algorithm. The pie chart in Fig. 5.7 shows the percentages of differentially bound transcripts in each class, computed as the ratio between the selected annotated transcripts and the tested annotated transcripts in that class.

When considering the proportions of transcripts with changing Xrn1 binding patterns out of *all* annotated transcripts available for each class, the largest changes were detected among the transport RNAs (tRNAs), ribosomal RNAs (rRNAs), small nuclear RNAs (snRNAs), and small nucleolar RNAs (snoRNAs). The changes associated with the tRNA class were especially prominent, with more than 100 transcripts selected as differentially bound by Xrn1 under stress. When considering only those transcripts in each class that were *tested* by the algorithm and computing the proportions of changing transcripts from those, large changes in Xrn1 binding were also observed among the cryptic unstable transcripts (CUTs). In general, the changes detected among the *tested* transcripts appeared to be fairly evenly spread between almost all classes, illustrating the diversity of Xrn1 as an interacting molecule.

Considerable changes in Xrn1 binding, demonstrated by the high numbers of selected targets, were also found among the Xrn1-sensitive unstable transcripts (XUTs), stable uncharacterised transcripts (SUTs), and anti-sense transcripts. Finally, the largest number of selected targets among all classes belonged to the class of protein-coding transcripts (298 transcripts), consistent with the role of Xrn1 in mRNA decay (Lebreton and Séraphin, 2008). The small percentage of the selected protein-coding transcripts could be potentially explained by either the selection of tested transcripts, the associated noise levels or the increased targeting of other RNA classes in stress response.

The results of the class-based differential binding analysis are supported by the suggested roles of Xrn1 in rapid degradation of mature tRNAs (Chernyakov et al., 2008) and in 5'-end processing of snoRNAs and rRNAs in yeast (Lee et al., 2003; Petfalski et al., 1998). Additionally, the characterisation of the Xrn1-dependent cryptic transcriptome previously described the XUTs, SUTs, and CUTs classes to have extensive interactions with Xrn1 (Van Dijk et al., 2011).

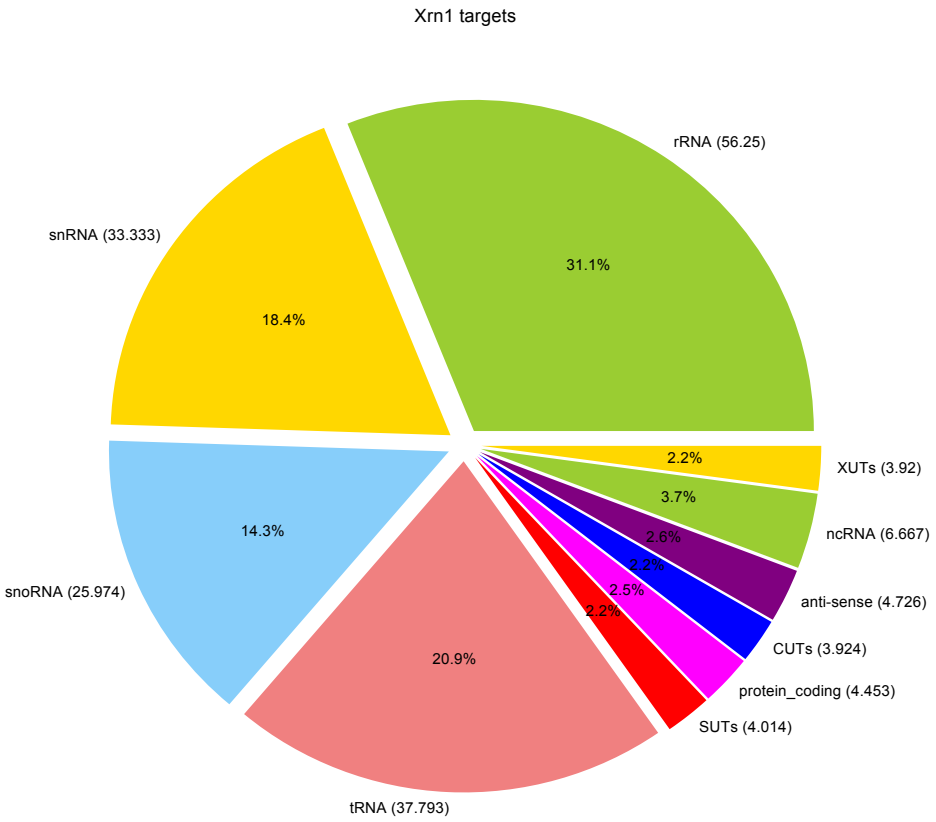


Figure 5.6: Percentages of RNA classes with detected changes in Xrn1 binding under stress. Numbers in brackets for each class indicate the percentage of the whole annotated class that was selected as targets showing differential Xrn1 binding. Numbers in slices of the chart (which sum to 100%) shows how big the differentially binding proportion of each class is compared to other classes.

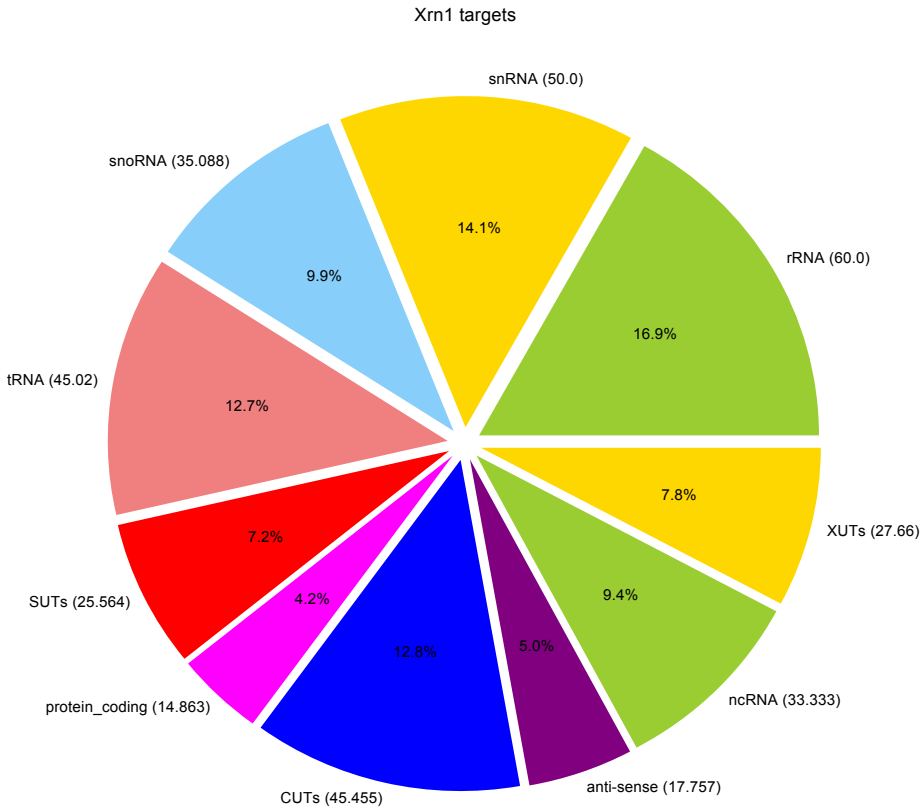


Figure 5.7: Percentages of RNA classes with detected changes in Xrn1 binding under stress. Numbers in brackets for each class indicate the percentage of the *tested* class that was selected as targets showing differential Xrn1 binding. Numbers in slices of the chart (which sum to 100%) shows how big the differentially binding proportion of each class is compared to other classes.

All annotations	
CUTs	2421
SUTs	847
XUTs	1658
anti-sense	402
ncRNA	15
protein-coding	6692
pseudogene	21
rRNA	16
snRNA	6
snoRNA	77
tRNA	299

Tested annotations	
CUTs	209
SUTs	133
XUTs	235
anti-sense	107
ncRNA	3
protein-coding	2005
pseudogene	1
rRNA	15
snRNA	4
snoRNA	57
tRNA	251

Selected annotations	
CUTs	95
SUTs	34
XUTs	65
anti-sense	19
ncRNA	1
protein-coding	298
pseudogene	0
rRNA	9
snRNA	2
snoRNA	20
tRNA	113

Figure 5.8: Left: number of annotated transcripts in each RNA class. Middle: number of annotated transcripts in each RNA class that were tested by the algorithm. Right: number of annotated transcripts in each RNA class that were selected as differentially bound by Xrn1.

5.7.6 *Example binding patterns of Xrn1*

Examples of stress-related Xrn1 regulation patterns included transient increase in binding. This increase happened very rapidly for some transcripts, within the first 5 minutes of stress induction. This is illustrated with the small nuclear RNA snR6 and the transcript YBR191W-A, an uncharacterised open reading frame (ORF), in top two rows of Fig. 5.9 (all curves in the figure show the polynomial fit to the data points for illustration purposes). These two transcripts, however, differed in the dynamics of their total abundance during nutrient stress. The abundance of snR6 was increasing during the first 15 minutes of the stress response and then remained roughly constant for the next 25 minutes, during which Xrn1 binding to snR6 reproducibly continued to increase. The interaction between YBR191W-A and Xrn1 showed a rapid increase in binding in both experimental replicates, however, varying in steepness (shown with blue and red curves). In both replicates, the change in Xrn1 binding was transient, reducing down to a steady value during the following 40 minutes. This behaviour was somewhat mirrored in the time-course of transcript abundance, which was going down during the first 5 minutes and then increasing again until around 15 minutes into the stress response. This example demonstrates that the change in Xrn1 regulation under stress was not simply the result of more transcripts present in the cell available for interaction.

The transcript abundance and its interaction with Xrn1 are undeniably tightly interconnected. More available transcript could result in increased binding with Xrn1, which in turn would lead to decreasing levels of the mature transcript. This relationship could be captured via dynamical modelling, which can be invaluably informed with the χCRAC technology. This will be further addressed in the final section of this chapter.

Another pattern of Xrn1 regulation included decreased interactions with transcripts. The decrease in binding varied in its duration. For the transfer RNA tD(GUC)I2, the nutrient shift led to a rapid decrease in Xrn1 binding during the first 5-8 minutes, followed by its recovery to a more stable level close to the initial condition (third row in Fig. 5.9). For another uncharacterised ORF YMR013W-A, the Xrn1 binding continued to decrease further into the adaptation response, following the initial rapid decline (last row in Fig. 5.9). The abundance of both of these transcripts was increasing for the first 15 minutes and reaching a more stable level after that. This once again demonstrates that changes in Xrn1 binding are not simply mirroring the increased or decreased presence of the transcript in the cell. Instead, Xrn1 regulation can act as a separate mechanism for controlling gene expression.

It should be noted that the differences between the Xrn1 binding time-series of the two experimental replicates are further amplified by the mismatching timepoints, at which the corresponding χCRAC experiments were performed. Specifically, the second replicate (shown in all plots in red) was only collected at 4 timepoints: 0, 4, 16, and 48 minutes after stress. Thus, it is possible that some very rapid interactions (during the first 4 minutes) or those taking place in the middle of the adaptation response (between 15 and 48 minutes after stress) would not be captured in the time-series of that experimental χCRAC replicate.

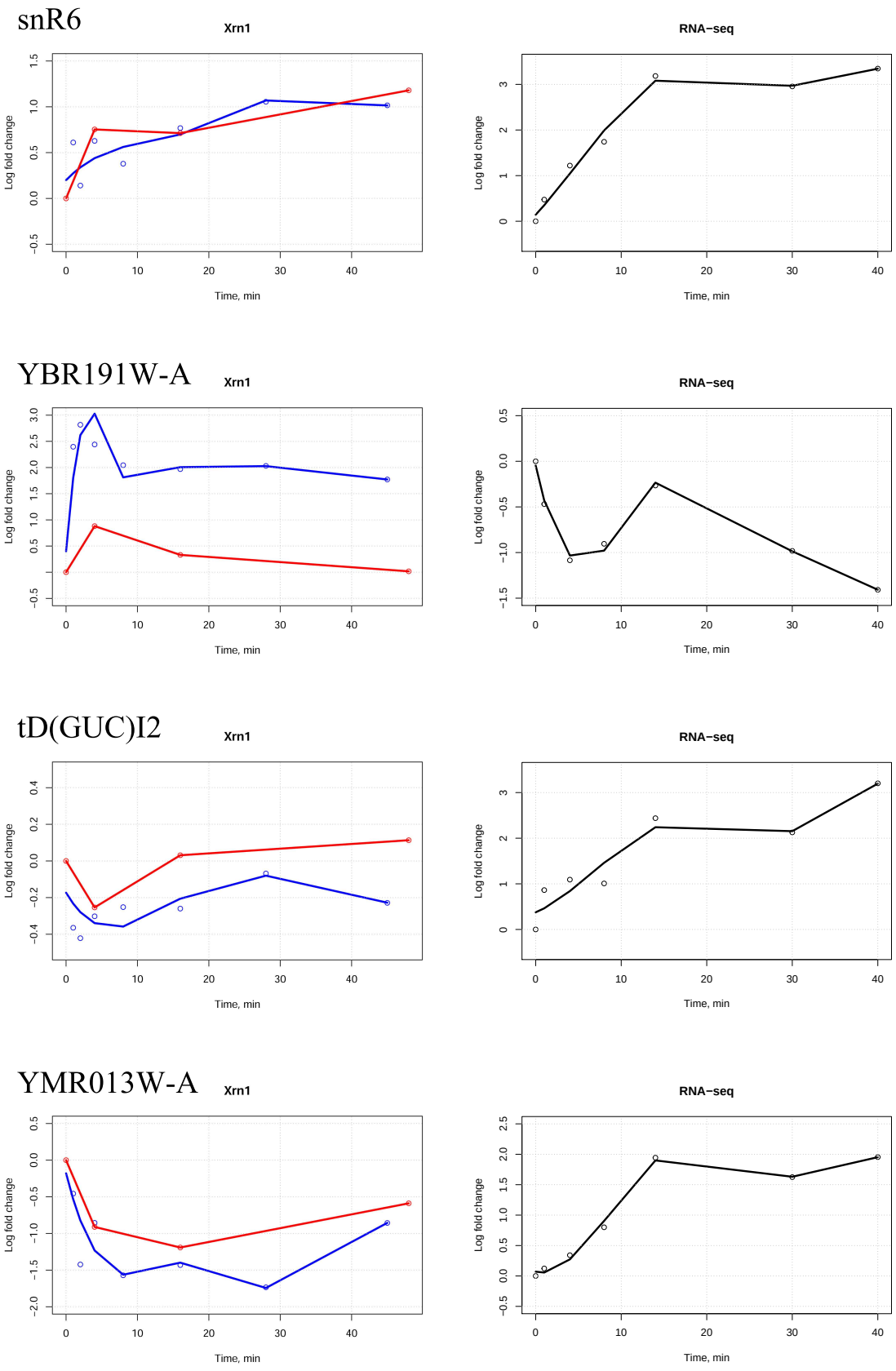


Figure 5.9: Examples of Xrn1 binding patterns. Top two transcripts demonstrate increase (and bottom two - decrease) in Xrn1 binding during stress. For each transcript, the Xrn1 binding time-series and the RNA-seq time-series in stress conditions are shown. Circles indicate time-series values, curves were plotted with polynomial regression. Blue and red curves show replicate time-series.

5.8 POISSON REGRESSION ANALYSIS OF NAB3 AND POL II TIME-SERIES

In order to compare the outputs of the analyses using Gaussian and Poisson noise models, it was decided to perform the Poisson regression analysis on the raw counts of the Nab3 and Pol II datasets presented in the paper. The Pol II and Nab3 datasets were also examined for their coverage levels. Reads were mapped and raw counts were generated as described in the Methods section "*Processing of raw sequencing data*", using the yeast genome annotation by Nagalakshmi et al.

The transcripts were selected for testing with the same normalisation procedure as described in Section 5.7.1, with the exception of generating the control time-series as real control data for Nab3 and Pol II binding was used instead. Briefly, the Nab3 and Pol II datasets were scaled such that the sum of counts was the same at each timepoint and the counts were linearly interpolated between non-equal timepoints. For the Nab3 analysis, the counts were normalised by the corresponding time-series of Pol II. Specifically, the control Nab3 time-series was normalised by the mean value of the Pol II control time-series, following the assumption that there should be no significant changes in the absence of stress. The stress time-series of Nab3 were normalised by the average time-course of the 3 Pol II stress replicates. This normalisation was performed in order to account for the co-transcriptional binding of Nab3. For both Nab3 and Pol II analyses, the resulting time-series were divided by their first value in each condition, ensuring that all time-series start at 1 a.u. at the timepoint of 0 minutes. Transcripts with too many zero values were excluded from the analysis as before.

The following sections examine the coverage levels of Pol II and Nab3 datasets, present the results of the Poisson regression analysis and compare them to the results obtained by the Gaussian regression analysis. In order to perform the comparison, the analysis using the Gaussian noise model was performed on the same raw count data, following the equivalent normalisation procedure as described above and in Section 5.7.1.

5.8.1 *Nab3 interacting partners*

After the normalisation procedure described above, 5134 Nab3-interacting transcripts were selected for differential binding testing. For these selected transcripts, the medians of mapped reads per timepoint were examined in each replicate (Fig. 5.10). The medians of the distributions corresponding to the control replicate and the stress replicate 2 were 17 reads and 19.5 reads, correspondingly. The median corresponding

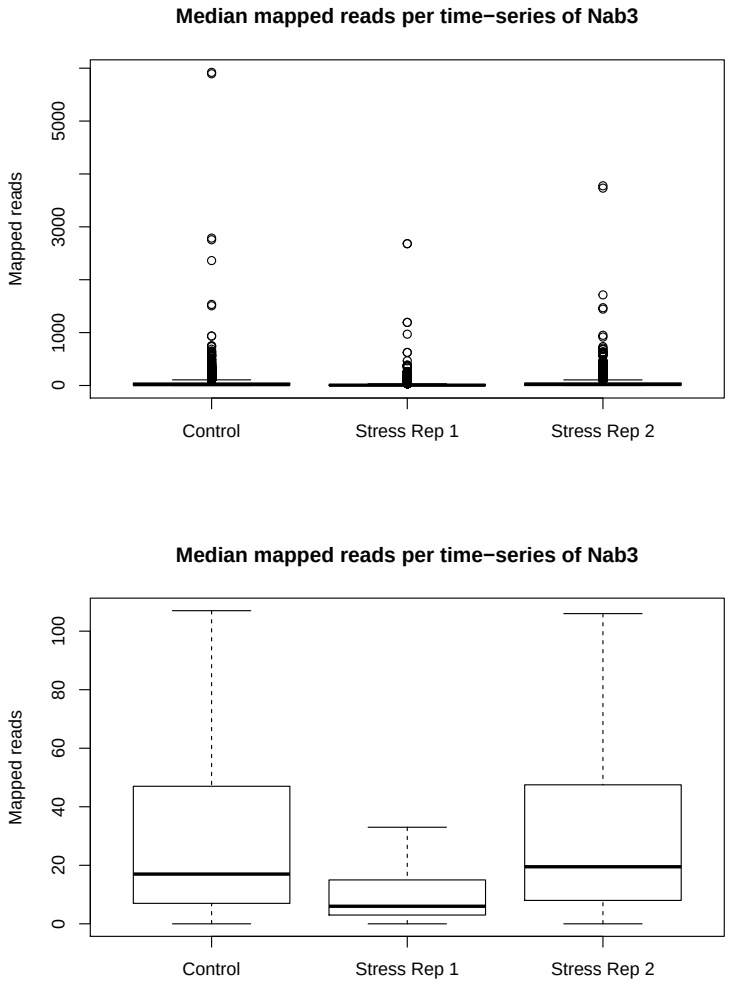


Figure 5.10: Boxplots of median values of mapped reads per Nab3 time-series of each transcript. Each boxplot corresponds to the specified experimental replicate. The bottom panel shows boxplots with removed outliers.

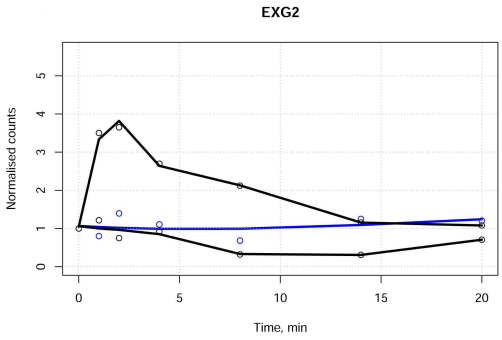
to the stress replicate 1 was lower (6 reads) and more comparable to the statistics of the Xrn1 datasets.

In order to perform the Poisson regression analysis for differential binding with Nab3, the time-series were normalised, rounded to integers, and filtered for zero values (the procedure was described in more detail in Section 5.7.5). Out of the 3163 transcripts selected for testing with the Poisson regression analysis, 731 were found to have significant changes in Nab3 binding under stress. 79% of these Nab3 regulatory targets were also selected by the Gaussian regression analysis, demonstrating a high degree of consensus of the two analyses on this dataset. Additionally, 153 transcripts were excluded from the Poisson regression analysis as they had at least one high count ($>$ 100 reads) in the time-series. They were tested with the Gaussian regression

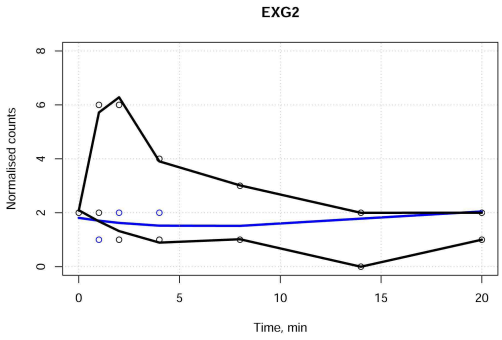
algorithm, identifying 147 transcripts with significant changes in binding. Thus, 26% of all tested transcripts (878 out of 3316) demonstrated stress-related changes in Nab3 binding patterns.

The results of the Gaussian regression analysis on the Nab3 raw count data were less stringent compared to the Poisson regression, selecting 2931 targets. This is likely the result of the noise handling by the two observation models. Under the Gaussian model, the noise with equal variance is added to explain all measurements, regardless of their magnitude, whereas under the Poisson model, higher measurement values can have larger associated noise. Thus, under the Gaussian noise model, high counts outside of the expected noise bands will be very improbable. In the Gaussian regression analysis, the noise variance was empirically set based on the variance of the control time-series. Therefore, stress time-series counts higher than what was expected based on the control binding profile even in one replicate is likely to lead the algorithm to call the time-series significantly different. This is illustrated with the transcript EXG2, selected by the Gaussian regression analysis but not by the Poisson regression analysis (Fig. 5.11(a), 5.11(b)). Another disadvantage arises from working with normalised values that do not preserve the original scale of the raw counts. In the case of EXG2 and another transcript SIP3 (Fig. 5.11(c), 5.11(d)), involved in sterol transfer (both selected by the Gaussian model but not the Poisson), the majority of timepoints only have 1 or 2 mapped reads.

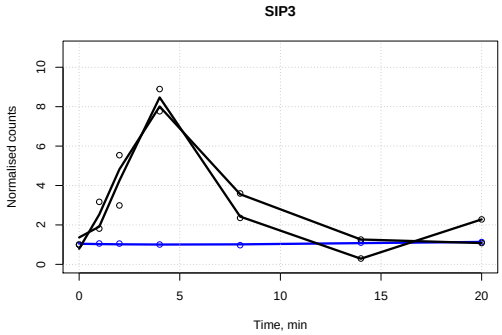
Setting the expected noise using the control time-series under the Gaussian model renders the algorithm unlikely to detect changes between time-series if the control time-course itself exhibits variation. While our assumption was that the time-series should be approximately unchanging in control conditions, it can be still beneficial for the algorithm to be stable against technical noise. This is illustrated with the cryptic unstable transcript CUT859, which was selected by the Poisson regression analysis but not the Gaussian regression analysis (Fig. 5.11(e), 5.11(f)). The transcript SPC2, belonging to a signalling complex, demonstrates an example binding profile selected by both analyses (Fig. 5.11(g), 5.11(h)).



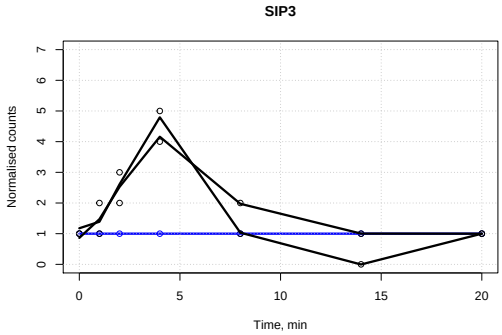
(a) Gaussian regression on EXG2



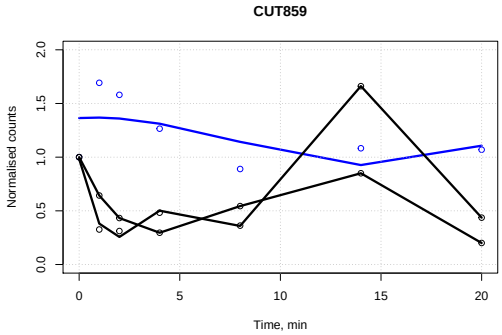
(b) Poisson regression on EXG2



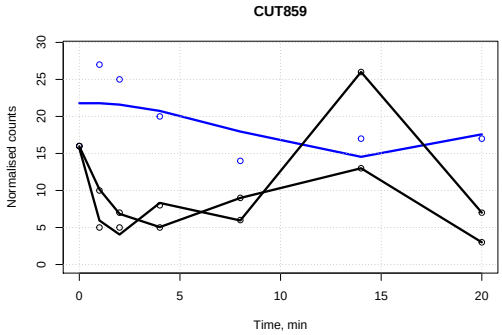
(c) Gaussian regression on SIP3



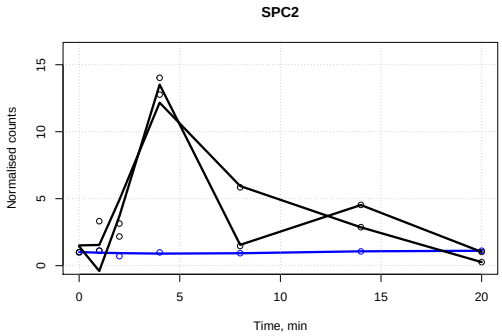
(d) Poisson regression on SIP3



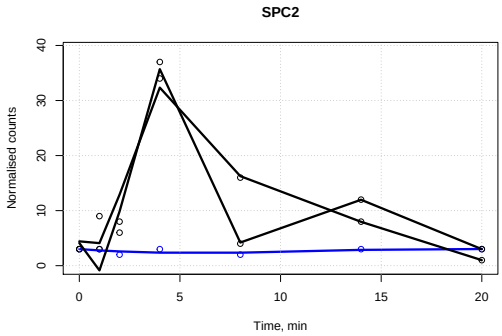
(e) Gaussian regression on CUT859



(f) Poisson regression on CUT859



(g) Gaussian regression on SPC2



(h) Poisson regression on SPC2

Figure 5.11: Nab3 binding profiles of transcripts selected by the Gaussian and Poisson regression analyses. Circles indicate time-series values, curves were plotted with polynomial regression. Blue corresponds to the time-series in control conditions, black corresponds to the replicate time-series in stress conditions. On the x-axis, minutes since the nutrient shift are shown. On the y-axis, the cross-linking values are shown, transformed according to the normalisation procedure corresponding to the Gaussian or Poisson observation model.

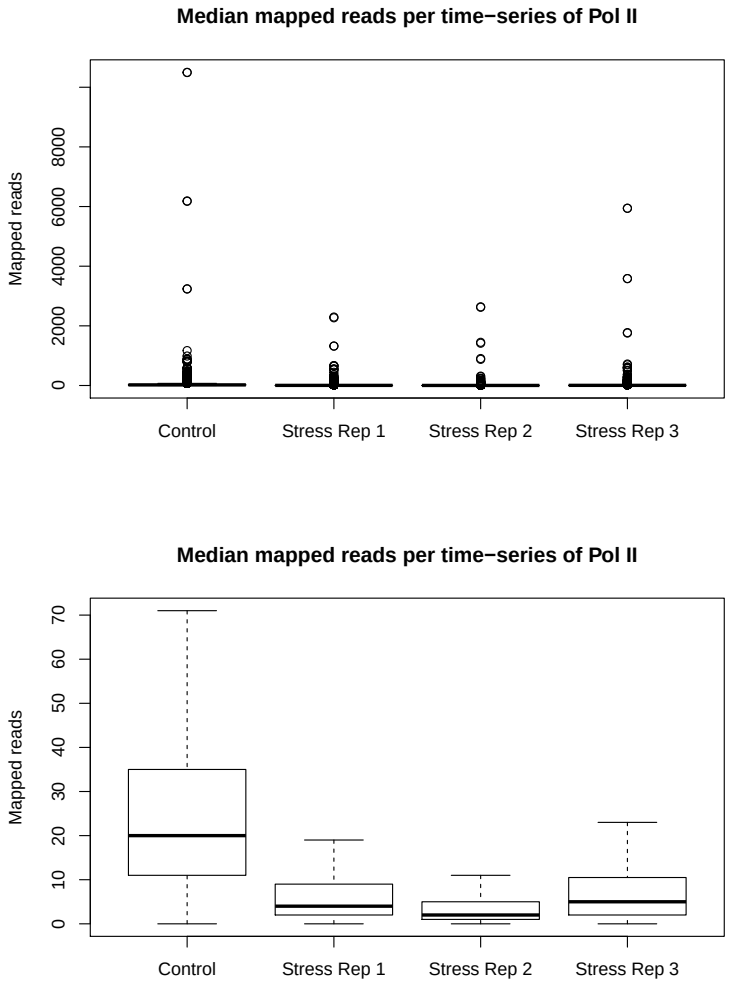


Figure 5.12: Boxplots of median values of mapped reads per Pol II time-series of each transcript. Each boxplot corresponds to the specified experimental replicate. The bottom panel shows boxplots with removed outliers.

5.8.2 *Pol II interacting partners*

The medians of mapped reads per timepoint were examined for the 7068 Pol II-interacting transcripts, selected for the differential binding testing with the normalisation procedure outlined at the beginning of this section. The median of the distribution for the Pol II control dataset was a reasonably high count of 20 reads (Fig. 5.12). In contrast, the medians for the 3 stress datasets were much lower: 4, 2, and 5 reads. As half of the considered transcripts had (in the worst case) 2 reads mapped on average at each timepoint, the Poisson distribution would provide a good candidate for modelling the likelihood of a measurement given the hidden function value, $p(y_i|f_i)$.

Following the normalisation procedure for the Poisson noise model and removing the transcripts with high counts, 2732 out of 6034 tested transcripts were selected by the Poisson regression analysis as demonstrating significant changes in Pol II binding under stress. In contrast, the Gaussian regression analysis identified 1782 targets, of which 45% agreed between the two analyses. This lower percentage of the overlapping targets selected by both analyses, compared to the 79% overlap in the analysis of the Nab3 binding data, is consistent with the lower average coverage of the Pol II dataset that was especially evident for stress replicates.

Additionally, 880 transcripts, which had at least one high count in their time-series, were processed with the Gaussian regression analysis, identifying 477 transcripts with significant changes in Pol II binding. Thus, 46% of all tested transcripts (3209 out of 6914) demonstrated changes in Pol II binding patterns during the adaptation response.

In order to further characterise the consensus between the two regression analyses, the coverage levels of the transcripts selected by both analyses were examined. As a Gaussian distribution provides an increasingly better approximation for a Poisson distribution as the Poisson rate grows, it could be plausible that the selections of the two regression analyses would agree on transcripts with higher coverage levels. Fig. 5.13 shows the histograms of the median coverage levels per gene for all genes that were tested with the Poisson regression analysis (left histogram) and for the consensus targets selected by both analyses (right histogram). It is evident from the histograms that the consensus targets span a similar range of coverage levels as does the set of the tested transcripts, with the majority of the consensus targets having a median coverage level of less than 10 reads per timepoint. Thus, a larger coverage level on average of a transcript was not enough to explain the consensus between the two observation models.

Another difference between the models comes with their ways of handling noise. As previously noted, under the Gaussian observation model, the variance of the noise parameter is defined using the variance of the control time-series. Thus, if the binding time-series in control conditions itself exhibited variation then it will be harder for the Gaussian regression analysis to detect changes between all time-series as equal noise is added to all measurements. To investigate the effect of the variation in transcript's binding in control conditions on the models' selections, I computed the variance of the control time-series for all transcripts that were tested with the Poisson regression analysis. Fig. 5.14 shows these variances on the log-scale, coloured in pink. On top of it, the green histogram shows the binding variances in control conditions for the targets selected by the Poisson regression analysis, while the overlapping targets selected by both analyses are shown in blue. The targets selected by the Poisson

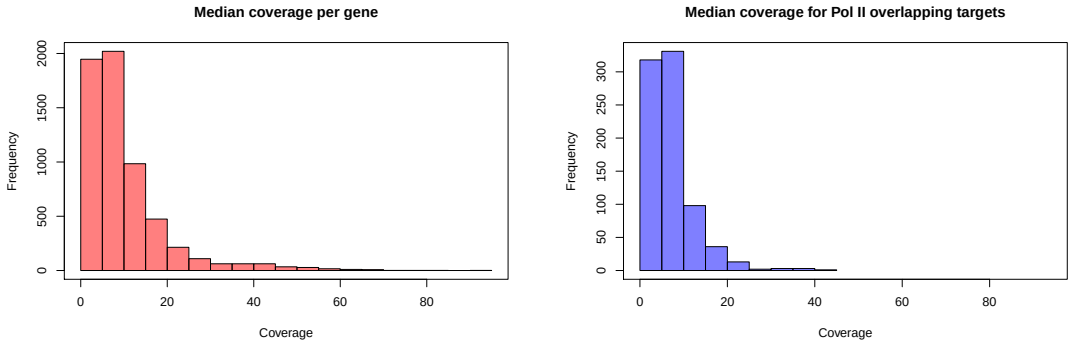


Figure 5.13: Left: a histogram showing median coverage levels for all genes that were tested with the Poisson regression analysis for differential binding with Pol II under stress. The medians were computed across each time-series in control and stress conditions and then averaged across conditions for each gene. Right: a histogram showing median coverage levels for all genes that were selected by both Poisson and Gaussian regression analyses as differentially bound by Pol II under stress.

regression analysis demonstrate various variance levels of their control time-series, spanning the whole scale exhibited by all tested transcripts. In contrast, those targets that were also selected by the Gaussian regression analysis systematically have lower variance of their control binding profiles (Fig. 5.14 and Fig. 5.15). This demonstrates that the consensus between the Poisson and Gaussian regression analyses is affected by the combination of the transcript's coverage level and the noise associated with its binding profile in control conditions. Similar analysis for the Nab3 dataset is provided in Section B.3 of Appendix B.

The examples in Fig. 5.16 illustrate the differences between the selections of the two regression analyses. Gaussian model analysis selects the transcript MTR10, involved in nuclear transport, while the Poisson regression analysis does not (Fig. 5.16(a), 5.16(b)). In one replicate, transcription of MTR10 is strongly increased, but in the other replicate, the changes are much closer to the control binding trace. In contrast, a metabolic transcript GLN4 is selected by the Poisson model, but not the Gaussian, likely due to the noise associated with the control time-series (Fig. 5.16(c), 5.16(d)). Finally, the ubiquitin gene transcript UBX6 provides an example selected by both analyses (Fig. 5.16(e), 5.16(f)).

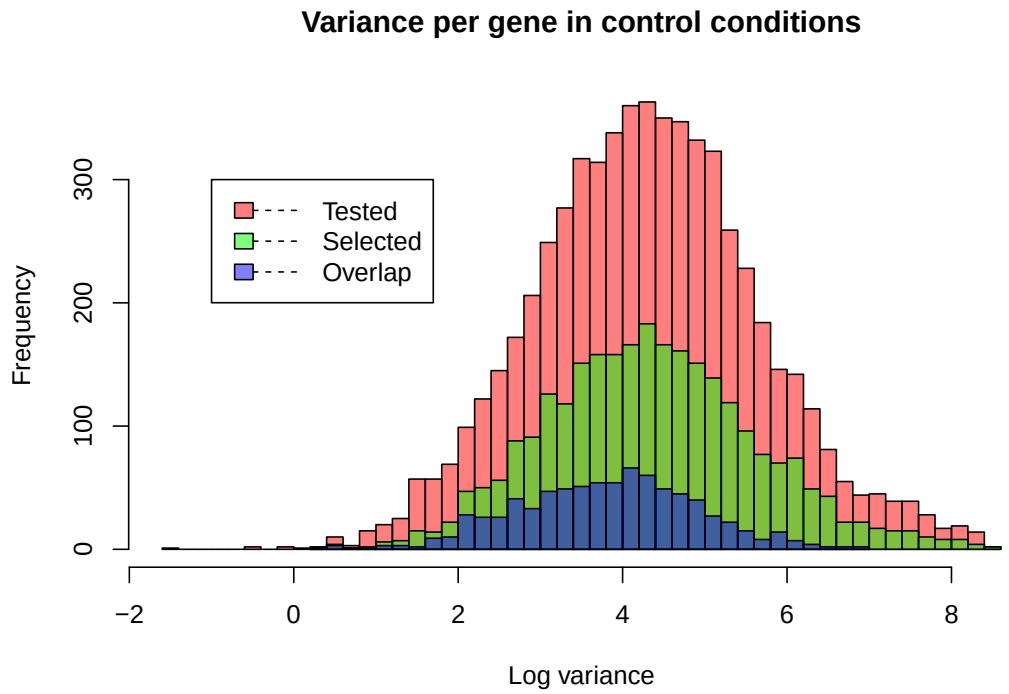


Figure 5.14: A histogram showing the log-variance of time-series in control conditions for all genes that were tested with the Poisson regression analysis for differential binding with Pol II under stress (coloured in pink). Overlaid are shown the corresponding histograms for the targets selected by the Poisson regression analysis (in green) and for the targets selected by both analyses (in blue).

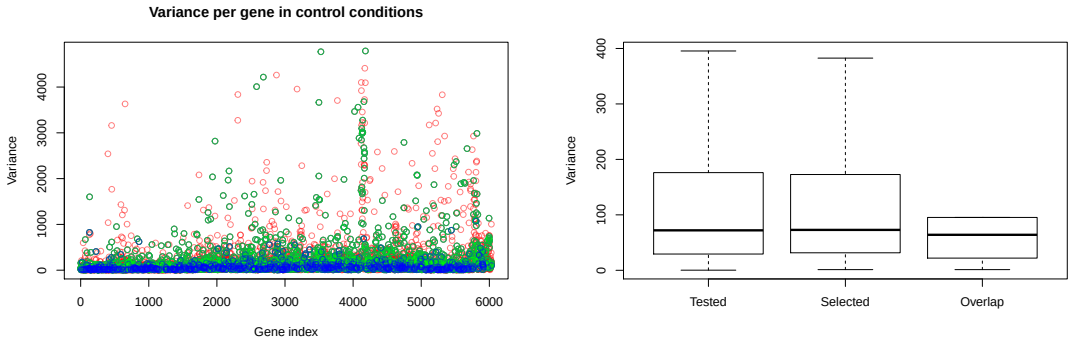
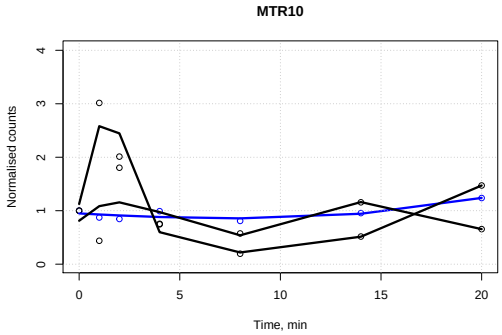
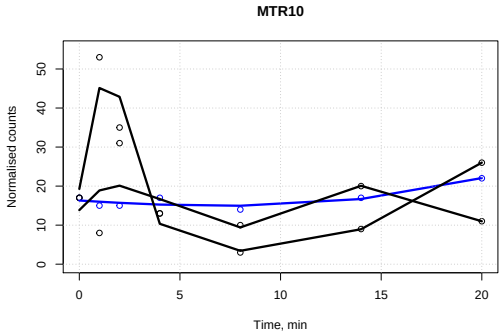


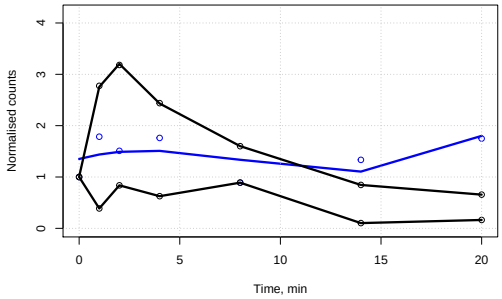
Figure 5.15: Further illustration of the distributions shown in Fig. 5.14. Left: variance of control time-series (shown on y-axis) for all tested transcripts (pink), for the transcripts selected by the Poisson regression analysis (green) and for the transcripts selected by both analyses (blue). X-axis indexes plotted transcripts. Right: boxplots of distributions of binding variance in control conditions for all tested transcripts, the transcripts selected by the Poisson regression analysis, and the transcripts selected by both analyses. Outliers are not shown.



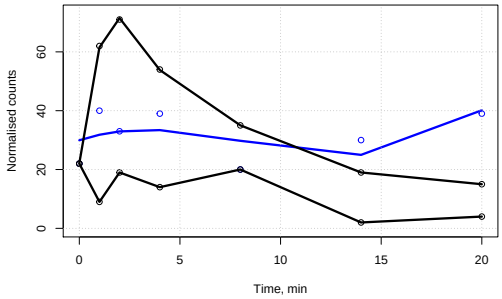
(a) Gaussian regression on MTR10



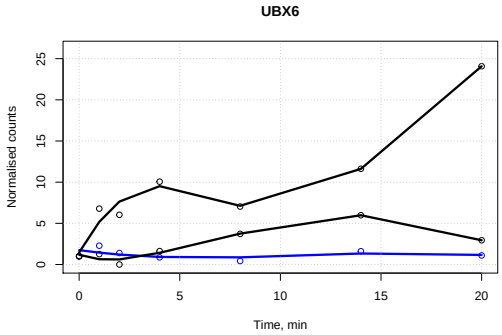
(b) Poisson regression on MTR10



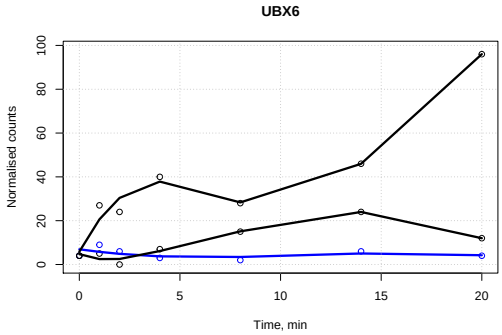
(c) Gaussian regression on GLN4



(d) Poisson regression on GLN4



(e) Gaussian regression on UBX6



(f) Poisson regression on UBX6

Figure 5.16: Pol II binding profiles of transcripts selected by the Gaussian and Poisson regression analyses. Circles indicate time-series values, curves were plotted with polynomial regression. Blue corresponds to the time-series in control conditions, black corresponds to the replicate time-series in stress conditions. On the x-axis, minutes since the nutrient shift are shown. On the y-axis, the cross-linking values are shown, transformed according to the normalisation procedure corresponding to the Gaussian or Poisson observation model.

5.9 CONCLUSIONS AND OUTLOOK

In this chapter, I introduced an improved UV cross-linking protocol χCRAC, which generates RNA-protein interaction maps at high temporal resolution (van Nues et al., 2017). I presented the algorithm for modelling χCRAC data and identifying target transcripts, which demonstrate significant changes in cross-linking profiles between conditions. This algorithm was applied in the context of studying the role of RNA degradation during nutrient stress.

First, we investigated the role of co-transcriptional degradation by mapping interactions of the yeast transcription termination factor Nab3. We detected widespread dynamic stress-related changes in Nab3 binding, often occurring within only a few minutes from the onset of the adaptation response. The changes in Nab3 were not a simple result of increased or decreased Pol II cross-linking, thus suggesting that transcription termination can provide a fast additional regulatory mechanism alongside transcriptional control.

It was then hypothesised that other RBPs influencing RNA expression could have similar dynamic behaviour in response to stress. This was investigated by analysing the cross-linking dynamics of the cytoplasmic degradation factor Xrn1. Closer examination of the Xrn1 dataset revealed that for many transcripts, a low number of reads was mapped at each timepoint. This motivated modifying the proposed algorithm to use a Poisson observation model and performing the analysis on raw sequencing counts. We found many transcripts (22% of all tested candidates) whose interaction patterns with Xrn1 was altered in response to stress, often involving fast and transient changes in binding. When testing for differential dynamic response, changes in overall transcript abundance was taken into account, exposing the control mechanism associated with the 5′-3′ decay pathway.

The proposed algorithm aimed at testing transcripts for differential binding by the protein of interest is powerful not only in its ability to model longitudinal data rather than individual data points due to being based on the Gaussian Process model, but also in its flexibility to incorporate different noise models. I presented two versions of the algorithm using a Gaussian noise model and a Poisson noise model, correspondingly, and compared the outputs of both analyses.

The advantage of the analysis using Gaussian regression comes with the ability to exactly compute the marginal likelihood of the data under competing models, providing a very fast computational method suitable for testing thousands of transcripts in seconds. However, the assumption of the model might not be suitable for a dataset with low counts. I presented an approximate method for computing the marginal likelihood of the data when using a Poisson regression model. As the approximation

involves optimising parameters in order to recover the mean and variance of the approximating Gaussian, this method is slower (with 5000 transcripts taking $\approx$ 12 minutes) and might fail to recover parameters for some transcripts. Other approaches for approximating the marginal likelihood are possible (apart from the Laplace's method presented in this chapter), e.g. expectation propagation (Cseke and Heskes, 2011).

Transcripts with high counts in their time-series were tested with Gaussian regression analysis, facilitating a combined approach that uses different models on the corresponding parts of the dataset, filtered by the coverage levels. This emphasises the importance of ensuring the agreement between the modelled data and the underlying assumptions of the used model. The agreement between the models' outputs was higher on the χCRAC datasets with larger coverage levels as the Gaussian distribution provides a reasonable approximation of the Poisson distribution as its rate increases.

The results obtained by the χCRAC data analysis suggested that the degradation process can act via different pathways and be highly dynamic, providing a versatile control mechanism of gene expression regulation. It is expected that applying χCRAC to individual nucleases or other RBPs mediating RNA expression will provide valuable insights in the determinants of transcription and degradation rates.

However, having collected interactions data from the main molecular actors of transcription and co-transcriptional and cytoplasmic degradation, we asked whether their dynamic behaviour is already enough to explain transcript abundance. Specifically, it would be of great interest to dissect the relative contributions of these pathways to the overall expression, uncovering the stress-specific regulation profile for each transcript. A useful metaphor to consider this idea uses the notion of "control knobs", which together determine the abundance of a mature transcript (Fig. 5.17). Each knob corresponds to the production rate, measured by Pol II, co-transcriptional degradation rate, corresponding to Nab3, and cytoplasmic degradation rate, informed by Xrn1. The question is, can one recover the settings of these knobs that give rise to the abundance dynamics of each transcript?

These considerations motivated the conception of a dynamical model for RNA expression, which aims to explain transcript abundance using the binding patterns of Pol II, Nab3, and Xrn1. The theoretical formulation of the proposed model was devised by myself, David Schnoerr, Edward Wallace, and Guido Sanguinetti. David Schnoerr advised me on the early exploratory studies of simultaneous dynamical modelling of the RNA and proteins' time-series. Further and ongoing work involving implementing the proposed model and applying it to the χCRAC and RNA-seq datasets is performed by David Schnoerr.

$$var(\rho_t) \cong$$

$$Avar(\pi_t) \quad + \quad Bvar(\nu_t) \quad + \quad Cvar(\chi_t)$$

Figure 5.17: Graphical illustration motivating the proposed dynamical model for RNA expression under stress. Denoting time-series data for RNA-seq as ρ_t, for Pol II χCRAC as π_t, for Nab3 χCRAC as ν_t, and for Xrn1 χCRAC as χ_t, we would like to seek a decomposition for the variance in transcript's abundance in terms of the variances of its binding with the mediators of transcription and degradation.

5.9.1 *Modelling dynamics of RNA expression*

We propose a dynamical model for RNA expression regulation under nutrient stress that aims to explain changes in a transcript's abundance using the dynamics of its transcription and degradation rates, which are estimated from the χCRAC datasets. Specifically, all rates are modelled as time-dependent and the goal is to decompose the contributions to RNA expression from the nuclear and cytoplasmic degradation pathways. This decomposition would recover a stress-related regulation profile for each modelled transcript.

The need to combine various datasets quantifying RNA abundance (either for total RNA or specific populations bound by the protein of interest) in successive experiments as variables in a unified global model recapitulates the importance of carefully normalising the datasets in question. Apart from the *within-sample* normalisation strategies used in RNA-seq data analyses (such as normalisation for gene length and library size), *between-sample* normalisation needs to be additionally considered, which enables capturing changes in the composition of the probed RNA samples. An effective normalisation method was proposed by Robinson and Oshlack, which recovers scaling factors that account for the sampling properties of the compared RNA-seq datasets.

Following a similar idea, we define a Poisson observation model, which relates the raw χCRAC counts ($\hat{\pi}$ for Pol II, $\hat{\nu}$ for Nab3, $\hat{\chi}$ for Xrn1) and RNA-seq counts ($\hat{\rho}$) to the hidden variables that represent the true numbers of transcripts interacting with the corresponding protein (π, ν, χ). Those true numbers and the true total cellular amount of RNA (ρ) are assumed to be independent of the library size and gene length. Below, the subscript t indicates minutes after the nutrient shift, g corresponds

to the modelled gene and l_g indicates its length, and s specifies the index of an experimental replicate.

$$\hat{\pi}_{gts} \sim \mathrm{Pois}(\alpha_{\pi st}\pi_{gt}l_g) \tag{5.30}$$

$$\hat{\nu}_{gts} \sim \mathrm{Pois}(\alpha_{\nu st}\nu_{gt}l_g) \tag{5.31}$$

$$\hat{\chi}_{gts} \sim \mathrm{Pois}(\alpha_{\chi st}\chi_{gt}l_g) \tag{5.32}$$

$$\hat{\rho}_{gts} \sim \mathrm{Pois}(\alpha_{\rho st}\rho_{gt}l_g) \tag{5.33}$$

In order to estimate the scaling factors α which would normalise sequencing counts between different χCRAC and RNA-seq experiments, we made the following simplifying assumption: the total number of molecules of each protein species interacting with transcripts stays constant in time (Eqs. 5.34, 5.35, 5.36). This assumption, while proving useful for our modelling strategy, finds support in the experimental data assessing the efficiency of protein recovery in χCRAC experiments. Fig. 3b in the paper shows that comparable amounts of Nab3 were recovered in experiments at different times after the shift. Similarly and for simplicity, we also assume that the true total cellular amount of all transcripts also does not change with time (Eq. 5.37).

$$P = \sum_g \pi_{gt} \tag{5.34}$$

$$N = \sum_g \nu_{gt} \tag{5.35}$$

$$X = \sum_g \chi_{gt} \tag{5.36}$$

$$R = \sum_g \rho_{gt} \tag{5.37}$$

We can then use this assumption to estimate the scaling factor for each sequencing dataset up to a constant, taking the expectation of the sum of the measured counts normalised to gene length, over all genes. The scaling factor is then proportional to that sum. Eqs. 5.38 and 5.39 show how to estimate the scaling factor $\alpha_{\pi st}$ for the dataset corresponding to the s-th replicate of the Pol II χCRAC experiment collected at t minutes after the shift. Other scaling factors can be estimated similarly.

$$\mathbf{E}\left[\sum_g \frac{\hat{\pi}_{gts}}{l_g}\right] = \sum_g \alpha_{\pi st}\pi_{gt} = \alpha_{\pi st}\sum_g \pi_{gt} = \alpha_{\pi st}P \tag{5.38}$$

$$\alpha_{\pi st} = \frac{1}{P}\sum_g \frac{\hat{\pi}_{gts}}{l_g} \tag{5.39}$$

Having estimated the scaling factors, the posterior distributions over the hidden variables π_t, ν_t, χ_t can be inferred with Gaussian Process regression using e.g. the Laplace's method to approximate the Poisson likelihood and the global constants. The global protein constants P, N, X can be optimised iteratively by maximising marginal likelihoods and inferring the posterior distributions.

The means of these posterior distributions can then be used to solve the ordinary differential equation (introduced in Section 3.3 of Chapter 3), which captures the relationship between the number of transcripts participating in transcription and degradation processes and the change in RNA abundance (Eq. 5.40). Here, the scaling factors A, B, C transform the posterior means π_t, ν_t, χ_t into reaction fluxes governing RNA expression. Each term in Eq. 5.40 corresponds to the rate of turnover of transcript molecules through the corresponding transcription or degradation pathway, regulated by Pol II, Nab3 or Xrn1. For simplification, we assume that the rate $A\pi_t$ encapsulates both the process of generating nascent transcripts and the process of their maturation. The term $B\nu_t$ describes the rate at which nascent transcripts are terminated via NNS regulation and $C\chi_t$ corresponds to degradation of mature transcripts. Their combined effort determines the rate of change of the total cellular amount of transcript.

$$\frac{d}{dt}\rho_t = A\pi_t - B\nu_t - C\chi_t \tag{5.40}$$

Other choices for the form of the differential equation are possible; one considered example includes non-linearly combining π_t and ν_t in an effort to reflect the effect of co-transcriptional degradation on transcription. This choice presents modelling complications such as non-linear data transformation and is complicated to justify biologically due to combining measurements from different populations of transcripts. Thus, we opted for a simple linear model, which facilitates straight-forward separation of pathways' contributions.

The parameters A, B, C can be globally optimised for all transcripts by maximising the product of marginal likelihoods $p(\hat{\rho}_0, ..., \hat{\rho}_T)$ over all transcripts and replicates, where timepoints range between 0 and T minutes. Alternatively, A, B, C could be

optimised independently for each transcript to see whether different genes cluster around similar parameter values, which control the contributions of the individual RBPs. Locally fitting parameters should be handled with care to prevent overfitting. The proposed model provides the means for deconvolving the contributions of transcriptional control and two separate degradation pathways to regulating RNA expression in stress response. The model allows extensions to more RBPs of interest, thereby investigating the roles of molecular actors shaping gene expression.

6

DISCUSSION

In this PhD project, I developed computational methods for the analysis and modelling of experimental data elucidating central aspects of RNA biology. RNA is intimately involved in all processes within the cell, facilitating the flow of genetic information and playing important regulatory roles in post-transcriptional control. Most, if not all of the RNA regulatory capabilities are affected by either (or both) its complex structural repertoire or its vast and dynamic interactome. The methods I developed are concerned with the analysis of next-generation sequencing data obtained in the context of probing RNA structure or its interactions with RNA-binding proteins.

6.1 MODELLING RNA STRUCTURE PROBING DATA

RNA structure can be indirectly probed with structure probing experiments, which arose and have been rapidly developing in the last decade. In these experiments, parts of the RNA molecule become chemically modified depending on the structural properties of those parts. A commonly used type of structure probing assays relies on the drop-off of a dedicated enzyme (reverse transcriptase, RT) at the chemically modified nucleotide positions. The enzyme implements reverse transcription of the probed RNAs into their complementary DNA fragments. Thus, the sequencing experiment generates a collection of DNA fragments of variable lengths, where the shorter length arises from the enzyme's drop-off, which could bear structural information (typically indicating a flexible single-stranded region).

The complexity is introduced through the stochasticity of the process as the enzyme can terminate randomly. To account for this, a control experiment is typically performed, sequencing transcripts that have not been in contact with the reagent implementing structure-dependent chemical modification. Additionally, the modification efficiency depends on the chemical concentration; determining the precise concentration level of the chemical reagent that is required for a successful structure probing experiment is a laborious task. Finally, other factors may influence the binding of the chemical to the transcript, especially probed *in vivo*, e.g. local presence of interacting proteins, multiple possible structural conformations of the same transcript or processes within the cell that dynamically affect the transcript's structure.

It is clear that such complicated interacting factors that introduce noise to the system have to be adequately taken into account. Despite this observation, only a few models for RNA structure probing data exist that handle the associated noise. Most methods simply subtract the numbers of modified nucleotides between the control and treated experiments, and thus are unable to model the biological variability of the probed process.

I developed BUM-HMM (beta-uniform mixture hidden Markov model), discussed in Chapter 4, which quantifies the variability of the enzyme's *drop-off rate* that arises by chance at each nucleotide position and uses that to compute the probability of chemical modification at each position in treated experiments. The method relies on multiple experimental replicates and assumes no parametric form for the distribution of observations given the true hidden values. The strength of the method comes with the choice of modelling *empirical p-values* instead of the actual observations, for which a statistically justified model can be used. The method was validated on known structures, demonstrating good agreement and a highly informative output by generating information about more positions than the existing methods. The method's output, while directly interpretable due to its probabilistic nature, proved to be useful for constraining structure prediction algorithms. Finally, the output remained highly consistent as the coverage of the data set was decreased in simulation studies, demonstrating valuable robustness to variations in coverage-related experimental conditions.

6.1.1 *Future work*

The BUM-HMM method can be extended in the following directions. The improvements on the existing functionalities could include modifications to parameter inference. The method presently features an EM optimisation of the shape parameters for the Beta distribution component of the mixture model. During validation experiments, it was found that the default parameter values gave good performance, while the optimisation appeared vulnerable to local optima. Further investigation of the parameter space and the effects the data have on it could provide insights in how to constrain the search space. Additionally, the transition probabilities of the HMM are currently inferred from empirical data; these parameters could also be optimised. However, this would also require analyses examining the robustness of the optimisation. As the true transition probability values are unknown, these analyses would be limited to simulation studies, where the lack of understanding of the underlying process could cast doubts on how representative the simulated data is of real measurements.

Another extension is concerned with data imputation, an idea briefly considered during the BUM-HMM development. Many nucleotide positions in structure probing data were assigned a value of 0 RT drop-off events, leading to a drop-off rate of 0, even if the numbers of mapped reads at these positions were positive. These positions were excluded from the analysis due to the definition of the variability measure LDR used within the model (the log-ratio of drop-off rates). A data imputation strategy could be considered, which would estimate the expected drop-off count from data or according to an observation model. The problem with parametric assumptions lies with their potential inability to generalise to technological modifications, e.g. the model by Aviran et al. and random-priming experiments. However, the model's formulation has since been modified to support this type of data (Li et al., 2017), so perhaps a similar model could be used for imputation. An empirical imputation strategy could estimate drop-off counts in control conditions from the coverage at each position and potentially, its neighbouring sequence. Such a strategy would, however, not apply to the data in treated conditions as it does not account for the structure-dependent stops.

Apart from data transformations, the underlying model itself could be extended to answer different biological questions. One such extension involves comparing structural properties between two datasets corresponding to different conditions (such as temperature). This would amount to modelling the hidden state not as a binary variable (taking values to be in a modified or unmodified state) but instead as a variable with 4 states: unmodified in condition 1 but modified in condition 2, modified in condition 1 but unmodified in condition 2 or unmodified/modified in both conditions. This would also enable the probing of dynamic changes in RNA structure if the datasets were collected at various temporal stages.

Another extension involves data integration. An obvious source of information, useful for inferring the structural state of a nucleotide, is its interaction status with other molecules. Methods mapping RNA-protein interactions can provide such information. In order to integrate it in the context of BUM-HMM, interaction sites must be recovered; a challenging problem in itself, recently addressed with Bayesian approaches (Drewe-Boss et al., 2017; Krakau et al., 2017). The two measurements, derived from structure probing and cross-linking experiments, could be combined to inform a single latent variable, representing the underlying RNA structure. Integrating various types of RNA-protein interaction data could enable one to study various aspects of changes in RNA structure. Examples of these are cross-linking data of proteins known to participate in different developmental stages or time-series cross-linking data collected in temporally successive experiments.

6.2 MODELLING DYNAMICS OF RNA-PROTEIN INTERACTIONS

One of the most prevalent experimental technologies to probe the RNA interactome is UV cross-linking, which creates bonds between RNA and proteins in close proximity and isolates the interaction partners for sequencing. The time it took to irradiate the cells to facilitate cross-linking used to be a rate-limiting step of the technology, rendering fast RNA-protein interactions impossible to detect experimentally. This limitation was addressed with a novel improvement of an existing cross-linking method called χCRAC (van Nues et al., 2017), which can generate interaction maps at a high temporal resolution of up to 1 minute.

χCRAC was applied to the proteins involved in different degradation pathways, creating rapid time-series of cross-linking events during the adaptation response of yeast cells to glucose starvation. I developed a non-parametric Bayesian method, discussed in Chapter 5, which identifies transcripts with significant changes in dynamic binding patterns to the protein of interest between different biological conditions. Using this method, we were able to investigate the role of the transcription termination factor Nab3 and the cytoplasmic degradation factor Xrn1 in stress response. We found that these proteins changed their binding to a large fraction of the transcriptome, often transiently and early into the stress response. These changes were not just a consequence of the increased amounts of transcribed RNA or mature transcripts available for interaction with Nab3 and Xrn1. Our results suggested that degradation processes play a regulatory role in shaping gene expression, particularly at the early stages of stress response. It follows that modelling the degradation rate as constant, as often done when explaining the dynamics of RNA expression, might not be enough to capture the changes to transcript abundance brought by the kinetics of decay during stress.

6.2.1 *Future work*

The current implementation of the algorithm testing for differential binding response uses a squared exponential kernel for defining the prior covariance of the hidden process. This covariance includes two hyperparameters that are currently fixed in a data-driven way. Specifically, the value of the lengthscale hyperparameter is set globally across all transcripts (reflecting our expectation based on the experimental design) and the value of the amplitude is computed heuristically from the time-series of each transcript. As we note in the Methods section of the paper (van Nues et al., 2017), presented in Chapter 5, a fully Bayesian treatment would involve integrating these parameters out when computing the marginal likelihood of the data under both mod-

els. It was decided to fix these parameters in the current implementation to facilitate exact computation of the marginal likelihood and lighten the computational burden, allowing fast testing of time-series for thousands of transcripts. It is not immediately trivial to choose suitable priors for the hyperparameters and using Markov Chain Monte Carlo methods to approximate the marginal likelihood can be a challenging problem. If these problems were addressed, it would be informative to examine the effects of these modifications in hyperparameter treatment on the algorithm's output.

Similarly, the Gaussian regression analysis is dependent on the kernel choice. The exponential kernel was chosen as a standard choice for modelling a function that varies smoothly in time. Other kernels could be useful for modelling the cross-linking profiles, however, it is not fully clear how to evaluate and compare the resulting performance, given the binary nature of the algorithm's output, lack of a complete "ground truth" list of protein's interacting partners, and the noise levels associated with the experimental technology.

Perhaps the richest direction for future work inspired by the results on differential binding involves dynamical modelling of RNA expression under stress. During adaptation, rapid reprogramming of gene expression might be crucial for the survival of an organism. Our results, uncovering rapid changes in the binding of proteins that mediate degradation, suggest that this could play an important role in controlling transcript abundance while cells are adapting to stress. We thus asked whether the cross-linking data of Nab3, Xrn1, and Pol II could be used for modelling time-varying transcription and degradation rates in a dynamical model for RNA expression. Another even bolder question asks whether the dynamic behaviours of these three proteins are *enough* to explain the changes in transcript abundance over time.

These questions were addressed with the proposed model for dynamical regulation of RNA expression during stress, which is outlined in Section 5.9.1. During the conceptual formulation of the model, a lot of consideration was put into understanding how to correctly normalise the χCRAC data, both between the experiments at different timepoints for the same protein (which represents the problem of between-sample RNA-seq normalisation) and between the time-series corresponding to different proteins. The χCRAC datasets for Pol II and Nab3 presented a particularly challenging case as Nab3 interactions should theoretically co-occur with Pol II binding events. In practice, all measurements are corrupted by noise and while their variance could be estimated from many replications, the complexity of the technology and hard-to-control conditions make it hard to generate more than only a small number of experimental replicates. In an attempt to solve this problem, we proposed to model raw cross-linking counts, infer scaling factors for all experiments, and use a Poisson

observation model to handle the noise associated with the measurements. In general, RNA-seq normalisation remains a challenging problem to be addressed.

In the differential equation, which describes the rate of change of transcript abundance, we include multiplicative factors for variables modelling the proteins of interest. A question remains whether these factors, aimed to transform the true binding response of each protein at a given time into a *rate* of transcription or degradation, should be optimised globally for all transcripts or on a per-transcript basis. If a good global fit exists, the first option would recover a common scale for the χCRAC technology, whereas the latter choice would attempt to characterise the gene-specific efficiencies for transcription and degradation. Further, non-linear forms of differential equations could be explored as they might be better suited to capture the interaction between the binding of Pol II and Nab3.

The literature characterising the previously measured transcription or degradation rates could be useful for both validating the model and potentially constraining the various optimisations within the analysis. However, this relies on the assumption that the reported range of rates is representative of the regulation processes studied within the model. Further, considering individual case studies of genes with well-characterised regulation profiles under stress can provide a validation strategy for the method. However, it is important to note that one of the aims of the model seeks to *collectively* describe the behaviour of many transcripts in terms of the pathways contributing to expression regulation. The metric for expressing the "goodness of fit" of the model to different transcripts also remains an open question; the percentage of the explained variance of transcript abundance is one candidate. The proposed model provides the flexibility to include the binding profiles of other proteins affecting gene expression. The relative contribution of each included pathway could be quantified using the optimised parameter values and the inferred cross-linking profiles included in the differential equation.

In general, methods for modelling stochastic dynamical processes, which underlie major biological events in the cell (e.g. modelling transcriptional control in single cells (Suter et al., 2011)), do not progress as fast as experimental data become available that directly or indirectly illuminate these processes. The analyses outlined and proposed in Chapter 5 of this thesis, made possible by the availability of the highly temporally resolved χCRAC data, take another step in the direction of the important field of dynamical modelling in RNA biology.

APPENDIX

APPENDIX A

Appendix A provides the definitions for the transition probabilities used in the BUM-HMM method and the Supplementary Figures and Tables for the paper "Robust statistical modeling improves sensitivity of high-throughput RNA structure probing experiments", presented in Chapter 4.

A.1 TRANSITION PROBABILITIES

Transition probabilities were defined through empirically derived lengths of single- and double-stranded stretches of nucleotides. The model assumes expected uninterrupted stretches of 20 double-stranded (or constrained) nucleotides and 5 single-stranded (or flexible) nucleotides.

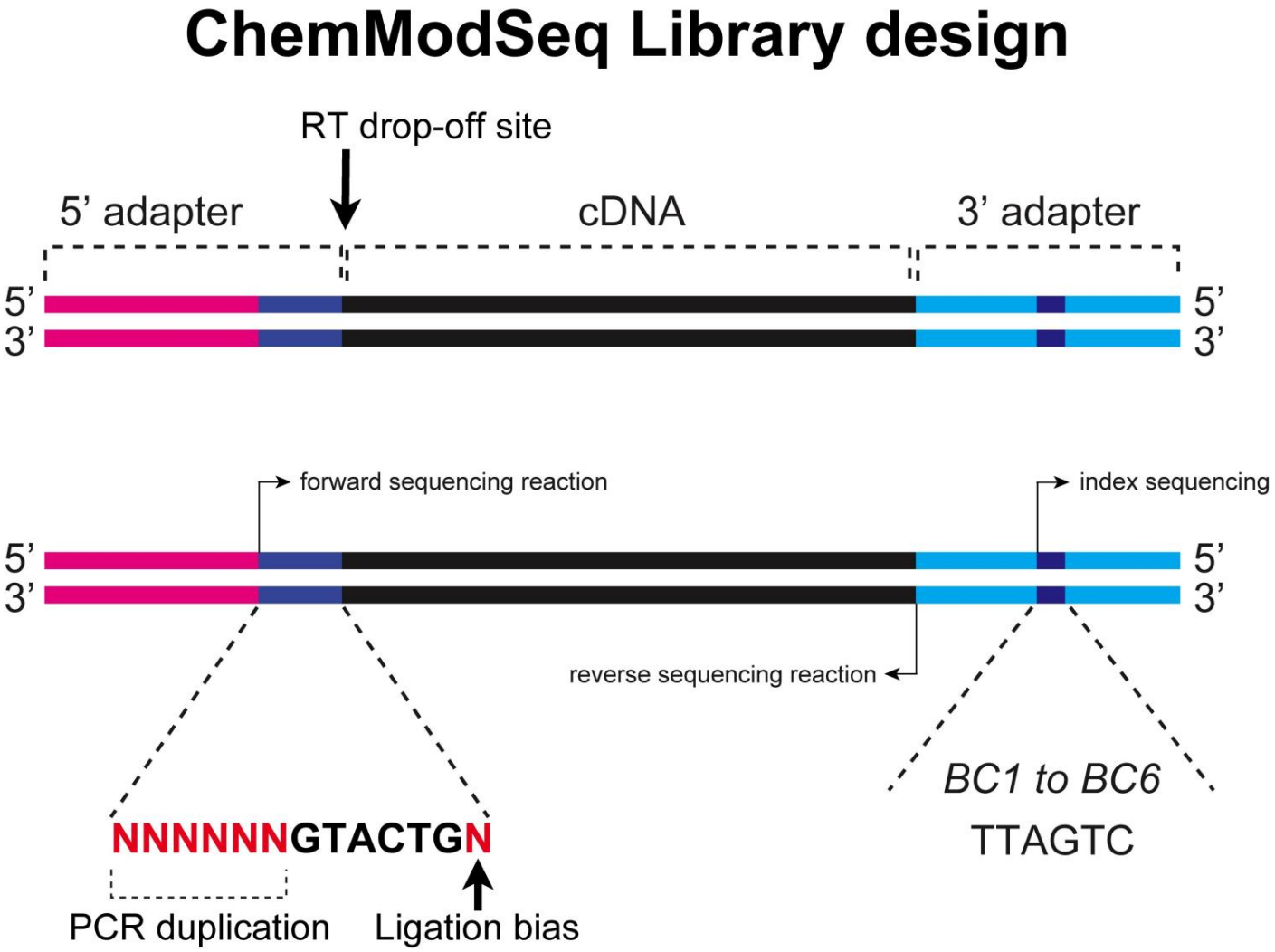
Below, h_t corresponds to the hidden state at the nucleotide position t, which can take the value U, corresponding to the unmodified true state, or the value M, corresponding to the modified true state.

$$p(h_{t+1} = U | h_t = U) = \frac{19}{20} = 0.95 \tag{A.1}$$

$$p(h_{t+1} = M | h_t = U) = \frac{1}{20} = 0.05 \tag{A.2}$$

$$p(h_{t+1} = M | h_t = M) = \frac{4}{5} = 0.8 \tag{A.3}$$

$$p(h_{t+1} = U | h_t = M) = \frac{1}{5} = 0.2 \tag{A.4}$$

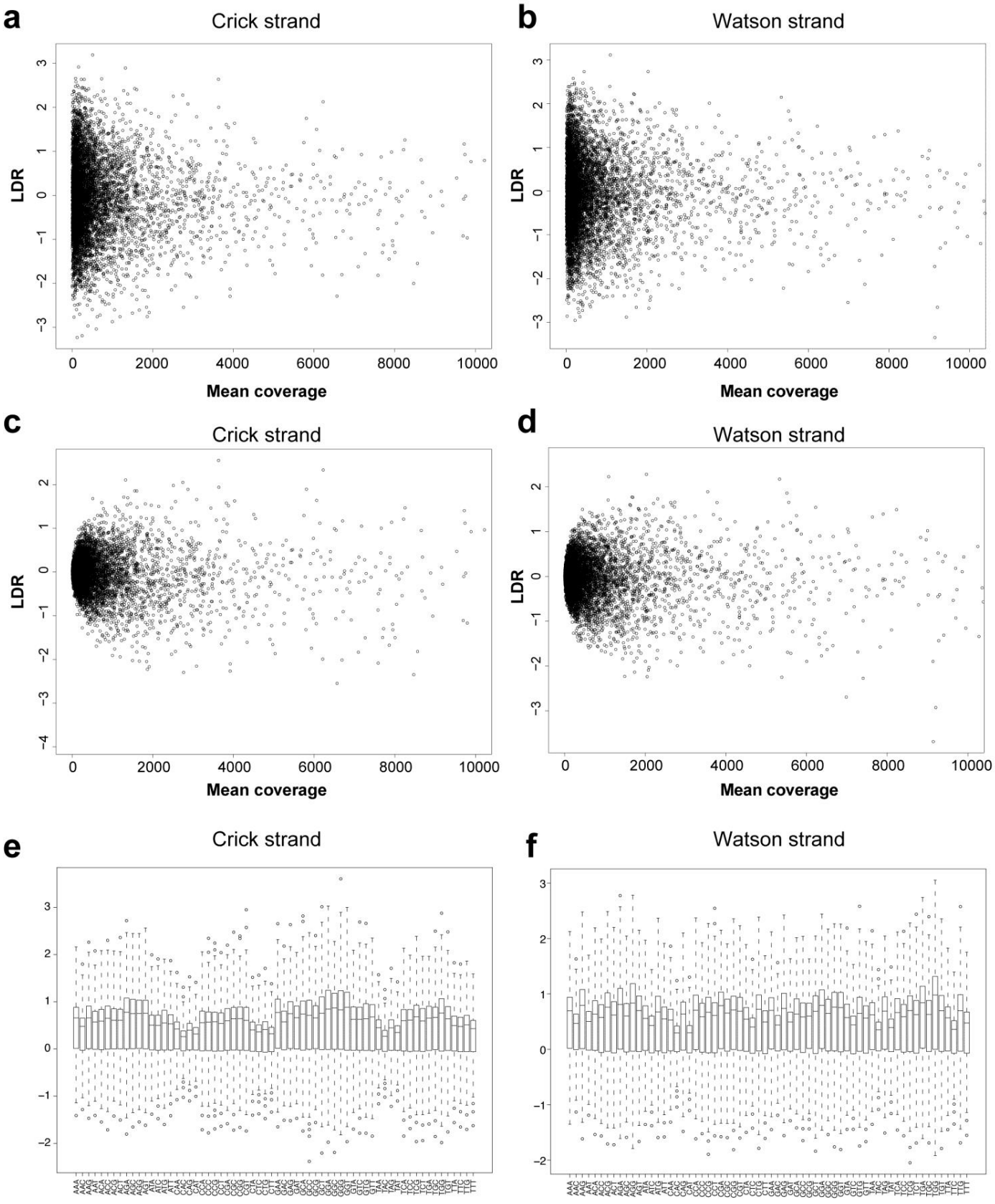


Supplementary Figure 1

ChemModSeq library preparation design.

Chemically probed RNAs were reverse transcribed with an oligonucleotide containing a random hexamer and an Illumina compatible sequence for PCR amplification. Subsequently, adapters were ligated to the 3' end of cDNAs that contained six random nucleotides and a six nucleotide barcode followed by another random nucleotide. The latter was introduced to minimize sequence bias representation introduced during the CircLigase ligation reaction. The six random nucleotides were used to eliminate potential PCR duplicates. Indexing barcodes were added to the 3' adapter sequence by PCR. The in-read barcodes in the 5' end of the PCR product were processed using pyBarcodeFilter.py and reads were collapsed using pyFastqDuplicateRemover.py from the pyCRAC package[1].

1. Webb, S., Hector, R. D., Kudla, G. & Granneman, S. "PAR-CLIP data indicate that Nrd1Nab3-dependent transcription termination regulates expression of hundreds of protein coding genes in yeast." Genome biology 15, R8 (2014).

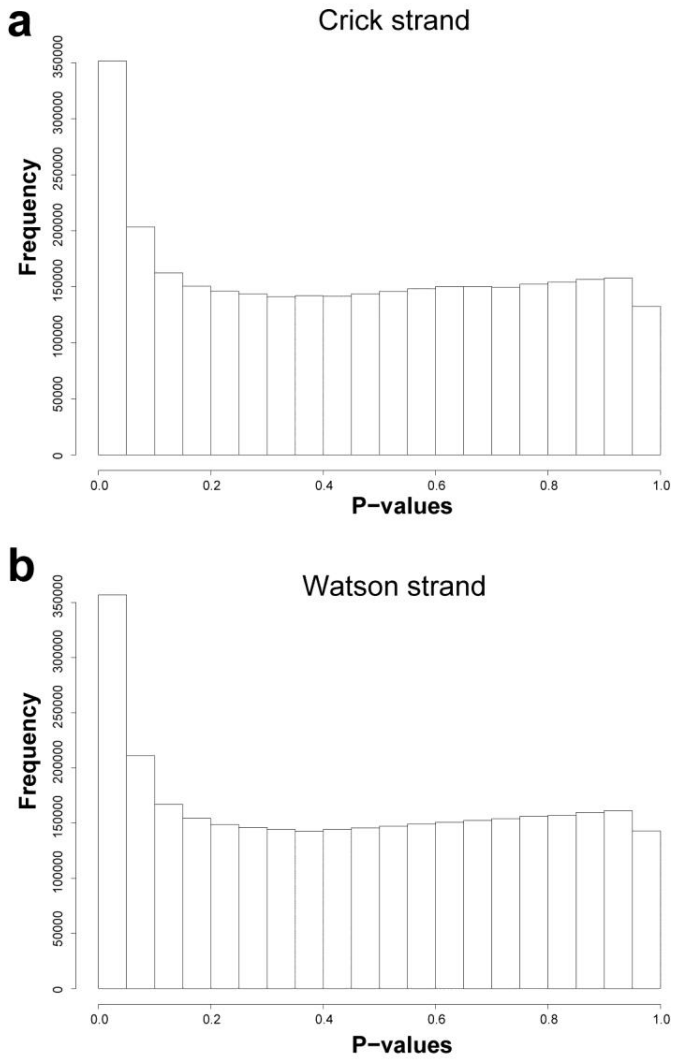

a
Crick strand
LDR
Mean coverage
b
Watson strand
LDR
Mean coverage
c
Crick strand
LDR
Mean coverage
d
Watson strand
LDR
Mean coverage
e
Crick strand
f
Watson strand

Supplementary Figure 2

Coverage- and sequence-dependent biases were identified in the transcriptome data set.

(a, b) Presence of a coverage-dependent bias, reflected by the dependency between the LDR and the mean coverage at each nucleotide position in a pair of control replicate samples, for all such pairs, computed from the yeast transcriptome-wide data set on both strands. (c, d) Same dependency plotted as in (a, b) after applying a bias-correcting strategy to the LDRs. (e, f) Presence of a sequence-dependent bias, reflected by differing null distributions of LDRs. Each boxplot represents the null distribution (y-axis shows LDR) computed only for the nucleotide positions corresponding to a given trinucleotide pattern (indicated on the x-axis).

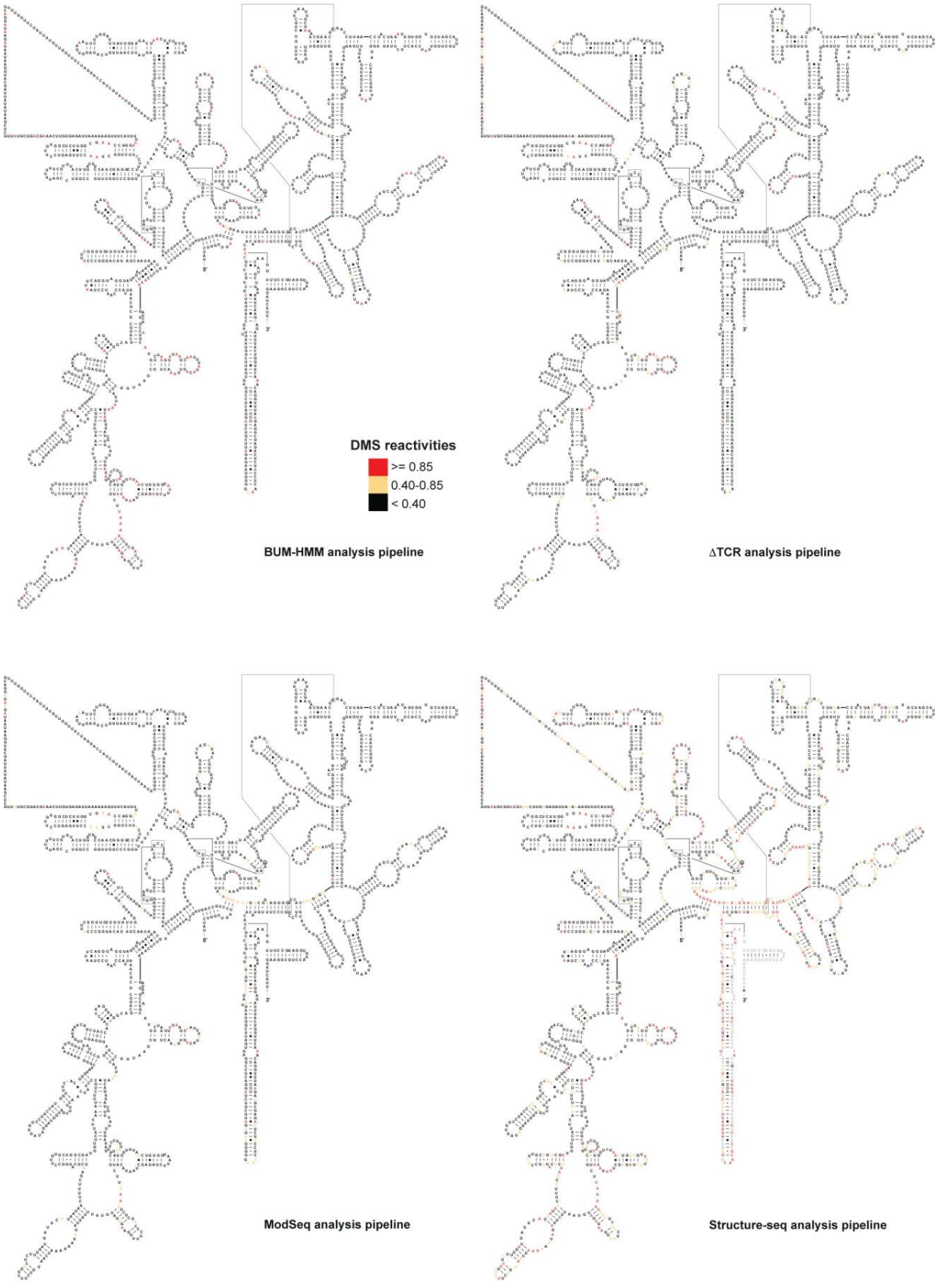
Nature Methods: doi:10.1038/nmeth.4068



Supplementary Figure 3

Distributions of empirical *P* values for the transcriptome data set closely follow the Beta-Uniform distribution on both strands.

The histograms show the distributions of empirical *P* values associated with LDRs between all combinations of treatment and control samples on the transcriptome data set for both strands.

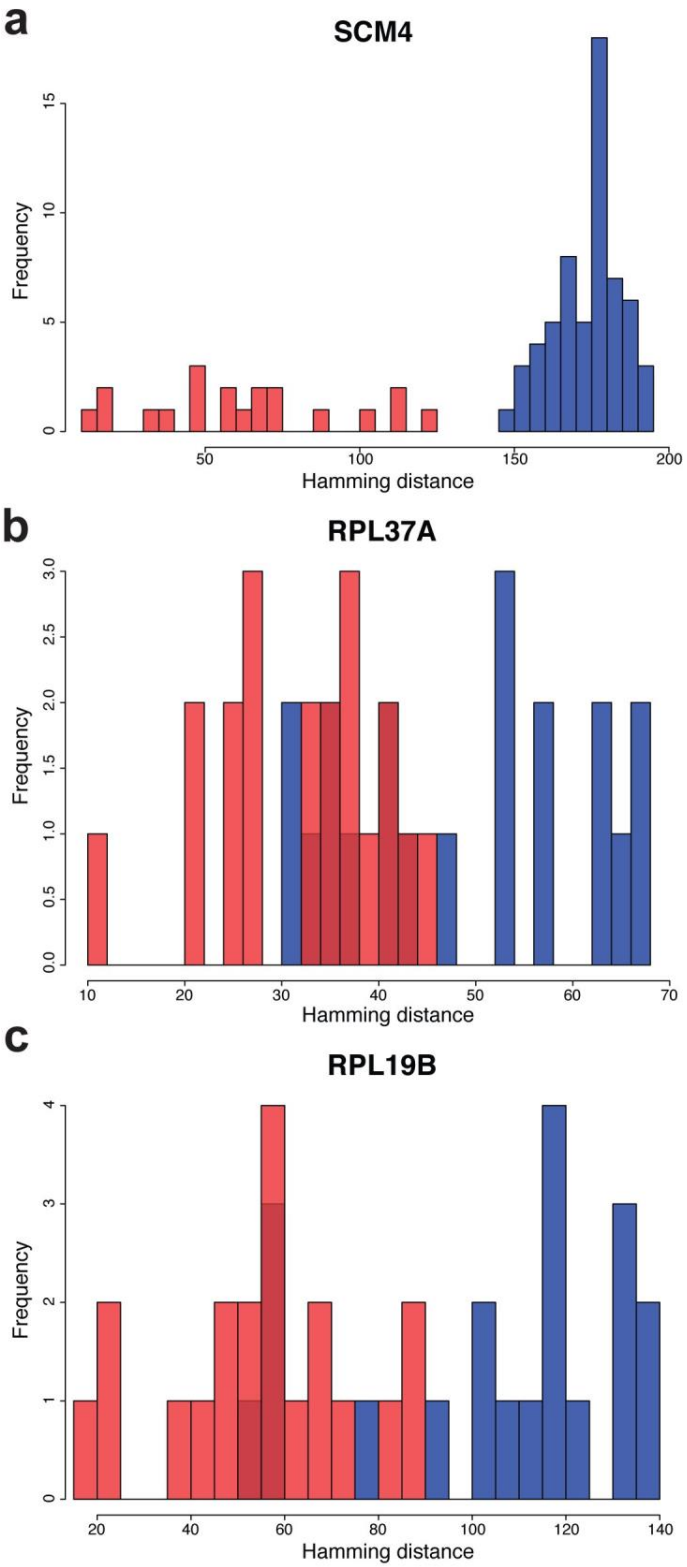

DMS reactivities
>= 0.85
0.40-0.85
< 0.40
BUM-HMM analysis pipeline
ΔTCR analysis pipeline
ModSeq analysis pipeline
Structure-seq analysis pipeline

Supplementary Figure 4

BUM-HMM correctly identifies many flexible A's and C's as modified nucleotides.

Secondary structures of the 18S ribosomal RNA with bases colored according to the reactivity score or posterior probability at the corresponding nucleotide position, generated by BUM-HMM, ΔTCR, Mod-seq, and structure-seq analysis pipelines on the data set using a DMS probe.

Nature Methods: doi:10.1038/nmeth.4068

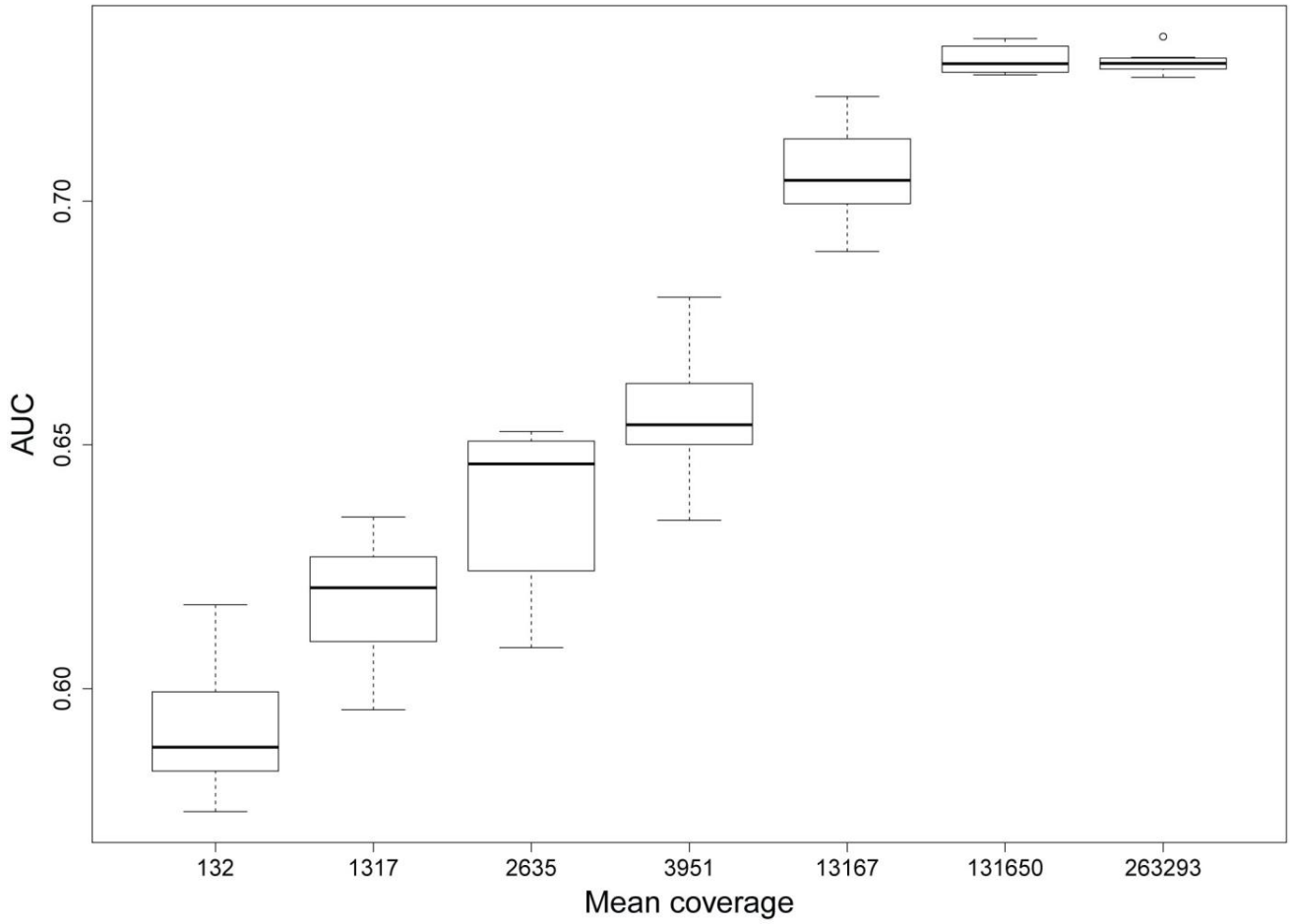


Supplementary Figure 5

Using BUM-HMM output as constraints results in more consistent secondary structure prediction across different methods.

(a) Distribution of Hamming distances between the structures predicted for SCM4 by *Fold* (n=20) and by *MaxExpect* (n=3 with sequence, n=1 with BUM-HMM) when using only sequence (blue) and when adding the BUM-HMM output as constraints (red). (b, c) Same as in (a), for RPL37A (b) and RPL19B (c) (with *Fold*, n=20 structures were generated, with *MaxExpect*, n=1 structure).

Nature Methods: doi:10.1038/nmeth.4068

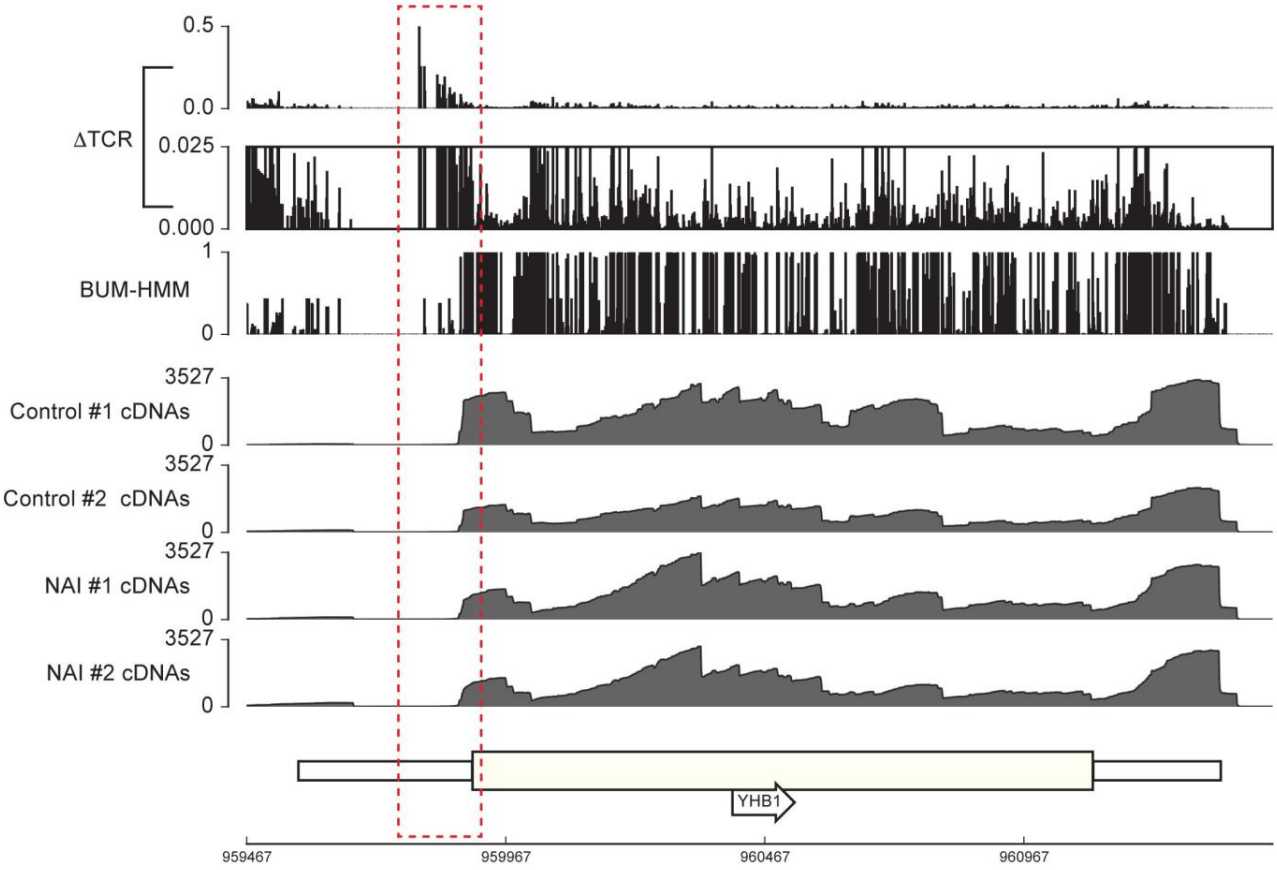


Supplementary Figure 6

BUM-HMM retains good accuracy at 18S secondary structure reconstruction at lower coverage levels.

Agreement with the 18S crystal structure of the posterior probabilities generated by BUM-HMM on data sets with progressively lower mean coverage (shown on the x-axis), synthesized from the DMS data set for the 18S ribosomal RNA. Agreement was measured with the AUC statistic (shown on the y-axis) between the binary 'ground truth' matrix derived from the crystal structure and the generated probabilities for each synthetic data set. The subsets of 2 million, 1 million, 100,000, 30,000, 20,000, 10,000, and 1,000 reads (corresponding to 7 progressively reducing coverage levels) were randomly selected from the full data set 10 times for each coverage level. The error bars quantify the variability in the agreement of the BUM-HMM predictions with the crystal structure across these 10 selections for each coverage level.

Nature Methods: doi:10.1038/nmeth.4068



Supplementary Figure 7

The ΔTCR algorithm produces very high numbers in regions with low coverage.

Shown is a genome browser image of a gene (YHB1) with an FPKM of 190. The red-dotted box shows a region near the 3' end of the gene where there is low coverage. The top two panels show the ΔTCR output, with the second panel displaying the same data but scaled to a maximum ΔTCR value of 0.025. The third panel shows the BUM-HMM posterior probabilities for the same region. The last four panels show the cDNA coverage over the gene from the two control RNA sequencing data and the two NAI treated sequencing data.

Nature Methods: doi:10.1038/nmeth.4068

18S DMS data

Sample	*Total reads*	*Total aligned*	*Paired reads*	*Pairs aligned*
Control 1	7298504	7281982	3649252	2832740
Control 2	6746926	6541815	3373463	2800015
Control 3	25186768	25044042	12593384	12363010
DMS 1	15961436	15922997	7980718	7340656
DMS 2	13876274	13406913	6938137	6478171
DMS 3	35111510	34672845	17555755	17083033

18S 1M7 data

Sample	*Total reads*	*Total aligned*	*Paired reads*	*Pairs aligned*
Control 1	13578078	13238681	6789039	6227292
Control 2	7585196	7461105	3792598	3706939
Control 3	6765306	6559927	3382653	2807798
1M7 1	12737362	12475601	6368681	6204239
1M7 2	5668572	5668568	3305227	2821236
1M7 3	10713590	10444647	5356795	4995539

Transcriptome data

Sample	*Total reads*	*Total aligned*	*Paired reads*	*Pairs aligned*
Control 1	77562940	76077320	38781470	36597927
Control 2	83454506	82021191	41727253	39975802
NAI 1	99164982	97509211	49582491	46831503
NAI 2	115420404	112932454	57710202	54645707

Supplementary Table 1: Overview of paired cDNA reads analyzed from each data set. All raw sequencing data have been collapsed before aligning to the reference sequences to remove potential PCR duplicates. Only properly paired reads were considered for the analyses.

Structure recovery of 18S rRNA on the 1M7 data set

Method	*AUC*
BUM-HMM	0.59
ΔTCR	0.62
Structure-seq	0.59
Mod-seq	0.58

Supplementary Table 2: Accuracy of reconstructing secondary structure of 18S ribosomal RNA from the 1M7 data set for all methods, measured with the AUC statistic against the known crystal structure of the rRNA.

Nature Methods: doi:10.1038/nmeth.4068

B

APPENDIX B

Appendix B provides the derivations for computing the marginal likelihood under the Gaussian and Poisson observation models and the Supplementary Information for the paper "Kinetic CRAC uncovers a role for Nab3 in determining gene expression profiles during stress", presented in Chapter 5.

B.1 DERIVATIONS FOR COMPUTING THE MARGINAL LIKELIHOOD UNDER THE GAUSSIAN OBSERVATION MODEL

B.1.1 *Completing the square*

Under the exponent in Eq. 5.14 (this and the following equations are given in Chapter 5), the terms quadratic in $\mathbf{f}$ are:

$$-\frac{1}{2}\left(\mathbf{f}^T K^{-1}\mathbf{f} + \frac{1}{\sigma^2}\sum_{k=1}^{M}\mathbf{f}^T I_N \mathbf{f}\right) \tag{B.1}$$

Equating these terms to the quadratic term from Eq. 5.15, we can compute the expression for the inverse covariance (or precision) C^{-1} as given in Eq. B.4:

$$-\frac{1}{2}\left(\mathbf{f}^T K^{-1}\mathbf{f} + \frac{1}{\sigma^2}\sum_{k=1}^{M}\mathbf{f}^T I_N \mathbf{f}\right) = -\frac{1}{2}\mathbf{f}^T C^{-1}\mathbf{f} \tag{B.2}$$

$$\mathbf{f}^T\left(K^{-1} + \sum_{k=1}^{M}\frac{1}{\sigma^2}\right)\mathbf{f} = \mathbf{f}^T C^{-1}\mathbf{f} \tag{B.3}$$

$$C^{-1} = K^{-1} + \frac{M}{\sigma^2} I_N \tag{B.4}$$

Similarly, equating the linear term in $\mathbf{f}$ under the exponent in Eq. 5.14 to the linear term of the general quadratic form from Eq. 5.15 enables us to compute the expression for $\boldsymbol{\mu}$ as given in Eq. B.7:

$$-\frac{1}{2\sigma^2}\sum_{k=1}^{M}\left(-2\mathbf{f}^T I_N \mathbf{y}^k\right) = \mathbf{f}^T C^{-1}\boldsymbol{\mu} \tag{B.5}$$

$$\mathbf{f}^T\left(\frac{1}{\sigma^2}\sum_{k=1}^{M} I_N \mathbf{y}^k\right) = \mathbf{f}^T C^{-1}\boldsymbol{\mu} \tag{B.6}$$

$$\boldsymbol{\mu} = \frac{C}{\sigma^2}\sum_{k=1}^{M}\mathbf{y}^k \tag{B.7}$$

We now have to add and subtract the $-\frac{1}{2}\boldsymbol{\mu}^T C^{-1}\boldsymbol{\mu}$ term from the expression under the exponent in Eq. 5.14 in order to complete the square with respect to $\mathbf{f}$. This term, quadratic in $\boldsymbol{\mu}$, simplifies to the expression shown in Eq. B.8, using our results in Eqs. B.7 and B.4:

$$\begin{aligned}-\frac{1}{2}\boldsymbol{\mu}^T C^{-1}\boldsymbol{\mu} = -\frac{1}{2}\left(\frac{C}{\sigma^2}\sum_{k=1}^{M}\mathbf{y}^k\right)^T C^{-1}\left(\frac{C}{\sigma^2}\sum_{k=1}^{M}\mathbf{y}^k\right) = \\ -\frac{1}{2\sigma^4}\left(\sum_{k=1}^{M}\mathbf{y}^k\right)^T C^T C^{-1} C\left(\sum_{k=1}^{M}\mathbf{y}^k\right) = \\ -\frac{1}{2\sigma^4}\left(\mathbf{y}^1+\ldots+\mathbf{y}^M\right)^T C\left(\mathbf{y}^1+\ldots+\mathbf{y}^M\right)\end{aligned} \tag{B.8}$$

Now, adding and subtracting it from the expression under the exponent of the integral yields:

$$\begin{aligned}\sum_{k=1}^{M}\left(-\frac{1}{2\sigma^2}(\mathbf{y}^k-\mathbf{f})^T I_N(\mathbf{y}^k-\mathbf{f})\right) - \frac{1}{2}\mathbf{f}^T K^{-1}\mathbf{f} - \\ -\frac{1}{2\sigma^4}\left(\sum_{k=1}^{M}\mathbf{y}^k\right)^T C\left(\sum_{k=1}^{M}\mathbf{y}^k\right) + \frac{1}{2\sigma^4}\left(\sum_{k=1}^{M}\mathbf{y}^k\right)^T C\left(\sum_{k=1}^{M}\mathbf{y}^k\right) = \\ = -\frac{1}{2\sigma^2}\sum_{k=1}^{M}\left((\mathbf{y}^k)^T I_N \mathbf{y}^k\right) + \frac{1}{2\sigma^4}\left(\sum_{k=1}^{M}\mathbf{y}^k\right)^T C\left(\sum_{k=1}^{M}\mathbf{y}^k\right) - \\ -\frac{1}{2}(\mathbf{f}-\boldsymbol{\mu})^T C^{-1}(\mathbf{f}-\boldsymbol{\mu})\end{aligned} \tag{B.9}$$

B.1.2 *Evaluating the integral*

We can now take the terms not containing $\mathbf{f}$ outside of the integral. And thus, the integral evaluates to the following expression:

$$\begin{aligned}
&\int \exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{M}\left((\mathbf{y}^k)^T I_N \mathbf{y}^k\right)+\frac{1}{2\sigma^4}\left(\sum_{k=1}^{M}\mathbf{y}^k\right)^T C\left(\sum_{k=1}^{M}\mathbf{y}^k\right)-\right.\\
&\qquad\left.-\frac{1}{2}(\mathbf{f}-\boldsymbol{\mu})^T C^{-1}(\mathbf{f}-\boldsymbol{\mu})\right)d\mathbf{f}=\\
&=\exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{M}\left((\mathbf{y}^k)^T I_N \mathbf{y}^k\right)+\frac{1}{2\sigma^4}\left(\sum_{k=1}^{M}\mathbf{y}^k\right)^T C\left(\sum_{k=1}^{M}\mathbf{y}^k\right)\right)\cdot\\
&\qquad\cdot\int\exp\left(-\frac{1}{2}(\mathbf{f}-\boldsymbol{\mu})^T C^{-1}(\mathbf{f}-\boldsymbol{\mu})\right)d\mathbf{f}=\\
&=\exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{M}\left((\mathbf{y}^k)^T I_N \mathbf{y}^k\right)+\frac{1}{2\sigma^4}\left(\sum_{k=1}^{M}\mathbf{y}^k\right)^T C\left(\sum_{k=1}^{M}\mathbf{y}^k\right)\right)\sqrt{(2\pi)^N\,|C|}
\end{aligned} \tag{B.10}$$

Finally, the expression for the marginal likelihood of the data $\mathbf{y}$ is given by:

$$\begin{aligned}
&p(\mathbf{y})=\frac{1}{(\sqrt{2\pi}\sigma)^{NM}}\cdot\frac{1}{(\sqrt{(2\pi)^N|K|})}\cdot\\
&\exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{M}\left((\mathbf{y}^k)^T I_N \mathbf{y}^k\right)+\frac{1}{2\sigma^4}\left(\sum_{k=1}^{M}\mathbf{y}^k\right)^T C\left(\sum_{k=1}^{M}\mathbf{y}^k\right)\right)\sqrt{(2\pi)^N\,|C|}
\end{aligned} \tag{B.11}$$

B.1.3 *Log-marginal likelihood*

The term $\sqrt{\frac{|C|}{|K|}}$ appears in the expression for the marginal likelihood of the data. Using the properties of a determinant, we get:

$$\frac{|C|}{|K|}=|C|\cdot|K|^{-1}=|C|\cdot|K^{-1}|=|CK^{-1}| \tag{B.12}$$

Using the definition of C, we can write down the expression for the matrix CK^{-1}:

$$\begin{aligned}
CK^{-1}&=\left(K^{-1}+\frac{M}{\sigma^2}I_N\right)^{-1}\cdot K^{-1}=\left(I_N+\frac{M}{\sigma^2}K\right)^{-1}=\\
&=\left(\frac{M}{\sigma^2}\cdot\left(\frac{\sigma^2}{M}I_N+K\right)\right)^{-1}=\frac{\sigma^2}{M}\cdot\left(\frac{\sigma^2}{M}I_N+K\right)^{-1}
\end{aligned} \tag{B.13}$$

Using the properties of a determinant, we can write down the expression for the determinant of CK^{-1}:

$$|CK^{-1}| = \left|\frac{\sigma^2}{M}\left(\frac{\sigma^2}{M}\cdot I_N + K\right)^{-1}\right| = \left(\frac{\sigma^2}{M}\right)^N \left|\frac{\sigma^2}{M}\cdot I_N + K\right|^{-1} \tag{B.14}$$

In the expression for the log-marginal likelihood, the term $|CK^{-1}|$ appears under the squared root and its logarithm is given by:

$$\begin{aligned}\log\sqrt{|CK^{-1}|} &= \frac{1}{2}\log\left(\left(\frac{\sigma^2}{M}\right)^N \left|\frac{\sigma^2}{M}\cdot I_N + K\right|^{-1}\right) = \\ &= N\log\sigma - \frac{N}{2}\log M - \frac{1}{2}\log\left|\frac{\sigma^2}{M}\cdot I_N + K\right|\end{aligned} \tag{B.15}$$

B.2 DERIVATIONS FOR COMPUTING THE MARGINAL LIKELIHOOD UNDER THE POISSON OBSERVATION MODEL

B.2.1 *Approximating the integral with Laplace's method*

The integral in Eq. 5.25 can be approximated as a Gaussian distribution using Laplace's method (MacKay, 2003). This method approximates any single-mode distribution with a Gaussian distribution with the mean set as its mode and the variance computed using the curvature at the mode.

Let's assume $P(x)$ to be an unnormalised density with the mode at x_0. Its normalisation constant is given by

$$Z = \int P(x)\,dx \tag{B.16}$$

Using Taylor expansion of $\log P(x)$ at $x = x_0$, $P(x)$ can be approximated as an unnormalised Gaussian with the mean x_0 and variance c^2:

$$P(x) \approx P(x_0)\exp\left(-\frac{1}{2c^2}(x-x_0)^2\right), \tag{B.17}$$

$$\text{where } c^2 = \frac{1}{-\frac{d^2}{dx^2}\log P(x_0)} \tag{B.18}$$

And the normalising constant Z, to which the integral evaluates, is given by:

$$Z = \int P(x)dx \approx P(x_0) \int \exp\left(-\frac{1}{2c^2}(x - x_0)^2\right) dx = P(x_0) \cdot \sqrt{2\pi c^2} \tag{B.19}$$

Or, generalising to K dimensions with the covariance C, $Z \approx P(x_0) \cdot \sqrt{(2\pi)^K |C|}$.

Thus, the integral of interest from Eq. 5.25 can be approximated by recovering the mode x_0 of the expression under the integral (which gives the mean of the approximating Gaussian) and the inverse of the negative Hessian of the log-expression evaluated at x_0 (which gives the covariance of the approximating Gaussian).

B.3 INVESTIGATION OF THE CONSENSUS BETWEEN THE DIFFERENTIAL BINDING ANALYSES WITH DIFFERENT NOISE MODELS ON THE NAB3 DATASET

The figures below illustrate the same analysis for the Nab3 χCRAC dataset as presented for Pol II in Section 5.8.2 (Chapter 5). Fig. B.1 shows the median coverage levels of the transcripts selected by both analyses in comparison to those of all transcripts tested with the Poisson regression analysis. Fig. B.2 shows the log-variance of the control time-series for all transcripts that were tested with the Poisson regression analysis and compares it with the corresponding log-variance of the Poisson-selected targets and the transcripts selected by both analyses.

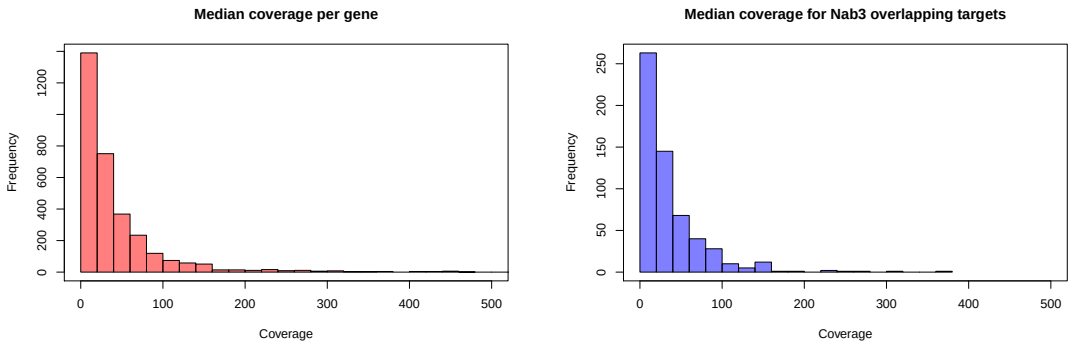


Figure B.1: Left: a histogram showing median coverage levels for all genes that were tested with the Poisson regression analysis for differential binding with Nab3 under stress. The medians were computed across each time-series in control and stress conditions and then averaged across conditions for each gene. 4 genes had a median coverage > 500 and are not shown. Right: a histogram showing median coverage levels for all genes that were selected by both Poisson and Gaussian regression analyses as differentially bound by Nab3 under stress.

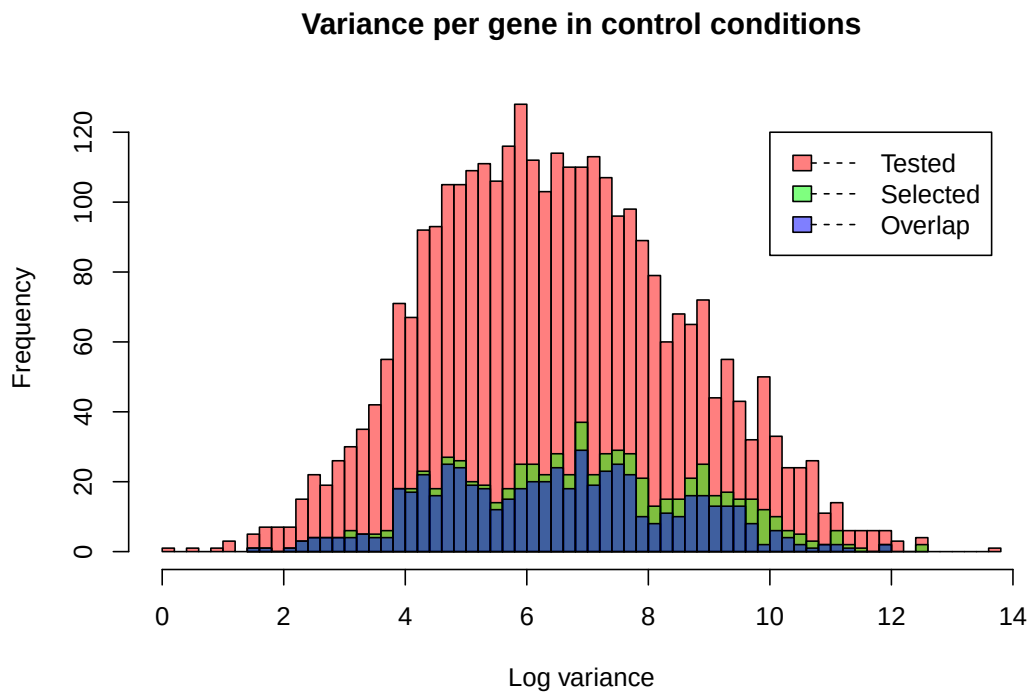


Figure B.2: A histogram showing the log-variance of time-series in control conditions for all genes that were tested with the Poisson regression analysis for differential binding with Nab3 under stress (coloured in pink). Overlaid are shown the corresponding histograms for the targets selected by the Poisson regression analysis (in green) and for the targets selected by both analyses (in blue).

Supplementary note 1

Rapid nuclear depletion of Nab3 using the anchor-away system

To rapidly deplete Nab3 from the nucleus we used the anchor-away system. Nab3 was fused to the FKBP12 rapamycin-binding domain (FRB) and depleted from the nucleus by rapamycin treatment, which forms a very stable complex with Nab3-FRB and the nuclear r-protein RPL13A-FKBP12 fusion associated with ribosome assembly intermediates, which are exported to the cytoplasm. Strains expressing Nab3-FRB grew comparable to the parental strain, but failed to grow in the presence of rapamycin (Supplementary Fig. 5a) A one-hour rapamycin treatment was sufficient to detect 3'-extended snR13 species, which are known to accumulate in NNS mutants[2] (Supplementary Fig. 5b). To measure changes in transcription profiles, we performed Pol II χCRAC on Nab3 depleted cells. This revealed an accumulation of 3'-extended CUTS and snoRNAs, both under normal (glucose) and stress (no glucose) conditions (Supplementary Fig. 6a; Nab3-FRB data). Importantly, when we incubated the parental strain with rapamycin for one hour, we did not observe these changes in snoRNA and CUT transcriptional profiles (Supplementary Fig. 6a; Nab3 data). *NRD1* transcript levels are auto-regulated by NNS-dependent attenuation of transcription[3]. In the rapamycin treated *nab3::FRB* cells, we observed a higher density of Pol II downstream of the Nab3 binding sites, suggesting NNS-dependent attenuation of *NRD1* is strongly diminished (Supplementary Fig. 6b; Nab3-FRB data). Increased Pol II transcription could also be detected downstream of the *snR13* gene in the Nab3 depleted cells, consistent with a defect in Nab3-dependent transcription termination (Supplementary Fig. 6c). Again, these changes in Pol II distribution were not observed in the rapamycin treated parental

strain (Nab3 panels), demonstrating that these changes are a direct result of Nab3 depletion.

We conclude that with our Nab3 depletion conditions we can faithfully reproduce previously published work.

Supplementary note 2

Nab3 and regulation of *ENO1* transcription initiation

As shown in a schematic overview of *ENO1* transcription-regulation (not to scale; Supplementary Fig. 12), two CUTs are detected upstream of the gene, which is activated by two upstream activating sequences (UAS1, UAS2) that bind multiple factors[4-6]. *ENO1* also contains an upstream repressor sequence element (URS) in the promoter, located between -226 and -125 bp from the TSS. This URS has a directionality as reverting this element relieves inhibition of *ENO1* expression when cells are grown on glucose[7]. Experimental evidence has been published for three URS-binding factors: i) Reb1, that associates with a region not essential for, but enhancing URS function[8] and binds the UAS2 as well[4,8]. Deletion of Reb1 had no major effect on *ENO1* expression[9,10]; ii) the BUF-complex, consisting of Rfa1 and Rfa2 – which participate in yeast DNA replication as ssDNA binding proteins – that recognizes the repressor-of-CAR1-like sequence (TaGCCaCCTC) at the 5' end of the region essential for URS activity[11,12]. Repression is possibly mediated by Ume6[13], which recruits histone deacetylase Rpd3p and chromatin-remodeling factor Isw2p[14]. Long-range DNA-looping via Ume6 is associated with transcriptional repression[15]. Although Ume6 had no influence, Isw2 was found to downregulate *ENO1* expression[9,10] and can repress

transcription of cryptic RNAs[16]. iii) The basic-domain helix-loop-helix (bHLH) protein Sgc1 (Tye7) binds to the E-box motif (CAnnTG) in the 3' region of the URS[17], possibly as a (hetero) dimer[18], and, depending on surrounding sequences, could bend DNA[19]. Sgc1, like Gcr1, Gcr2 and Rap1, down-regulates *ENO1*[9,10]. In cells expressing the dominant *SGC1-1* allele (E→Q, ten residues upstream of the bHLH DNA-binding/dimerization domain), *ENO1* mRNA levels become comparable to those of *ENO2* in wild-type cells[20] during growth in glucose.

As shown in Supplementary Fig. 8c and Fig. 6d, we observed increased Nab3 binding to RNAs overlapping the region spanning the TATA-box, a long pyrimidine-rich region including three Nab3 binding motifs (UCUU; CUUG) and up to the TSS when cells were grown on glucose, but significantly less so after removal of the carbon source. These products increased in number after depletion of Nab3 (Fig. 6e) and carry poly-A tails (Fig. 6d), indicative for poly-adenylation by the TRAMP-complex and needed for degradation by the exosome. The 5'UTR-derived products also overlap the two CUTs, that are only observed in the absence of exosome components[21,22]. *CUT166*, which initiates from the 5' end of the URS has also been detected by CRAC using Cbc1, a component of the cap-binding complex[23] and by transcript isoform sequencing (TIFseq)[24] (Supplementary Fig. 8).

The role of the URS and the associated protein factors suggest that these are the primary regulators of *ENO1* transcription by controlling productive transcription initiation from its TSS, ~40 nt upstream of the ATG. They appear to do this (possibly indirectly), by stimulation of *CUT166* transcription, but not by promoting alternative TSS choice. For cells growing on glucose we did not observe (also after depleting Nab3) a marked

change in the distribution of Pol II; all the associated RNAs had their 5' end around the mapped TSS[25,26] as confirmed by TIFseq[24] (Supplemental Fig. 8). Furthermore, after removal of glucose, the levels of transcripts derived from the 5'UTR did not alter over time (also when Nab3 had been depleted; Fig. 6e), whereas those reflective of *ENO1* transcription increased more than 2-fold (Fig. 6f). Overall, our data suggest that Pol II still finds the TSS but that under repressive conditions this happens slowly.

Formation of productive transcription initiation complexes at the *ENO1* TSS appear impeded on glucose. This is corroborated by mapping of pre-initiation complexes (PICs) of general transcription factors and Pol II[27], which – compared to *ENO2* – are not very abundant for *ENO1* around the respective TATA boxes and also locate around the TSS of *CUT166* in the URS (Supplementary Fig. 8c). Their low abundance is reflected in the low levels of transcription on glucose which after depletion of Nab3 only enhance slightly (Fig. 6e-f), in line with reduced NNS activity on the 5'UTR RNAs and the 5' end of the *ENO1* transcript. Protein factors involved with transcription initiation shape the DNA[28]. When repressive factors like Sgc1, that bind the URS just upstream of the TATA-box, would counteract this, formation of a productive transcription initiation complex would be delayed and thereby the association of Pol II with the TSS. When repression of *ENO1* is lifted, although transcription of *CUT166* still occurs, formation of an active transcription initiation complex at the TSS of *ENO1* is promoted. This could be realized by the release of repressive factors such as Sgc1. The level of induced transcription increases so rapidly that NNS-guided degradation – although increasing as well according to enhanced Nab3-crosslinking – does not have an overall impact on steady state levels. Possibly, NNS-mediated termination of transcripts will

happen on Pol II complexes that are slowed down or do not elongate properly, so that ongoing *ENO1* transcription will not be affected.

Supplementary Methods

Yeast strains and media

Saccharomyces cerevisiae strain BY4741 (MATa; *his3Δ1; leu2Δ0; met15Δ0; ura3Δ0*) was used as the main parental strain[29]. The HTP (HIS6-TEV-2xProtA) carboxyl-tagged strains (Calmodulin binding peptide-TEV-2xProtA) were generated by PCR as described[30,31]. Strains used in this study are listed in Supplementary Table 1. For the anchor away experiments we fused Nab3 to the FRB domain and integrated a HTP-tag at the 3' end of the Rpo21 gene in the HHY168 strain[1]. Rapamycin was added to the media at a final concentration of 1 µg/ml. For the glucose deprivation experiments, strains were grown in synthetic medium lacking tryptophan (Formedium) in the presence (SD-TRP) or absence of glucose (S-TRP). For the PAR-CLIP experiments, cells were grown in synthetic glucose-containing medium lacking tryptophan and uracil (SD-URA-TRP). Strains carrying changes in cross-linked Nab3-motifs were generated by site-directed mutagenesis in two steps using the Delitto-Perfetto system[32]; first a small deletion was generated covering the *YBR085C-A* Nab3 motifs by insertion of the cassette from plasmid pGSHU, which was then replaced by a gBlock (Integrated DNA Technologies) containing the mutations. Resultant strains were checked by sequencing and for normal growth on glycerol.

Western blot analyses

Western blot analysis was performed using the polyclonal rabbit anti-TAP antibody from Thermo Fisher (CAB1001), which recognizes the spacer between the TEV cleavage site and the six histidines. Blots were incubated with the antibody (1:5000 dilution) in blocking buffer (5% nonfat milk powder, 0.1% Tween-20 and phosphate buffer saline (PBS)) for one hour at room temperature (diluted 1:5000 in blocking buffer. Following two five minute washes with PBS-0.1% Tween, the blots were then incubated with goat anti-rabbit antibodies (Thermo Fisher (31466) 1:5000 in blocking buffer) for one hour at room temperature. Proteins were visualized using the Pearce enhanced chemiluminescence solutions as described by the manufacturer's procedures.

Quantitative RT-PCR

Cells were grown in SD-TRP to an OD_{600} of 0.4, harvested by filtration and then shifted to S-TRP. Cells were harvested before the shift (0) and 20, 40 minutes after the shift. RNA was extracted using the Guanidium thiocyanide method[33] or the masterpure yeast RNA purification kit (Epicentre) and quantitative RT-PCR was carried out using the Agilent Brilliant III SYBR master mix, using oligonucleotides listed in Supplementary Table 3. In an end-volume of 10 µl, 12 µg total RNA was treated with 2 units Turbo DNAseI (Ambion, ThermoFisher) in the presence of 4 units RNAsin (Promega) at 37ºC for 1 hour, followed by 15 min. at 65ºC to inactivate the enzyme. After addition of 2.5 µl 2.5 µM reverse PCR primers the nucleic acids are denatured at 85ºC for 3 min and then cooled on ice for 5 min. Of the annealed RNA/oligo mixture, 5 µl is added to a tube

containing 5 µl 2*RT-mix (100 u Superscript III reverse transcriptase (Invitrogen), 2 µl 5xFirstStrandBuffer, 0.5 µl 0.1 M DTT, 1.5 µl 5 mM dNTP mix, 1 u RNAsin, H_2O to 5 µl) and another 5 µl to a 2*NoRT-mix (as 2*RT mix but with H_2O replacing the enzyme) and incubated for 90 min at 55ºC. The enzyme is inactivated for 20 min at 65ºC and RNA/cDNA hybrids are resolved by digestion with 5 units RNAse H for 30 min at 37ºC. The mixture is diluted by adding 200 µl H_2O and 3 µl is added to 7 µl QPCR mix (3 µl water with 1µM of each primer, 4 µl 2*Brilliant III SYBR master) and amplified on a LightCycler 96 or 480 (Roche): 1 cycle 95ºC; 40 cycles (plate reading at end of each cycle) 95ºC 5s; 60ºC 10s; 72ºC 15s. A melt curve was generated by ramping to 95ºC (0.11; 5/°C) with continuous reading. QPCR reactions on each plate were done in triplicate.

Northern blot analysis

Total RNA was resolved on a 1.25% Agarose Bis-Tris (pH 7) gel and transferred to nitrocellulose. Northern blotting was performed using UltraHyb hybridization buffer according to the manufacturer's procedures (Ambion). The snR13 oligo sequence used for hybridization is provided in Supplementary Table 3.

Supplementary Tables

Strain	Genotype	Reference
BY4741	MATa; *his3Δ1; leu2Δ0; met15Δ0; ura3Δ0*	[29]
HHY168	MATα; *tor1-1; fpr1::NAT RPL13A-2×FKBP12::TRP*	[1]
HHY110	MATα; *tor1-1; fpr1::NAT PMA1-2×FKBP12::TRP*	[1]
YSG882	As BY4741 but with *nab3::HTP::K.l.URA3*	This study
YSG1013	As BY4741 but with *rpo21::HTP::K.l.URA3*	This study
YSG1010	As HHY110 but with *rpo21::HTP::K.l.URA3*	This study
YSG1042	As HHY168 but with *nab3::FRB::KAN and rpo21::HTP::K.l.URA3*	This study
YSG1051	As BY4741 but with allele *YBR085C*-B1547*	This study

Supplementary Table 1. *Saccharomyces cerevisiae* strains used in this study

5' adapters	Sequence (5'-3')
L5Aa	invddT-ACACrGrArCrGrCrUrCrUrUrCrCrGrArUrCrUrNrNrNrUrArArGrCrN
L5Ab	invddT-ACACrGrArCrGrCrUrCrUrUrCrCrGrArUrCrUrNrNrNrArUrUrArGrCrN
L5Ac	invddT-ACACrGrArCrGrCrUrCrUrUrCrCrGrArUrCrUrNrNrNrGrCrGrCrArGrCrN
L5Ad	invddT-ACACrGrArCrGrCrUrCrUrUrCrCrGrArUrCrUrNrNrNrCrGrCrUrUrArGrCrN
L5Ba	invddT-ACACrGrArCrGrCrUrCrUrUrCrCrGrArUrCrUrNrNrNrArGrArGrCrN
L5Bb	invddT-ACACrGrArCrGrCrUrCrUrUrCrCrGrArUrCrUrNrNrNrGrUrGrArGrCrN
L5Bc	invddT-ACACrGrArCrGrCrUrCrUrUrCrCrGrArUrCrUrNrNrNrCrArCrUrArGrCrN
L5Bd	invddT-ACACrGrArCrGrCrUrCrUrUrCrCrGrArUrCrUrNrNrNrUrCrUrCrUrArGrCrN
L5Ca	invddT-ACACrGrArCrGrCrUrCrUrUrCrCrGrArUrCrUrNrNrNrCrUrArGrCrN
L5Cb	invddT-ACACrGrArCrGrCrUrCrUrUrCrCrGrArUrCrUrNrNrNrUrGrGrArGrCrN
L5Cc	invddT-ACACrGrArCrGrCrUrCrUrUrCrCrGrArUrCrUrNrNrNrArCrUrCrArGrCrN
L5Cd	invddT-ACACrGrArCrGrCrUrCrUrUrCrCrGrArUrCrUrNrNrNrGrArCrUrUrArGrCrN
L5Da	invddT-ACACrGrArCrGrCrUrCrUrUrCrCrGrArUrCrUrNrNrNrCrGrUrGrArUrN
L5Db	invddT-ACACrGrArCrGrCrUrCrUrUrCrCrGrArUrCrUrNrNrNrGrCrArCrUrArN
L5Dc	invddT-ACACrGrArCrGrCrUrCrUrUrCrCrGrArUrCrUrNrNrNrUrArGrUrGrCrN
L5Dd	invddT-ACACrGrArCrGrCrUrCrUrUrCrCrGrArUrCrUrNrNrNrArUrCrArCrGrN
L5Ea	invddT-ACACrGrArCrGrCrUrCrUrUrCrCrGrArUrCrUrNrNrNrCrArCrUrGrUrN
L5Eb	invddT-ACACrGrArCrGrCrUrCrUrUrCrCrGrArUrCrUrNrNrNrGrUrGrArCrArN
L5Ec	invddT-ACACrGrArCrGrCrUrCrUrUrCrCrGrArUrCrUrNrNrNrUrGrUrCrArCrN
L5Ed	invddT-ACACrGrArCrGrCrUrCrUrUrCrCrGrArUrCrUrNrNrNrArCrArGrUrGrN
3' adapter	**Sequence (5'-3')**
App-PE	App-NAGATCGGAAGAGCACACG
RT primer	**Sequence (5'-3')**
PE-RT	CAGACGTGTGCTCTTCCGATCT
PCR primers	**Sequence (5'-3')**
Forward primer:	
P5	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
Reverse indexing primers:	
BC1	CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
BC2	CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
BC3	CAAGCAGAAGACGGCATACGAGATGCCTAAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
BC4	CAAGCAGAAGACGGCATACGAGATTGGTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
BC5	CAAGCAGAAGACGGCATACGAGATCACTGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
BC6	CAAGCAGAAGACGGCATACGAGATATTGGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
BC7	CAAGCAGAAGACGGCATACGAGATCAGATCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
BC8	CAAGCAGAAGACGGCATACGAGATTAGCTTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
BC9	CAAGCAGAAGACGGCATACGAGATGATCAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
BC10	CAAGCAGAAGACGGCATACGAGATATCACGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

Supplementary Table 2 - Oligonucleotides for χCRAC library preparation.

Lowercase ‘r’ indicates RNA nucleotides, ‘rN’ indicate random RNA nucleotides and blue sequences are barcodes.

qPCR primer	Sequence (5'-3')
ACT1-exon2-F	TGTTTTGGATTCCGGTGATGG
ACT1-exon2-R	ACCGGCCAAATCGATTCTCA
PIC2-qFwd	CCGCTGAATTCCTCGCTGAT
PIC2-qRev	TTGCAAAAGGGCGGCATAGT
NRD1-qFwd	TCCAGCCAAACAGCAAGCAG
NRD1-qRev	TGTATGGGCTTGGGGAGGTG
MAL33-qFwd	CGGACCCTCGACTTTCTTTGG
MAL33-qRev	GACTCCGTAGCCAAAGCATCAAA
IMD3-qFwd	TCCAATGGACACCGTGACAGA
IMD3-qRev	GTCAGCTTGGTCCTCTGGGGTA
ENO1-5UTR-Fwd	CTGTGGCCTTTTCTGGCACA
ENO1-5UTR-Rev	AGGGATTACAAGAGAGATGTTACAAGAAAGAA
ENO1-Fwd	TGCCGCTGCTGAAAAGAATG
ENO1-Rev	CGTGGGAACCACCGTTCAA
YBR085C-A-Fwd	CGACAGTCAGTGGGTACCAGGA
YBR085C-A-Rev	CGATGACTTCGCCGTCCTTT
TY1-2Fwd	ACGCTACACACGTCATCGACATC
TY1-2Rev	TGCTGTTTCTCGATCCCTGTTG
TY1-3Fwd	CGCCTCTGAGCACTCCATCA
TY1-3Rev	GGTGAGGTTAACATTGGTGGTGGT
TY1-4Fwd	CACCTGGGCCACAATCACAG
TY1-4Rev	AGAGTCCGCTGAGGATGAATCAGT
TY2-1Fwd	CTCGCACATCTCCAAACACGA
TY2-1Rev	TGTGAGCTTTTGCTGCTCTTGG
TY2-3Fwd	AAACGTCACCTGCGTATTATCAACC
TY2-3Rev	TCGATTGGATCAGGTGAGGAAGT
TY2-4Fwd	CTGAGTACTTCCTCACCTGATCCAATC
TY2-4Rev	TTCTGATGTTAAAGTGTGTGGTGGTAAGA

Northern primer	Sequence (5'-3')
snr13_Northern	GCCAAACAGCAACTCGAGCCAAATGCACTCATATTCATC

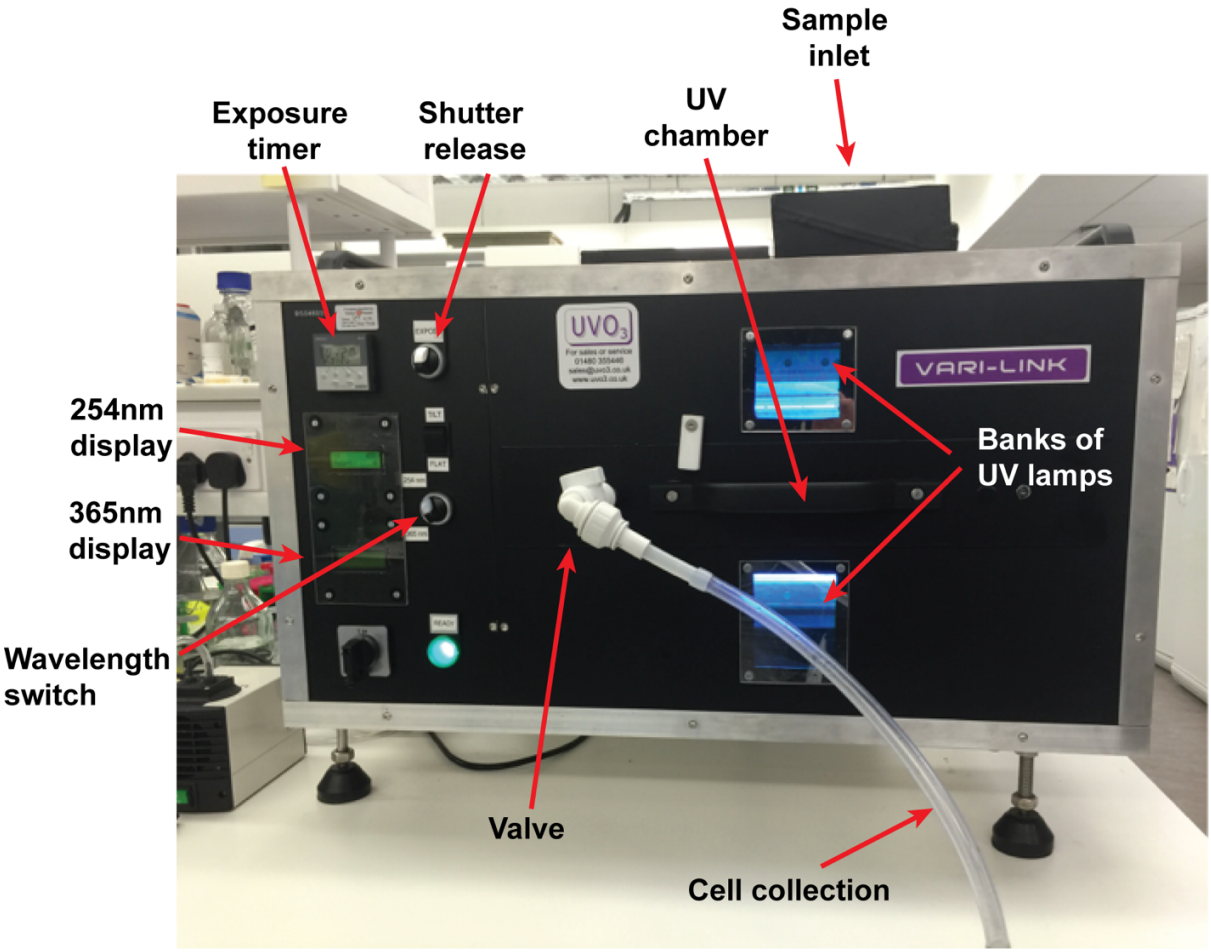
Supplementary Table 3 - Oligonucleotides used for qRT-PCR and Northern blot analysis

Glucose*	No Glucose (4 minutes)	No Glucose (18 minutes)
CUP1-1	AHA1	ACA1
FLX1 (snR68)	ARG82	BAP3
GRE3	ATP1	GUT2
HEM4 (snR5)	DIP5	MAL31
HIS1	ECL1	MAL33
ICT1	ERO1	NRD1
IMD2	GNP1	PIC2
IMD3	GRE3	RHO5
NRD1	GUT2	UIP4
SEN2 (snR79)	HIS4	YLR171W
TRS31 (snR13)‡	NRD1	YLR410W-B
URA8	RPN4	YOR059C
YCL007C (snR43)	TRS31	
	URA8	
	WTM1	
	YAR010C	
	YTM1	

*) Genes in italics might be identified due to transcripts originating from a preceding snoRNA gene (in brackets) that is not properly terminated in absence of Nab3.
‡) See also Supplementary Fig. 5b (Northern blot probed for snR13) and Supplementary Fig. 6c.

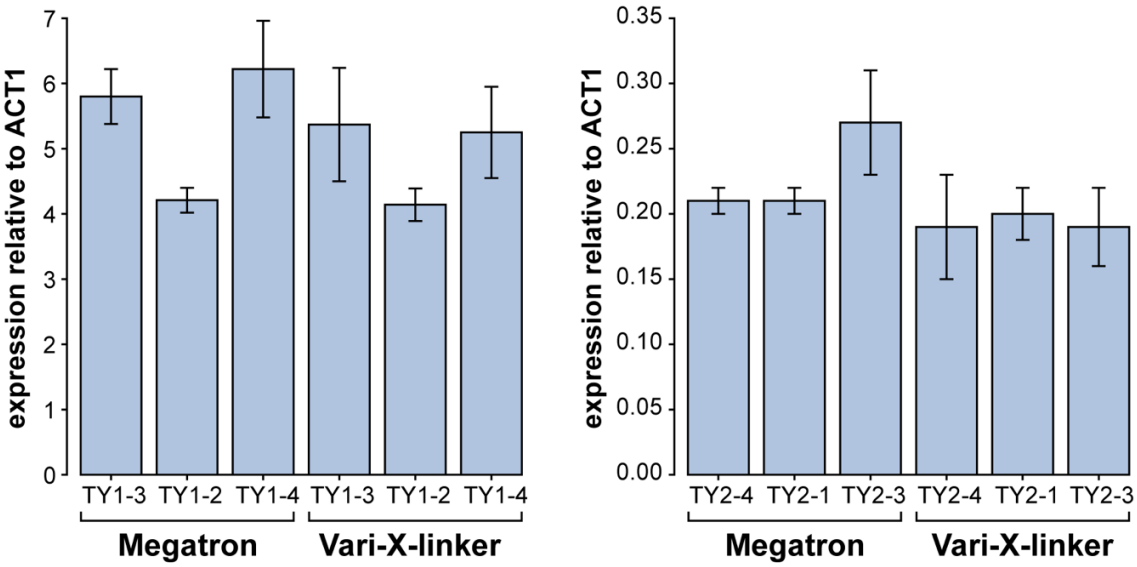
Supplementary Table 4 - List of genes controlled by Nab3 attenuation

Supplementary Figure Legends



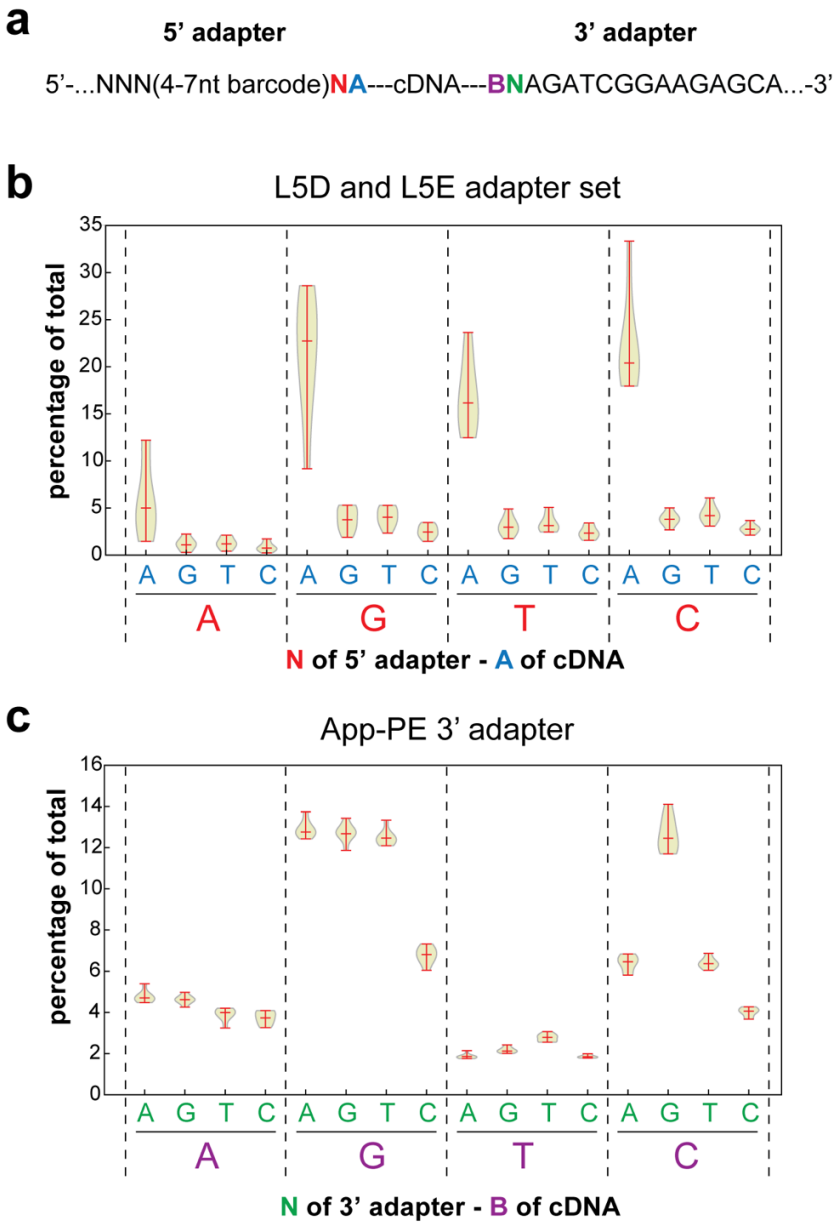
Supplementary Figure 1. The Vari-X-linker

Picture of the latest Vari-X-linker prototype. Cells are poured into the machine on the top of the unit where the sample enters the UV-irradiation chamber. Sample is UV irradiated from both sides and irradiation length can be controlled via a shutter system. Indicated are some of the features of the machine. The machine can be loaded with 254nm and 365 nm lamps, which can easily be swapped. The cells can be extracted by attaching a pump to the tube connected to the valve. The cells are loaded into the UV chamber on the top of the unit. Separate UV sensors were built in for 254nm and 365mm lamps, which shows at what percentage of the maximum output the lamps are. The machine also has a tilting mechanism to allow the bag in the UV chamber to empty completely.



Supplementary Figure 2. Ty1 and Ty2 transcript levels do not significantly change during UV irradiation.

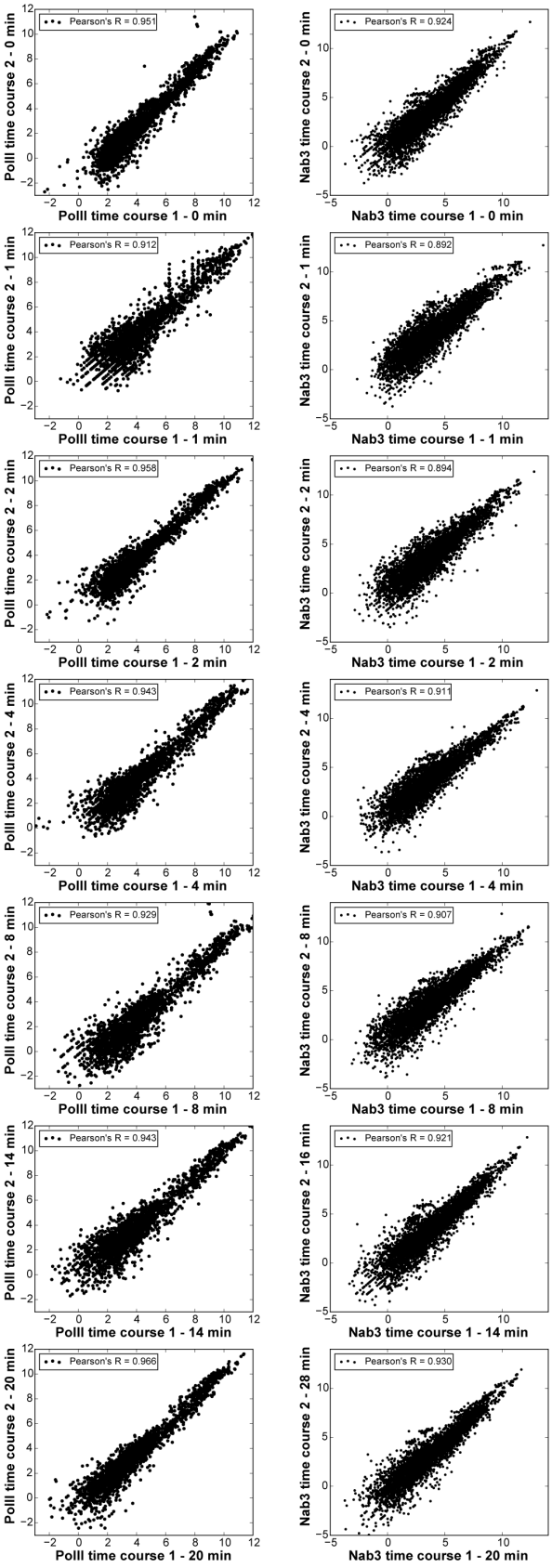
Cells were cross-linked in the Megatron (100 seconds) and the Vari-X-linker (12 seconds) and qRT-PCR was performed with three different primer sets on total RNA to measure levels three variants of Ty1 and Ty2 retrotransposons. Transcript levels were normalized to those of the *ACT1* gene (y-axis). Error bars indicate s.d. from four experimental replicates.



Supplementary Figure 3. T4 RNA ligase has preference for specific nucleotide donor-acceptor nucleotide combinations.

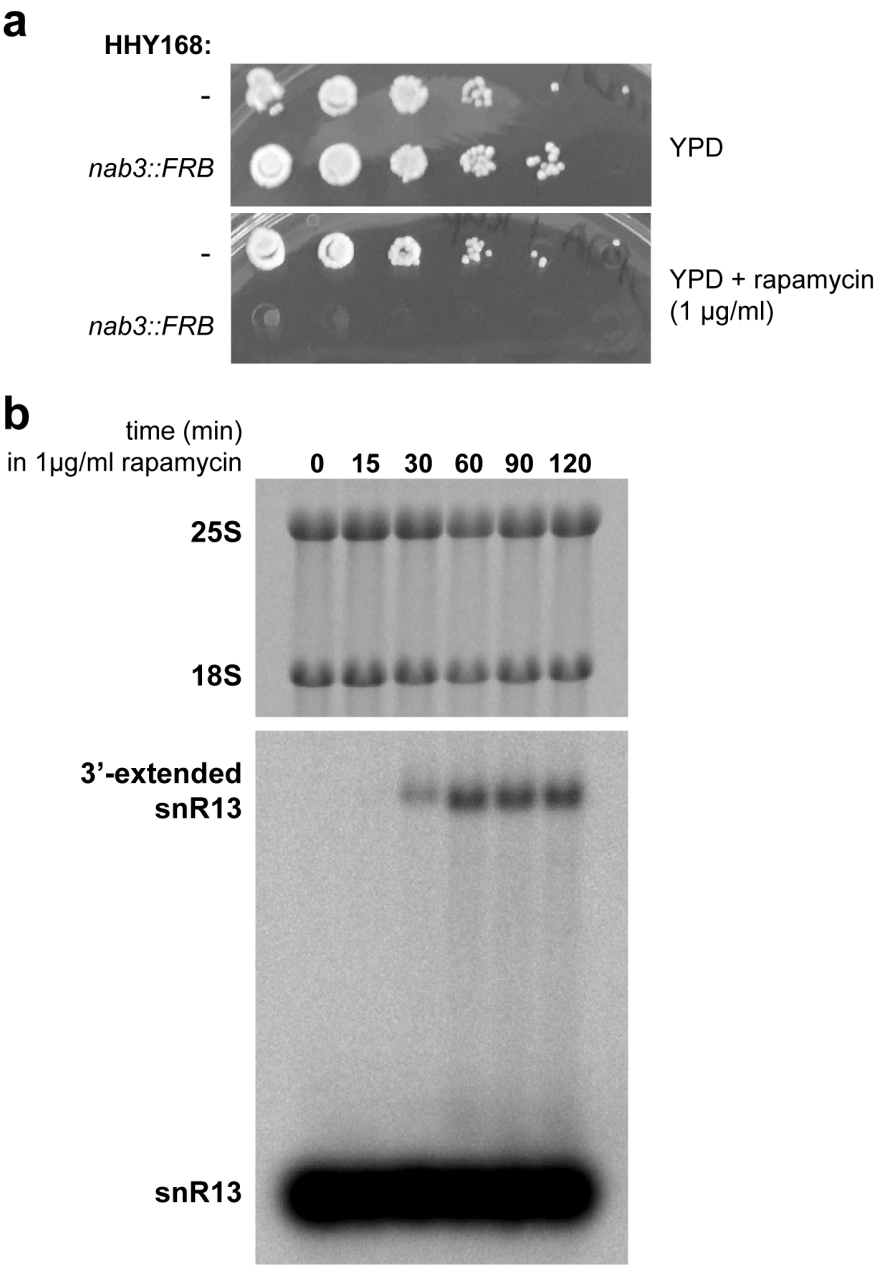
(**a**) Adapter strategy. Both 5' and 3' adapters contain a random nucleotide at the termini that ligate to the cDNA (red colored and green colored "N"). The blue "A" in the sequence indicates the first nucleotide of the cDNA that ligates to the 5' adapter. The purple "B" indicates the last nucleotide that ligates to the 3' adapter. All adapter sequences are provided in Supplementary Table 2.

(**b**-**c**) T4 RNA ligase has a preference for specific nucleotide combinations. The violin plots show the distribution of donor-acceptor nucleotides found in a Pol II χCRAC dataset using the L5D and L5E series of 5' adapters. The data show that T4 RNA ligase has a clear preference for specific donor-acceptor nucleotide combination, with an A being the preferred nucleotide at the 5' end of cDNAs. The variability in some of the data points indicates that the surrounding nucleotides can influence the ligation efficiency. However, there was not a specific 5' adapter barcode sequence that showed significantly better or worse ligation efficiencies. The violin plot in (**c**) shows that the truncated T4 RNA ligase (NEB) has a strong preference for a G or a C at the 5' end of the 3' adapter sequence, although GC and CC combinations are less enriched.


Pearson's R = 0.951
PolII time course 2 - 0 min
PolII time course 1 - 0 min
Pearson's R = 0.924
Nab3 time course 2 - 0 min
Nab3 time course 1 - 0 min
Pearson's R = 0.912
PolII time course 2 - 1 min
PolII time course 1 - 1 min
Pearson's R = 0.892
Nab3 time course 2 - 1 min
Nab3 time course 1 - 1 min
Pearson's R = 0.958
PolII time course 2 - 2 min
PolII time course 1 - 2 min
Pearson's R = 0.894
Nab3 time course 2 - 2 min
Nab3 time course 1 - 2 min
Pearson's R = 0.943
PolII time course 2 - 4 min
PolII time course 1 - 4 min
Pearson's R = 0.911
Nab3 time course 2 - 4 min
Nab3 time course 1 - 4 min
Pearson's R = 0.929
PolII time course 2 - 8 min
PolII time course 1 - 8 min
Pearson's R = 0.907
Nab3 time course 2 - 8 min
Nab3 time course 1 - 8 min
Pearson's R = 0.943
PolII time course 2 - 14 min
PolII time course 1 - 14 min
Pearson's R = 0.921
Nab3 time course 2 - 16 min
Nab3 time course 1 - 14 min
Pearson's R = 0.966
PolII time course 2 - 20 min
PolII time course 1 - 20 min
Pearson's R = 0.930
Nab3 time course 2 - 28 min
Nab3 time course 1 - 20 min

Supplementary Figure 4. χCRAC generates highly reproducible data.

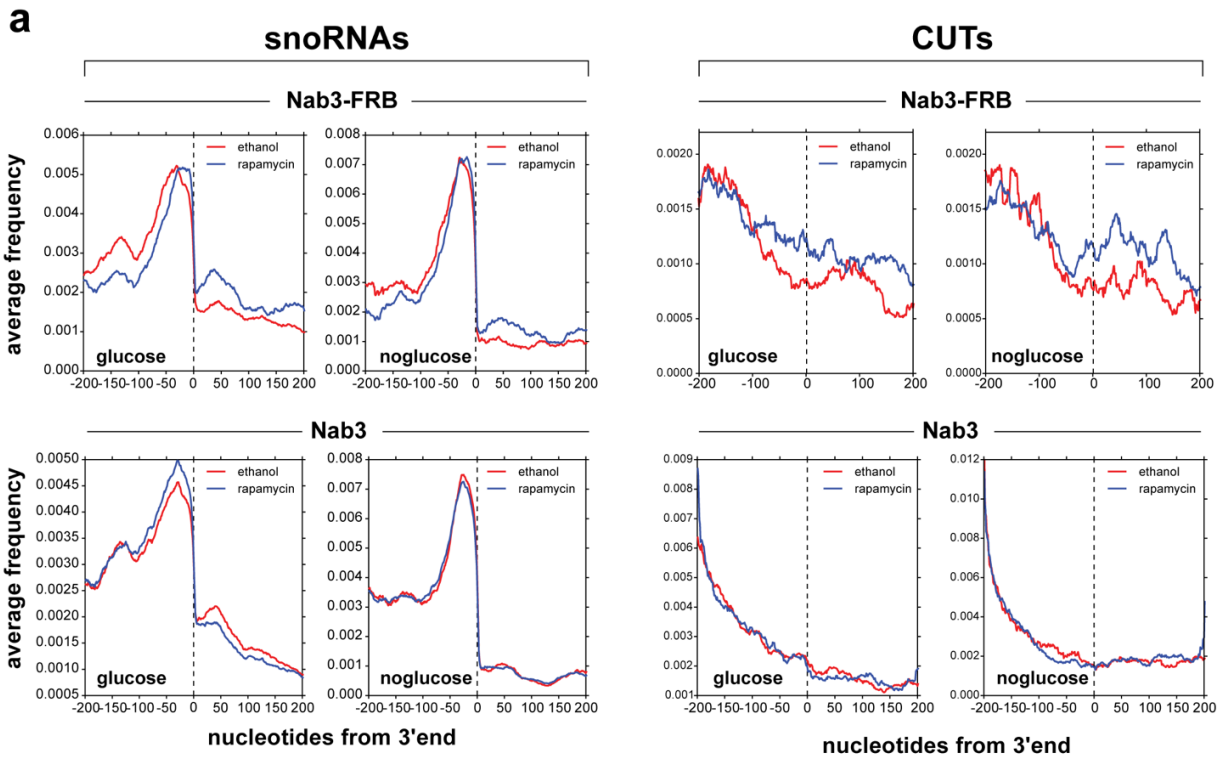
The scatter plots show the pairwise comparison of Pol II (left) and Nab3 (right) biological replicates for each time point (0-20 min). All datasets were log2 transformed before Pearson correlations were calculated. Pearson's R correlations between the time-points are shown in the top left corner.

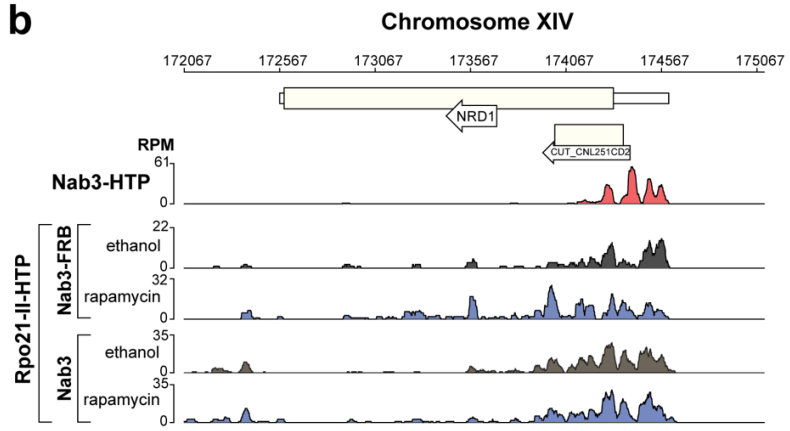


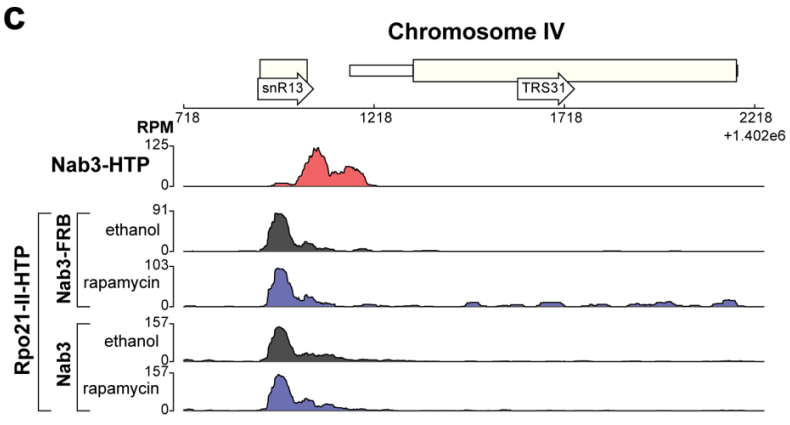
Supplementary Figure 5. The Nab3-FRB strain can be effectively used to deplete Nab3 from the nucleus.

(**a**) Cells expressing Nab3-FRB die on plates with rapamycin. Shown is a serial dilution assay where growth of the parental strain (HHY168[3]) is compared to the strain expressing Nab3-FRB. Cells were spotted on YPD and YPD supplemented with rapamycin (1µg/µl). The results show that rapamycin treatment of the *nab3::frb* strain is lethal.

(**b**) Northern blot analysis of 3' extended snR13 species during a rapamycin time-course. The HHY168 *nab3::FRB* strain was grown in YPD to exponential phase. Cells were harvested shortly at exponential phase (t=0). Subsequently, rapamycin was added to a final concentration of 1µg/ml and cells were harvested 15, 30, 60, 90 and 120 minutes after adding the drug. Total RNA was extracted and resolved on a 1.25% agarose gel. Ribosomal RNAs (25S and 18S) were detected by SyBr safe staining. After transferring the RNA to a nitrocellulose membrane, the blot was probed with an anti-sense snR13 oligonucleotide (Supplementary Table 2). After about 60 minutes of rapamycin treatment the amount of 3' extended snR13 species reached its maximum level. Therefore, for subsequent depletion experiments a 60-minute rapamycin incubation was used.


a
snoRNAs
CUTs
Nab3-FRB
Nab3
ethanol
rapamycin
glucose
noglucose
average frequency
nucleotides from 3'end

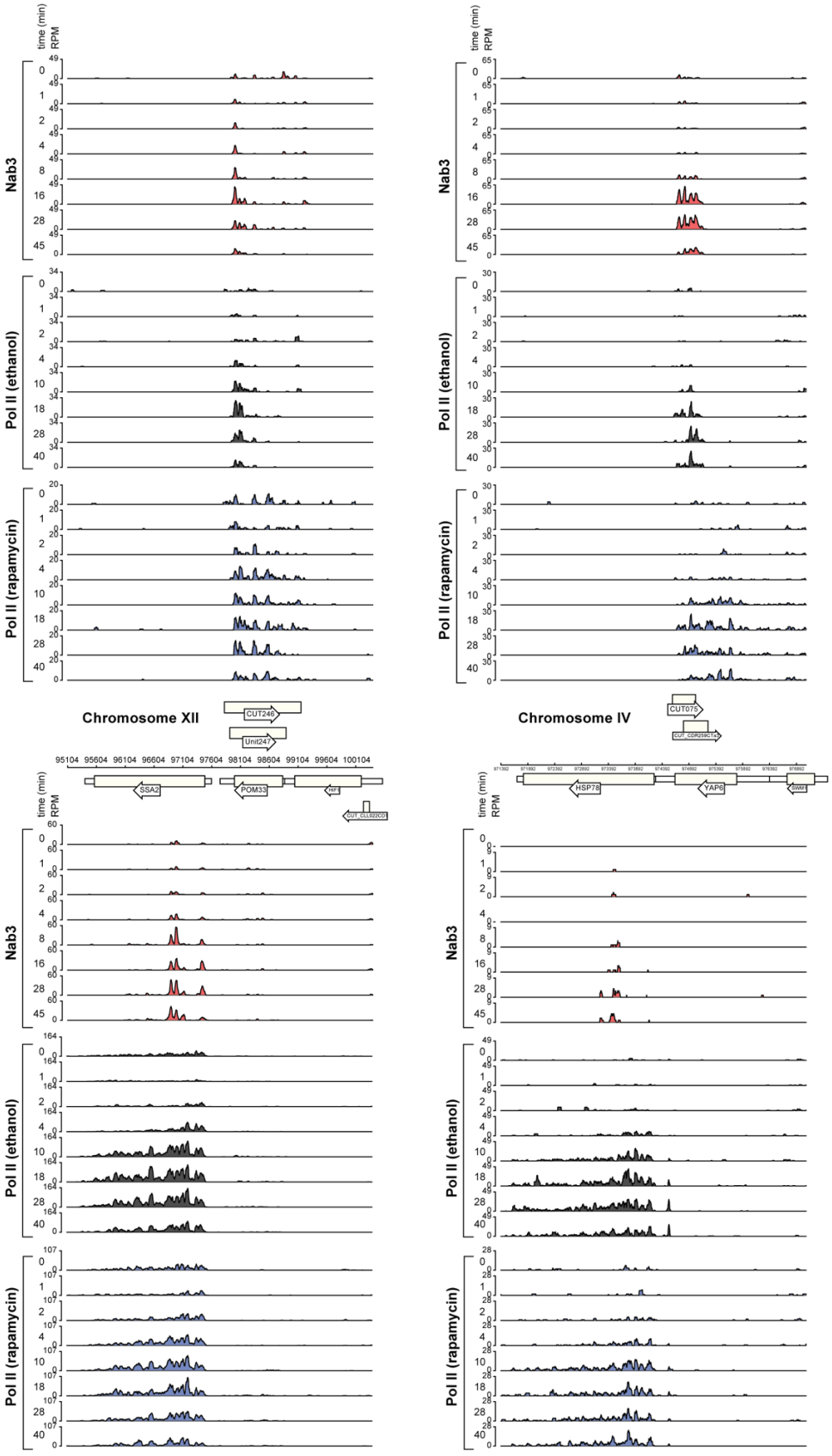

b
Chromosome XIV
172067
172567
173067
173567
174067
174567
175067
NRD1
CUT_CNL251CD2
RPM
Nab3-HTP
Rpo21-II-HTP
Nab3-FRB
Nab3
ethanol
rapamycin


c
Chromosome IV
snR13
TRS31
718
1218
1718
2218
+1.402e6
RPM
Nab3-HTP
Rpo21-II-HTP
Nab3-FRB
Nab3
ethanol
rapamycin

Supplementary Figure 6. Nuclear depletion of Nab3 results in the accumulation of 3' extended CUTs and snoRNAs.

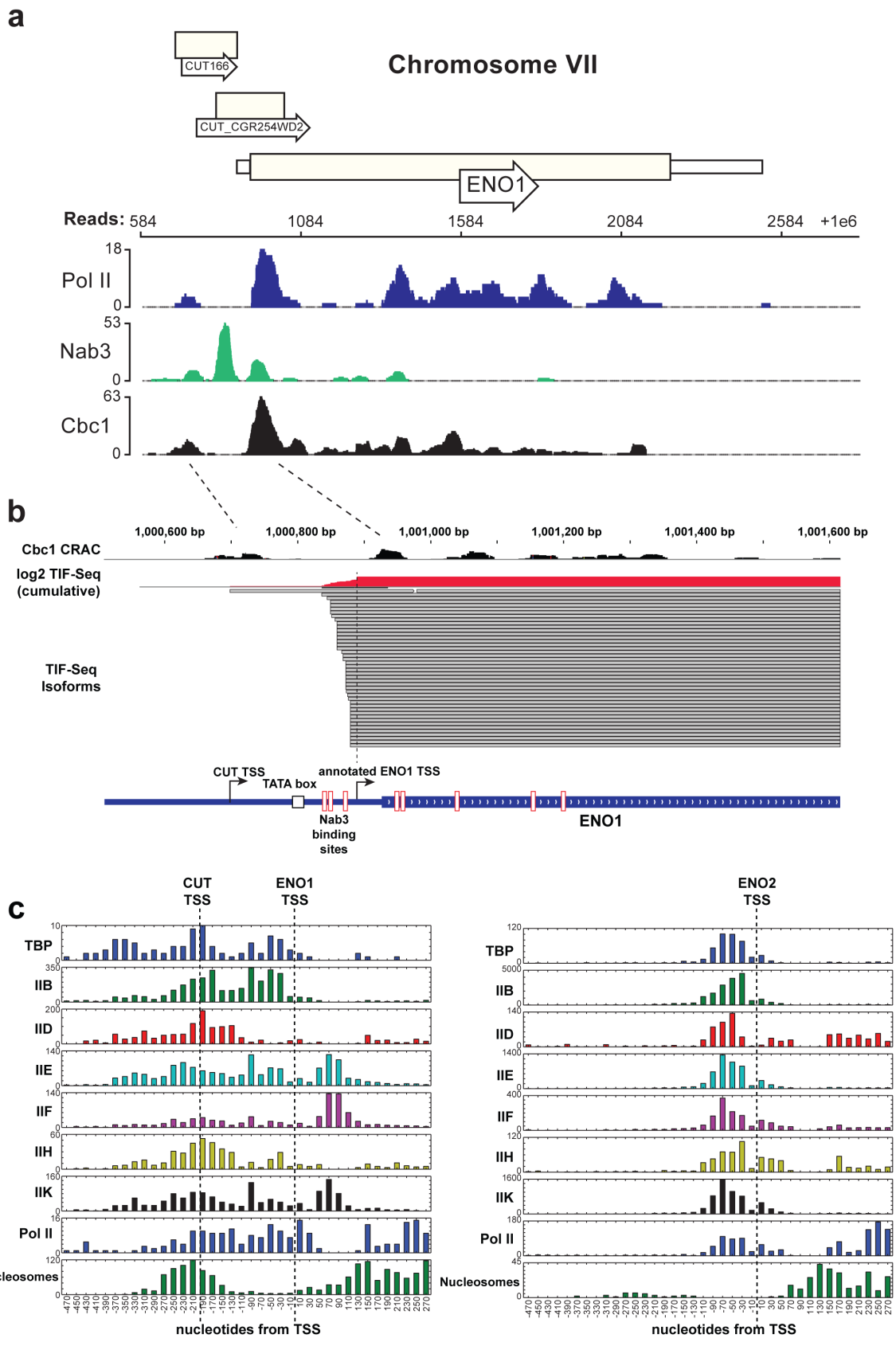
(**a**) Distribution of reads that mapped to CUTs and snoRNAs around the 3' end of the features (x-axis). Cells expressing the FRB-tagged Nab3 and the parental strain (Nab3) were grown in glucose to exponential phase and rapidly shifted to medium lacking glucose. Cells were harvested before the shift (glucose panels) or 14 (Nab3) to 18 minutes (Nab3-FRB) after the shift (no glucose panels). Reads mapped to each features were divided over 400 bins (1nt per bin) and for each bin the fraction of total reads that mapped to each bin was calculated. These numbers were then averaged (y-axis) to generate the plots. The reads for the rapamycin treated cells are represented as a blue line, whereas the reads for the ethanol (control) experiment are represented as a red line. The results show an increase in read density downstream of the annotated 3' ends of CUTs and snoRNAs in rapamycin treated *nab3::frb* cells (but not the parental strain), indicative of defects in transcription termination.

(**b**-**c**) Nab3-dependent termination of known targets are detectable in Pol II χCRAC experiments. Shown are genome browser images of the Nrd1 and snR13 genes. On the y-axis the number of reads per million mapped reads (RPM) is plotted. The top track shows the Nab3 cross-linking data, the second and third tracks show the Pol II cross-linking data in the ethanol and rapamycin treated *nab3::frb rpo21::HTP* strain, respectively. The fourth and fifth track shows the results from the ethanol and rapamycin treated *rpo21::HTP* anchor-away strain.


time (min)
RPM
Nab3
Pol II (ethanol)
Pol II (rapamycin)
Chromosome XII
CUT246
Unit247
95104 95604 96104 96604 97104 97604 98104 98604 99104 99604 100104
SSA2
POM33
Chromosome IV
CUT075
HSP78
YAP6
SWI1

Supplementary Figure 7. Nab3 terminates anti-sense divergent cuts

Genome browser images for the *SSA2* and *HSP78* region showing Pol II χCRAC data from the solvent (ethanol; black), rapamycin treated cells (blue) and data from a Nab3 χCRAC experiment (red) for both strands. On the y-axis of each track the reads per million (RPM) mapped reads is plotted. The time points at which cells were harvested (in minutes) is indicated left of each track.

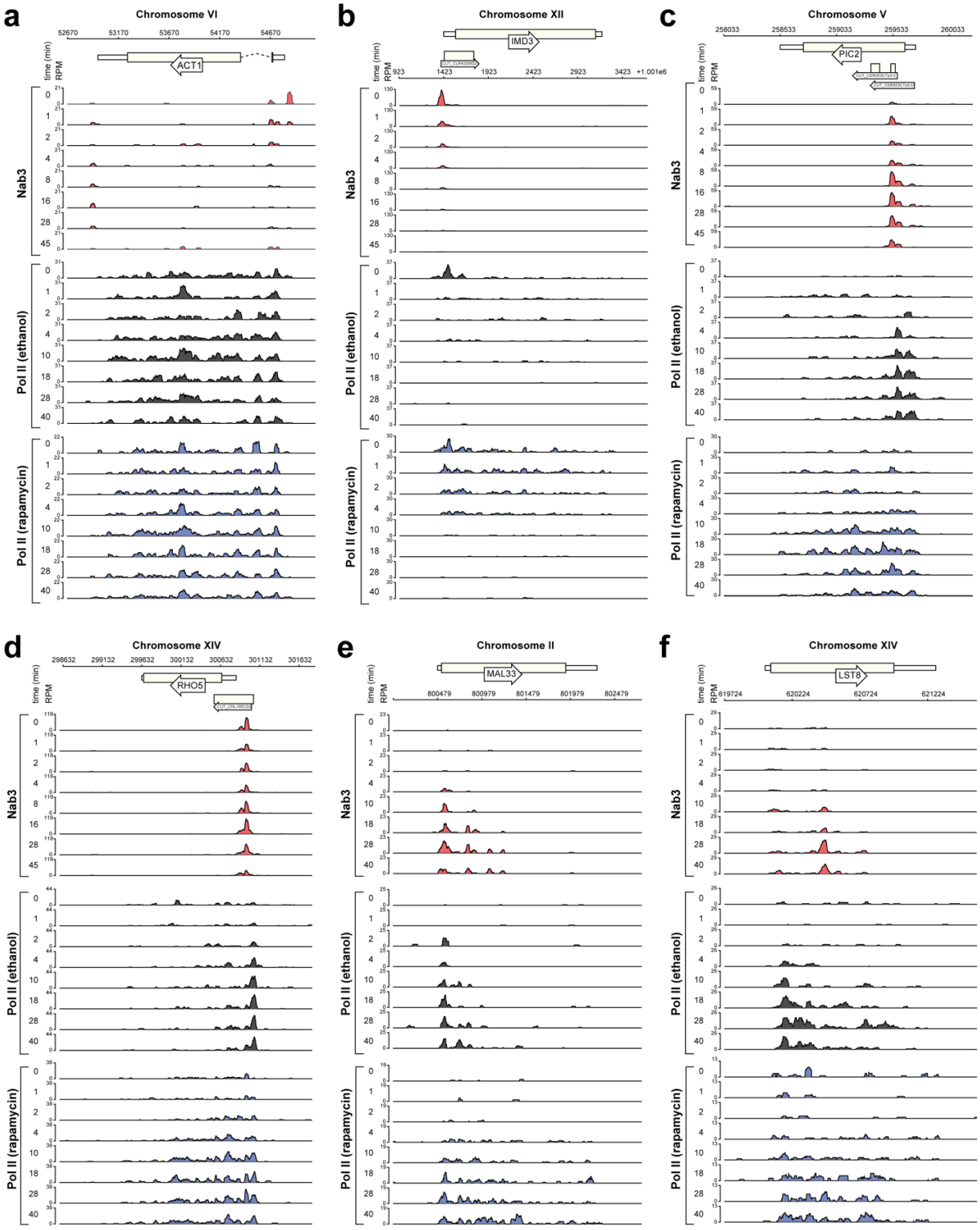

a
Chromosome VII
CUT166
CUT_CGR254WD2
ENO1
Reads: 584
1084
1584
2084
2584
+1e6
Pol II
Nab3
Cbc1
b
1,000,600 bp
1,000,800 bp
1,001,000 bp
1,001,200 bp
1,001,400 bp
1,001,600 bp
Cbc1 CRAC
log2 TIF-Seq (cumulative)
TIF-Seq Isoforms
CUT TSS
TATA box
annotated ENO1 TSS
Nab3 binding sites
ENO1
c
CUT TSS
ENO1 TSS
ENO2 TSS
TBP
IIB
IID
IIE
IIF
IIH
IIK
Pol II
Nucleosomes
nucleotides from TSS

Supplementary Figure 8. Cryptic transcription upstream of *ENO1*

(**a**) Genome browser snapshot of Pol II (blue), Nab3 (green) and cap-binding complex protein Cbc1 (black) CRAC data. The Cbc1 CRAC data was obtained from[23].

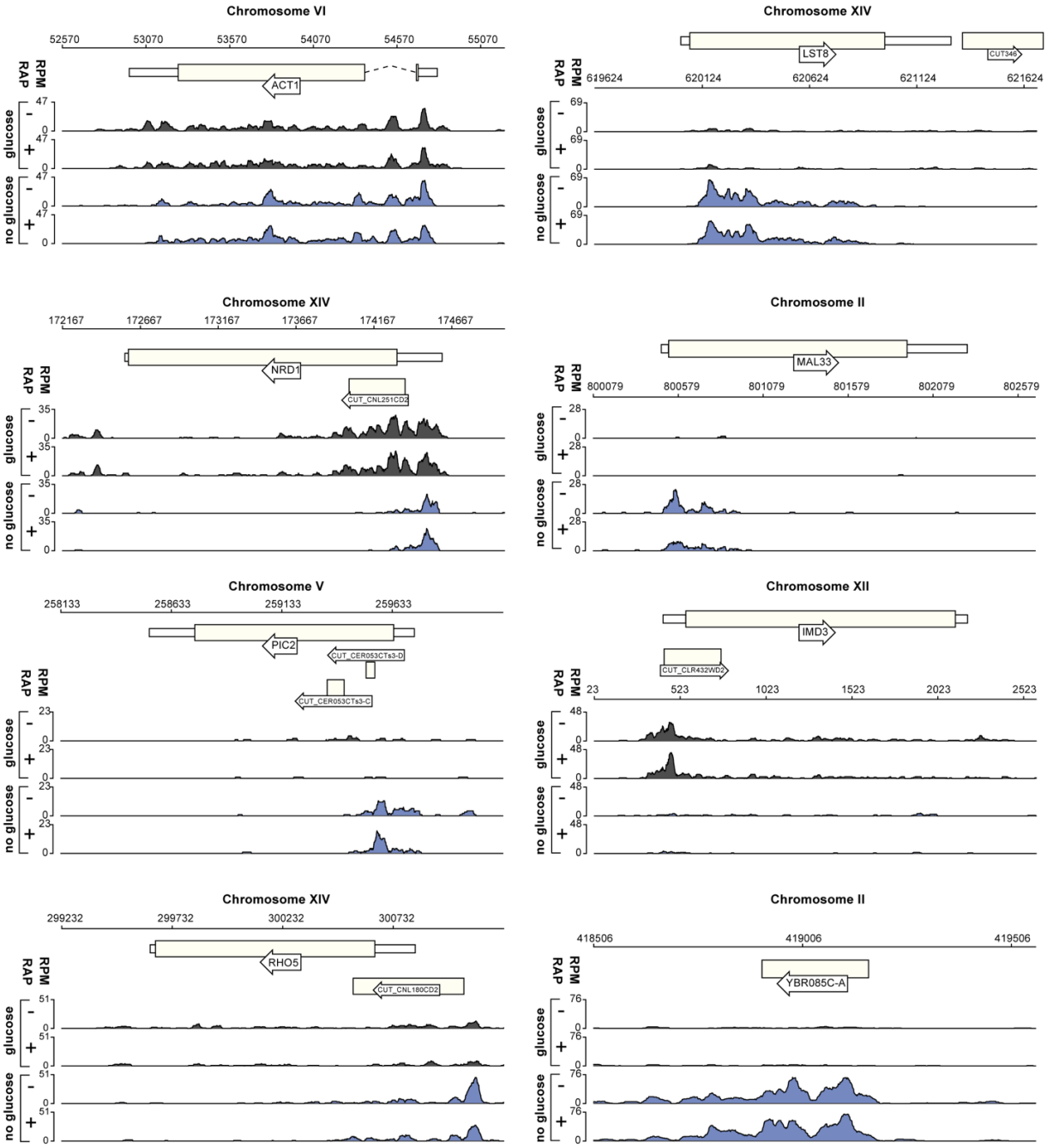
(**b**) Transcription can initiate upstream of the *ENO1* promoter. Genome browser snapshot of Cbc1 CRAC and transcription isoform sequencing data (TIFseq[24]. The first track shows the Cbc1 CRAC data. The second track shows the log2 of the cumulative density of all isoforms (red). The third track shows some of the transcript isoforms that mapped to *ENO1*. The position of the annotated *ENO1* TSS, Nab3 binding sites, *ENO1* TATA box and the TSS of the upstream CUT is shown in the fourth track.

(**c**) Chip-exo sequencing data of Pol II transcription factors[27] for *ENO1* and the orthologous *ENO2*. The y-axis shows the read density and the names of individual transcription factors (TFs). The x-axis shows the nucleotides away from the *ENO1* or *ENO2* transcription start site (TSS). Individual TSSs for *CUT166* and *ENO1* or *ENO2* are indicated with dashed lines. Note the at least 10-fold scale differences between *ENO2* and *ENO1* TF-data, reflecting the difference of transcription on glucose.


a
Chromosome VI
52670
53170
53670
54170
54670
ACT1
b
Chromosome XII
IMD3
923
1423
1923
2423
2923
3423
+1.001e6
c
Chromosome V
258033
258533
259033
259533
260033
PIC2
d
Chromosome XIV
298632
299132
299632
300132
300632
301132
301632
RHO5
e
Chromosome II
MAL33
800479
800979
801479
801979
802479
f
Chromosome XIV
619724
620224
620724
621224
LST8
time (min)
RPM
Nab3
Pol II (ethanol)
Pol II (rapamycin)

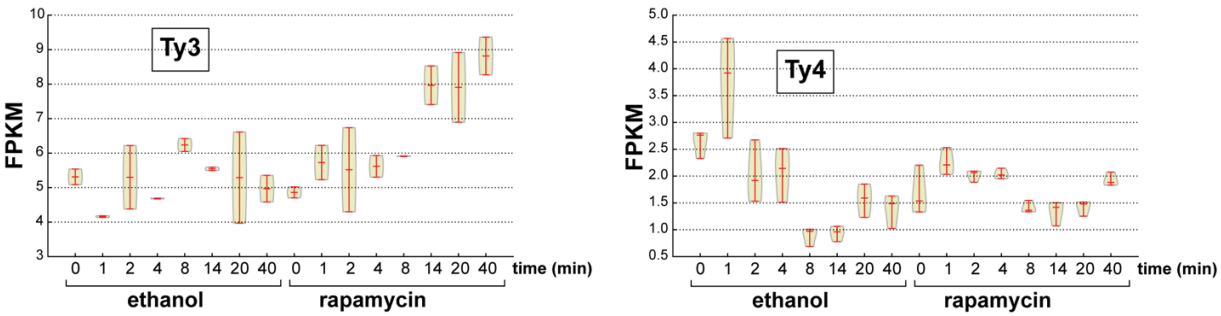
Supplementary Figure 9. Nab3 regulates the induction of stress-responsive protein-coding genes during glucose deprivation.

(**a**-**f**) Genome browser image showing the Pol II χCRAC from the solvent (ethanol; black), rapamycin treated cells (blue) and data from a Nab3 χCRAC experiment (red) for selected genes. On the y-axis of each track the reads per million (RPM) mapped reads are shown. The time points at which cells were harvested (in minutes) is indicated left of each track.


Chromosome VI
52570
53070
53570
54070
54570
55070
ACT1
Chromosome XIV
619624
620124
620624
621124
621624
LST8
CUT346
Chromosome XIV
172167
172667
173167
173667
174167
174667
NRD1
CUT_CNL251CD2
Chromosome II
800079
800579
801079
801579
802079
802579
MAL33
Chromosome V
258133
258633
259133
259633
PIC2
CUT_CER053CTs3-D
CUT_CER053CTs3-C
Chromosome XII
23
523
1023
1523
2023
2523
IMD3
CUT_CLR432WD2
Chromosome XIV
299232
299732
300232
300732
RHO5
CUT_CNL180CD2
Chromosome II
418506
419006
419506
YBR085C-A
RAP
RPM
glucose
no glucose

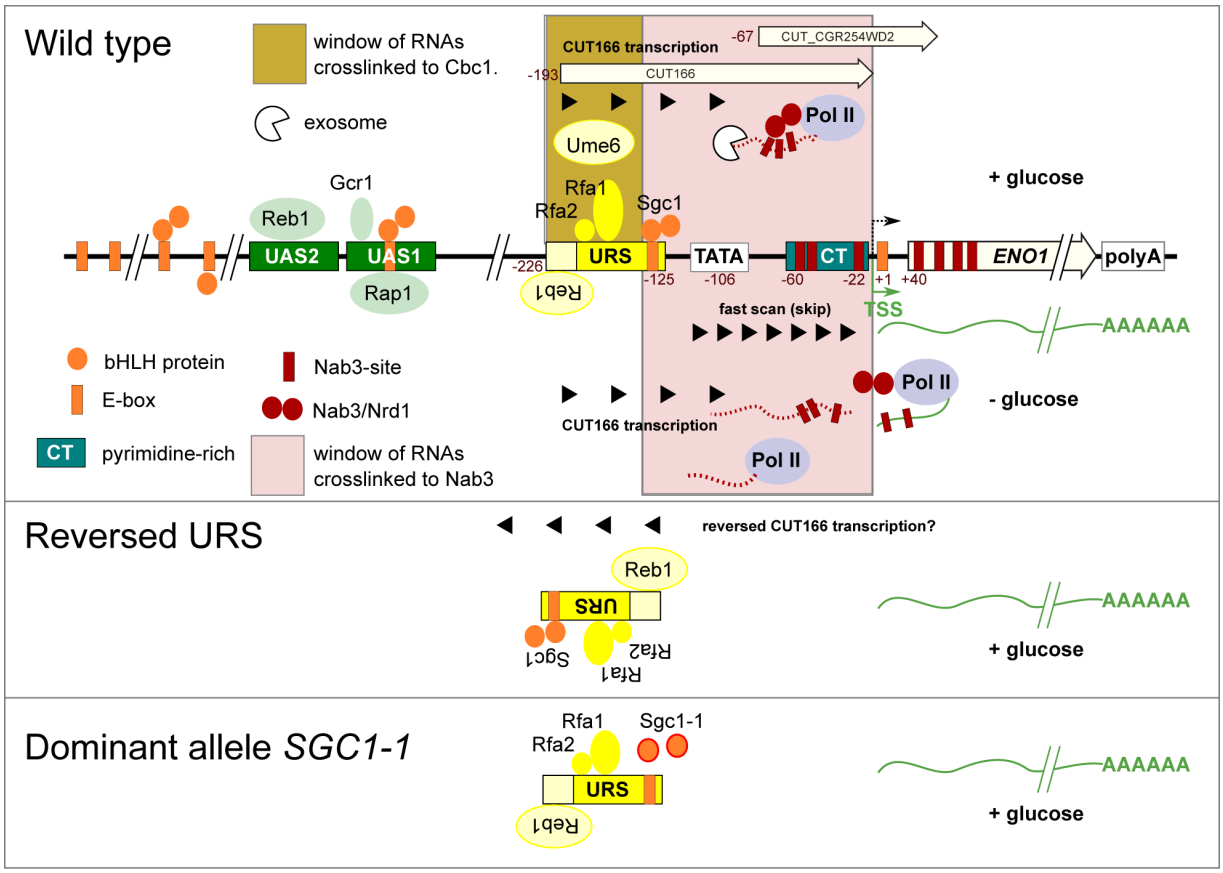
Supplementary Figure 10. The changes in Pol II transcription profiles are not induced by the drug rapamycin itself.

Shown are genome browser images of CRAC data generated using the anchor-away strain expressing an HTP-tagged Rpo21. Cells were grown in glucose to exponential phase and incubated with ethanol (-) or rapamycin (RAP; +) for one hour. A fraction of the cells was harvested (glucose samples) and the rest was shifted to medium lacking glucose for 14 minutes. These data demonstrate that the drug rapamycin does not influence Pol II transcription of these genes.



Supplementary Figure 11. Nab3 regulates the expression of Ty3 retrotransposons during glucose deprivation.

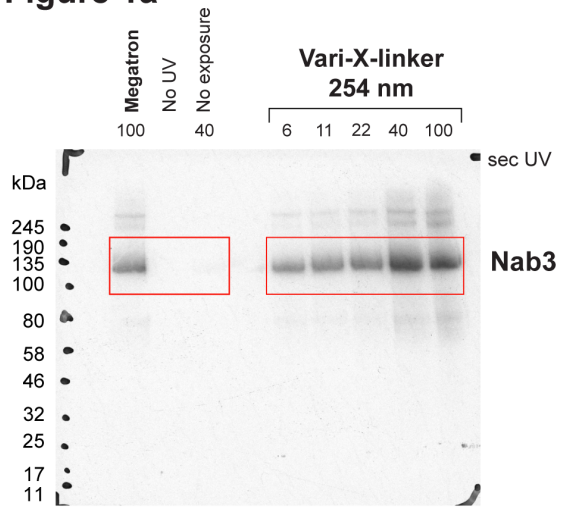
Violin plots showing the Ty3 and Ty4 pol II FPKM distribution from the *nab::frb* Rpo21-HTP χCRAC data generated in the presence of solvent (ethanol) or rapamycin. Shown are the averaged data from two independent experiments. Time (min) indicates the number of minutes in medium lacking glucose.



Supplementary Figure 12. Regulation of *ENO1* transcription is controlled by a repressive element.

Schematic overview of the *ENO1* promoter and control of transcription in the presence or absence of glucose. See Supplementary note 2 for details.

Figure 1a



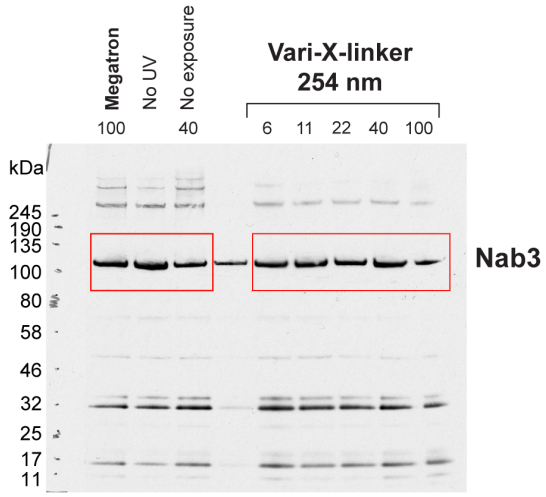
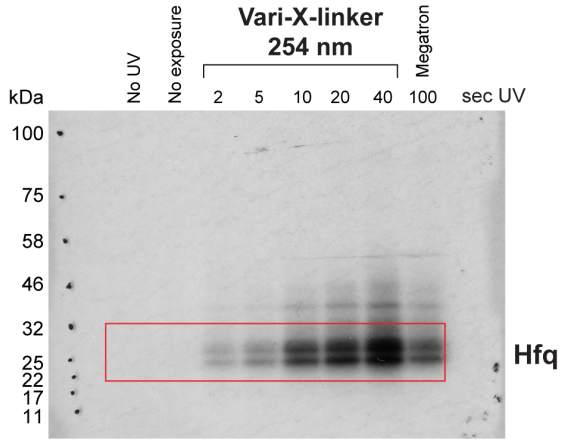


Figure 1b



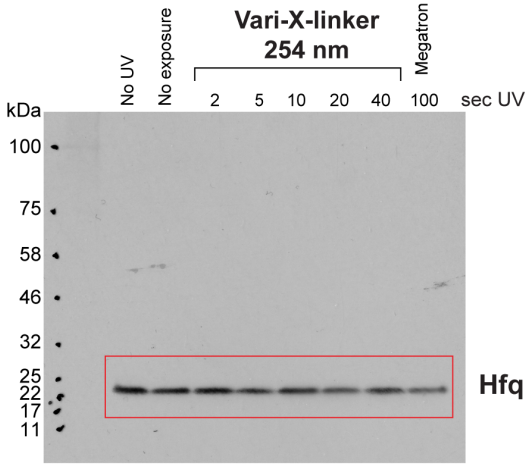
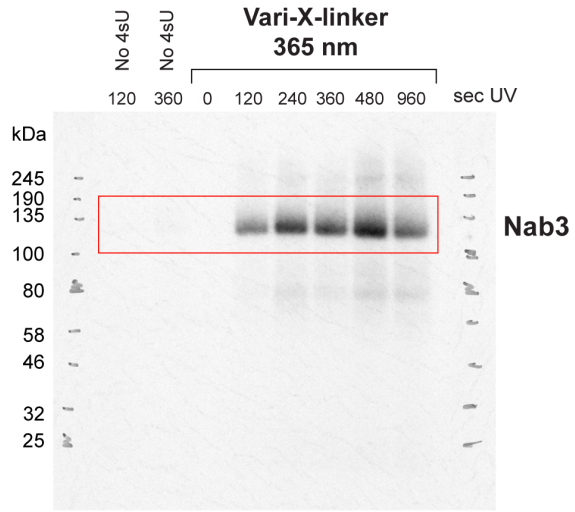
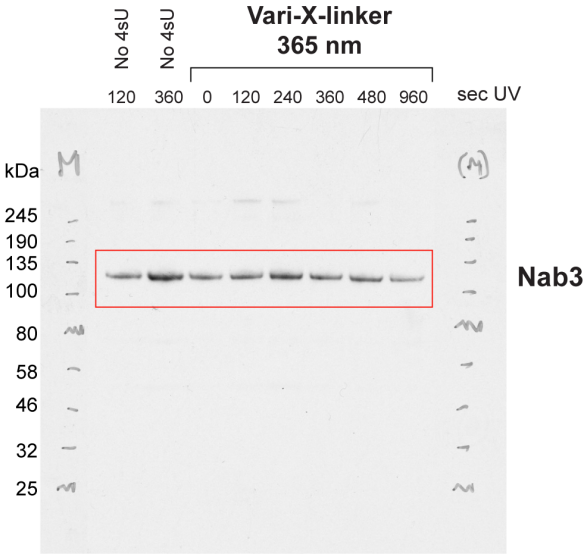


Figure 1c

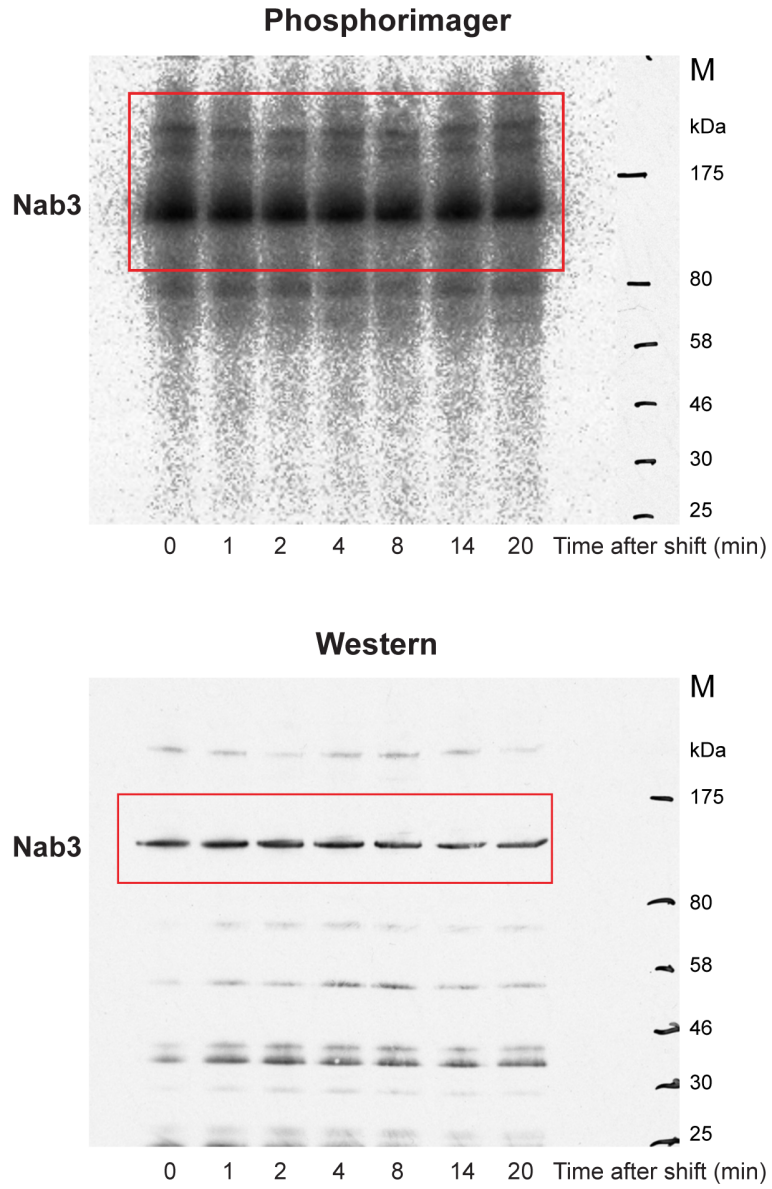




Supplementary Figure 14. Uncropped images.

Uncropped images of films or phosphoimager scans show in Figure 3b. The red boxes indicate the cropped regions.

Figure 3b



Supplementary Figure 14. Uncropped images.

Uncropped images of films or phosphoimager scans show in Figure 3b. The red boxes indicate the cropped regions.

Supplementary References

1. Haruki, H., Nishikawa, J. & Laemmli, U. K. The anchor-away technique: rapid, conditional establishment of yeast mutant phenotypes. *Mol. Cell* **31,** 925–932 (2008).
2. Steinmetz, E. J., Conrad, N. K., Brow, D. A. & Corden, J. L. RNA-binding protein Nrd1 directs poly(A)-independent 3'-end formation of RNA polymerase II transcripts. *Nature* **413,** 327–331 (2001).
3. Arigo, J. T., Carroll, K. L., Ames, J. M. & Corden, J. L. Regulation of yeast NRD1 expression by premature transcription termination. *Mol. Cell* **21,** 641–651 (2006).
4. Brindle, P. K., Holland, J. P., Willett, C. E., Innis, M. A. & Holland, M. J. Multiple factors bind the upstream activation sites of the yeast enolase genes ENO1 and ENO2: ABFI protein, like repressor activator protein RAP1, binds cis-acting sequences which modulate repression or activation of transcription. *Mol. Cell. Biol.* **10,** 4872–4885 (1990).
5. Machida, M., Jigami, Y. & Tanaka, H. Purification and characterization of a nuclear factor which binds specifically to the upstream activation sequence of Saccharomyces cerevisiae enolase 1 gene. *FEBS J.* **184,** 305–311 (1989).
6. Machida, M., Uemura, H., Jigami, Y. & Tanaka, H. The protein factor which binds to the upstream activating sequence of Saccharomyces cerevisiae ENO1 gene. *Nucleic Acids Res.* **16,** 1407–1422 (1988).
7. Carmen, A. A., Brindle, P. K., Park, C. S. & Holland, M. J. Transcriptional regulation by an upstream repression sequence from the yeast enolase gene ENO1. *Yeast* **11,** 1031–1043 (1995).
8. Carmen, A. A. & Holland, M. J. The upstream repression sequence from the yeast enolase gene ENO1 is a complex regulatory element that binds multiple trans-acting factors including REB1. *J. Biol. Chem.* **269,** 9790–9797 (1994).
9. Reimand, J., Vaquerizas, J. M., Todd, A. E., Vilo, J. & Luscombe, N. M. Comprehensive reanalysis of transcription factor knockout expression data in Saccharomyces cerevisiae reveals many new targets. *Nucleic Acids Res.* **38,** 4768–4777 (2010).
10. Harbison, C. T. *et al.* Transcriptional regulatory code of a eukaryotic genome. *Nature* **431,** 99–104 (2004).
11. Luche, R. M., Smart, W. C. & Cooper, T. G. Purification of the heteromeric protein binding to the URS1 transcriptional repression site in Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci. U.S.A.* **89,** 7412–7416 (1992).
12. Luchelli, L., Thomas, M. G. & Boccaccio, G. L. Synaptic control of mRNA translation by reversible assembly of XRN1 bodies. *J. Cell. Sci.* **128,** 1542–1554 (2015).
13. Park, H.-D., Luche, R. M. & Cooper, T. G. The yeast UME6 gene product is required for transcriptional repression mediated by the CAR1 URS1 repressor binding site. *Nucleic Acids Res.* **20,** 1909–1915 (1992).
14. Fazzio, T. G. *et al.* Widespread collaboration of Isw2 and Sin3-Rpd3 chromatin remodeling complexes in transcriptional repression. *Mol. Cell. Biol.* **21,** 6450–6460 (2001).

15. Yadon, A. N., Singh, B. N., Hampsey, M. & Tsukiyama, T. DNA looping facilitates targeting of a chromatin remodeling enzyme. *Mol. Cell* **50,** 93–103 (2013).
16. Yadon, A. N. *et al.* Chromatin remodeling around nucleosome-free regions leads to repression of noncoding RNA transcription. *Mol. Cell. Biol.* **30,** 5110–5122 (2010).
17. Sato, T. *et al.* The E-box DNA binding protein Sgc1p suppresses the gcr2 mutation, which is involved in transcriptional activation of glycolytic genes in Saccharomyces cerevisiae. *FEBS Lett.* **463,** 307–311 (1999).
18. Chen, M. & Lopes, J. M. Multiple basic helix-loop-helix proteins regulate expression of the ENO1 gene of Saccharomyces cerevisiae. *Eukaryotic Cell* **6,** 786–796 (2007).
19. Gordân, R. *et al.* Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep* **3,** 1093–1104 (2013).
20. Nishi, K. *et al.* The GCR1 requirement for yeast glycolytic gene expression is suppressed by dominant mutations in the SGC1 gene, which encodes a novel basic-helix-loop-helix protein. *Mol. Cell. Biol.* **15,** 2646–2653 (1995).
21. Neil, H. *et al.* Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature* **457,** 1038–1042 (2009).
22. Xu, Z. *et al.* Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457,** 1033–1037 (2009).
23. Tuck, A. C. & Tollervey, D. A transcriptome-wide atlas of RNP composition reveals diverse classes of mRNAs and lncRNAs. *Cell* **154,** 996–1009 (2013).
24. Pelechano, V., Wei, W. & Steinmetz, L. M. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* **497,** 127–131 (2013).
25. Jigami, Y. *et al.* Analysis of expression of yeast enolase 1 gene containing a longer pyrimidine-rich region located between the TATA box and transcription start site. *J. Biochem.* **99,** 1111–1125 (1986).
26. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320,** 1344–1349 (2008).
27. Rhee, H. S. & Pugh, B. F. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature* **483,** 295–301 (2012).
28. Sainsbury, S., Bernecky, C. & Cramer, P. Structural basis of transcription initiation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.* **16,** 129–143 (2015).
29. Brachmann, C. B. *et al.* Designer deletion strains derived from Saccharomyces cerevisiae S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast* **14,** 115–132 (1998).
30. Longtine, M. S. *et al.* Additional modules for versatile and economical PCR-based gene deletion and modification in Saccharomyces cerevisiae. *Yeast* **14,** 953–961 (1998).
31. Granneman, S., Kudla, G., Petfalski, E. & Tollervey, D. Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 9613–9618 (2009).
32. Storici, F. & Resnick, M. A. The delitto perfetto approach to in vivo site-directed mutagenesis and chromosome rearrangements with synthetic oligonucleotides in

yeast. *Meth. Enzymol.* **409,** 329–345 (2006).
33. Tollervey, D. & Mattaj, I. W. Fungal small nuclear ribonucleoproteins share properties with plant and vertebrate U-snRNPs. *EMBO J* **6,** 469–476 (1987).

LIST OF FIGURES

BIBLIOGRAPHY

Adams, R. L. (2012). *The biochemistry of the nucleic acids*. Springer Science & Business Media. (Cited on page 11.)

Äijö, T., Butty, V., Chen, Z., Salo, V., Tripathi, S., Burge, C. B., Lahesmaa, R., and Lähdesmäki, H. (2014). Methods for time series analysis of RNA-seq data with application to human Th17 cell differentiation. *Bioinformatics*, 30(12):i113–i120. (Cited on pages 35 and 74.)

Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., and Walter, P. (2014). *Molecular Biology of the Cell, Sixth Edition:*. Taylor & Francis Group. (Cited on page 4.)

Alon, U. (2006). *An introduction to systems biology: design principles of biological circuits*. CRC press. (Cited on page 36.)

Andronescu, M. S., Pop, C., and Condon, A. E. (2010). Improved free energy parameters for RNA pseudoknotted secondary structure prediction. *RNA*, 16(1):26–42. (Cited on page 13.)

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, 181(4096):223–230. (Cited on page 16.)

Arigo, J. T., Eyler, D. E., Carroll, K. L., and Corden, J. L. (2006). Termination of cryptic unstable transcripts is directed by yeast RNA-binding proteins Nrd1 and Nab3. *Molecular cell*, 23(6):841–851. (Cited on page 23.)

Aviran, S., Trapnell, C., Lucks, J. B., Mortimer, S. A., Luo, S., Schroth, G. P., Doudna, J. A., Arkin, A. P., and Pachter, L. (2011). Modeling and automation of sequencing-based characterization of RNA structure. *Proceedings of the National Academy of Sciences*, 108(27):11069–11074. (Cited on pages 16, 18, 63, 66, 68, and 133.)

Barenco, M., Brewer, D., Papouli, E., Tomescu, D., Callard, R., Stark, J., and Hubank, M. (2009). Dissection of a complex transcriptional response using genome-wide transcriptional modelling. *Molecular systems biology*, 5(1):327. (Cited on page 21.)

Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The annals of mathematical statistics*, 37(6):1554–1563. (Cited on page 25.)

Beckmann, B. M. (2017). RNA interactome capture in yeast. *Methods*, 118:82–92. (Cited on page 20.)

Berg, J., Tymoczko, J., and Stryer, L. (2002). *Biochemistry, Fifth Edition*. W.H. Freeman. (Cited on page 6.)

Berkovits, B. D. and Mayr, C. (2015). Alternative 3'UTRs act as scaffolds to regulate membrane protein localization. *Nature*, 522(7556):363. (Cited on page 10.)

Bernabo, P., Tebaldi, T., Groen, E. J., Lane, F. M., Perenthaler, E., Mattedi, F., Newbery, H. J., Zhou, H., Zuccotti, P., Potrich, V., et al. (2017). In vivo translatome profiling in spinal muscular atrophy reveals a role for SMN protein in ribosome biology. *Cell*

Reports, 21(4):953–965. (Cited on page 3.)

Berretta, J. and Morillon, A. (2009). Pervasive transcription constitutes a new level of eukaryotic genome regulation. *EMBO reports*, 10(9):973–982. (Cited on page 23.)

Black, D. L. (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annual review of biochemistry*, 72(1):291–336. (Cited on page 10.)

Bluewhale22 (2014). File:basic principle of CLIP.jpg. https://commons.wikimedia.org/wiki/File:Basic_Principle_of_{CLIP}.jpg. Accessed: 2017-11-21. (Cited on page 20.)

Boukouvalas, A., Hensman, J., and Rattray, M. (2017). Bgp: Branched gaussian processes for identifying gene-specific branching dynamics in single cell data. *bioRxiv*, page 166868. (Cited on page 35.)

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of molecular biology*, 268(1):78–94. (Cited on page 30.)

Butcher, J. C. (2016). *Numerical methods for ordinary differential equations*. John Wiley & Sons. (Cited on page 36.)

Bystroff, C., Thorsson, V., and Baker, D. (2000). HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *Journal of molecular biology*, 301(1):173–190. (Cited on page 30.)

Carlile, T. M., Rojas-Duran, M. F., Zinshteyn, B., Shin, H., Bartoli, K. M., and Gilbert, W. V. (2014). Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. *Nature*, 515(7525):143. (Cited on page 66.)

Carroll, K. L., Ghirlando, R., Ames, J. M., and Corden, J. L. (2007). Interaction of yeast RNA-binding proteins Nrd1 and Nab3 with RNA polymerase II terminator elements. *Rna*, 13(3):361–373. (Cited on page 23.)

Cech, T. R. (2012). The RNA worlds in context. *Cold Spring Harbor perspectives in biology*, 4(7):a006742. (Cited on page 6.)

Cheong, H.-K., Hwang, E., Lee, C., Choi, B.-S., and Cheong, C. (2004). Rapid preparation of RNA samples for NMR spectroscopy and X-ray crystallography. *Nucleic acids research*, 32(10):e84–e84. (Cited on page 11.)

Chernyakov, I., Whipple, J. M., Kotelawala, L., Grayhack, E. J., and Phizicky, E. M. (2008). Degradation of several hypomodified mature tRNA species in Saccharomyces cerevisiae is mediated by Met22 and the 5′–3′ exonucleases Rat1 and Xrn1. *Genes & development*, 22(10):1369–1380. (Cited on page 111.)

Choudhary, K., Deng, F., and Aviran, S. (2017a). Comparative and integrative analysis of RNA structural profiling data: current practices and emerging questions. *Quantitative Biology*, pages 1–22. (Cited on page 68.)

Choudhary, K., Ruan, L., Deng, F., Shih, N., and Aviran, S. (2017b). SEQualyzer: interactive tool for quality control and exploratory analysis of high-throughput RNA structural profiling data. *Bioinformatics*, 33(3):441–443. (Cited on page 68.)

Cleary, M. D., Meiering, C. D., Jan, E., Guymon, R., and Boothroyd, J. C. (2005). Biosynthetic labeling of RNA with uracil phosphoribosyltransferase allows cell-specific microarray analysis of mRNA synthesis and decay. *Nature biotechnology*, 23(2):232. (Cited on page 22.)

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome biology*, 17(1):13. (Cited on page 10.)

Creamer, T. J., Darby, M. M., Jamonnak, N., Schaughency, P., Hao, H., Wheelan, S. J., and Corden, J. L. (2011). Transcriptome-wide binding sites for components of the Saccharomyces cerevisiae non-poly (A) termination pathway: Nrd1, Nab3, and sen1. *PLoS genetics*, 7(10):e1002329. (Cited on page 23.)

Cseke, B. and Heskes, T. (2011). Approximate marginals in latent Gaussian models. *Journal of Machine Learning Research*, 12(Feb):417–454. (Cited on page 126.)

Darby, M. M., Serebreni, L., Pan, X., Boeke, J. D., and Corden, J. L. (2012). The Saccharomyces cerevisiae Nrd1-Nab3 transcription termination pathway acts in opposition to Ras signaling and mediates response to nutrient depletion. *Molecular and cellular biology*, 32(10):1762–1775. (Cited on page 73.)

Darnell, R. B. (2010). HITS-CLIP: panoramic views of protein–RNA regulation in living cells. *Wiley Interdisciplinary Reviews: RNA*, 1(2):266–286. (Cited on page 19.)

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38. (Cited on page 29.)

Deneke, C., Lipowsky, R., and Valleriani, A. (2013). Complex degradation processes lead to non-exponential decay patterns and age-dependent decay rates of messenger RNA. *PloS one*, 8(2):e55442. (Cited on page 24.)

Deng, F., Ledda, M., Vaziri, S., and Aviran, S. (2016). Data-directed RNA secondary structure prediction using probabilistic modeling. *RNA*. (Cited on page 68.)

Dimon, M. T., Sorber, K., and DeRisi, J. L. (2010). HMMSplicer: a tool for efficient and sensitive discovery of known and novel splice junctions in RNA-seq data. *PloS one*, 5(11):e13875. (Cited on page 30.)

Ding, Y. and Lawrence, C. E. (2003). A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic acids research*, 31(24):7280–7301. (Cited on page 13.)

Ding, Y., Tang, Y., Kwok, C. K., Zhang, Y., Bevilacqua, P. C., and Assmann, S. M. (2014). In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature*, 505(7485):696–700. (Cited on pages 16, 18, 61, and 62.)

Drewe-Boss, P., Wessels, H.-H., and Ohler, U. (2017). omniCLIP: Bayesian identification of protein-RNA interactions from CLIP-seq data. *bioRxiv*, page 161877. (Cited on page 133.)

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press. (Cited on page 30.)

Eddy, S. R. (2014). Computational analysis of conserved RNA secondary structure in transcriptomes and genomes. *Biophysics*, 43. (Cited on page 68.)

Elkon, R., Zlotorynski, E., Zeller, K. I., and Agami, R. (2010). Major role for mRNA stability in shaping the kinetics of gene induction. *BMC genomics*, 11(1):259. (Cited on page 21.)

EMBL-EBI (2017). Illumina sequencing. https://www.ebi.ac.uk/training/online/course/ebi-next-generation-sequencing-practical-course/what-next-generation-dna-sequencing/illumina-. Accessed: 2017-11-28. (Cited on page 8.)

Ernst, J., Vainas, O., Harbison, C. T., Simon, I., and Bar-Joseph, Z. (2007). Reconstructing dynamic regulatory maps. *Molecular systems biology*, 3(1):74. (Cited on page 21.)

Fedor, M. J. and Williamson, J. R. (2005). The catalytic diversity of RNAs. *Nature reviews. Molecular cell biology*, 6(5):399. (Cited on page 6.)

Forney, G. D. (1973). The Viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278. (Cited on page 28.)

Freyhult, E., Gardner, P. P., and Moulton, V. (2005). A comparison of RNA folding measures. *BMC bioinformatics*, 6(1):241. (Cited on page 13.)

Friedel, C. C. and Dölken, L. (2009). Metabolic tagging and purification of nascent RNA: implications for transcriptomics. *Molecular Biosystems*, 5(11):1271–1278. (Cited on page 22.)

Gardner, P. P. and Giegerich, R. (2004). A comprehensive comparison of comparative RNA structure prediction approaches. *BMC bioinformatics*, 5(1):140. (Cited on page 14.)

Glisovic, T., Bachorik, J. L., Yong, J., and Dreyfuss, G. (2008). RNA-binding proteins and post-transcriptional gene regulation. *FEBS letters*, 582(14):1977–1986. (Cited on pages 11 and 19.)

Granneman, S., Kudla, G., Petfalski, E., and Tollervey, D. (2009). Identification of protein binding sites on u3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. *Proceedings of the National Academy of Sciences*, 106(24):9613–9618. (Cited on pages 19, 20, and 23.)

Granneman, S., Petfalski, E., and Tollervey, D. (2011). A cluster of ribosome synthesis factors regulate pre-rRNA folding and 5.8 S rRNA maturation by the Rat1 exonuclease. *The EMBO journal*, 30(19):4006–4019. (Cited on page 21.)

Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N., and Altman, S. (1983). The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, 35(3):849–857. (Cited on page 13.)

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A.-C., Munschauer, M., et al. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1):129–141. (Cited on pages 20 and 23.)

Hector, R. D., Burlacu, E., Aitken, S., Le Bihan, T., Tuijtel, M., Zaplatina, A., Cook, A. G., and Granneman, S. (2014). Snapshots of pre-rRNA structural flexibility reveal eukaryotic 40S assembly dynamics at nucleotide resolution. *Nucleic acids research*, page gku815. (Cited on pages 17, 18, and 62.)

Hofacker, I. L., Fekete, M., and Stadler, P. F. (2002). Secondary structure prediction for aligned RNA sequences. *Journal of molecular biology*, 319(5):1059–1066. (Cited on page 14.)

Hogg, R. V. and Craig, A. T. (1995). *Introduction to mathematical statistics.(5"" edition)*. Upper Saddle River, New Jersey: Prentice Hall. (Cited on page 56.)

Holmes, I. (2005). Accelerated probabilistic inference of RNA structure evolution. *BMC bioinformatics*, 6(1):73. (Cited on page 14.)

Honkela, A., Peltonen, J., Topa, H., Charapitsa, I., Matarese, F., Grote, K., Stunnenberg, H. G., Reid, G., Lawrence, N. D., and Rattray, M. (2015). Genome-wide modeling of transcription kinetics reveals patterns of RNA production delays. *Proceedings of the National Academy of Sciences*, 112(42):13115–13120. (Cited on pages 24 and 36.)

Houseley, J. and Tollervey, D. (2009). The many pathways of RNA degradation. *Cell*, 136(4):763–776. (Cited on page 72.)

Incarnato, D., Anselmi, F., Morandi, E., Neri, F., Maldotti, M., Rapelli, S., Parlato, C., Basile, G., and Oliviero, S. (2017). High-throughput single-base resolution mapping of RNA 2´-O-methylated residues. *Nucleic acids research*, 45(3):1433–1441. (Cited on page 66.)

Irimia, M., Weatheritt, R. J., Ellis, J. D., Parikshak, N. N., Gonatopoulos-Pournatzis, T., Babor, M., Quesnel-Vallieres, M., Tapial, J., Raj, B., O'Hanlon, D., et al. (2014). A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell*, 159(7):1511–1523. (Cited on page 10.)

Jeffreys, H. (1998). *The theory of probability*. OUP Oxford. (Cited on page 76.)

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–1502. (Cited on page 7.)

Kalaitzis, A. A. and Lawrence, N. D. (2011). A simple approach to ranking differentially expressed gene expression time courses through Gaussian process regression. *BMC bioinformatics*, 12(1):180. (Cited on page 35.)

Kertesz, M., Wan, Y., Mazor, E., Rinn, J. L., Nutter, R. C., Chang, H. Y., and Segal, E. (2010). Genome-wide measurement of RNA secondary structure in yeast. *Nature*, 467(7311):103–107. (Cited on pages 14, 15, 18, 61, and 62.)

Kim, H. D., Shay, T., O'Shea, E. K., and Regev, A. (2009). Transcriptional regulatory circuits: predicting numbers from alphabets. *Science*, 325(5939):429–432. (Cited on page 21.)

Kloeden, P. E., Platen, E., and Schurz, H. (2012). *Numerical solution of SDE through computer experiments*. Springer Science & Business Media. (Cited on page 36.)

Knudsen, B. and Hein, J. (2003). Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic acids research*, 31(13):3423–3428. (Cited on pages 13 and 14.)

Kondo, J., Sauter, C., and Masquida, B. (2014). RNA crystallization. *Handbook of RNA Biochemistry: Second, Completely Revised and Enlarged Edition*, pages 481–498. (Cited on page 11.)

König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M., and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature structural & molecular biology*, 17(7):909–915. (Cited on pages 20 and 66.)

Korf, I. (2004). Gene finding in novel genomes. *BMC bioinformatics*, 5(1):59. (Cited on page 30.)

Krakau, S., Richard, H., and Marsico, A. (2017). PureCLIP: Capturing target-specific protein-RNA interaction footprints from single-nucleotide CLIP-seq data. *bioRxiv*,

page 146704. (Cited on page 133.)

Kubota, M., Tran, C., and Spitale, R. C. (2015). Progress and challenges for chemical probing of RNA structure inside living cells. *Nature chemical biology*, 11(12):933–941. (Cited on page 39.)

Kuhn, A., Thu, D., Waldvogel, H. J., Faull, R. L., and Luthi-Carter, R. (2011). Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. *Nature methods*, 8(11):945–947. (Cited on page 2.)

Kwok, C. K., Ding, Y., Tang, Y., Assmann, S. M., and Bevilacqua, P. C. (2013). Determination of in vivo RNA structure in low-abundance transcripts. *Nature communications*, 4. (Cited on pages 16, 18, 61, and 62.)

Larimer, F. W. and Stevens, A. (1990). Disruption of the gene XRN1, coding for a 5′-3′ exoribonuclease, restricts yeast cell growth. *Gene*, 95(1):85–90. (Cited on page 23.)

Lawrence, N. D., Girolami, M., Rattray, M., and Sanguinetti, G. (2010). *Learning and inference in computational systems biology*. MIT press. (Cited on page 36.)

Lawrence, N. D., Sanguinetti, G., and Rattray, M. (2007). Modelling transcriptional regulation using Gaussian processes. In *Advances in Neural Information Processing Systems*, pages 785–792. (Cited on page 34.)

Lebreton, A. and Séraphin, B. (2008). Exosome-mediated quality control: substrate recruitment and molecular activity. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1779(9):558–565. (Cited on page 110.)

Lee, B., Matera, A. G., Ward, D. C., and Craft, J. (1996). Association of RNase mitochondrial RNA processing enzyme with ribonuclease P in higher ordered structures in the nucleolus: a possible coordinate role in ribosome biogenesis. *Proceedings of the National Academy of Sciences*, 93(21):11471–11476. (Cited on page 13.)

Lee, C. Y., Lee, A., and Chanfreau, G. (2003). The roles of endonucleolytic cleavage and exonucleolytic digestion in the 5′-end processing of S. cerevisiae box C/D snoRNAs. *RNA*, 9(11):1362–1370. (Cited on page 111.)

Li, B., Tambe, A., Aviran, S., and Pachter, L. (2017). PROBer provides a general toolkit for analyzing sequencing-based toeprinting assays. *Cell Systems*, 4(5):568–574. (Cited on pages 66 and 133.)

Licatalosi, D. D., Mele, A., Fak, J. J., Ule, J., Kayikci, M., Chi, S. W., Clark, T. A., Schweitzer, A. C., Blume, J. E., Wang, X., et al. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456(7221):464. (Cited on page 19.)

Little, R. J. and Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons. (Cited on page 30.)

Long, D., Lee, R., Williams, P., Chan, C. Y., Ambros, V., and Ding, Y. (2007). Potent effect of target structure on microRNA function. *Nature structural & molecular biology*, 14(4):287–294. (Cited on page 11.)

Long, J. C. and Caceres, J. F. (2009). The SR protein family of splicing factors: master regulators of gene expression. *Biochemical Journal*, 417(1):15–27. (Cited on page 19.)

Low, J. T. and Weeks, K. M. (2010). SHAPE-directed RNA secondary structure prediction. *Methods*, 52(2):150–158. (Cited on pages 17, 61, and 62.)

Lucks, J. B., Mortimer, S. A., Trapnell, C., Luo, S., Aviran, S., Schroth, G. P., Pachter, L., Doudna, J. A., and Arkin, A. P. (2011). Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-seq). *Proceedings of the National Academy of Sciences*, 108(27):11063–11068. (Cited on pages 16, 17, 18, and 62.)

Lyngsø, R. B. and Pedersen, C. N. (2000). RNA pseudoknot prediction in energy-based models. *Journal of computational biology*, 7(3-4):409–427. (Cited on page 13.)

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press. (Cited on pages 34, 108, and 154.)

Marguerat, S. and Bähler, J. (2010). Rna-seq: from technology to biology. *Cellular and molecular life sciences*, 67(4):569–579. (Cited on page 9.)

Marguerat, S., Lawler, K., Brazma, A., and Bähler, J. (2014). Contributions of transcription and mRNA decay to gene expression dynamics of fission yeast in response to oxidative stress. *RNA biology*, 11(6):702–714. (Cited on pages 22, 36, and 73.)

Markov, A. A. (1906). Rasprostranenie zakona bol'shih chisel na velichiny, zavisyaschie drug ot druga. *Izvestiya Fiziko-matematicheskogo obschestva pri Kazanskom universitete*, 15(135-156):18. (Cited on page 25.)

Mathews, D. H. (2006). Revolutions in RNA secondary structure prediction. *Journal of molecular biology*, 359(3):526–532. (Cited on pages 12 and 13.)

Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M., and Turner, D. H. (2004). Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences of the United States of America*, 101(19):7287–7292. (Cited on page 12.)

Mathews, D. H. and Turner, D. H. (2002). Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *Journal of molecular biology*, 317(2):191–203. (Cited on page 14.)

Mattick, J. S. and Makunin, I. V. (2006). Non-coding RNA. *Human molecular genetics*, 15(suppl_1):R17–R29. (Cited on page 6.)

Matzke, M. A. and Matzke, A. J. (2004). Planting the seeds of a new paradigm. *PLoS biology*, 2(5):e133. (Cited on page 6.)

Mauger, D. M. and Weeks, K. M. (2010). Toward global RNA structure analysis. *Nature biotechnology*, 28(11):1178–1179. (Cited on page 15.)

McGinnis, J. L., Dunkle, J. A., Cate, J. H., and Weeks, K. M. (2012). The mechanisms of RNA SHAPE chemistry. *Journal of the American Chemical Society*, 134(15):6617–6624. (Cited on page 15.)

Miller, C., Schwalb, B., Maier, K., Schulz, D., Dümcke, S., Zacher, B., Mayer, A., Sydow, J., Marcinowski, L., Dölken, L., et al. (2011). Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Molecular systems biology*, 7(1):458. (Cited on page 22.)

Modrek, B. and Lee, C. (2002). A genomic view of alternative splicing. *Nature genetics*, 30(1):13–19. (Cited on page 10.)

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nature methods*, 5(7):621–628. (Cited on page 9.)

Mortimer, S. A. and Weeks, K. M. (2007). A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *Journal of the American Chemical Society*, 129(14):4144–4145. (Cited on page 15.)

Murdoch, D. J., Tsai, Y.-L., and Adcock, J. (2012). P-values are random variables. *The American Statistician*. (Cited on page 42.)

Nachman, I., Regev, A., and Friedman, N. (2004). Inferring quantitative models of regulatory networks from expression data. *Bioinformatics*, 20(suppl_1):i248–i256. (Cited on page 21.)

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., and Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881):1344–1349. (Cited on pages 102 and 116.)

Norris, M., Kwok, C. K., Cheema, J., Hartley, M., Morris, R. J., Aviran, S., and Ding, Y. (2017). FoldAtlas: a repository for genome-wide RNA structure probing data. *Bioinformatics*, 33(2):306–308. (Cited on page 18.)

Nudler, E. and Mironov, A. S. (2004). The riboswitch control of bacterial metabolism. *Trends in biochemical sciences*, 29(1):11–17. (Cited on pages 6 and 13.)

Nussinov, R. and Jacobson, A. B. (1980). Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proceedings of the National Academy of Sciences*, 77(11):6309–6313. (Cited on page 13.)

Pachter, L. (2011). Models for transcript quantification from RNA-seq. *arXiv preprint arXiv:1104.3889*. (Cited on page 9.)

Pelechano, V. and Pérez-Ortín, J. E. (2008). The transcriptional inhibitor thiolutin blocks mRNA degradation in yeast. *Yeast*, 25(2):85–92. (Cited on page 22.)

Petfalski, E., Dandekar, T., Henry, Y., and Tollervey, D. (1998). Processing of the precursors to small nucleolar RNAs and rRNAs requires common components. *Molecular and cellular biology*, 18(3):1181–1189. (Cited on page 111.)

Poblete, S., Bottaro, S., and Bussi, G. (2015). A nucleobase-centric coarse-grained model for structure prediction of RNA fragments. *Biophysical Journal*, 108(2):235a. (Cited on page 14.)

Puglisi, J. D., Tan, R., Calnan, B. J., Frankel, A. D., and Williamson, J. R. (1992). Conformation of the TAR RNA-arginine complex by NMR spectroscopy. *Science*, 257(5066):76–80. (Cited on page 11.)

Rabani, M., Levin, J. Z., Fan, L., Adiconis, X., Raychowdhury, R., Garber, M., Gnirke, A., Nusbaum, C., Hacohen, N., Friedman, N., et al. (2011). Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nature biotechnology*, 29(5):436–442. (Cited on pages 22, 24, and 36.)

Rasmussen, C. E. and Williams, C. K. (2006). Gaussian processes for machine learning. 1. (Cited on page 31.)

Reece, J. B., Urry, L. A., Cain, M. L., Wasserman, S. A., Minorsky, P. V., Jackson, R., et al. (2014). *Campbell biology*. Pearson Boston. (Cited on page 5.)

Reuter, J. S. and Mathews, D. H. (2010). RNAstructure: software for RNA secondary structure prediction and analysis. *BMC bioinformatics*, 11(1):1. (Cited on pages 14 and 68.)

Rivas, E. and Eddy, S. R. (1999). A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of molecular biology*, 285(5):2053–2068. (Cited on page 13.)

Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology*, 11(3):R25. (Cited on pages 9, 10, and 127.)

Rouskin, S., Zubradt, M., Washietl, S., Kellis, M., and Weissman, J. S. (2014). Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*, 505(7485):701–705. (Cited on pages 16, 18, 61, and 62.)

Sakurambo (2006). File:stem-loop.svg. https://en.wikipedia.org/wiki/File:Stem-loop.svg. Accessed: 2017-11-21. (Cited on page 12.)

Sandberg, R., Neilson, J. R., Sarma, A., Sharp, P. A., and Burge, C. B. (2008). Proliferating cells express mRNAs with shortened 39 untranslated regions and fewer microRNA target sites. *Science*, 320(5883):1643–1647. (Cited on page 10.)

Sankoff, D. (1985). Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM journal on applied mathematics*, 45(5):810–825. (Cited on page 14.)

Schuster, S. C. (2007). Next-generation sequencing transforms today's biology. *Nature*, 200(8):16–18. (Cited on page 6.)

Selega, A., Granneman, S., and Sanguinetti, G. (2016). BUMHMM: Computational pipeline for computing probability of modification from structure probing experiment data. Available at http://www.bioconductor.org/packages/BUMHMM. (Cited on pages 2 and 65.)

Selega, A. and Sanguinetti, G. (2016). Trends and challenges in computational RNA biology. *Genome biology*, 17(1):253. (Cited on pages 3 and 11.)

Selega, A., Sirocchi, C., Iosub, I., Granneman, S., and Sanguinetti, G. (2017). Robust statistical modeling improves sensitivity of high-throughput RNA structure probing experiments. *Nature methods*, 14(1):83–89. (Cited on pages 2, 18, 30, 38, 44, 57, 66, and 68.)

Shalem, O., Dahan, O., Levo, M., Martinez, M. R., Furman, I., Segal, E., and Pilpel, Y. (2008). Transient transcriptional responses to stress are generated by opposing effects of mRNA production and degradation. *Molecular systems biology*, 4(1):4. (Cited on page 21.)

Sharma, S., Ding, F., and Dokholyan, N. V. (2008). iFoldRNA: three-dimensional RNA structure prediction and folding. *Bioinformatics*, 24(17):1951–1952. (Cited on page 14.)

Shiroguchi, K., Jia, T. Z., Sims, P. A., and Xie, X. S. (2012). Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proceedings of the National Academy of Sciences*, 109(4):1347–1352. (Cited on page 17.)

Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A., and Weeks, K. M. (2014). RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nature methods*, 11(9):959–965. (Cited on pages 17, 18, and 61.)

Speir, J. A., Munshi, S., Wang, G., Baker, T. S., and Johnson, J. E. (1995). Structures of the native and swollen forms of cowpea chlorotic mottle virus determined by

X-ray crystallography and cryo-electron microscopy. *Structure*, 3(1):63–78. (Cited on page 11.)

Staple, D. W. and Butcher, S. E. (2005). Pseudoknots: RNA structures with diverse functions. *PLoS biology*, 3(6):e213. (Cited on page 13.)

Stegle, O., Denby, K. J., Cooke, E. J., Wild, D. L., Ghahramani, Z., and Borgwardt, K. M. (2010). A robust Bayesian two-sample test for detecting intervals of differential gene expression in microarray time series. *Journal of Computational Biology*, 17(3):355–367. (Cited on page 35.)

Suter, D. M., Molina, N., Gatfield, D., Schneider, K., Schibler, U., and Naef, F. (2011). Mammalian genes are transcribed with widely different bursting kinetics. *Science*, 332(6028):472–474. (Cited on page 136.)

Svoboda, P. and Cara, A. D. (2006). Hairpin RNA: a secondary structure of primary importance. *Cellular and molecular life sciences*, 63(7):901–908. (Cited on page 11.)

Tang, Y., Bouvier, E., Kwok, C. K., Ding, Y., Nekrutenko, A., Bevilacqua, P. C., and Assmann, S. M. (2015). StructureFold: genome-wide RNA secondary structure mapping and reconstruction in vivo. *Bioinformatics*, page btv213. (Cited on page 68.)

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28(5):511–515. (Cited on pages 17 and 66.)

Ule, J., Jensen, K., Mele, A., and Darnell, R. B. (2005). CLIP: a method for identifying protein–RNA interaction sites in living cells. *Methods*, 37(4):376–386. (Cited on pages 7 and 19.)

Underwood, J. G., Uzilov, A. V., Katzman, S., Onodera, C. S., Mainzer, J. E., Mathews, D. H., Lowe, T. M., Salama, S. R., and Haussler, D. (2010). FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nature methods*, 7(12):995–1001. (Cited on pages 15, 18, and 62.)

Van Dijk, E., Chen, C., d'Aubenton Carafa, Y., Gourvennec, S., Kwapisz, M., Roche, V., Bertrand, C., Silvain, M., Legoix-Né, P., Loeillet, S., et al. (2011). XUTs are a class of Xrn1-sensitive antisense regulatory non-coding RNA in yeast. *Nature*, 475(7354):114. (Cited on pages 23 and 111.)

van Nues, R., Schweikert, G., de Leau, E., Selega, A., Langford, A., Franklin, R., Iosub, I., Wadsworth, P., Sanguinetti, G., and Granneman, S. (2017). Kinetic CRAC uncovers a role for Nab3 in determining gene expression profiles during stress. *Nature Communications*, 8. (Cited on pages 2, 20, 21, 23, 24, 35, 70, 74, 101, 125, and 134.)

Vasiljeva, L., Kim, M., Mutschler, H., Buratowski, S., and Meinhart, A. (2008). The Nrd1–Nab3–Sen1 termination complex interacts with the Ser5-phosphorylated RNA polymerase II C-terminal domain. *Nature structural & molecular biology*, 15(8):795–804. (Cited on page 23.)

Wan, Y., Qu, K., Zhang, Q. C., Flynn, R. A., Manor, O., Ouyang, Z., Zhang, J., Spitale, R. C., Snyder, M. P., Segal, E., et al. (2014). Landscape and variation of RNA secondary structure across the human transcriptome. *Nature*, 505(7485):706–709. (Cited on page 67.)

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63. (Cited on pages 7, 9, and 17.)

Webb, S., Hector, R. D., Kudla, G., and Granneman, S. (2014). PAR-CLIP data indicate that Nrd1-Nab3-dependent transcription termination regulates expression of hundreds of protein coding genes in yeast. *Genome biology*, 15(1):R8. (Cited on pages 23, 72, and 73.)

Weinberg, Z. and Ruzzo, W. L. (2005). Sequence-based heuristics for faster annotation of non-coding RNA families. *Bioinformatics*, 22(1):35–39. (Cited on page 30.)

Weinreb, C., Riesselman, A. J., Ingraham, J. B., Gross, T., Sander, C., and Marks, D. S. (2016). 3D RNA and functional interactions from evolutionary couplings. *Cell*, 165(4):963–975. (Cited on page 14.)

Wilkinson, D. J. (2011). *Stochastic modelling for systems biology*. CRC press. (Cited on page 36.)

Wilkinson, K. A., Merino, E. J., and Weeks, K. M. (2006). Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. *Nature protocols*, 1(3):1610–1616. (Cited on page 15.)

Wu, Y., Shi, B., Ding, X., Liu, T., Hu, X., Yip, K. Y., Yang, Z. R., Mathews, D. H., and Lu, Z. J. (2015). Improved prediction of RNA secondary structure by integrating the free energy model with restraints derived from experimental probing data. *Nucleic acids research*, 43(15):7247–7259. (Cited on page 39.)

Xia, T., SantaLucia Jr, J., Burkard, M. E., Kierzek, R., Schroeder, S. J., Jiao, X., Cox, C., and Turner, D. H. (1998). Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson- Crick base pairs. *Biochemistry*, 37(42):14719–14735. (Cited on page 12.)

Yeo, G. W., Coufal, N. G., Liang, T. Y., Peng, G. E., Fu, X.-D., and Gage, F. H. (2009). An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nature structural & molecular biology*, 16(2):130–137. (Cited on page 19.)

Yikrazuul (2010). File:50s-subunit of the ribosome 3cc2.png. https://en.wikipedia.org/wiki/File:50S-subunit_of_the_ribosome_3CC2.png. Accessed: 2017-11-21. (Cited on page 12.)

Yoon, B.-J. and Vaidyanathan, P. P. (2008). Structural alignment of RNAs using profile-csHMMs and its application to RNA homology search: overview and new results. *IEEE Transactions on Automatic Control*, 53(Special Issue):10–25. (Cited on page 30.)

Zemora, G. and Waldsich, C. (2010). RNA folding in living cells. *RNA biology*, 7(6):634–641. (Cited on page 16.)

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic acids research*, 31(13):3406–3415. (Cited on page 13.)

Zuker, M. and Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic acids research*, 9(1):133–148. (Cited on page 13.)

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

http://code.google.com/p/classicthesis/

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

http://postcards.miede.de/

Final Version as of February 21, 2018 (`classicthesis` version 1.0).