Computational methods for exploring gene regulation mechanisms using high-throughput sequencing data

Hao Wang

CMU-CB-16-102

September 29, 2016

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Carl Kingsford, Chair
James Faeder
Joel McManus
Sridhar Hannenhalli

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

This research is funded in part by the Gordon and Betty Moore Foundation?s Data-Driven Discovery Initiative through Grant GBMF4554 to Carl Kingsford, by the US National Science Foundation (CCF-1256087, CCF-1319998), and by the US National Institutes of Health (R21HG006913, R01HG007104). C.K. received support as an Alfred P. Sloan Research Fellow.

The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, donors or the U.S. Government.

Keywords: Algorithms, Computational Biology, Chromosome Conformation Capture, Three-Dimensional Structure, Ribosome Profiling, Translation, Ribosome Load, Isoform-level Estimation, Codon Bias, Wobble Pairing, Ribosome Collision

To those who struggle to unveil insights from noisy observations.

Abstract

Gene expression has been studied extensively on the transcript level with the help of RNA-seq technology, however less attention has been paid to gene regulation pre-transcription and post-transcription. For example, it is not clear whether genome structure plays an important role in gene functionality, nor is it clear how gene expression is regulated by translational speed on a codon basis. Recently, several high-throughput sequencing techniques have been developed to help answer these questions. Specifically, Chromosome Conformation Capture (3C) was developed to capture spatially close chromatin loci in cell nuclei and enables whole-genome structure studies, and ribosome profiling (ribo-seq) is developed to study ribosome location preferences during translation and enables genome-wide translational studies. However, the complicated experimental pipelines make these data inherently noisy, and typical approaches to process these data are prone to errors and computationally expensive. We developed various computational pipelines to fundamentally process these data to advance downstream analysis regarding gene regulation. Specifically, we developed a graph-based test to identify sets of functionally related genomic loci that are statistically spatially closer than expected by chance using 3C data. Compared to typical methods, our approach is computationally inexpensive and more robust to unmeasured interactions and the inclusion of non-associated loci. We also developed a pipeline to estimate ribosome occupancy preferences on a transcript level from ribo-seq data. This is the first systematic approach to address the ubiquitous multi-mappings in ribo-seq data and quantify ribosome loci on a transcript level. It results in better estimations of both ribosome profiles and ribosome loads. In addition, we designed a mathematical model and algorithm to recover ribosome positions from ribo-seq data. Unlike existing simple heuristics that make inaccurate assumptions on ribo-seq read digestions, our approach captured the complicated digestion pattern in a flexible and data-driven way, and outputs better ribosome profiles that help reveal biologically reasonable observations on translation patterns. Using these improved preprocessing pipelines above, we estimated the codon decoding time in yeast, and showed that both codon usage and wobble pairing play a role in regulating translational speed. Lastly, we performed the first genome-wide analysis on ribosome collisions with the help of a modified ribosome profiling protocol. Our preliminary results indicate that extreme slow-down of local ribosome movements during translation is likely to be random and rare, and the identification of programmed ribosome stalling requires further experiments with deeper sequencing. Together, our algorithms and analysis have helped to build the foundation for exploring pre- and post-transcriptional regulation in gene expression, which will help us understand the mechanism of cell growth and death, the differential gene expression across conditions and cell types, and the development and causes of diseases.

Acknowledgments

Call me narcissistic, but the first person I would like to thank is myself. I'm very grateful that I kept myself together during the long journey of getting a Ph.D. Thanks Hao, for the strong determination of finishing what you started, despite all the obstacles, such as switching majors twice, language inefficiencies, and inexperience in basic prerequisite skills like programming and writing; thanks for the courage for facing uncertainties, thanks for the monologues for searching for potential solutions, and thanks for the pep talks during seemingly hopeless conditions. I made it to the end because I did not give up on me. I am grateful for this experience. It made me stronger and tougher, made me develop a vision in conducting research, and made me less afraid of clearing up technical difficulties.

I want to thank my advisor Dr. Kingsford for supporting me financially so that I can focus on solving important biological problems, for giving me enough freedom to explore various solutions, for commenting on my work critically, and for hand-holding me going over depression. I also want to specially thank Rob Patro, Geet Duggal, and Darya Filippova for their belief in me more than myself, for their patience in helping me overcoming technical difficulties, and for their passion and curiosity for problem solving.

I want to thank other fellows both within and outside of Kingsford group for providing feedbacks during group presentations and commenting on my drafts. They are Heewook Lee, Brad Solomon, Tim Wall, Natalie Sauerwald, Cong Ma, Mingfu Shao, Jing Xiang, Hongyi Xin, David Pellow, Pieter Spealman, Saket Navlakha, and Chengxi Ye.

I want to thank my committee for their service and other co-authors for jointly publishing our findings: James Faeder, Joel McManus, Sridhar Hannenhalli, and Michelle Girvan.

Lastly I want to thank my parents for their continuous emotional support.

This research is funded in part by the Gordon and Betty Moore Foundation's Data-Driven Discovery Initiative through Grant GBMF4554 to Carl Kingsford, by the US National Science Foundation (CCF-1256087, CCF-1319998), and by the US National Institutes of Health (R21HG006913, R01HG007104). C.K. received support as an Alfred P. Sloan Research Fellow.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Proteins are the fundamental building blocks of all living cells. Comprising over half of the dry mass of the cell, they serve as enzymes to catalyze chemical reactions, as passengers to transmit small particles, and as specialized molecules to fight disease and regulate functions. In order for proteins to be synthesized, first, DNA segments that encode proteins (genes), need to be transcribed into messenger RNAs (mRNAs), and second, mRNAs need to be translated into proteins. The whole process of converting DNA segments into protein products is called gene expression. Since cells adjust protein levels according to needs, understanding the process and abundance of gene expression is therefore important in understanding how cells function fundamentally. The RNA-seq technology enabled quantifying transcript abundance at a low cost, therefore many gene expression studies are conducted at the transcription level. Recent high-throughput sequencing technologies such as Chromosome Conformation Capture and ribosome profiling have allowed more studies on gene regulation pre-transcriptionally and post-transcriptionally. They provide a better view on the whole process of gene expression from DNA to protein. However, the complicated protocols of these techniques often introduce internal noise that biases the view of meaningful biological signals, and computational approaches are therefore needed to interpret observations from these technologies with caution. In this thesis, we developed several computational methods and techniques to study the mechanism of gene regulation both pre-transcriptionally and post-transcriptionally. Our methods help to answer the following two fundamental questions:

1.1 Question 1: What role does genome structure play in gene expression regulation?

1.1.1 Chromosome Conformation Capture enables studies of genome strcture

DNA molecules are not randomly packed into the cell. Gene-rich regions tend to reside in the center of the nucleus, while gene-poor regions tend to reside on the peripheral of the nucleus. Activator proteins can bind to enhancer regions, which causes the DNA to bend and form long range loops, placing the promoter regions of these genes close to the activator proteins. These genes can then be activated to be transcribed. On the other hand, if a repressor protein binds to the insulator regions near a gene, it can prevent the activator protein from binding, de-activating gene transcription. Some co-regulated genes are also observed to be spatially co-localized in the genome [20].

Chromosome Conformation Capture (3C) (see review in [41]) is a recently developed technique to study 3D genome structure. It allows two genomic loci to be cross linked together if they are spatially close, and the subsequent high-throughput sequencing step quantitatively summarizes the observed instances of such cross linking interactions.

To examine whether spatial proximity can explain certain types of gene co-regulation, valid statistical tests need to be developed to confidently identify functionally important genome structures. Genomic proximity and global chromosomal patterns need to be controlled for such tests to rule out less interesting features in the 3C experiments that cause the significance of the tests. Validation needs to be performed to determine whether significant patterns of 3C interactions actually correspond to significant spatial proximity.

1.1.2 Our contributions

We developed a rigorous statistical framework to identify whether a given set of functionally related genomic locations are spatially compact. We proposed to use novel topological properties of chromatin interaction graphs as proxies of spatial closeness measurements. We demonstrated these properties are good approximations of spatial volumes within spatially compact regions. In addition, these properties, such as shortest paths, maximum flows, and dense cores, can be efficiently computed directly from the 3C interactions, therefore avoiding the computationally costly step of building a three-dimensional embedding of the genome. Furthermore, these properties also make better use of the raw interaction frequencies, allowing inference of indirect and transitive spatial closeness events, and are therefore more robust to noise from raw 3C interactions. Lastly, our method based on maximum density subgraph has the additional benefit of identifying the spatially compact cores from a given set of genomic locations, therefore is capable of detecting spatial closeness masked by outliers. To sum up, our proposed methods are robust and unbiased, and can be used to identify spatially compact sets involving multiple chromosomes. Our methods serve as the foundation to study spatial preferences of co-expressed or functionally related genes, and therefore help answer question one.

1.2 Question 2: How is translation speed regulated on a per-codon basis?

The steady-state mRNA abundances measured by RNA-seq [124] or microarrays have been widely used to infer gene expression levels. One of the reasons to study gene expression is to better understand the process of protein synthesis. A highly expressed gene tends to produce more protein. High-throughput sequencing technologies such as RNA-seq are prevalent, in part because they are easier to obtain than other abundance measurements such as mass spectrometry, which measure protein abundances more directly. While largely effective, mRNA abundances are imperfectly correlated with protein abundances. Moreover, the RNA-seq technique alone cannot capture the dynamic aspect of gene regulation.

1.2.1 Ribosome profiling helps bridge the gap between transcription measurement and protein quantification

Ribosome abundance measurements may lead to a better understanding of the translation mechanism, and ribosome profiling (ribo-seq) is developed to quantify ribosome abundance. It is like taking snapshots of ribosome locations during translation. Therefore, ribosome profiling captures the ribosome density of mRNAs with sub-codon resolution [79]. So far ribosome profiling has been performed on many species in various conditions, including plant [31, 91, 103, 108, 186, 190], virus [11, 85, 95, 110, 146, 158, 183], bacteria [14, 50, 70, 87, 93, 105, 106, 123, 125, 128, 173, 178], yeast [8, 12, 13, 22, 59, 60, 61, 69, 79, 86, 98, 115, 126, 135, 148, 174, 176, 185, 189], *C. elegans* [71, 126, 156, 157, 172], Drosophila [49, 122], zebrafish [17, 27, 100], mouse [72, 80, 139, 140, 163], and human across many cell and tissue types [9, 21, 64, 67, 73, 101, 111, 140, 144, 145, 157, 158, 160].

Ribosome profiling experiments produce ribosome footprints (RPF) on the transcripts, represented by short read sequences, from which a *ribosome profile* can be generated for every mRNA to reflect the probability of observing a ribosome bound to any specific location on the transcript. Such a probability vector is summarized from counts of ribosomes that were inferred to bind to a location along the mRNA. Since the translation process can be viewed as a queue of ribosomes moving along a mRNA chain while recruiting the matched tRNAs to form a peptide, every ribosome footprint along the chain indicates a translation event, and a larger observed count of ribosome footprints at a certain location over many copies of the same mRNA can be interpreted as a longer average duration of a ribosome translating at that location.

Such positional information has been widely used to experimentally define the translatome — the set of translated mRNAs — and proteome — the set of expressed proteins — of many organisms. For instance, ribosome profiles have been used to identify small open reading frames, or micropeptides, outside of known open reading frames [18, 88, 90, 133, 154]. They have also been used to refine the existing knowledge of functional annotations on various species [27, 52, 80, 83, 121]. Ribosome profiling has also been adapted to block initiating ribosomes instead of elongating ribosomes to identify alternative initiation sites [57, 101]. Together, ribosome profiles lead to new discoveries of alternative initiation and alternative translations, which in turn broaden and adjust our view on protein production.

The total number of ribosome footprints are used to study genome-wide translational trends. Since each ribosome footprint represents a translation event and leads to a protein product, the total number of ribosome footprints are usually used to infer the translation level of a gene [79], and sometimes even the relative gene expression level [19]. The ratio between the ribosome abundance for the ribosome profiling experiment and the transcript abundance from the RNA-seq experiment, which is called *translational efficiency*, is also used to detect differential translational events [67]. *Meta-gene analysis* is another measurement from the ribosome profiling to study the overall trend of the ribosome density [79], where the average count of ribosomes on a single location over all genes is computed, and a meta ribosome profile is generated.

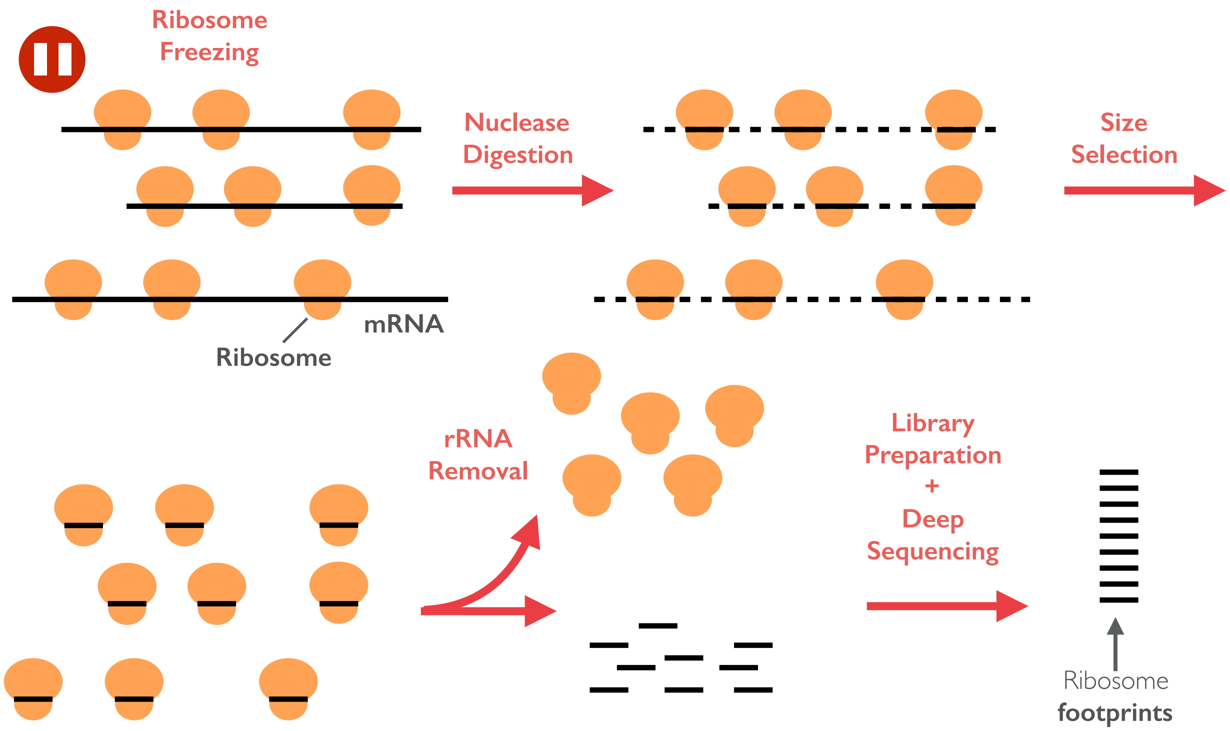


Figure 1.1: Experimental pipeline of ribosome profiling.

These measurements from ribosome profiling data are essential to quantitatively analyze genome-wide translational regulation. For example, time series of ribosome profiling data have been used to study the dynamic process of yeast meiotic program [22], gene activation and lncRNA regulations in zebrafish development [27, 100], and larval development of C. elegans [71, 157]. Different cell conditions like stress or starvation alter gene expression, which are then reflected in ribosome profile patterns [10, 61, 79, 103, 140, 167]. Ribosome profiles have also been used to study translation evolution [12, 115], the development of diseases [64, 73, 163], the progression of aging and cell death [97, 175], and the role of specific translational regulation factors [67, 69, 178]. Thorough reviews on the application of ribosome profiling can be found in [23, 78, 81, 84, 97, 117]. In summary, ribosome profiling is a powerful tool to study translational regulation. It bridges the gap between transcript measurement and protein quantification.

1.2.2 Complicated experimental pipeline outputs noisy ribo-seq data

Briefly, ribosome profiling works in five broad steps (Figure 1.1): First, the ongoing translation process needs to be stopped. That is, the moving ribosomes along mRNAs are suddenly stopped. Second, those mRNA portions not protected by ribosomes are digested away, leaving only ribosomes and the underlying mRNA fragments. Third, these ribosome-mRNA complexes are purified, and other byproducts are discarded. Fourth, ribosomal RNA fragments are removed from such complexes, and only the mRNA fragments — the ribosome footprints, are kept. Lastly, cDNA library preparation is conducted on these ribosome footprints, and they are finally deep sequenced [81, 84].

Biases and noise can be introduced to the final ribo-seq output from the complicated experimental pipeline. First, the picture of ribosome pileups from ribo-seq data might be blurry. This is because, to stop ribosome movements, polysome stablizers, such as cycloheximide, are generally used. But the reversible binding between the ribosome and cycloheximide does not entirely stop ribosome movements, thus it leads to a blurry picture of ribosome pileups [60, 76, 81]. Second, the end product of ribo-seq reads is different from the true ribosome footprints. This is caused by nuclease imperfect digestion [78]. Under-digestion keeps mRNA portions not protected by ribosomes, and over-digestion makes ribosome footprints too short. A more detailed discussion about imperfect digestion can be found in chapter 4. Third, the raw ribo-seq reads are enriched with other contaminated sequences. This is caused by two main reasons: Most contaminants are often removed by a physical size selection step, such as sucrose cushion [155]. Therefore contaminants with a similar size of ribosome footprints are still kept during the library preparation. Further, ribosomal RNA removal step is substantially different among different labs [82, 155, 174], and the efficiency of such a step will largely influence the sequencing depth of the real ribosome footprints.

Bias sources in ribosome profiling experiments have been well documented and summarized [15, 81]. Alternative protocols are compared to minimize biases introduced in rRNA removal, size selection, and library preparation [7, 31, 174]. However, very few methods are proposed to correct for these biases computationally. Artieri and Fraser propose to normalize ribosome occupancies with the parallel RNA-seq level [13]. Hussmann et al. specifically address positional biases caused by cycloheximide, but inevitably lose ribosome footprint enrichment resolution on specific active sites, since the correction requires accumulation of ribosome footprint enrichment on all three active sites [76]. To sum up, ribosome profiling data are inherently noisy due to complicated experiment pipelines, developing computational methods to clean up these data are challenging, yet these biases might distort true ribosome profiles and bury biologically meaningful insights.

Table 1.1: Computational methods for ribo-seq data

Task	**List of methods**
data preprocessing	RPFdb [180], GWIPS-viz [118, 119, 120], RiboTools [102], riboseqR [31], RiboProfiling [136]
data normalization	RUST [127]
profile properties	PTS [121], RRS [68], FLOSS [83], ORFscore [18], Zupanic et al. [192], Liu et al. [109]
ORF detection	TOC [27, 133], ORF-RATER [52], RibORF classifier [88], RiboTaper [24], SPECtre [30], RiboHMM [138]
differential translation	Xtail [179], Babel [129], RiboDiff [187]
translation rate	Siwiak and Zelenkiewicz [153], Huang et al. [75], Pop et al. [135], Gritsenko et al. [66], Shah et al. [150]
bias correction	Artieri and Fraser [13], Hussmann et al. [76]

1.2.3 Computational work on ribosome profiling data reveals genome-wide translational patterns

Many computational methods are developed to process ribo-seq data, identify common features of ribosome profiles, and model translation rate. Table 1.1 is a summary of the major computational problems regarding ribo-seq data and existing methods. First, many efforts have been put to design computational pipelines to preprocess ribo-seq data before downstream analysis. RPFdb [180], GWIPS-viz [118, 119], its galaxy-based version [120], and RiboTools [102] were developed to visualize ribosome profile tracks, riboseqR [31] was developed to visualize reading frame usage, RiboProfiling [136] was developed to visualize summary information on ribosome profiling data, such as amino acid stalling strength and ribo-seq read distributions. RUST [127] was developed to normalize ribo-seq reads. While largely useful, these visualization tools do not systematically handle multi-mappings in ribo-seq reads, which, as will be discussed in chapter 3, is very common. Handling these multi-mappings without caution might induce inaccurate estimation of transcript-level ribosome profiles, and might lead to faulty conclusions on translational regulation.

The common patterns of known ORFs are often extracted to help discover regions with high coding potentials. Since ribosomes move in units of codons (3-nt long), the 3-nt periodicity is recognized as a representative feature in coding regions. A Periodicity Transition Score (PTS) [121] was developed to identify dual-coding regions by spotting frame preference changes in ribosome profiles. A ribosome release score (RRS) [68] was developed to detect the termination of translation at the end of an ORF. A Fragment Length Organization Similarity Score (FLOSS) [83] has been developed to quantify the magnitude of disagreement between read length distributions from two sets of profiles. FLOSS is thus used to identify coding potentials of a set of profiles. An ORFscore [18] is designed to quantify the biased distribution of ribosome footprints toward the first frame of a given region.

Besides these specially defined statistics for capturing ribosome density patterns within ORFs, many machine learning and signal processing approaches have been recently applied to classify novel ORFs from ribo-seq data. A Translated ORF Classifier (TOC) [27, 133] was built to classify ORF regions based on different ribosome density features with a random forest classifier. An ORF Regression Algorithm for Translational Evaluation of RPFs (ORF-RATER) [52] was developed to identify novel ORFs with a linear regression approach by comparing candidate regions with meta-gene profiles from known coding regions. RibORF classifier [88] identifies translated ORFs with a support vector machine classifier. RiboTaper [24] uses a multitaper spectral analysis based on Fourier transform to capture 3-nt periodicity patterns from raw ribo-seq reads and identify novel ORFs. SPECtre [30] identify potential ORFs by calculating the spectral coherence on sliding windows over a candidate region. RiboHMM [138] identifies ORFs with a Hidden Markov Model approach to capture ribosome density features on CDS and UTR regions. Zupanic et al. [192] apply change point detection to identify translational regulations from change of ribosome density. Most recently, Liu et al. [109] successfully predict ribo-seq count from sequence content with a wavelet analysis. This is evidence that sequence content play a role in translational regulation, and the predicted marginal density can serve as a prior in probabilistically mapping reads in the inference of isoform-specific ribosome footprints.

Several methods have also been proposed to identify differentially translated genes from ribo-seq data. Xtail [179] identifies significantly differentially translated genes by estimating a posterior distribution of fold change of translational efficiencies given the translational efficiencies under both conditions, Babel [129] identifies differential translational regulation based on an errors-in-variables regression model by combining ribosome loads with mRNA abundance, RiboDiff [187] models ribosome loads from mRNA abundance with a generalized linear model (GLM). These methods all take into account mRNA abundance level estimated from RNA-seq data while identifying translational changes between conditions.

As described above, currently computational methods on ribo-seq data focus on extracting common features from ribosome profiles, and summarizing transcript-level statistics such as ribosome loads. However, unique features on individual ribosome profiles are not yet fully explored.

1.2.4 Ribosome profiling advances translation models

Translation mechanisms were often studied theoretically prior to ribosome profiling. One type of approach is to divide the entire translation into a list of very fine-grained chemical reactions describing the formation of the ribosome and the process of ribosome elongation. A translation model with a list of ODEs is formed based on the Michaelis–Menten kinetics of the reactions [116, 137, 184, 191]. The variables of the ODEs are the change of the concentrations of the products for a chain of reactions. Such models are very descriptive, yet a prior knowledge of the reaction rates are required. Further, such models can only describe a single ribosome interacting with the mRNA, and are not scalable to include multiple ribosomes on a mRNA.

Translation can also be modeled as a Totally Asymmetric Simple Exclusion Process (TASEP). A stochastic continuous-time Markov chain process is usually performed to simulate the translation process [32, 38]. Here mRNA is modeled as a two-dimensional lattice, and ribosomes move from one location to another with specified transition rates that are given as inputs. More details about TASEP are described in section 3.4.1, where we use this model to generate synthetic test sets for mapping. The translation of each mRNA is considered independently. A similar approach is to treat ribosomes as agents that keep records of information about the particular transcript position to which they are bound and the tRNA with which they are interacting [28, 29, 150]. The agent-based model considers the translation of the entire transcriptome all together. However, modeling the entire ribosome population and the locations on all the available mRNAs is computationally intensive. Both the TASEP model and the agent-based method are capable of quantitatively simulating the genome-wide translation dynamics. Although the full vector of transition rates can be estimated using the TASEP model given the ribosome profile, if we assume a unique transition rate for every position along the mRNA, the model is likely to overfit the data without bringing too much insight about the general knowledge of the rates.

Existing ribosome profile analyses do not take full advantage of codon-specific ribosome occupancy. Earlier works on incorporating ribo-seq data into translational model often focus on usings the overall ribosome abundance estimation. For example, Siwiak and Zelenkiewicz [153] modeled the tRNA-ribosome binding as a diffusion process, and estimated properties of translation such as the translation and initiation rate by incorporating the overall ribosome density from the ribosome profiling data. However their model is deterministic and cannot describe the dynamics of the translation. Huang et al. [75] applied a machine learning approach to predict whether a gene is highly translated or not, where the ‘ground truth’ of the translation rates are also estimated by the overall ribosome density [79]. Hundreds of candidate features that describe both the sequence and the structure properties of the mRNA molecules and the corresponding proteins are selected. While it is desirable to include a large variety of biologically meaningful features, the model itself is a simple binary classifier. Moreover, treating the overall ribosome density as the translation rate might be oversimplified.

Bridge the gap between theory and experiment. Recently, more methods are emerging to model translation using transcript-level ribosome profiles in a more data-driven way. Many methods study the relationship between codon type and elongation speed [13, 39, 40, 59, 98]. Typically, ribosome profiles are normalized by the average ribosome density of a transcripts, and these normalized footprint counts are grouped by codon types, and some statistics are computed based on these codon count distributions to represent codon decoding time. Either the distribution means are used [98], or the distributions are fit to some statistical model, and specific model parameters are used as the codon decoding time estimates [39, 40]. These approaches assume that ribosome footprint counts among different transcripts can be brought to the same level by proper normalization, and the effect of transcript abundance and initiation rates on ribosome footprint counts can be canceled out. They also assume that codon type is the main factor that regulates translational speed, and this speed can be estimated from codon count distributions. However, they might overlook the effect of transcript abundance and initiation rate on translation speed of individual transcript, where these effects might play a larger role in the observed ribosome footprint counts, and the effect of codon type on translational speed might be buried by other confounding factors.

Alternatively, Pop et al. [135] model the codon-specific elongation rate on each transcript individually. The observed ribosome footprint counts are assumed to follow a poisson distribution. Their model also assumes that the observed ribosome profiles are at steady states, so a constant ribosome flow assumption must hold at each transcript location. They iteratively estimate the elongation rates and the ribosome flows so that the observed ribosome footprint counts are best explained. Unintuitively, their estimated elongation rates do not correlate with tRNA abundance estimations. Gritsenko et al. [66] take another approach by assuming universal elongation rate per codon type, but transcript-specific initiation rates. They use TASEP to estimate the rate parameters that fit well with the ribosome footprint observation on multiple scales. Shah et al. [150] propose another approach (which is recently applied in [174]) to estimate the initiation probability from ribosome profiles with a continuous-time, discrete-state Markov model of translation, assuming other parameters in the model are perfectly known.

To validate the model assumptions, the estimated codon decoding time is often compared with tRNA abundance estimates [39, 40, 66, 98, 135], codon usage [59, 125], and mRNA secondary structure [135, 182]. However, controversial conclusions about factors that regulate translation rate are often derived from these analysis. For example, Pop et al. do not find tRNA abundance to be correlated with elongation rates, while Dana et al. find tRNA abundance to be significantly correlated with several definitions of elongation rates in yeast [39, 40]. It might be that multiple factors are acting together to regulate translation speed [65]. It might also be that noise and biases introduced from the experiment pipeline distort the true ribosome occupancy [13, 76]. A recent study compared the reproducibility of ribosome profiles among biological replicates, and showed that ribosome profiles lack consistency between replicates at a codon level, yet the consistency increases as ribosome counts are estimated with a coarser resolution [42]. This is another line of evidence that ribosome profiling data are inherently noisy. This also shows that exploring the mechanism of translational speed requires a better designed protocol both experimentally and computationally.

1.2.5 Challenges in ribosome profiling analysis

In summary, ribosome profiling is a useful relatively recent sequencing technique to study genome-wide translational patterns. Many methods have been developed to extract common features of ribosome profiles. However, reliably identifying unique patterns from individual ribosome profiles and quantify translational speed have remained challenging. Raw ribo-seq reads need to be handled with caution to produce accurate transcript-level ribosome profiles, since all downstream analysis relies on such profiles. Meanwhile, methods need to be developed to filter noise and reveal true ribosome footprint locations.

1.2.6 Our contributions

In this thesis we introduce two methods to tackle two challenges in ribo-seq data processing. We first resolve multi-mappings in ribo-seq data and output transcript-level ribosome profiles in chapter 3. Our method is the first approach to systematically handle multi-mappings in ribo-seq data. We show that our method accurately recovers isoform-level ribosome profiles by assigning multi-mapping reads guided by transcript abundance. Our method also leads to a more reasonable estimation of ribosome loads. Both ribosome profiles and ribosome loads are crucial for downstream analysis such as differential translational analysis and translation speed estimation, and our computational pipeline prepares ribo-seq data for an unbiased translational analysis.

We then propose a mathematical model and algorithm to infer ribosome footprint locations from ribo-seq reads in chapter 4. This is a necessary step to convert raw ribo-seq reads into meaningful ribosome profiles that reflect the ribosome occupancy preferences over time. We show that such challenge arises due to incomplete digestion during the experimental procedure, as mentioned above. Our method successfully reduces noise from ribo-seq data and outputs crisper ribosome profile estimations. Using these better estimated profiles, our estimated codon decoding time illustrates that both tRNA abundance and wobble pairing influence translational speed.

Lastly, we provide a genome-wide analysis on ribosome collision quantification. This is the first approach to quantify extreme local slow-downs of ribosome movements that result in two ribosomes sitting adjacently on the transcript. With the help of a modified ribosome profiling protocol that experimentally captures ribosome collisions and a rigorous computational analysis framework, we show that extreme ribosome stallings are consistent with the model that they occur randomly. Our study compliments previous findings by providing experimental evidence for ribosome stallings. However, we consistently observe the libraries for capturing ribosome collisions are dominated by contaminated sequences from ribosomal RNAs, and the coverage for the real ribosome collision sequences from mRNAs are extremely low. Therefore our study is only preliminary and our observations are suggestive but not conclusive. Additional experiments that can effectively remove contaminants from biological samples are needed, and biological replicates are required for future validation of our findings.

Together, we provide multiple computational methods to guarantee an accurate ribosome profile estimation from raw ribo-seq reads, which further promotes our understanding of translation mechanisms and translational regulation. Our methods provide the groundwork to preprocess ribo-seq data and output unbiased ribosome profiles at a codon level, whose accuracy is critical in answering question two.

1.3 Summary

In this thesis, we developed novel computational methods to use next-generation-sequencing technologies to understand gene expression. These methods are based on graph theories, signal processing, and statistical tests, and they focus on pre- and post-transcriptional aspects of gene expression, two important areas that are under studied compared with transcription. Our methods reliably preprocess the raw sequencing data and produce biologically meaningful results. They will be useful for researchers to better understand the workings of the cell.

Chapter 2

Topological properties of chromosome conformation graphs reflect spatial proximities within chromatin

This chapter describes a statistical framework to detect sets of spatially closed regions from Chromosome Conformation Capture data. The validity of the proposed method is crucial in examining the relationship between genomic function and its spatial arrangement. We start with a brief introduction of the Chromosome Conformation Capture technique and a survey of existing approaches to identify spatially closed regions. We then provide a formal problem definition and give a design of the tested framework using topological properties from chromatin interaction graph. Finally, we end the chapter with validations of our method for detecting spatial closeness on both simulated test sets and actual genomic annotations in yeast.

The content of this chapter was originally presented at ACM Conference on Bioinformatics, Computational Biology, and Biomedical Informatics (ACM BCB) in 2013, and was published in the Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics [169]. It was a joint work with Geet Duggal, Rob Patro, Michelle Girvan, Sridhar Hannenhalli, and Carl Kingsford.

2.1 Background

Chromosome Conformation Capture (3C) [41] is an experimental method used to study the spatial structure of chromosomes by observing pairwise spatial contacts between regions of chromatin. Such experiments provide counts of observed instances of cross-linking between pairs of genomic segments flanked by restriction enzyme sites, with the interpretation that pairs of segments with high counts were often in close proximity among the population of cells assayed. The 3C technique and its subsequent refinements (4C [47, 151], 5C [16, 46, 166], Hi–C [43, 107, 149, 161], TCC [92]) have been used as tools to explore the genomic structures and features of bacteria [166], yeast [47, 161], fruit fly[149], mouse [43], and human [16, 43, 107]. These interactions have been used to verify the large-scale organization of chromatin territories [107], to investigate cancer and disease related genome alternations [55, 114], and to confirm and postulate instances of long-range regulation [16].

Embeddings of three-dimension genomic structure are expensive to compute. In order to identify spatially colocated genomic features, a three-dimensional model of the chromatin is often computed to study the genome structure [16, 20, 47, 161, 166]. The model is usually built to satisfy as many of the 3C interactions as possible, while respecting a variety of other established properties of chromatin, including the volume constraint of the nucleus, the physical constraints of the DNA molecules, and some known biological preferences of the chromosome structure. A three-dimensional model of the chromatin structure is useful, and it can incorporate biological constraints on highly repetitive sequence regions such as centromeres and ribosomal DNAs [47] that are not available in the raw 3C data. However, requiring the computation of an embedded chromatin structure is computationally complex and expensive. In the yeast 4C experiment, for example, there are 4,053 genomic fragments and 914,746 spatial constraints between pairs of fragments, and it takes more than one day to optimize this complex objective function [47]. Further, a large amount of uncertainty exists in the embedded structure. In order to build the constraints among genomic loci, the observed 3C interaction frequencies need to be mapped to distances. Different frequency-to-distance mappings will result in very different constraints, which will then lead to different structures. Moreover, the constraints from the 3C interactions are usually metrically inconsistent [48]. Such inconsistencies are a result of both the noisy 3C data and the nature of the 3C experiment. The interactions come from an aggregate population of cells which may have different and incompatible geometries. Only a small subset of the constraints can be satisfied to some degree, and different subsets of constraints will result in different 3D embeddings [48].

Existing tests for spatial enrichment might be sensitive to missing measurements and outliers in the test set. Several approaches have been proposed to infer spatial colocalization of the chromatin structure without computing a chromosome embedding. Earlier methods assume that the number of observed 3C interactions given the number of possible interactions follows a hypergeometric distribution [37, 47]. Such a parametric approach, as pointed out by Witten and Noble [177], makes the inaccurate assumption that every observation of an interaction is an independent event. This assumption does not hold because 3C interactions involving the same or nearby fragments of genes are strongly correlated. To address this, Witten and Noble [177] proposed a non-parametric procedure to evaluate the spatial closeness for sets of genes by randomly resampling sets of restriction fragments as the background distribution. Kruse et al. [96] subsequently proposed a rewiring procedure, randomly shuffling the interactions while preserving the degrees of restriction fragments and the transitivity of the entire graph, to sample from background distribution. All of the methods above are based on the (sometimes weighted [161]) fraction of observed interactions (edge-fraction) between the loci of a given set of genomic features.

2.2 Contributions

Here, we propose and compare a variety of topological metrics as measurements of spatial proximity without the need to compute a three-dimensional embedding of the genome (or any subset of it). Given a set of genomic loci, we explore the properties of all pairwise shortest paths, network flow, and the maximum density subgraph to evaluate spatial proximity of the chromosome structure directly from the 3C graph. Our methods are able to make better use of the observed frequencies of each interaction (edge weights) than the previously proposed edge-fraction approach. For example, our approach based on shortest paths accounts for transitivity of distance constraints and ought to reveal some information about non-observed 3C interactions due to the absence of cross-linking. The maximum density subgraph approach can extract the densest core in a graph and can reveal surprisingly compact regions masked by outliers.

We investigate the topological properties on 4C measurements in budding yeast *Saccharomyces cerevisiae* [47]. These data have been widely studied for colocalization tests on genomic features [20, 37, 47, 96, 177]. We show that all tested topological properties correlate well with the spatial proximity measured by the volume of the convex hull of the dense subgraph of a set of positions. The volume of the convex hull is computed based on an existing three-dimensional model of the yeast genome [47].

We apply our methods to both synthetic feature sets selected from the yeast three-dimensional embedding and to real genomic features [47]. To test the intuition that richer graph properties will more accurately identify spatially close sets, we introduce a new framework for systematically evaluating a method's false positive rate, true positive rate, and ability to handle outliers when estimating spatial enrichment.

We show that under a reasonable resampling scheme that controls for chromosome-specific interaction patterns, our colocalization statistical analyses are both robust and unbiased. Moreover, all methods perform well on test sets that contain a variety of chromosomes. For such sets of genomic loci, the method of finding the maximum density subgraph has the added benefit that it finds dense regions in the graph that overlap well with the true spatially compact regions. Lastly, by incorporating interaction frequencies into the tests, we find the telomere sets, which were previously thought to be significantly colocalized [47], are likely to be not. Overall, we illustrate that these proposed graph-theoretic measures can identify spatial closeness well without the need to compute an embedding and can be an alternative indicator of spatial functional enrichment.

2.3 Problem specification

Given a set F of genomic loci (representing genes or other features) and a collection of observed 3C interactions G, we would like to test whether the points in F are significantly spatially close as implied by the 3C interactions. We compute a statistic f on G and F, and we argue that statistical significance of f likely indicates statistically significant spatial proximity. This leads to the following problem:

Problem 1 (Spatial Proximity Test) *Given a set of genomic loci F and a weighted graph $G = (V, E, d)$ of 3C interactions where V is the set of genomic segments produced by the 3C experi-*

*ment, E is the set of 3C interactions, and $d(e)$ is the weight of interaction e, return **YES** if F is statistically significantly spatially close in three dimensions. Otherwise, return **NO**.*

The input loci F here is a subset of V. Input sets consisting of genes or genomic ranges should be mapped to genomic fragments first, and the test statistic is computed using the fragments. The given edge weight $d(e)$ is typically an estimate of pairwise spatial proximity derived from observed 3C interaction frequency. Problem 1 does not address the issue of outliers within the provided set F. To handle these, we introduce the following problem:

Problem 2 (Compact Core Finding) *Given input as in Problem 1, return **YES** if some* subset *of F is spatially close in three-dimensions. If so, return the subset.*

2.4 Methods

2.4.1 Graph-based proxies for spatial closeness

We evaluate the following topological properties for their statistical correlation with spatial proximity. Given a set of restriction fragments $F \subset V$, we compute the following topological properties $f(F)$:

(a) $f_{\text{edge\_fraction}}(F) = \dfrac{|E(F)|}{|E_a(F)|}$.
Here, $E(F)$ is the set of observed edges with both endpoints in F. $E_a(F)$ is the set of all possible edges among the given set of nodes. If only inter-chromosomal edges are included,

$$|E_a(F)| = \sum_{\substack{i,j,i\neq j\\ i,j\in \text{chromosomes in F}}} c_i c_j,$$

where c_i is the number of fragments in F on chromosome i. $f_{\text{edge\_fraction}}$ is widely used as a proxy for spatial proximity [47, 177], however it is likely sensitive to the effects of outliers or missing measurements in the 3C experiments. This method does not use edge weights directly, but rather considers an edge present only if the false discovery rate (FDR) derived from its observed frequency is less than 0.01.

(b) $f_{\text{sp\_mean}}(F) =$ the mean of all pairwise shortest path lengths between nodes in F. The weight on each edge e here is the distance $d(e)$ computed using a frequency-to-distance mapping. Computing shortest paths instead of using just observed edge weights addresses the issue of missing 3C interactions in connected triples and longer paths in the 3C graph. $f_{\text{sp\_mean}}$ is therefore robust to this kind of incomplete experimental data.

(c) $f_{\text{flow\_mean}}(F) =$ the mean of the maximum flow value between pairs of nodes in F. The maximum flow on an edge u, v can be thought of as the largest amount of water that can be sent from u to v by treating the edge as a pipe with the capacity of the observed 3C interaction frequency. Unlike when computing shortest paths, the edge weight here is the interaction frequency. Maximum flow avoids the problem that shortest paths can be significantly lengthened or shorted by a single edge deletion or addition. All pairwise maximum flows are efficiently computed via a Gomory-Hu tree [63].

(d) $f_{\text{max\_dense}}(F) =$ the density of the maximally dense subgraph D contained in F, where

$$\text{density} = \frac{\sum_{e \in E(D)} w(e)}{|V(D)|},$$

and $w(e)$ is the interaction frequency of edge e. The unweighted density can be computed by setting $w(e) = 1$. The definition of density used in $f_{\text{max\_dense}}$ has been widely studied and admits maximization via a polynomial-time algorithm [62]. This statistic emphasizes the importance of a compact core and helps to eliminates the effects of outliers. A density definition like (a) above — the portion of observed edges — is not applicable to maximization since there exists a trivial solution that can maximize the density: a graph with one edge.

2.4.2 Scheme for spatial enrichment tests

We obtain a p-value for the statistic $f(F)$ in a non-parametric manner similar to that proposed by Witten and Noble [177]. The procedure is described below:

(1) Resample a set $\mathcal{B}$ of 1000 sets of from V. How these sets are sampled depends on whether the input features are fragments, genes, or genomic ranges. In the case that they are fragments, then $|F|$ random restriction fragments are chosen. If the original input is a set of genes, then the same number of genes are randomly selected, and the selected genes are converted into fragments by choosing fragments whose midpoint lies within the selected gene. If the inputs are genomic regions, then new random starting coordinates for the regions are chosen, keeping the length of each region unchanged; we then choose those fragments whose midpoint lies within the regions. In all three cases, we keep the number of elements selected from each chromosome the same as in the input F. Such a procedure controls the fact that different chromosomes may interact with each other quite differently due to the tethered nature of the yeast genome and due to the differences in the chromosomal lengths [164].

(2) For each $B \in \mathcal{B}$ compute $f(B)$ to get a background distribution of values.

(3) Compute the empirical p-value as the fraction of examples $B \in \mathcal{B}$ where we count $f(B) \geq f(F)$ (except for shortest path, where we count $f(B) \leq f(F)$). A set is called statistically significant if this p-value ≤ 0.05.

In order to generate the background distribution, our resampling procedure randomly samples nodes in the graph while keeping the graph topology fixed. An alternative approach would be to randomly rewire the interactions of the network by fixing the nodes of interest [96]. However, generating a set of random graphs as the null model that preserves the topological structure of the original graph without introduing any artificial bias is quite challenging. Kruse et al. [96] propose a Markov-chain procedure for reshuffling 3C edges until the rewired graph reaches or exceeds the transitivity of the original graph. This is a computationally intensive procedure since it requires many reshuffling steps to obtain a graph with transitivity comparable to that of the observed 3C graph. Further, there might exist other properties, or a combination of properties, that can better describe the topological structure of the yeast network. We thus chose the procedure above in consideration of efficiency, simplicity, and generality.

2.5 Data and test sets

2.5.1 Yeast 3C interaction data

We use the *S. cerevisiae* 4C measurements based on the HindIII restriction enzyme library from Duan et al [47]. The 3C experimental procedure may introduce systematic biases that distort the true frequency of the data. We therefore applied the same false discovery rate (FDR) cutoff of 0.01 to pre-filter the noisy 4C data, and we applied the same frequency-to-distance mapping to convert the interaction frequencies to distances. Other normalization methods have been proposed for Hi-C experiments [74, 77, 181]. Some are not directly applicable to 4C data because the assumptions for normalization are specific to Hi-C data [77]. Kruse et al. [96] applied the normalization method proposed by Yaffe and Tanay [181] to get an interaction probability for every fragment pair. They then chose a different FDR cutoff to filter the 3C interactions. However, Yaffe and Tanay's normalization method [181] does not take into account the circulation bias [34] that is specific to 4C experiments. Here, in order to compare directly with Duan et al. [47] and Witten and Noble [177], we use their data processing framework.

We test our methods on 3C graphs considering only the inter-chromosomal (fragments from different chromosomes) interactions. Including intra-chromosomal (fragments from the same chromosome) data ought to be beneficial since more information is incorporated for evaluating spatial enrichment. It can reveal unique spatial structure patterns of a specific chromosomes like zippering [47] and long range looping [16]. However, intra-chromosomal interactions are strongly influenced by linear genomic proximity. A high frequency intra-chromosomal interaction can be caused by genomic closeness or spatial closeness, and it is difficult to distinguish the two. Furthermore, intra-chromosomal interactions have higher frequencies than inter-chromosomal interactions, and thus can carry more weight in spatial enrichment estimation. We therefore consider only inter-chromosomal interactions following Witten and Noble [177] and Kruse et al. [96]. We validate this choice further in section 2.6.5.

2.5.2 Sets of yeast genomic loci of interest

We use the genomic feature sets from Duan et al. [47]. These features include: centromeres, telomeres, breakpoints (including the ancestor breakpoints of *S. cerevisiae* and the evolutionary breakpoints between *S. cerevisiae* (Scer) and *Kluyveromyces waltii* (Kwal)), transfer RNAs (tRNAs) (including the entire tRNA set, two sub-clusters of the tRNA set, and tRNAs outside the two clusters), and origins of early and late DNA replication (including two sets with different identification mechanism). These features were chosen by Duan et al. [47] with both theoretical and experimental support of their clustering behavior.

2.5.3 Selection of chromatin regions for spatial enrichment test

Generation of synthetic cores

To test each method's ability to recognize truly spatially close cores masked by outliers, we generate synthetic sets of features with different sizes (20, 50 and 100 fragments) by choosing

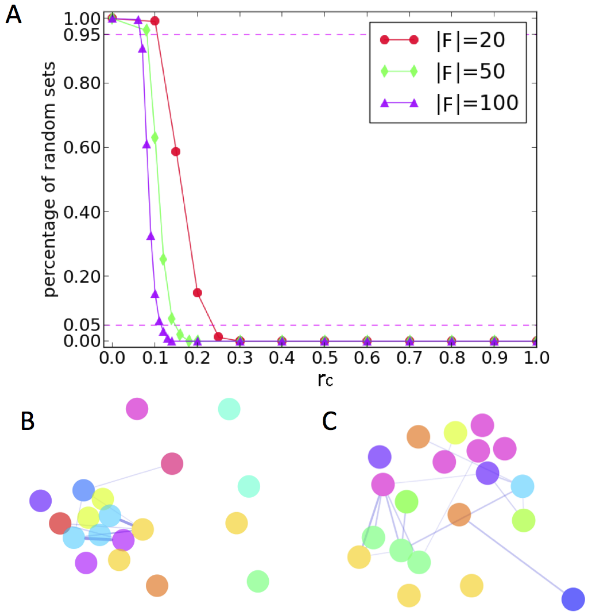


Figure 2.1: Defining positive examples and negative examples. (A) Cumulative distribution of 1-r_c. (r_c here is the relative size of the largest spatially-close subset of the instance.) None of the random sets contain a spatially close set with $r_c > 0.5$. (B) An example of true positive set ($r_c = 0.6, |F| = 20$). (C) An example of random set ($|F| = 20$). (B) and (C) are drawn using a spring layout. Node colors represent different chromosomes. A less transparent and wider edge represents a higher 3C interaction frequency between fragment pairs.

random segments on the embedded yeast chromatin structure computed by Duan et al. [47]. We sample from Duan et al.'s [47] embedding since it is built based on real 3C data, and it incorporated a variety of known biological constraints. We define a set of segments to be spatially close if they are within a diameter of 400nm, and we define them to be not spatially close otherwise. This diameter is chosen by observing that 400nm is 1/5 of the diameter of the yeast nucleus (2000nm) [47], and such a diameter results in a volume $< 1\%$ of the entire nucleus volume. Additionally, in the histogram of the pairwise Euclidean distances between beads in the Duan et al. [47] yeast embedding, only 11.7% of the distances are within 400nm.

To construct a synthetic set with a spatially close core and with some fraction of outliers, we first pick a certain percent (r_c) of the synthetic set as *core segments*. To do this, we choose a center segment $u \in V$ and then search for restriction fragments that fall within the sphere centered at that point with a radius of 200nm. All fragments inside this sphere will be at most 400nm away from each other. We define $C_a(u)$ as the set of all segments within a 200nm radius of u. We discard $C_a(u)$ if $|C_a(u)| < r_c|F|$. We then randomly pick $C(u) \subset C_a(u)$ as the set of core segments such that $|C(u)| = r_c|F|$. Secondly, we randomly choose the rest of the nodes in F outside the 200nm radius from the center point to be outliers. The reason for choosing some spatially-not-close fragments to add to the synthetic core instead of some random fragments is to enlarge the effect of the outliers.

If all sampled fragments are from the same chromosome, any method based solely on inter-chromosomal interactions will fail to detect the spatial proximity by construction, since no intra-chromosomal interactions are included. We therefore discard sets only containing fragments from a single chromosome since our test data is inter-chromosomal. (See discussion in section 2.6.5.)

Constructing spatially close sets (positive examples)

In order to evaluate a statistic's power for detecting spatial enrichment, a positive example—a set of fragments that can be called significantly more spatially close than expected by chance, and a true negative—a set that cannot be called significantly compact, should be well defined.

Intuitively, a set containing a large compact core (r_c is big) is likely to be called spatially enriched as a whole. We therefore find a r_c cutoff such that a set with a core at least that large can be called significantly spatially compact. To find such a cutoff, we estimate the size of the largest set with a diameter of 400nm in 1000 randomly selected fragment sets with different set sizes ($|F| = 20, 50, 100$). This distribution is an estimation of the probability of observing a compact core of a particular size in randomly chosen samples (Figure 2.1A). None of the samples in the random sets contains a compact core with $r_c > 50\%$. We thus define a set of embedded fragments generated as in the previous section to be a positive example if the largest close core within the set has a size $\geq 0.5|F|$. An example of a true positive set is shown in Figure 2.1B. For different set sizes $|F| = 20, 50, 100$, we generate 1000 positive sets with r_c varying from 0.5 to 1.0.

Constructing negative examples

Analogous to the definition of positives, we could define the negative set with another cutoff of r_c such that a set containing a core with fewer than r_c nodes is not significantly spatially close. However, such a filtering scheme makes the example less random. This introduces the new problem that a method that can reject this 'far' set is not necessarily capable of rejecting the true null hypothesis of random loci. Therefore we define the negative set as a set of randomly chosen fragments (or genes). An example of a random set is shown in Figure 2.1C.

2.6 Results

2.6.1 Graph-based statistics correlate well with embedded distances on dense cores

We randomly sample 1000 sets ($|F| = 20, 50, 100$) of fragments from G. As an indication of the true spatial proximity, we compute the volume of the convex hull [25] using the embeddings from Duan et al. [47] for every random set. We then compute the correlation between the topological properties and the true spatial proximity. We take the cubic root of the volume so that the unit of the topological properties and the unit of the true spatial proximity are on the same scale.

Table 2.1: Pearson correlation coefficient between the tested topological properties and the cubic root of the volume of the covex hull on the entire randomly sampled fragment sets and on the maximum density subgraphs within in the sets.

	$\lvert F\rvert = 20$		$\lvert F\rvert = 50$		$\lvert F\rvert = 100$	
	whole set	dense core	whole set	dense core	whole set	dense core
edge-fraction	-0.08	-0.79	-0.11	-0.88	-0.10	-0.94
average shortest path	0.05	0.79	0.03	0.91	0.07	0.96
average max flow	-0.05	-0.16	-0.09	-0.58	-0.11	-0.74
weighted density	-0.06	0.05	-0.05	-0.35	-0.02	-0.51
unweighted density	-0.07	0.41	-0.08	0.20	-0.07	0.16

Although no strong correlations are observed when the properties are computed on the entire set (Table 2.1), if the maximum density subgraphs within the sets are found first, and the embedded distances and the topological properties are computed on these subgraphs, strong correlations appear between the embedded distance and the edge-fraction, average shortest path, and average maximum flow (Figure 2.2, Table 2.1). (Here property (d), the density is computed on the entire maximum density subgraph, not on the maximum density subgraph of the maximum density subgraph). Edge-fraction, max flow and density are all inversely correlated with the embedded distance, while shortest path positively correlates with the embedded distance. Intuitively, we expect a spatially compact set to be denser, to have a larger average max flow, and to have shorter average shortest path, and we find the sets returned by maximum density subgraph to have these properties. These correlations increase as $|F|$ increases. This is because not enough 3C interactions are included in smaller sets to accurately evaluate the density and edge fraction, and noisy data has a larger effect.

The density of the maximum density subgraph has a weaker correlation compared to other tested properties. As we observe, the density grows as the graph size increases. For instance, a complete non-weighted graph of size n has a density of $(n-1)/2$. Edge-fraction, on the other hand, assigns all complete graphs of all different sizes the same score of 1. Therefore, the positive correlation between the density and the graph size weakens the correlation between the density and the embedded distance.

The weighted density correlates better with real spatial proximity than the unweighted density (Table 2.1). The unweighted density does not correlate as expected with spatial proximity. For set size 50 and 100, weighted density is inversely correlated with the cubic root of the volume, while the unweighted density are positively correlated with the cubic root of the volume. The inverse correlation is expected since denser regions should have a smaller volume. The correlation for both weighted density and non-weighted density are positive when set size is 20, which is probably due to the sparsity of the small graph as discussed above. These results illustrate that a more precise evaluation of the spatial proximity within set can be achieved by considering the interaction frequencies rather than just edge presence or absence as done by Witten and Noble [177].

High correlations can occur if the densest subgraphs of all sets strongly overlap with each other. Under such circumstances, the conclusion that the topological properties drive the high

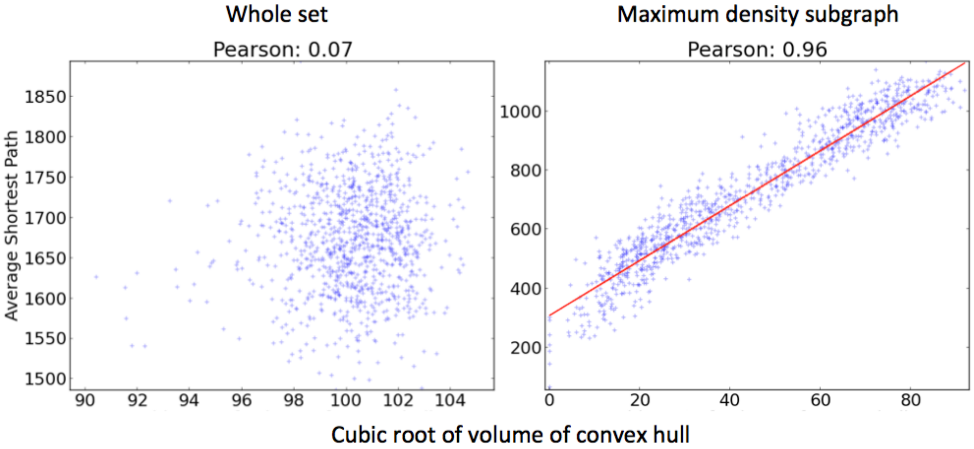


Figure 2.2: Scatter plot relating average shortest paths and the cubic root of volume of convex hull (nm) on the entire randomly sampled fragment sets and on the maximum density subgraph within the sets ($|F| = 100$). Average shortest path on the maximum density subgraph strongly correlates with the cubic root of the volume, while no correlation is observed on the entire set.

correlation is not valid. To make sure the correlation is not caused by dense, highly overlapping regions, we compute the Jaccard similarity coefficient on nodes in maximum density subgraphs for all pair of sets. More than 99% of the pairwise Jaccard scores are less than 0.5 for all set sizes. Moreover, 99.1% pairs share zero nodes in their maximum density subgraphs when $|F| = 20$. The proportion of zero overlap is 95.3% for $|F| = 50$ and 82.4% for $|F| = 100$. This test illustrates that we have covered distinct regions on the chromosome and that the correlation between the approximate embedded distance and the tested topological properties holds in general.

The results above not only demonstrate that maximum density subgraphs correspond to spatially compact cores, but they also indicate that the tested topological properties are a good approximation for spatial proximity of these cores.

2.6.2 The maximum density approach identifies true compact cores

Further, to evaluate the ability of the maximum density subgraph approach to find spatially compact cores, we want to show both that the nodes inside the maximum density subgraph (dense core) overlap well with the nodes in a known true spatial compact core (true core), and that the volume of the dense core agrees well with the volume of the true core. We use a spatial Jaccard score to measure a combination of both properties. The spatial Jaccard score J_{vol} between nodes in the dense core D and nodes in the true core C is defined as:

$$J_{\text{vol}}(D, C) = \frac{\text{volume}(D \cap C)}{\text{volume}(D \cup C)},$$

where volume(X) is the volume of the convex hull of a set X of points.

We compute the spatial Jaccard score on all positive sets and observe that the spatial Jaccard score increases if the portion of the fragments from the most common chromosome decreases and

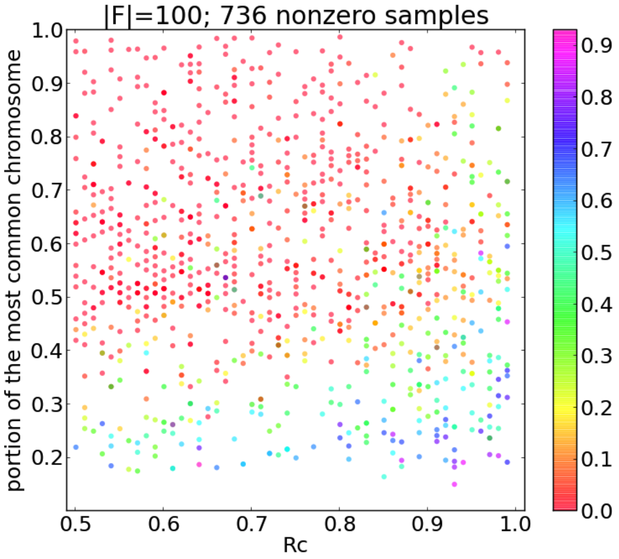


Figure 2.3: Scatter plot of the spatial Jaccard score between the true compact core and the dense core on the positive sets with size $|F| = 100$ (similar results observed for $|F| = 20, 50$). The x axis is r_c, y axis is the portion of the most common chromosome, and the color represents the value of the spatial Jaccard score. The spatial Jaccard score is high when r_c approaches 1 and when the test sets contains fragments from a variety of chromosomes.

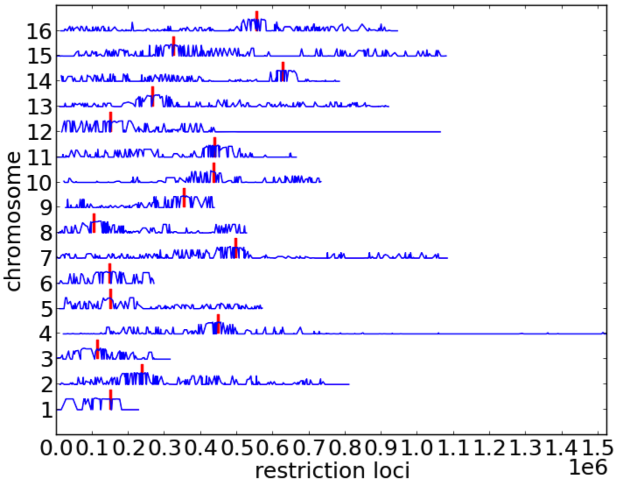


Figure 2.4: Maximum spatial Jaccard scores of fragments (J_f) observed in the maximum density subgraphs of all test set with $|F| = 100$. Centromere regions for different chromosomes are marked with red rectangles. J_f is high near centromere regions.

if r_c increases. The score is generally between 0.5 and 0.9 when the portion of the most common chromosome is around 20% (Figure 2.3). This indicates that the dense core overlaps well with the true core in terms of volume when the set is not denominated by a single chromosome.

The spatial Jaccard score drops down to near zero when the portion of the most common chromosome is over 50%. The reason that maximum density subgraph cannot extract the most dense regions for such test sets is due to the absence of intra-chromosomal interactions. A detailed discussion about our reason to exclude intra-chromosomal interactions is in section 2.6.5.

To determine whether certain chromosomal regions correspond to dense regions that overlap well with true compact regions, we look at the maximum spatial Jaccard score of a fragment within the maximum density subgraph for every test set. Formally, every test set F contains a true compact core C_F and a maximum dense subgraph D_F; these result in a spatial Jaccard score $J_{\text{vol}}(D_F, C_F)$. Let $\mathcal{F}$ represent all test sets. For a restriction fragment $r \in \bigcup_{F\in\mathcal{F}} D_F$, the maximum spatial Jaccard score is defined as:

$$J_f(r) := \max_{F\in\{S|S\in\mathcal{F},\, r\in D_S\}} J_{\text{vol}}(D_F, C_F).$$

A high J_f indicates that there exists some highly overlapping cores involving this fragment. We observe fragments with a high spatial Jaccard scores often locate near centromere regions (Figure 2.4). More than half of the chromosomes have a average non-zero J_f scores greater than 0.4 within a 20,000 bp window of the centromere. Moreover, 100% of the non-zero J_f scores of short chromosomes (such as chromosome 1, 3 and 9) near the centromere are > 0.5, and the average non-zero J_f scores are greater than 0.6. Yet there are 432 fragments outside of the centromere regions (100,000 bp window) with $J_f > 0.5$. In summary, fragments with high Jaccard scores mainly locate near centromere regions, but can be also found in other areas on the chromosome.

2.6.3 An unbiased null hypothesis

When tested on randomly generated gene sets containing randomly selected genes without any functional relationships or colocalized properties [177], all tested topological statistics produced a uniform p-value distribution (Figure 2.5). Evaluating the p-value distribution on the null sets is a standard approach to check whether a statistic has good control for type I error [96, 134, 177]. A uniform distribution of p-values is expected if the statistic is valid. Similar results are observed on sets of randomly chosen fragments.

2.6.4 Topological properties as spatial proxies for spatial enrichment test

We test the power of each topological property as spatial proxies for evaluating spatial enrichment of a given set on the positive sets with different set sizes. We observe all methods correctly call compact cores significant when the portion of fragments from the most common chromosome is less than 30% (Figure 2.6). All methods except for edge-rewiring achieve a true positive rate of 100% when the portion of fragments from the most common chromosome is less than 20%, and a positive rate above 80% when the portion of the most common chromosome is less than 30%. Similar to the result of the dense core overlap, the spatial enrichment evaluation is more accurate when sets of fragments are from several chromosomes.

The true positive rates of all methods decrease when the portion of the most common chromosome increases, and it reaches a low level when the majority of the test set consists of fragments from the same chromosome. If most of the fragments are from one or two chromosomes, there

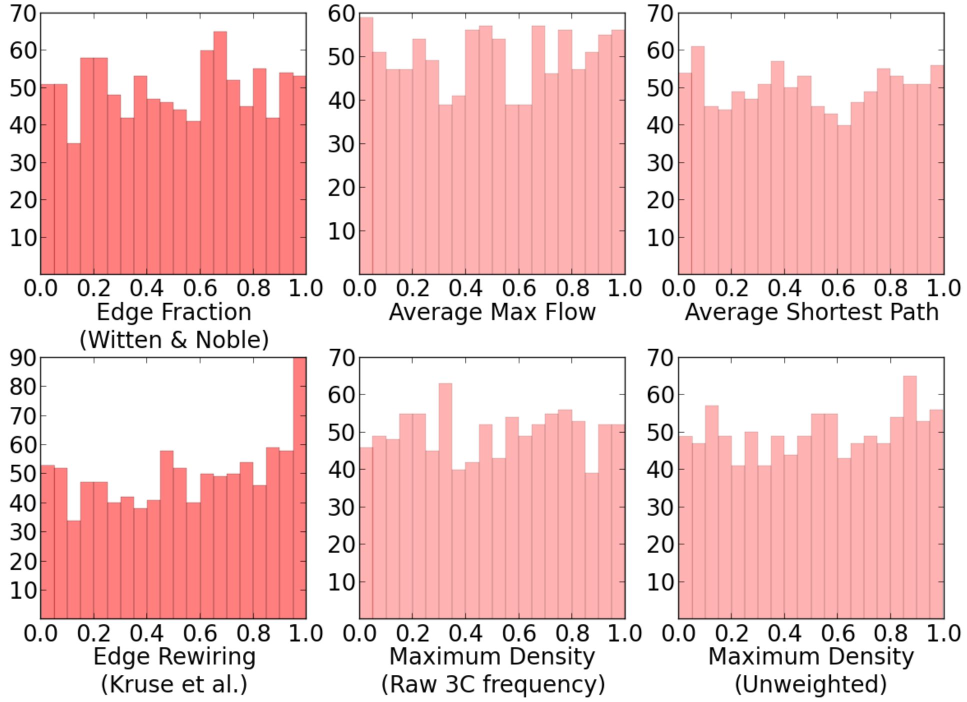


Figure 2.5: Histogram of p-values for different methods on random gene sets [177]. All sets of genes are chosen from a list of target genes of all known transcription factors, the size of the gene set is determined by the number of target genes of a randomly selected transcription factor. All statistics in section 3.2 achieve a near uniform distribution on this null set, indicating the methods are not biased.

will be very few inter-interactions among the set to accurately estimate whether the set is more spatially close than expected by chance.

The edge rewiring method proposed by Kruse et al. [96] is the most conservative among all tested statistics: The true positive rate on $|F| = 20$ is below 20%. This is probably because edge rewiring controls for the global transitivity of the entire graph in the random rewiring procedure, while the transitivity in local subgraphs in yeast might vary radically.

2.6.5 Rationale for including only inter-chromosomal edges

In line with previous studies [47, 96], our tests are on the set of inter-chromosomal edges. Ideally, both intra-chromosomal and inter-chromosomal edges would be used. However, testing for spatial enrichment including intra-chromosomal interactions remains a challenge.

First, a cutoff of 400nm is potentially unsuitable for a primarily intra-chromosomal set of fragments. Based on the embedding, more than 50% of the intra-chromosomal distances are less than 400nm, while less than 10% of the inter-chromosomal distances are less than 400nm. Thus 400nm is not a ‘surprisingly close’ cutoff for intra-chromosomal distances. However, it is difficult to find a distance cutoff that captures sets that contains both significantly spatially close

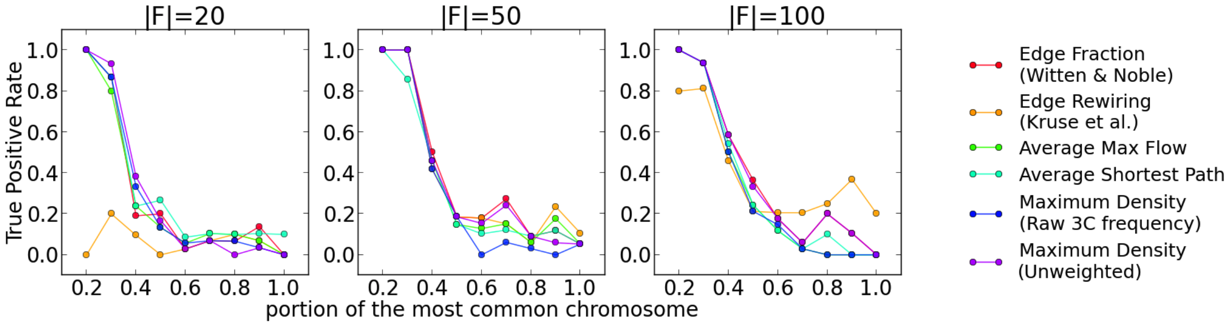


Figure 2.6: True positive rate of different methods on positive examples with different set sizes ($|F| = 20, 50, 100$). The x axis is the portion of the most common chromosome. The true positive rate is high when the fragments of the test sets are from different chromosomes, and is low when the fragments are mainly from the same chromosome. Edge rewiring is the most conservative method.

intra-chromosomal and inter-chromosomal structures.

Second, the close spatial proximity between intra-chromosomal pairs is due in large part to the genomic proximity between these pairs of loci. Distinguishing spatially close sets caused by polymer packing from otherwise more interesting close sets such as fragments involving long range loops is not straightforward.

Finally, inter-chromosomal and intra-chromosomal interaction frequencies are not on the same scale: intra-chromosomal interactions have a much higher expected frequency than inter-chromosomal interactions. It is thus necessary to place inter-chromosomal interactions and intra-chromosomal interactions on the same scale, and to place intra-chromosomal interactions with different genomic distances on the same scale. One such approach is to set q-values as the edge weights, where a null model has already taken genomic proximity into consideration. Another approach is to set the edge weight as z-scores conditioned on different genomic distances [134]. Experiments run on the current test sets do not provide evidence that these approaches perform well in estimating spatial enrichment. Methods of averaged shortest path and average maximum flow only achieve a true positive rate of $< 10\%$ when the edge weights are either q-values or z-scores. It is possible that these approaches are too conservative and are not sensitive enough to detect spatially enrichment. It is also possible that since the embedding is computed on the constraints based on frequencies (mapped to distances), not on z-scores or q-values, setting edge weights as raw frequencies will thus perform better by construction.

We thus construct the 3C graph by only including inter-chromosomal interactions. Good estimations are achieved in sets containing fragments from a variety of chromosomes. As mentioned, such regions often locate near centromeres (Figure 2.4). These regions are known to be spatially compact and clustered around the spindle pole body (SPB) with multiple chromosomes [188]. The fact that our tests perform well in these cases helps to validate that the sets we identify as spatially close correspond to truly spatially close sets.

Table 2.2: P-values (before Bonferroni correction) of different methods on the yeast feature sets from Duan et al. [47] Numbers marked in red bold are p-values considered significant. Asterisks after the numbers indicate significance after Bonferroni correction.

features	edge-fraction	mean shortest path	mean flow	unweighted maximum density subgraph	weighted maximum density subgraph
centromeres	**0.00E+00***	**0.00E+00***	**2.67E-02**	**0.00E+00***	**0.00E+00***
telomeres all	**1.23E-02**	8.56E-01	1.00E+00	9.97E-01	9.72E-01
early firing CIB5-independent origins	**0.00E+00***	**0.00E+00***	**0.00E+00***	**0.00E+00***	**2.00E-03***
late firing CIB5-dependent origins	1.66E-01	7.72E-01	8.42E-01	3.38E-01	6.30E-01
early firing Rad53-regulated origins	**3.80E-03***	**1.20E-02**	5.40E-02	**2.00E-03***	**4.00E-03***
late firing Rad53-regulated orgins	4.04E-01	5.48E-01	6.96E-01	6.90E-01	8.40E-01
breakpoints (Scer)	**0.00E+00***	**0.00E+00***	**0.00E+00***	**0.00E+00***	**2.00E-03***
breakpoints (Scer and Kwal)	**3.80E-02**	1.2E-01	5.40E-02	**4.00E-03***	**3.20E-02**
trnas	**2.00E-03***	**1.60E-02**	**2.20E-02**	**1.00E-02**	**4.00E-02**
trna cluster bright	**0.00E+00***	**0.00E+00***	**0.00E+00***	**0.00E+00***	**0.00E+00***
trna cluster dim	**0.00E+00***	**6.00E-03**	**1.00E-02**	**3.40E-02**	2.34E-01
trna cluster other	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E00

2.6.6 Evaluation of the spatial closeness of various yeast feature sets

For the yeast feature sets [47], most statistics tested here agree with the results presented by Witten and Noble [177]. The edge-fraction method using only the inter-chromosomal interaction data is equivalent to the resampling method proposed by Witten and Noble [177]. Before Bonferroni correction, this test finds that all tested features except the two sets of late-firing origins (late firing CIB5-dependent origins, late firing Rad53-regulated origins) and the tRNA outside two clusters (trna cluster other) are statistically co-located. Among all the methods tested, the non-weighted maximum density approach agreed with the edge-fraction approach in the most instances (Table 2.2). This is not surprising as these are the methods that ignore edge weights. On the other hand, the mean flow statistic is more conservative.

Witten and Noble [177] correctly identified the telomere set as not spatially close only after Bonferroni correction. In contrast, the other methods do not rely on multiple hypothesis correction to get the correct answer, and the p-values from the other methods are all close to 1.0. Telomeres tend to form five to eight foci inside the nucleus during interphase [47]. However, most interactions within the telomere set are low-frequency interactions. Thus edges of the subgraph of the telomere set are long-distance edges, and the distribution of the pairwise shortest paths is not significantly smaller when compared to a random set (Figure 2.7). The fact that

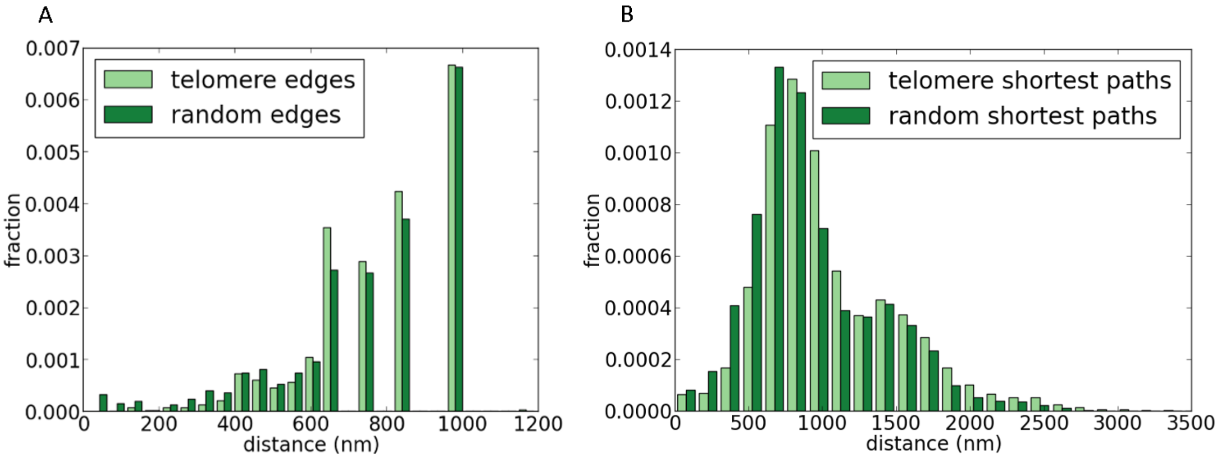


Figure 2.7: (A) Edge lengths and (B) pairwise shortest-path distribution between points in the telomeres set. Interactions among telomeres are overall low-frequency (long-distance) interactions and the distribution of the neither $d(e)$ nor shortest paths are significantly different from a randomly generated set.

$f_{\text{edge\_fraction}}$ cannot exploit edge weights may have lead to a false indication that this feature is statistically significantly colocalized.

tRNAs are also observed to have clustering behavior in the nucleolus [47]. Duan et al. [47] found two clusters of tRNAs with 3C interaction data: one colocalized with centromeres (trna cluster bright), and the other colocalized with rDNAs (trna cluster dim). Although the trna cluster dim is considered significantly spatially enriched by the method of edge-fraction, the weighted maximum density subgraph does not identify it as spatially close. Again, it is plausible that they are not significantly colocalized. The interactions between points in this set are of lower frequency and thus the cluster appears to be 'dim' in the heat map. Taking into account the edge frequencies, as in the weighted maximum density subgraph approach, might lead to a more accurate estimate.

2.7 Discussion

We proposed several novel topological properties as proxies for testing for spatially compact regions of chromatin. These methods avoid the costly process of computing a 3D embedding. The shortest path and the maximum flow approach implicitly apply inferred information from the 3C graph, while the maximum density subgraph approach reduces the effect of outliers. The topological properties we chose here are directly related to spatial proximity of the 3C structures and are easy to compute. Alternative properties of the 3C graph could result in an equally good or better estimation of spatial proximity. One particularly interesting approach for future work is to measure proximity via a diffusion process on the 3C graph, providing robust estimates of node proximity in the graph.

We illustrate that the tested topological properties can be used to infer true spatial proximities in the chromosome structure by first showing in section 2.6.1 that graphical proximities within dense cores are strongly correlated with proximities in their corresponding embeddings. We then show in section 2.6.2 that dense regions found by the maximum density subgraph overlap well with true spatial compact cores when the test sets contains fragments from several chromosomes.

To evaluate the power of the graphical properties for testing spatial enrichment, we first show in section 2.6.3 that all methods result a uniform p-value distribution on the true negative sets and are thus unbiased and valid. We then systematically evaluated the performance of all methods by testing them on both synthetic test sets (section 2.6.4) and yeast feature sets (section 2.6.6). We have shown that Problem 1 (Spatial Proximity Test) can be solved equally well by many statistics based on different topological properties when test sets involve fragments from several chromosomes. We also demonstrate that, under such circumstances, the weighted maximum density method is a good solution to both Problem 1 and Problem 2 (Compact Core Finding) since the cores it finds overlap significantly with synthetically generated cores. The maximum density subgraph solves a slightly different problem from the other methods and previous approaches. It finds the densest subset of a given set of fragments, while the other methods evaluate the given set as a whole.

As mentioned in section 2.5, to validate our methods with true spatial proximity of the chromatin, we use an embedding that is partially computed based on 3C data. This introduces circularity in the validation: Both our methods and the compared ground truth use the knowledge of raw 3C data, which might contribute to the good agreement between the two. However, there is no ground truth of the genome structure that is constructed without 3C data, we therefore chose such an embedding because it is based on both experimental observation (real 3C data) and theoretical knowledge of the genome.

While the framework for using topological properties to infer spatial enrichment is generally effective, the proposed methods cannot accurately evaluate the spatial enrichment in regions that fragments mainly come from the same chromosome due to the absence of the intra-chromosomal interactions when constructing the 3C graph. We discussed in section 2.6.5 the challenge of including intra-chromosomal interactions. A more comprehensive test set that includes single chromosomal examples that cannot be simply explained by genomic proximity is yet to be developed.

Overall, we show that incorporating richer topological features such as flow, shortest path, and maximum density subgraphs provides insight into finding regions that are truly spatially enriched when the 3C graph contains sufficient interactions. These topological features can be efficiently computed using well-known graph algorithms.

Chapter 3

Isoform-level ribosome occupancy estimation guided by transcript abundance with Ribomap

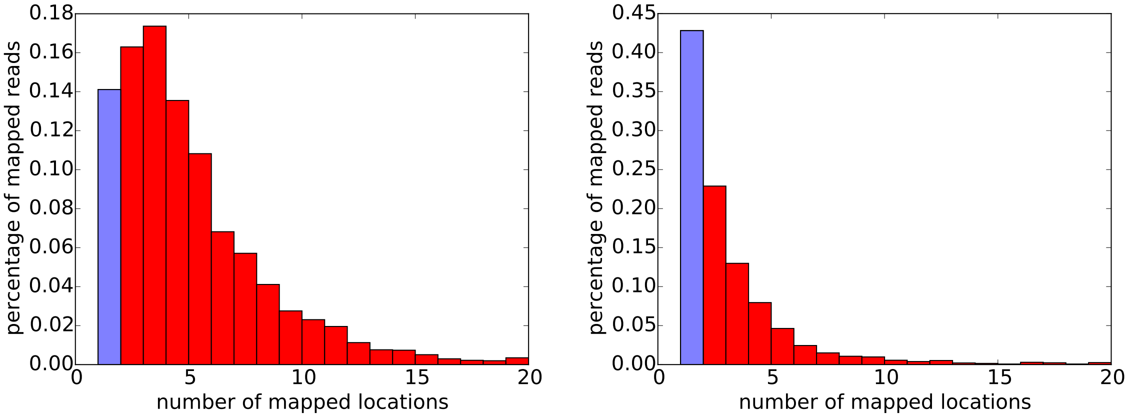
Estimating ribosome profiles is essential to study translational dynamics from ribo-seq data: all downstream analysis depends on an accurate estimation of such profiles. This chapter presents a pipeline to preprocess ribo-seq and output isoform-level ribosome profiles. It tackles one of the biggest problems in ribosome profile estimation: ambiguous mappings. We begin this chapter by quantifying the severeity of ambiguous mappings in ribo-seq data. We then described our solution on handling such multi-mappings. Finally, we validated our methods on both real ribosome profiling data and synthetic data with known ground truth, and show that our method produces accurate estimations of ribosome profiles on a transcript level.

The content of this chapter was originally published in Bioinformatics [171]. It was a joint work with Joel McManus and Carl Kingsford.

3.1 Background

Ribosome profiling (ribo-seq) provides snapshots of the positions of translating ribosomes by sequencing ribosome-protected fragments [79, 84]. The distribution of ribo-seq footprints along a transcript, called the ribosome profile, can be used to analyze translational regulation and discover alternative initiation [58], alternative translation and frameshifting [121], and may eventually lead to a better understanding of the regulation of cell growth, the progression of aging [97] and the development of diseases [73, 163]. Different environmental conditions such as stress or starvation alter the ribosome profile patterns [61, 79], indicating possible changes in translational regulation.

In higher eukaryotes, alternative transcription initiation, pre-mRNA splicing, and 3’ end formation result in the production of multiple isoforms for most genes. The resulting isoforms can have dramatically different effects on mRNA stability [99] and translation regulation [159]. However, to date ribosome profiling analyses have been conducted at the gene, rather than isoform level, due to the absence of necessary bioinformatic tools.



(a) Hela data from [67]

(b) mouse data from [80]

Figure 3.1: Histogram of the number of mappings for all mapped ribo-seq reads in two data sets. Reads are mapped to the human and mouse transcriptome respectively. The proportion of uniquely mapped reads are marked in blue. More than 50% of the reads in both data sets are ambiguously mapped, and ambiguous mappings are extremely common in human reads.

The challenge in estimating isoform ribosome profiles is that a short ribo-seq read may map to many different transcripts. Ambiguous mappings are not rare in ribo-seq data and can be caused by either repetitive sequences along the genome or alternative splicing [78]. For example, in the human Hela cell ribo-seq data (GSM546920, [67]), among all mapped reads (about 50% of all reads), only 14% (Figure 3.1a) can be uniquely mapped to a single location of a single mRNA isoform, 22% can be mapped to multiple regions on the reference genome due to repetitive sequences, and 64% can be mapped to multiple mRNAs due to alternative splicing. Even for a mouse ribo-seq (GSM765301 [80]), where ambiguous mappings are less frequent (Figure 3.1b), 68% of the multimappings are caused by alternative splicing. This indicates that ambiguous mappings are very common in ribo-seq data and are mainly caused by alternative splicing.

Current approaches generally avoid dealing with ambiguous mapping caused by alternative splicing by mapping reads to the genome. The remaining ambiguously mapped reads are either discarded [67, 79, 128, 163] or randomly assigned to one of the candidate regions [115, 144]. Reads are therefore generally mapped to genes by choosing a single 'representative' isoform (e.g. [67]), or by using the union of all possible exons [39, 129, 160]. Although there is a pipeline built to process ribo-seq reads for identifying protein sequences [35], a method proposed to resolve multi-mapping problems caused by repetitive sequences [38], and software developed to align short reads to splice junctions [118], none of these approaches so far handles multi-mapping problems caused by alternative splicing, which, as is shown above, is the major cause of ambiguous mappings.

These unprincipled heuristics can lead to an inaccurate estimations of ribosome profiles. First, only a small portion of reads can be correctly used. Second, regions without estimated ribosome footprints are indistinguishable from read-free regions caused by either discarded or wrongly-assigned multi-mapping reads. Third, randomly assigned reads might cause some regions to have a faulty peak in the profile, while leaving other regions footprint-free (see one

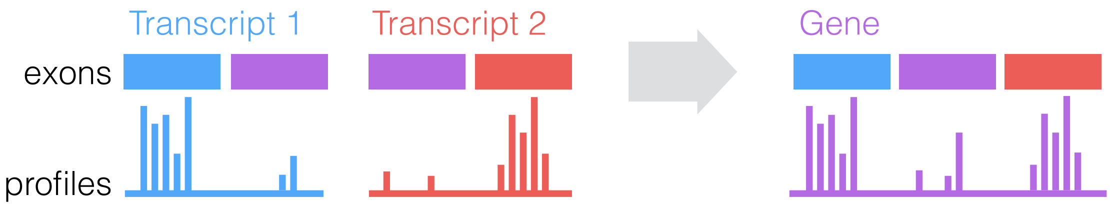


Figure 3.2: An illustration of when estimating ribosome profile on a gene level will fail. One such case is the pileup does not happen in all isoforms, where merging the isoform-level profile will result in multiple peaks being present simultaneously. Another case is that the pileup in an exon location is not significant in any isoforms, but accumulating them together might produce a faulty peak.

example in Section 3.5.2). Without dealing with multi-mappings caused by alternative splicing, these simplistic approaches can only be used to estimate the overall ribosome abundance of a given gene, and are incapable of estimating ribosome profiles of different mRNA isoforms. Further, they can even lead to incorrect gene-level estimates — see Figure 3.2 for some examples.

Another related problem where ambiguous mappings also need to be addressed is the estimation of mRNA isoform abundance from RNA-seq data [89, 124, 130]. However, unlike in RNA-seq, coverage in ribo-seq is highly non-uniform regardless of sequencing bias since ribosomes move along mRNAs at non-uniform rates, and it is in fact the non-uniformities that are of interest [78]. Further, ambiguous mappings are much worse for ribo-seq data since the read length cannot exceed the ribosome size (approximately 30bp), while paired-end and longer reads can be generated from RNA-seq experiments to reduce the problem of ambiguous mappings. Methods developed for transcript abundance are therefore not applicable to assigning ribo-seq reads.

3.2 Contributions

Here, we present a conceptual framework and software (Ribomap) to quantify isoform-level ribosome profiles. By accounting for multi-mapped reads using RNA-seq estimates of isoform abundance, Ribomap produces accurate isoform-specific ribosome profiles. Ribomap is the first approach to estimate ribosome profiles on an isoform-level. It systematically handles both types of ambiguous mappings, does not discard multi-mapped reads, and therefore results in more of the data to be used. Specifically, by observing that ambiguous mappings are mainly caused by multiple isoforms, Ribomap assigns ribo-seq reads to locations using estimated transcript abundance of the candidate locations. On synthetic data, our approach yields a more precise estimation of ribosome profiles compared with a pure mapping-based approach. Further, the ribosome abundance derived using our method correlates better with the transcript abundance on real ribo-seq data. This indicates that Ribomap not only produces better estimates of ribosome profiles, but also results in better ribosome load estimation.

Besides deconvolving multi-mapped reads, Ribomap also addresses other challenges of generating ribosome profiles. For example, Ribomap attempts to select the correct codon location that the P-site maps to, a problem we will return to in the next chapter, with a read-length specific

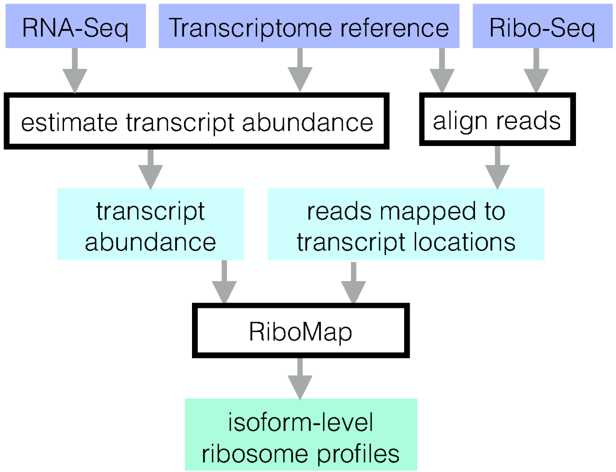

RNA-Seq
Transcriptome reference
Ribo-Seq
estimate transcript abundance
align reads
transcript abundance
reads mapped to transcript locations
RiboMap
isoform-level ribosome profiles

Figure 3.3: Ribomap pipeline for estimating ribosome profiles.

P-site offsets and a prioritized read-assignment heuristic. Ribomap also attempts to correct for sequencing bias by normalizing ribo-seq counts with RNA-seq counts.

Other than isoform-level ribosome profile estimation, Ribomap also outputs various transcript-level estimates from ribo-seq experiments. These estimates include: ribosome loads, translational efficiencies, transcript abundance, and transcripts with the top rank differences between ribosome loads and transcript abundance. In short, Ribomap is the first pipeline that automatically handles several challenges in isoform-level ribosome profile estimation.

3.3 Implementation

3.3.1 Rationale of the Ribomap footprint assignment scheme

Two conditions must hold for a ribosome footprint read to come from a transcript location: First, the transcript has to be present in the cell; second, a ribosome has to be translating the current codon. Therefore, in order to quantify the observed ribosome pileup, we make the following two assumptions: First, identical transcripts have identical translation dynamics; second, each codon location shares a unique pileup behavior affected by how fast the current codon can be elongated and how ribosomes are accumulated in surrounding codons. This means the final observation of the number of ribosome footprints c_{mi} from transcript m at location i is proportional to both the chances of observing the specific transcript codon fragment in the cell α_m (transcript abundance per base) and the chances of a ribosome occupying such a codon location p_{mi} given that the ribosome is from this transcript: $c_{mi} \propto \alpha_m \times p_{mi}$.

The transcript abundance can be estimated from the RNA-seq reads. However, there is no prior knowledge of the per-codon-location specific pileup. We therefore assume the per-codon ribosome abundance is uniformly unbiased. This leads to the assumption that the final ribosome abundance of a transcript location is determined by the transcript abundance: $c_{mi} \propto \alpha_m$.

Intuitively, the transcript abundance is like the outline of the profile, and the location-specific ribosome abundance is like the detail of the profile. Our method tries to grasp the outline of the profile first and then let the read sequences themselves take care of the profile details.

3.3.2 Resolving multi-mappings in ribo-seq data

Ribomap de-convolves multi-mappined reads via three steps (Figure 3.3):

Step I: Transcript abundance estimation Since RNA-seq experiments should always be performed in parallel with ribo-seq [78], the abundance α_t per base of each transcript t can be estimated from the RNA-seq data using Sailfish [132], an ultra-fast mRNA isoform quantification package. Ribomap also accepts transcript abundance estimations from cufflinks [165] and eXpress [143].

Step II: Mapping ribo-seq reads to the reference transcriptome We obtain all the transcript-location pairs L_r where the read sequence r matches the transcript sequence by aligning the entire set of ribo-seq reads R to the transcriptome with STAR [44].

Step III: Ribosome profiling estimation Let c_r be the number of ribo-seq reads with sequence r. Ribomap sets the number of footprints c_{rmi} with sequence r that originate from a specific location i on transcript m to be proportional to the transcript abundance α_m of transcript m: $c_{rmi} = c_r\alpha_m / \sum_{(m',i')\in L_r} \alpha_{m'}$, where the denominator is the total transcript abundance with a sequence matching r. The total number of reads c_{mi} that are assigned to transcript m, location i, is then $c_{mi} = \sum_{r\in R} c_{rmi}$. Essentially c_{mi} makes up the profile for each transcript. The sum is needed here because there can exist multiple read sequences being mapped to the same transcript location due to sequencing errors, so the final estimated ribosome count for a transcript location should be the sum of the estimated count for all matched read sequences.

3.3.3 Ribomap pipeline

Ribomap first preprocesses the raw ribosome profiling reads and the RNA-seq reads by trimming the 3' adapter portion of the reads and discarding contaminated reads. Ribomap ensures the quality of the footprint set by excluding reads that are too short or too long. Ribomap then covert candidate mapping locations to their P-site locations dynamically based on the read length and the start of the mapping location, and resolves multi-mappings and estimates ribosome P-site profiles. Ribomap also corrects for sequencing bias by normalizing a transcript's ribosome profile with its mRNA profile (as done in [13] and [192]). The Ribomap pipeline follows these steps:

Filtering contaminated reads with STAR Reads that can be mapped to rRNA, tRNA, snoRNA, may not be representative of ribosome protection *per se* and, in fact, may be contaminants merely associated with the ribosome. As such, they can be filtered out by mapping the reads to the ribosome RNA sequences and transfer RNA sequences of the organism in study and keeping only unmapped reads for downstream analysis. This is done via the STAR aligner [44]. Detailed commands and specific options to run STAR is described in section 3.7.

Mapping the remaining reads to the transcriptome After removing reads that may have been the result of contamination, the remaining reads are aligned to the transcriptome. This step is also accomplished by STAR. Different options are chosen to handle building the transcriptome index and recording the alignments (details in section 3.7).

Calculating transcript abundance with Sailfish The transcript abundance estimation is used to guide the isoform-level ribosome profile estimation. We use the latest version of Sailfish (called Salmon) that supports the read alignment results as the input.

Estimating Isoform-level ribosome profiles This is the last step of the Ribomap analysis pipeline. It further filters out ribo-seq reads if their lengths is not within a reasonable range of true ribosome footprint size. It also selects the P-site of ribo-seq reads based on read lengths, and it resolves multi-mapping with a heuristic via several iterations.

Specifically, only reads with size between 25 and 36 are kept for estimating the ribosome profiles. Alignments with RC flag set (reads are reverse complemented and then aligned to the transcriptome) are discarded due to the single strandedness of the ribosome profile protocol. Following [80], we assign the P-site of reads with length < 30 to be 12, with length between 31 and 33 to be 13, and with length > 33 to be 14. Only transcripts with abundance greater than zero are considered to be expressed and are included for ribosome profile estimation.

Footprints are assigned to candidate locations with three iterations: First, only candidate locations with a frame-0 P-site mapping are considered; second, the remaining reads are assigned to frame 1 and 2 locations; third, the rest of the unassigned reads — reads that cannot be mapped to any CDS regions, are mapped to UTR regions. For all three iterations, reads are mapped to candidate locations proportional to the transcript abundance if multiple locations are presented for one read.

3.3.4 Output file format

The final output of Ribomap is a vector of read counts per codon position for each transcript isoform. It also outputs sub-codon resolution, nucleotide-level ribosome profiles including the UTR regions, along with the total ribosome loads, translation efficiency, and the relative abundance for each transcript. Moreover, Ribomap reports transcripts in order of the rank difference between the relative transcript abundance and ribosome load, to help identify isoforms with different translation efficiency.

Ribomap outputs five files:

File 1: `XXX.base` gives the sub-codon resolution, nucleotide-level ribosome profiles including the UTR regions. Only transcripts with a non-zero total ribosome count are reported. Each entry of a specific transcript looks like this:

```
refID: 0
tid: YAL001C
ribo profile: 0 0 0 74 68 ...
mRNA profile: 31 35 50 73 87 96 104 ...
normalized ribo profile: 0 0 0 1.0137 0.781609 0.0208333 ...
```

where:

refID is the transcript fai index in the transcript fasta file.

tid is the transcript header name in the transcript fasta file.

ribo profile nucleotide level ribosome profile including the UTR regions.

mRNA profile RNA-seq read coverage profile.

normalized ribo profile is the ribosome profile after bias correction. Each number in the vector is the ratio between the ribo profile count and the mRNA profile count.

File 2: `XXX.codon` gives the in-frame ribosome profiles for each transcript within the CDS region. The file format is the same as the `XXX.base` file.

File 3: `XXX.stats` is the summarized statistics for each transcript. Each entry of a specific transcript looks like this:

```
refID: 0
tid: YAL001C
rabd: 3959
tabd: 0.000209384
te: 1.89078e+07
```

where:

rabd is the total ribosome loads, which is the sum of the `ribo profile` vector in `XXX.base`.

tabd is the relative transcript abundance from Sailfish's result.

te is the relative translational efficiency, which is the ratio between `rabd` and `tabd`.

File 4: `XXX_abundant.list` gives a list of transcripts whose total ribosome abundance is more than expected given the transcript abundance. Such a list explores the difference between the ribosome abundance ranking and the transcript abundance ranking. A higher rank of total ribosome footprint count compared to the transcript abundance might indicate that the transcript is more highly packed with ribosomes and this might suggest that there is a translational amplification regulation for this transcript. Each line in the file looks like this:

```
ENST00000340756.2 2.81302e-09 1046.59 0 96 -96
```

It is the transcript header name followed by several statistics, in the following order:

1. relative transcript abundance,
2. total ribosome footprint count,
3. the percentile ranking of the transcript abundance,
4. the percentile ranking of the total ribosome loads (transcripts with zero total ribosome loads are excluded from the analysis),
5. difference between the transcript abundance rank and the total ribosome footprint count rank.

The list of entries in this file is ordered by the ranking difference between the transcript abundance and the ribosome abundance. Only transcripts with ranking differences greater than 10 are listed.

File 5: `XXX_scarce.list` gives a list of transcripts whose total ribosome abundance is less than expected given the transcript abundance. Analogous to the abundant list, this list include transcripts that might have a translational buffering regulation. The file format is the same as `XXX_abundant.list`.

3.4 Validation

3.4.1 Generating synthetic footprints

To evaluate the performance of the ribosome footprint read assignment, we generated synthetic footprint data with known ribosome profiles. Let N be the total number of transcripts, α_m be the per-base relative abundance of a transcript m, and p_{mi} be the ribosome occupancy probability of a location i on transcript m. The number of synthetic footprints from location i on transcript m is set to be $N \times \alpha_m \times p_{mi}$, where N is set so that a total of 20,000,000 footprints from all locations of all transcripts are generated. How α_m and p_{mi} are set is described below.

Ribosome occupancy probability

We model the movement of ribosomes on a given transcript with a ribosome flow model based on a total asymmetric exclusion process (TASEP) [141]. In this process, each mRNA is modeled as a sequence of codons, with ribosomes moving along it with codon-specific elongation rates. The model is “asymmetric” because the ribosome can only move from the $5'$ end to the $3'$ end of the mRNA. The model has “exclusion” because a location can only be occupied by one ribosome at a time; a ribosome can only move from location i to the next one when it is currently at location i and the next location is not occupied by another ribosome.

For a mRNA m of length n, the entire translation process is modeled in three steps: First, the ribosome binds to the the start codon of the mRNA with initiation rate λ_{m0}. Second, the ribosome moves along the mRNA from codon c_i to c_{i+1} with a elongation rate λ_{mi} $(i = 1 \ldots n-1)$. Third, the ribosome terminates when it reaches the the stop codon c_n with a termination rate λ_{mn}, and the whole peptide chain of the protein will be formed. The λ vector thus describes the transition rates of the ribosome moving from one location to another.

According to TASEP, for a given a transcript m, that the change of the probability of observing a ribosome on location i (p_{mi}) is the difference between the incoming and the outgoing flow of the ribosome, modeled using the differential equations:

$$\frac{dp_{m1}(t)}{dt} = \lambda_{m0}[1 - p_{m1}(t)] - \lambda_{m1}p_{m1}[1 - p_{m2}(t)] \tag{3.1}$$

$$\frac{dp_{mi}(t)}{dt} = \lambda_{m,i-1}p_{m,i-1}(t)[1 - p_{mi}(t)] - \lambda_{mi}p_{mi}(t)[1 - p_{m,i+1}(t)] \quad 1 < i < n, \tag{3.2}$$

$$\frac{dp_{mn}(t)}{dt} = \lambda_{m,n-1}p_{m,n-1}(t)[1 - p_{m,n}(t)] - \lambda_{mn}p_{mn}(t) \tag{3.3}$$

Equation (3.1) and (3.3) describe the boundary case of the process, and equation (3.2) describes the intermediate case.

The ribosome profile under this model is the ribosome occupancy probability distribution when the steady state of the model is reached, during which the probability of observing a ribosome at any location will not change over time. This stationary distribution probability can be solved by setting the left hand side of the equations above to zero.

To generate synthetic profiles, we solve the above equations to find the steady state for all transcripts. This provides the probabilities p_{mi} that a read is drawn from location i on one copy of transcript m. Selection of the elongation rates (λ_{mi}) and initiation rates (λ_{m0}) is described below.

Elongation rate Following Reuveni et al. [141], we assume the elongation rate is codon-specific and is proportional to the tRNA abundance of a given codon. The tRNA gene copy number used as an approximation to the tRNA abundance in the cell, and the elongation rate is set to be the absolute adaptiveness value (W_i) of a given codon i:

$$W_i = \sum_{j=1}^{n_i} (1 - S_{ij}) t_{CGN_{ij}}. \tag{3.4}$$

Here n_i is the number of tRNA isoacceptors recognizing codon i. $t_{CGN_{ij}}$ is the gene copy number of the jth tRNA that recognizes the ith codon, and S_{ij} is the selective constraint on the efficiency of the codon-anticodon coupling by considering the wobble paring [45]. The absolute adaptiveness of a given codon gives an estimate for the total amount of tRNAs that can match to a specific codon. The larger the adaptiveness value is, the more efficient a codon can be translated, and thus the faster the elongation rate will be. The elongation rates range between 3.5 and 33 time steps from this calculation.

Initiation rate The initiation rate is thought to be the rate limiting step during the translation process [79]. We set the initiation rate range to be approximately 100 times smaller than the elongation rate, as in [150], by uniformly sampling the rates for each transcript between 0.03 and 0.3 timesteps.

Transcript abundance

The transcript abundance is estimated from the RNA-seq data paired with the ribosome profiling experiments (GSM546921) via Sailfish [132]. Only transcripts with TPM value (transcripts per million) greater than 1 are included, which results in a total of 39,414 transcripts. The relative transcript abundance per base α_m is computed as follows: Let t_m be the transcript abundance estimated from Sailfish for transcript m, let l_m be the transcript length, let T be the transcriptome, $\alpha_m = t_m / \sum_{m' \in T} t_{m'} \times l_{m'}$.

The rationale of using transcript abundance per base for assigning the reads is that each transcript codon location can be seen as a type of ball with a unique color, and the probability of observing such a transcript codon fragment in the cell can be compared as randomly picking a ball of a specific color from a pool of balls. We assume codon locations on the same transcript have equal visibilities. For the extreme case where there is only one type of transcript in the cell with l_m bases, the probability of seeing any codon is $1/l_m$. For the case where there are more than one type of transcript, if there are t_m transcripts with type m, then there will be t_m copies of each codon positions, and the probability of seeing a specific codon position on transcript m over all possible codon positions is: $t_m / \sum_{m' \in T} t_{m'} \times l_{m'}$.

Ribosome footprint generation

We calculate α_m and p_{mi} for every transcript as described above and then sample reads by selecting a random transcript with probability proportional to its α_m and then selecting ribosome position i proportional to p_{mi}. We then extract a 30-bp read with the 12^{th} position set to the ribosome P-site.

Introducing sequencing errors in the read set

We estimate the frequency of sequencing errors from ribosome profiling data (GSM546920 [67]) to be the total number of mismatches of read assignments in the data over the total number of aligned bases (total number of aligned reads × read length). The error rate estimated on the ribosome profiling data GSM546920 is 0.5%.

To add simulated sequencing errors to our simulated reads, we apply a Poisson process with the rate parameter set to the error rate for mutating the bases in the synthetic read set. We tried in total three different levels of error rates: 0.5%, 1%, and 2%.

3.4.2 RNA-seq data generation

We synthetically generated RNA-seq reads with the following procedure: The transcript abundance estimation from Section 3.4.1 is used as a prior to generate 20,000,000 synthetic RNA-seq fragments with default parameters using rlsim [152]. Consistent with the ribosome footprint generation settings, only transcripts with TPM value (transcripts per million) greater than 1 are included. We then truncate the fragments into 36bp-long single-end reads.

When applying Ribomap, we re-estimate the transcript abundance as usual from the synthetic RNA-seq reads via Sailfish [132] without looking at the original estimates from the true data.

The analysis shown in Figure 3.4 only uses transcripts with non-zero total ribosome counts, which results in 26,297 transcripts included.

3.4.3 Settings of Star prime

We include a baseline read assignment approach as comparison. This approach only maps reads to one candidate location, we call it *Star prime*. By default, STAR only marks one alignment for multi-mapping reads as primary (FLAG 0x100 unset), such an alignment “is randomly selected from the alignments of equal quality.” And this primary alignment is used to estimate ribosome profiles in Star prime. In addition, no prior transcript abundance knowledge is used in Star prime, and therefore all transcripts in the transcriptome are taken into account. All other settings (i.e. contaminated read filtering, adapter clipping, read size selection, dynamic P-site assignment) are kept the same as the Ribomap pipeline described above.

Table 3.1: Data used in Ribomap validation

Name	**Description**
Hela ribosome footprint data [67]	Human Hela cell ribo-seq mock 32hr runs1-2 (GSM546920).
Hela RNA-seq data [67]	Human Hela cell RNA-seq mock 32hr runs1-3 (GSM546921).
Human Transcriptome reference fasta [3]	The human transcriptome reference is downloaded from the Gencode ftp server. The CDS region information used in our analysis is obtained from the headers of the fasta sequence entries.
Human Transcriptome gene annotation gtf [4]	The human annotations are also downloaded from the Gencode ftp server. This file is used to obtain the frame information of the CDS regions of the transcripts.
Mouse Ribosome footprint data [80]	ES cell feeder-free, w/ LIF 60 s CYH (100 ug/ml) ribo_mesc_yeslif Illumina GAII (GSM765301).
Mouse RNA-seq data [80]	ES cell feeder-free, w/ LIF 60 s CYH (100 ug/ml) mrna_mesc_yeslif Illumina GAII (GSM765289).
Mouse Transcriptome reference fasta [5]	The mouse transcriptome reference is downloaded from the Gencode ftp server. The CDS region information used in our analysis is obtained from the headers of the fasta sequence entries.
Mouse Transcriptome gene annotation gtf [6]	Mouse annotations are also downloaded from the Gencode ftp server. This file is used to obtain the frame information of the CDS regions of the transcripts.
Human tRNA gene copy number [1]	tRNA gene copy numbers are obtained from the gtrna database [26].
Contaminated sequences	Both ribosomal sequences and tRNA sequences are included in the contaminated sequences. Ribosomal sequences are downloaded from Ensemble's [53] ncRNA database with `gene_biotype` as `rRNA`.
tRNA sequences [2]	tRNA sequences are downloaded from the gtrna database [26].

3.4.4 Data used in this study

The ribo-seq and RNA-seq data sets used in the study can be found in Table 3.1, along with the origin of the contaminant and transcriptome references.

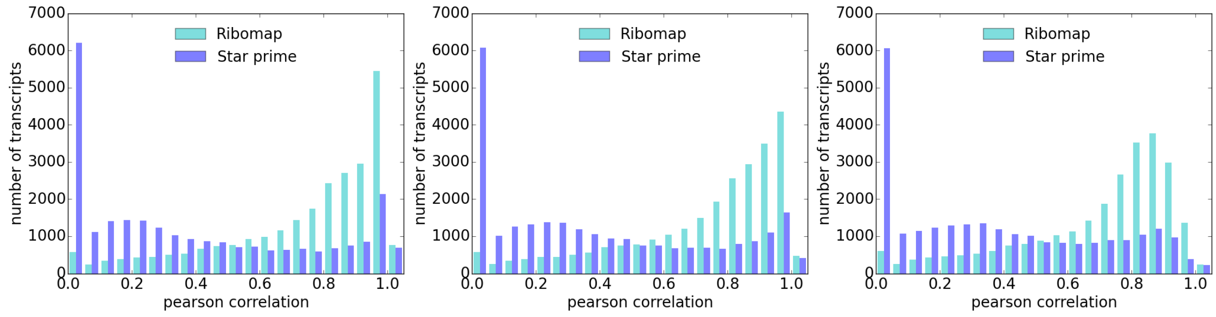
3.5 Results and discussion

3.5.1 Performance on synthetic ribo-seq data with known ground truth

To evaluate the performance of Ribomap, we synthetically generated ribo-seq reads with known ground truth profiles using transcript abundance of GSM546921 RNA-seq data [67] and a dynamic range of initiation rates. Ribosome occupancy probabilities for locations on a given transcript were simulated using the ribosome flow model [141]. Errors were added to the reads using a Poisson process with a rate of 0.5%, which was estimated from the ribo-seq data GSM546920 [67]. For comparison, we also test a naïve approach, called *"Star prime"*, that maps each read to a single candidate location. More details about the synthetic data generation and settings of Star prime can be found in Section 3.4.

The Pearson correlation coefficients between Ribomap's ribosome profiles and the ground truth is significantly higher than that of Star prime (Figure 3.4a): 81% of our profiles have a higher Pearson correlation (Mann–Whitney U test $p < 3 \times 10^{308}$) and 68% have a smaller root mean square error (Mann–Whitney U test $p = 3.3 \times 10^{221}$). This suggests that Ribomap more accurately recovers the ribosome profiles than the standard mapping procedure applied to isoforms.

We also test our method on synthetic data with different error rates (Figure 3.4). The Pearson correlation coefficients between Ribomap's estimated profiles and the ground truth are consistently higher than that of Star prime for all tested error rates (Mann-Whitney-U p $< 3 \times 10^{308}$). The distributions of the Pearson correlations are qualitatively the same among different error rates (Figure 3.4), and the median pearson correlation coefficients from Ribomap are significantly higher than Star prime (Table 3.2). It is worth noting that the spike at 0 of Star prime is due to STAR not assigning footprints to transcripts that are estimated to be present. Together,



(a) error rate = 0.5%

(b) error rate = 1%

(c) error rate = 2%

Figure 3.4: Histogram of the Pearson correlation between the footprint assignments and the ground truth ribosome profiles on read sets with different error rates.

Table 3.2: Median Pearson correlation between estimated profiles and ground truth profiles

	error rate = 0.5%	error rate=1%	error rate = 2 %
Ribomap	0.83	0.81	0.78
Star prime	0.28	0.30	0.31

these results show that Ribomap robustly outperform Star prime under various realistic error rates.

Admittedly both our methods and synthetic data make the same assumption that transcript abundance play a role in ribosome occupancy, which makes such a validation circular. However, to generate ribosome profiles with a known ground truth, a model is needed. Unfortunately there is no existing model for ribosome footprints. Therefore, we proposed a model for generating ribo-seq reads. Our model takes into account both transcript abundance and ribosome motion, and we believe such assumptions are reasonable.

3.5.2 Ribomap generates more reasonable profiles

We test our ribomap pipeline on a Human Hela cell ribosome profiling data (GSM546920 [67]). Figure 3.5 shows one example of the isoform-level ribosome profile estimated from both Ribomap and Star Prime. This is gene RPL37A (ENSG00000197756.5). It is a ribosomal protein, and is part of the 60S unit. As shown in Figure 3.5, it has 6 isoforms, which are sorted by their transcript abundance levels. While Ribomap maps the footprints to the more abundant transcripts, Star prime assigns footprint reads to a single candidate location at random, leaving two of the expressed transcripts almost ribosome-free (Figure 3.5A, C), two of the unexpressed transcripts with the highest ribosome loads (Figure 3.5E, F), and a surprisingly huge pile-up at the end of transcript ENST00000491306, inconsistent with its surrounding ribosome pile-ups (Figure 3.5D). Therefore, it is likely that Ribomap produces a more reasonable profile estimation.

3.5.3 Ribomap results in better ribosome load estimations

Ribomap's better read assignment mechanism results in a more precise estimation of the per-mRNA ribosome profiles. These better estimated profiles can also lead to a better estimation of the total ribosome loads on a transcript. For example, Ribomap's ribosome loads estimated from the Hela cell data (GSM546920, [67]) correlates well with the estimated transcript abundance (Pearson $r = 0.71$). We do not expect a perfect correlation due to isoform-specific translational regulation. On the other hand, the pure mapping-based approach of Star prime does not correlate as well ($r = 0.28$). Even for the case of the mouse data from [80], where approximately only 50% of the reads have multi-mappings, the estimated ribosome loads correlate better to the estimated transcript abundance (Pearson r=0.56) than the naïve Star prime approach (Pearson r=0.45).

It is not surprising that Ribomap improves the correlation between ribosome loads and transcript abundance: Ribomap assigns multi-mapping reads to candidate locations proportional to

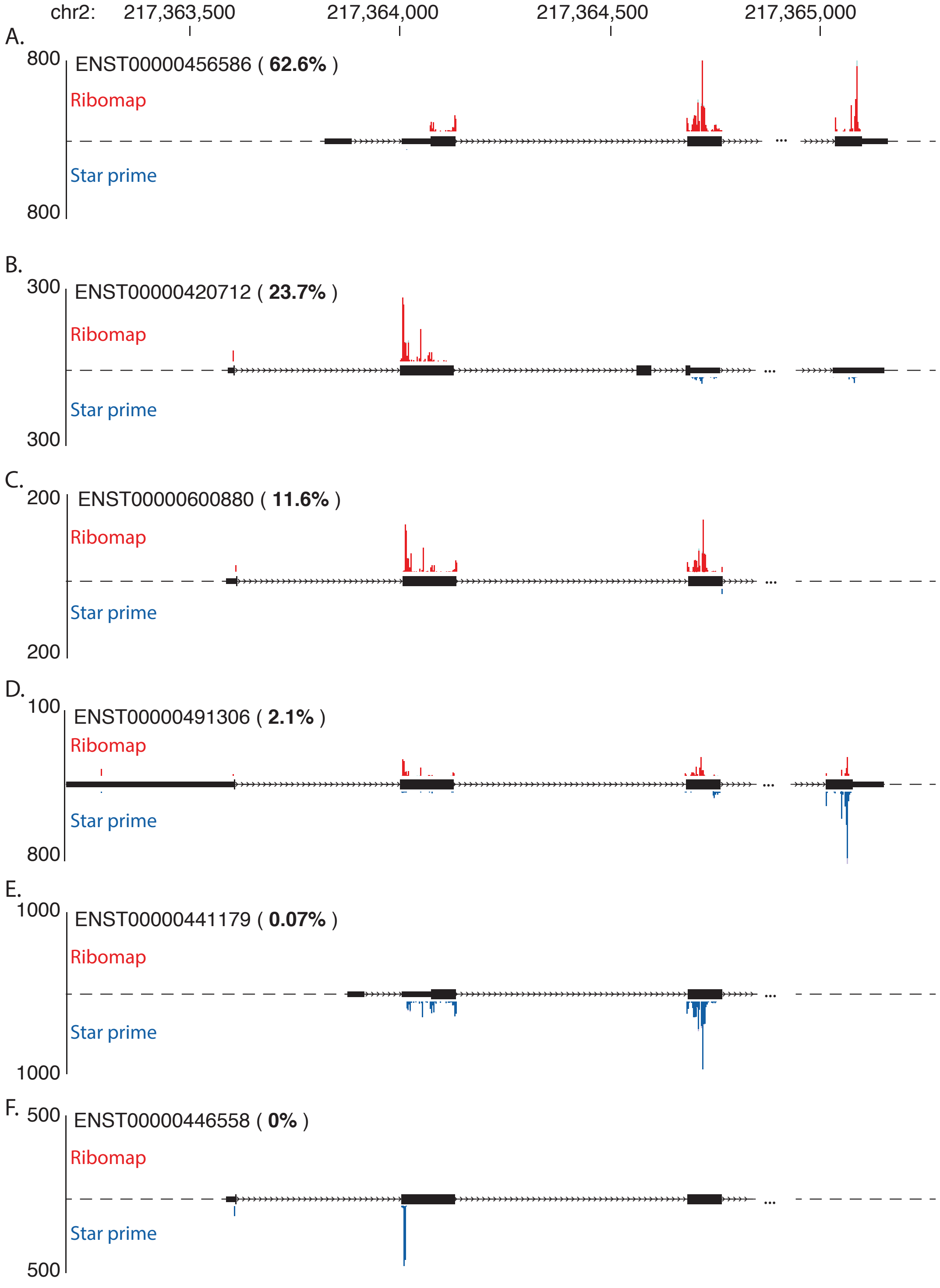


Figure 3.5: An example of isoform level ribosome profiles for gene RPL37A. Both Ribomap and Star Prime profiles are shown. Isoforms are listed in order of their relative abundance estimated from Sailfish (abundance values next to transcript IDs). Profile plot is shown in UCSC Genome Browser [94].

their transcript abundance. If all reads are uniquely mapped, there would be no difference between Ribomap and Star prime. Under such case, the imperfect correlation between ribosome loads and transcript abundance is entirely caused by translational regulation. On the other hand, if all reads are multi-mappings caused by alternative splicing, we should expect a better correlation between transcript abundance and ribosome-loads from Ribomap. In fact, if there is no translational regulation, we should expect a perfect correlation between ribosome loads and transcript abundance under such case. The current correlation result for Star prime is therefore a combination of ambiguous mapping and translational regulation. We believe that Ribomap at least partially resolves the problem of ambiguous mapping, and produces a better estimation of the isoform level ribosome profiles. Consequently, these profiles can lead to a better understanding of translational regulation.

3.5.4 Running time and memory usage of Ribomap

Ribomap runs for about 15 minutes on the ribo-seq data GSM546920 with 18 million reads on 15 threads. The running time includes the time to build the STAR index for both the contaminated sequences and the transcriptome, filtering the contaminated reads and aligning the remaining reads to the transcriptome for both RNA-seq and ribo-seq data, transcript abundance estimation, and estimating isoform level ribosome profiles. Memory usage is about 8.6G.

3.6 Conclusion

Quantifying ribosome occupancy correctly is the first step of analyzing the ribo-seq data. Measurements of protein translation efficiencies [67], ribosome loads [79], pileups and stalling [80] are all derived from ribosome profiles. The challenges of generating such a profile includes deconvolving multi-mapped reads, selecting the correct codon location that the P-site maps to, and bias correction [78]. While none of these issues has a standard protocol, we describe Ribomap, an automatic pipeline that addresses the above challenges and outputs isoform-level ribosome profiles and other ribo-seq analyses. Through two lines of evidence, on real and synthetic ribo-seq data, we show that Ribomap produces useful, high-quality ribosome profiles along individual isoforms. It can serve as a useful first step for downstream analysis of translational regulation from ribo-seq data.

3.7 Appendix: Specific commands for running Ribomap

The command to run Ribomap is:

```
run_ribomap.sh --rnaseq_fq rnaseq.fq.gz --riboseq_fq riboseq.fq.gz
--contaminant_fa contaminant.fa --transcript_fa transcript.fa
--cds_range cds_range.txt
```

Ribomap uses state-of-the-art read-processing tools for several of its steps. We below list in detail each step of the Ribomap pipeline and the command for executing them, in case the user wants to skip some intermediate steps or make adjustments on individual steps.

3.7.1 Contaminated reads filtering

The first step, which must be done only once per organism, is to create a STAR index for the contaminant RNA sequences with the following command, assuming the sequences of the unwanted molecules are in the file *contamination.fa*:

```
STAR --runThreadN nproc --runMode genomeGenerate
--genomeDir rrna_idx --genomeFastaFiles contamination.fa
--genomeSAindexNbases 5 --genomeChrBinNbits 11
```

nproc specifies the number of threads to run STAR. Since STAR treats every sequence entry in the reference FASTA as a 'genome', the option `--genomeSAindexNbases 5` forces STAR to build the index properly for small molecule sequences. We include both tRNA and rRNA sequences as contaminants in our analysis.

Once this index is created, it can be used to filter the contaminated reads as follows, assuming the zipped FASTQ file of raw sequencing reads is in *riboseq.fq.gz*:

```
STAR --runThreadN nproc --genomeDir rrna_idx
--readFilesIn riboseq.fq.gz
--readFilesCommand zcat --outFileNamePrefix riboaligned
--outStd SAM --outReadsUnmapped Fastx --outSAMmode NoQS
--clip3pAdapterSeq adapter --seedSearchLmax 10
--outFilterMultimapScoreRange 0 --outFilterMultimapNmax 255
--outFilterMismatchNmax nmismatch
--outFilterIntronMotifs RemoveNoncanonical > /dev/null
```

where *adpater* is the adapter sequence (`TCGTATGCCGTCTTCTGCTTG` for the Hela data set, and `CTGTAGGCACCATCAATTCGTATGCCGTCTTCTGCTTGAA` for the mouse data set), and *nmismatch* = 1 is the number of allowed mismatches for the alignment. The option `--outReadsUnmapped Fastx` causes STAR to output the unmapped reads to a FASTA file called *riboaligned*`Unmapped.out.mate1`; `--seedSearchLmax 10` increases the sensitivity of STAR for aligning short reads; `--outFilterMultimapScoreRange 0` guarantees that only the alignments with the best scores are reported; `--outFilterMultimapNmax 255` filters out reads that can be mapped to more than 255 locations. STAR by default clips off read ends if a better local alignment score can be achieved. Such a procedure, called 'soft clipping', is very useful for handling ribo-seq reads since the first couple of bases in the 5' end are likely to be contaminated, and an adapter sequence is usually attached to the 3' end.

3.7.2 Read mapping

After removing reads that may have been the result of contamination, the remaining reads are now ready to be aligned to the transcriptome. This step is also accomplished by STAR. Assuming the transcriptome sequences are in *transcript.fa*, the command to generate the transcriptome index is:

```
STAR --runThreadN nproc --runMode genomeGenerate
--genomeDir transcript_idx --genomeFastaFiles transcript.fa
--genomeSAindexNbases 11 --genomeChrBinNbits 12
```

Again, `--genomeSAindexNbases 11` insures that the index is built properly for shorter molecules (compared to chromosomes), and `--genomeChrBinNbits 12` reduces the memory consumption when many reference sequences are provided.

The command to align the reads to the transcriptome is:

```
STAR --runThreadN nproc --genomeDir transcript_idx
--readFilesIn riboalignedUnmapped.out.mate1
--outFileNamePrefix riboaligned_transcript
--clip3pAdapterSeq adapter --seedSearchLmax 10
--outFilterMultimapScoreRange 0 --outFilterMultimapNmax 255
--outFilterMismatchNmax nmismatch
--outFilterIntronMotifs RemoveNoncanonical
--outSAMtype BAM Unsorted --outSAMmode NoQS
--outSAMattributes NH NM
```

The aligned reads will be stored in *riboaligned_transcript*`Aligned.out.bam`. It includes two SAM attributes for each alignment record: `NH` is the number of reported alignments for the read, and `NM` is the number of mismatches in the current alignment.

3.7.3 Transcript abundance estimation

Assuming that the read alignments of the RNA-seq data are in `rnaaligned_transcriptAligned.out.bam`, the command to perform the transcript abundance estimation is:

```
salmon quant -t transcript.fa -l SF -a rnaaligned_transcriptAlig-
ned.out.bam
-o sm_quant -p nproc --bias_correct
```

The flag `--bias_correct` allows Salmon to correct for sequencing biases in the RNA-seq reads. The transcript abundance estimation is in `sm_quant/quant_bias_corrected.sf`.

3.7.4 Isoform-level ribosome profile estimation

This is the last step of the Ribomap analysis pipeline, and it is automatically handled by an executable (developed by us) called `riboprof`. It takes in the transcriptome fasta file, a CDS range file, the ribo-seq and RNA-seq alignment bam files (produced above), the transcript abundance estimation file (produced as above; or it also supports abundance estimations from eXpress [143] or Cufflinks [165] if those estimates are preferred and available). The CDS range file (assume its name is `cds_range.txt`) gives the coding region for each transcript. The command to perform an isoform-level ribosome profile estimation is:

```
riboprof --fasta transcript.fa --cds_range cds_range.txt
--mrnabam rnaaligned_transcriptAligned.out.bam
--ribobam riboaligned_transcriptAligned.out.bam
--min_fplen min_fplen --max_fplen max_fplen --offset offset.txt
--sf sm_quant/quant_bias_corrected.sf --tabd_cutoff tabd_cutoff
--out ribomap_out
```

As described before, *min_fplen*=25, *max_fplen*=36, and *tabd_cutoff*=0. *offset.txt* provides the P-site offset of a read given the read length, and the estimated ribosome profiles and other analysis will be written to the directory *ribomap_out*.

More information about the options and input file formats for Ribomap can be found in the README file in Ribomap's Github page (`https://github.com/Kingsford-Group/ribomap/blob/master/README.md`).

Chapter 4

Accurate recovery of ribosome positions reveals slow translation of wobble-pairing codons in yeast

In this chapter, we address another challenge in ribosome profile estimation: identifying the active translating sites. This chapter describes a novel algorithm and mathematical model to recover the ribosome A-site positions from high-coverage ribosome profiling reads. We begin the chapter by explaining the cause of the challenge of identifying A-sites from ribo-seq data: imperfect digestion. We then provide a detailed model to handle imperfect digestions and recover A-site positions from ribo-seq data. Finally, we show that our method produces better ribosome profile estimations. Using these refined profiles, we observe that both tRNA concentration and wobble pairing play roles in translation speed regulation in yeast.

The content of this chapter was originally presented at RECOMB in 2016, and was published in Research in Computational Molecular Biology [170]. It was a joint work Joel McManus and Carl Kingsford.

4.1 Background

4.1.1 Identifying the active site from ribo-seq reads is a necessary first step

Ribosome profiling is an important sequencing technique that enables various genome-wide translational studies, including on translational response to stress [61, 79, 167], protein synthesis rate [105], alternative translation initiation [57, 101], translation evolution [12, 115], cell development [22, 157], and the role of specific translation regulation factors [67, 69, 178]. The experiment extracts mRNA fragments protected by bound ribosomes (also called ribosome footprints) from RNase I digestion [79]. The technique is analogous to taking snapshots of ribosome locations during translation. Therefore the ribosome footprint counts at codon locations should be related to the elongation time [78, 80]. The vector of footprint counts at codon locations of a mRNA is called a ribosome profile, and each individual count is called a ribosome pileup. To date, ribosome profiles are generally used to qualitatively visualize ribosome pauses (e.g.

[69, 80]), translation initiation, and translation termination (e.g. [8, 49]). Yet attempts to quantify translation speed, even from the same experiment, often result in controversial conclusions on the determinants of translation rate [13].

One of the challenges in translation speed quantification is accurate measurement of ribosome decoding locations. Currently, there is no method to extract the precise ribosome decoding locations when the snapshots are taken [13, 112]. The ribosome P-site or A-site is usually considered the active decoding site [13, 59, 80, 98, 112, 121, 135, 156, 178]. This is because the A-site is where the aminoacyl-tRNA enters the ribosome, and the P-site is the position of peptide bond formation. Only the location of either the P-site or the A-site needs to be estimated from the experiment data, and the other one can be inferred.

4.1.2 Imperfect digestions complicate the view of ribosome pileups

Ideally, we expect the read length of ribo-seq data to be identical. For example, the typical ribosome footprint size for yeast is about 28 nt [79], so we would expect most of ribo-seq in yeast experiments to be 28 bases long. In addition, since ribosomes move in units of codons (3 nt), we expect the ribosome pileups to have a 3-nt periodicity, with most of the reads concentrated on a single base (an in-frame location).

However, ribosome footprint reads do not always share a typical length. For example, in a recently published yeast ribosome profiling data with deep coverage (GSM1335348 [8]), fewer than 60% of uniquely mapped reads have length 28. Further, ribosome footprint reads are not always highly concentrated on a single reading frame. For the experiment above, in order to examine frame distributions by read lengths, we group reads by their lengths, and tallied them on each location from all transcripts – we called these read count vectors meta profiles. Figure 4.1 shows the meta profiles on different read lengths: For each profile, reads are divided into 3 frames corresponding to the $1^{st}, 2^{nd}$ and 3^{rd} nucleotide of codons, and read length and the distribution of these frames are noted above each profile. We observe although 96% of the reads with length 28 are skewed towards one frame, the highest frame portion for reads with a length not equal to 28 vary from 60% to 80%. While it is possible for reads from the other two frames to be caused by frameshifts, frameshifts are generally rare and thus cannot explain all of the off-frame reads.

In fact, reads with various lengths show unique shapes of 3-nt periodicity. That is, for each of the 3 bases within a codon, the order of the read abundance often persists across codon locations. For example, if the first frame is the most abundant frame for the first codon, it is also likely to be the most abundant frame for subsequent codons. It is more plausible that the various lengths from ribo-seq reads and the complicated read pileup patterns are caused by imperfect digestions of RNAse I during the experimental procedure. As a result, the read start positions are 'redistributed' to off-frame locations. In other words, it is likely that the A-site locations the the majority of the reads are indeed highly concentrated on a single reading frame, but imperfect digestions alter the read start locations, making the observed read pileups to be inconsistent with the actual A-site locations. To sum up, the observations from meta profiles show that imperfect digestions are very common in ribo-seq data, and they complicate the view of ribosome pileups.

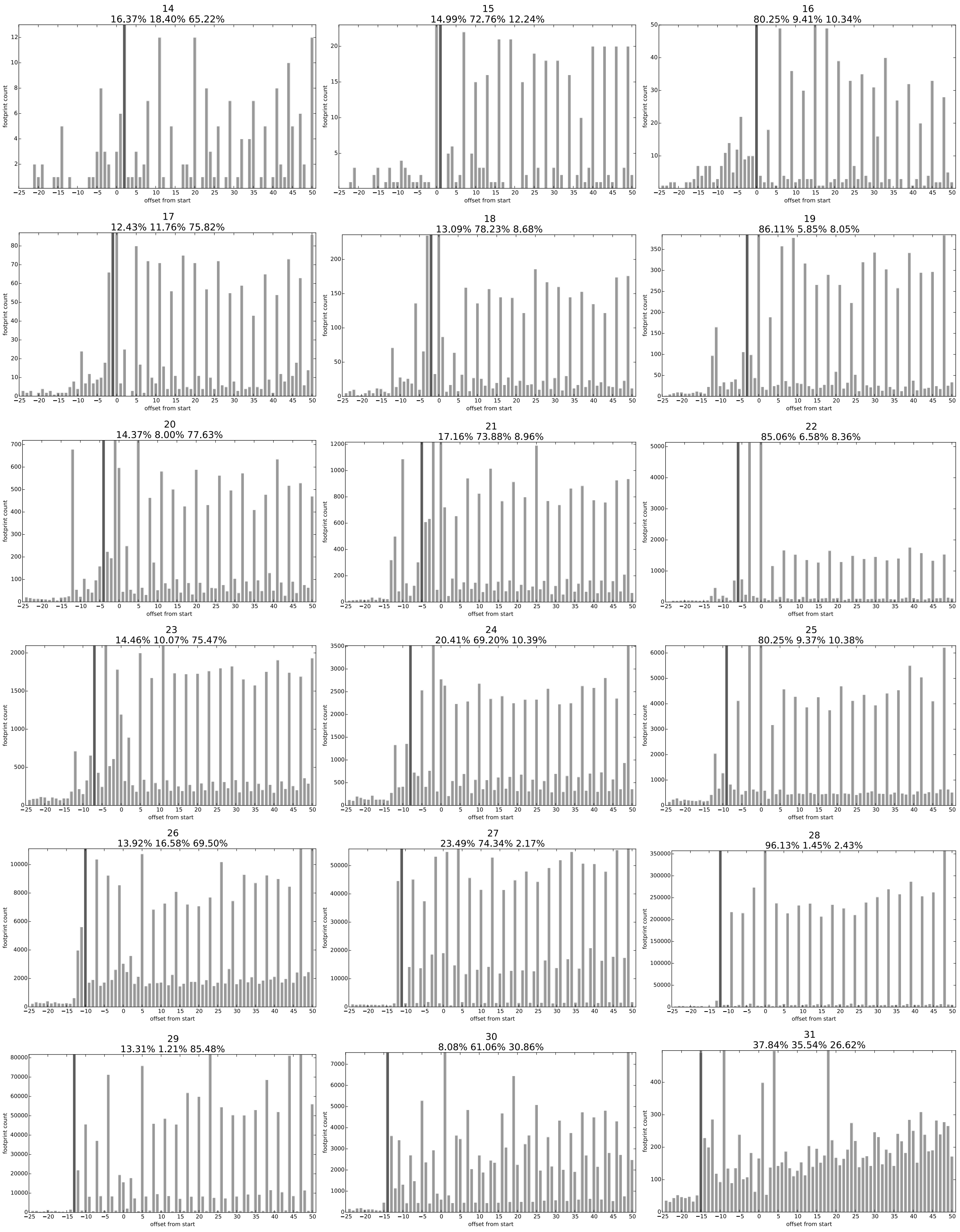


Figure 4.1: Meta-profiles on reads with different lengths near the start codon.

4.1.3 Existing A-site assignment heuristics cannot explain the complicated read digestion patterns

In past analyses, the A-site location estimation is usually based on simple heuristics. One widely used strategy is that the A-site is simply placed at 15 bases away from the 5' end of the footprint read [79, 121, 147, 156]. This is shown to be accurate for the typical ribosome footprint size [79]. However, as illustrated above, the read length from ribosome profiling experiments can span a wide range [69, 98, 112, 178], with as little as 40% being 28-nt reads [115]. The A-site position for 28-nt reads might not be suitable for other read lengths.

Since a read length not equal to the typical footprint size is mainly caused by incomplete RNase digestion during the experimental procedure [78], an alternative strategy is to use a constant A-site offset for a given read length [40, 80, 98]. This assumes that the digested portion is always the same for all reads with the same length. Such a strategy also implies a 3-nt periodic ribosome position pileups with a highly skewed frame distribution. However, as is observed above, such a frame distribution is not always presented in read pileups for all read lengths. Thus, a large fraction of ribosome footprints have under-complete or over-digestion (length $\neq 28$), and the simple offset heuristic is insufficient to explain the observed complex frame distribution pattern caused by various nuclease digestion possibilities.

In short, ribosome profiling is a powerful technique to study genome-wide translation mechanisms, but ribosome profiling data are inherently noisy due to complicated experiment pipelines. Specifically, imperfect RNase digestions distort true ribosome profiles and might bury biologically meaningful insights. Such complicated non-universal digestions vary between replicates and laboratories and cannot be well captured by existing simple heuristics of A-site assignments.

4.2 Contributions

We introduce a new model and computational method to recover the A-site positions from ribosome profiling data. Our method does not make the incorrect assumption that all reads with the same size are digested to the same extent. Instead, we systematically remove the distortion caused by imperfect digestions and retrieve true ribosome positions. Our procedure results in better A-site position estimation, which enables comparisons of ribosome profiling data from different replicates, conditions, and labs, and will hopefully lead to a better understanding of translation speed and regulation.

Observing that read pileups for each read length have a unique start for the 3-nt periodicity, we assume that there is a predominant digestion pattern for each read length. However, individual reads can be over-digested or under-digested to a certain amount centered around this major digestion pattern. Such an imperfect digestion causes the ribosome A-site to be a variable distance away from the read start. We also assume that there is an unknown underlying true A-site profile consistent across all read lengths. We define this true A-site profile as the ribosome position signal. Such a signal at a particular location is blurred to its surrounding neighborhood due to imperfect RNase digestions. We therefore model the observed read pileups as a blurring of the unknown ground truth positions. We then recover the ground truth positions by combining read pileups from different lengths and allowing the reads to be re-allocated with a non-universal

A-site offset (deblur).

Compared to previous work, our procedure does not assume any specific prior distribution of RNase digestion patterns, nor do we assume the imperfect digestion is limited to a 3-nt window [40, 192]. Rather, we learn the probabilities of the digestion for each read length from the observed data, enabling a more flexible model to explain the ribosome read pileups. Also, unlike heuristics that discard the off-frame reads [156] or take the sum of reads in all three frames [59, 135], we do not assume all ribosome reads are from a single reading frame, nor do we need to distinguish reads from different frames. Instead, we re-distribute reads to their nearby loci, naturally causing the ribosome pileups to be concentrated towards a single frame within a codon. Our approach therefore preserves the sub-codon resolution in the estimated A-site positions. We show that on a synthetic frameshift test set, our method retains the frame preferences and strengthens the frame skewness in the estimated A-site profiles.

We showcase our method by estimating codon decoding time (CDT) [40] in yeast ribosome profiling data [8]. Although abundant tRNAs are expected to speed up codon decoding, the naïve global offset heuristic only recovers a weak negative correlation between the tRNA abundance estimates and the CDT. This correlation improves after using our deblurred profiles. Also, for codons decoded by the same tRNA, our estimated CDT shows that the less stable wobble pairing codons generally translate more slowly than their synonymous codons with Watson-Crick pairing. We find that the difference in decoding time between Watson-Crick-paired codons and wobble-paired codons is generally larger than the difference between two wobble-paired codons. Such phenomena was previously only observed in metazoans [156]. This observation is consistent with the expectation that wobble pairing is likely to be delayed by the higher probability of tRNA rejection [162]. Our result therefore provides evidence for the first time in yeast to support such a mechanism. Together, our analysis gives further evidence that frequent codons translate faster than rare codons, and that both tRNA abundance and wobble pairing play roles in elongation speed.

4.3 Methods

4.3.1 Algorithm overview

For a given transcript and each read length l, let $P_{obs}(l)$ be the observed ribosome distribution from ribosome profiling reads. We model $P_{obs}(l)$ as the result of a blurring effect on an unknown, lengh-specific clear ribosome position signal $P_{true}(l)$. We assume such a position signal is consistent across all read lengths, and is deviated from an unknown consensus position signal P_{true} (Figure 4.2). We aim to recover the clear position signal from the observed blurred version of the read positions across all read lengths. The length-specific clear signals $P_{true}(l)$ should be consistent with each other, and our modeled positions $\widehat{P}_{obs}(l)$ should agree well with the observed read positions $P_{obs}(l)$. We formulate this task as a total least square optimization problem, where the difference between P_{true} and $P_{true}(l)$ and the difference between $P_{obs}(l)$ and $\widehat{P}_{obs}(l)$ are simultaneously minimized. We develop an EM-like procedure to optimize the objective and to extract the hidden clear position signal P_{true} concurrently. One example of our deblur result is shown in Figure 4.4.

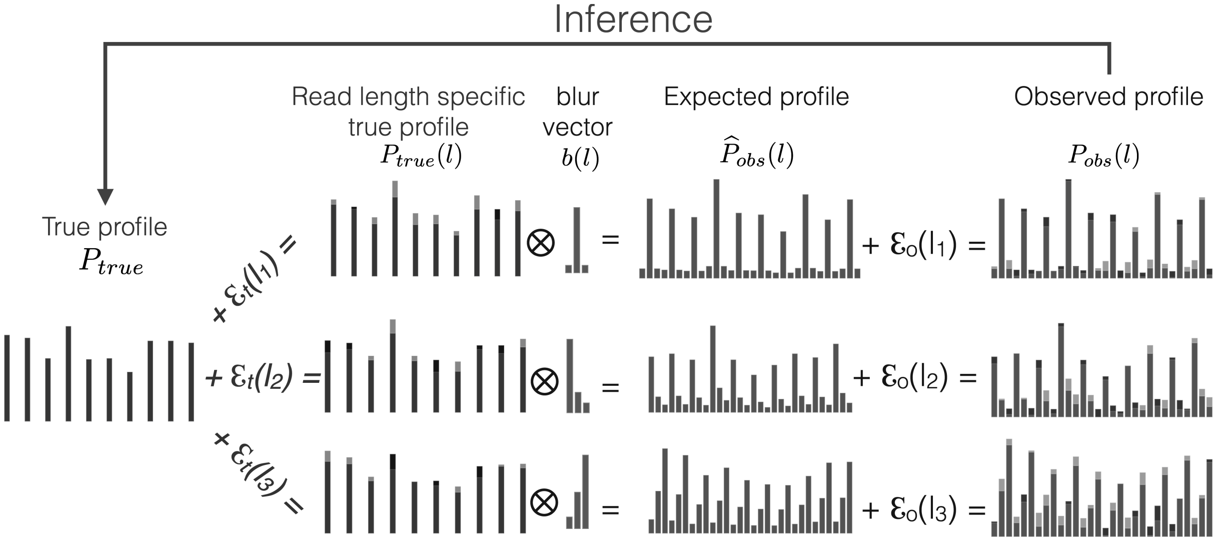


Figure 4.2: Model of the observed ribosome profiling read pileups. The observed read pileups $P_{obs}(l)$ for read length l are modeled as a convolution effect between a blur vector $b(l)$ and a clear ribosome position signal $P_{true}(l)$. The blur vector diffuses a signal to its nearby locations. The clear signal is somewhat consistent across all read lengths, and can be captured by a consensus clear signal P_{true}. An additive slack variable ε_t is used to match $P_{true}(l)$ with P_{true}, and an additive error ε_o is used to match the modeled pileups with the observed pileups. Our goal is to extract the consensus clear ribosome positions P_{true} from the observed ribosome pileups for all read lengths ($P_{obs}(l)$). We call such a clear profile extraction process *Deblur*.

4.3.2 Modeling observed profiles as blurred ribosome position signals

We model the observed ribosome read distribution $P_{obs}(l)$ for read length l as a convolution between an unknown clear ribosome position distribution $P_{true}(l)$ and an unknown blur probability vector $b(l)$: $\widehat{P}_{obs}(l) = b(l) * P_{true}(l)$, where $*$ is the convolution operator. The blur vector diffuses the position signal to its neighbor areas. This means, for location i on a transcript, the estimated observed ribosome abundance is a linear combination of the nearby true signals:

$$\widehat{P}_{obs}(l)[i] = \sum_{j=-w}^{w} b(l)[j] \times P_{true}(l)[i-j] \ ,$$

where w is the width of the blurring effect. The notation $x[i]$ indicates the ith element of vector x.

We require $P_{true}(l)$ to be as consistent as possible across all read lengths. Specifically:

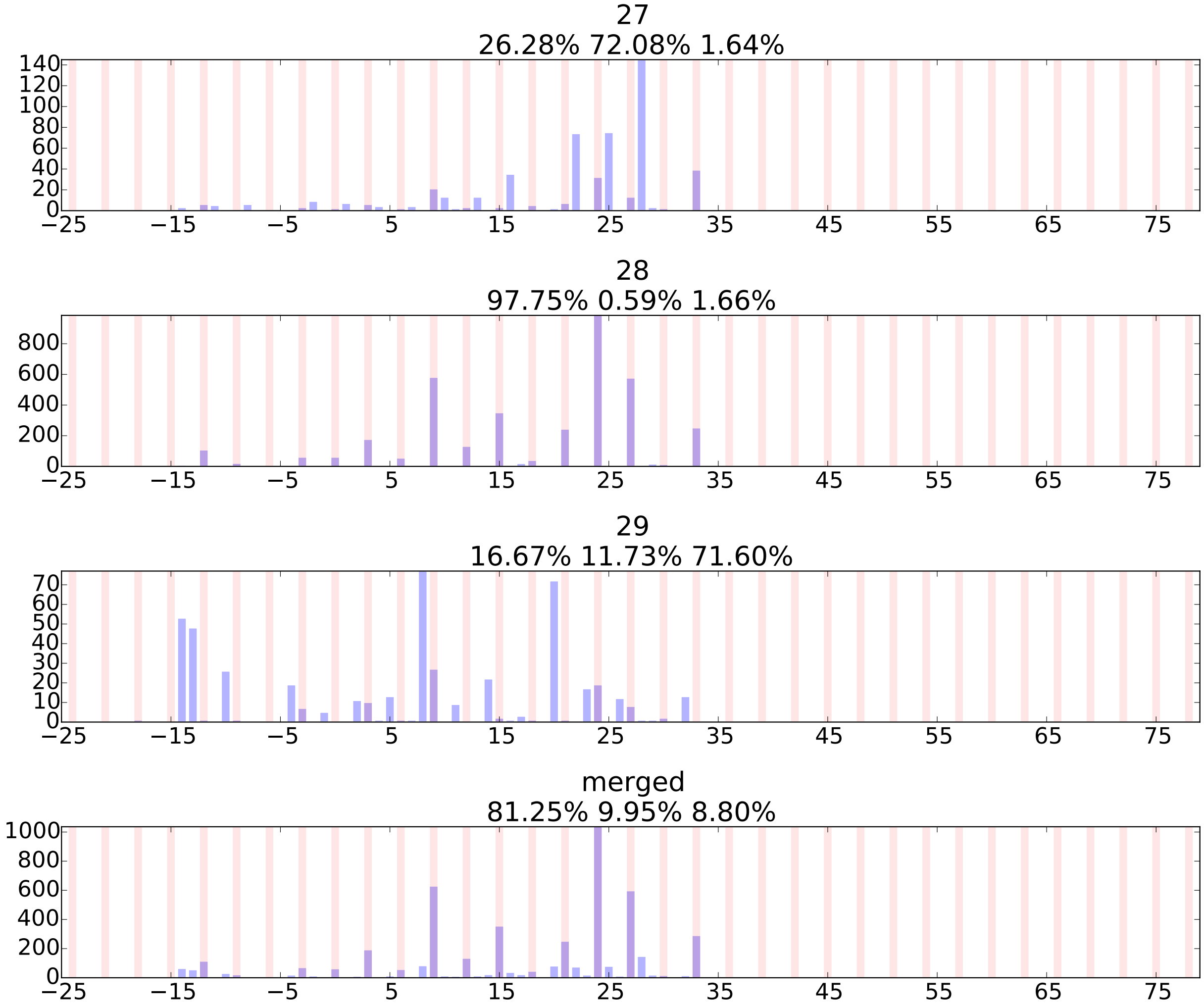
$$P_{true}(l)[i] = P_{true}[i-k_l] - \varepsilon_t(l)[i-k_l] \ ,$$

where P_{true} is the consensus position signal consistent across all read lengths, $\varepsilon_t(l)$ is the deviation of $P_{true}(l)$ from P_{true} due to length-specific digestion preferences, and k_l is a shift to align profiles with different lengths.

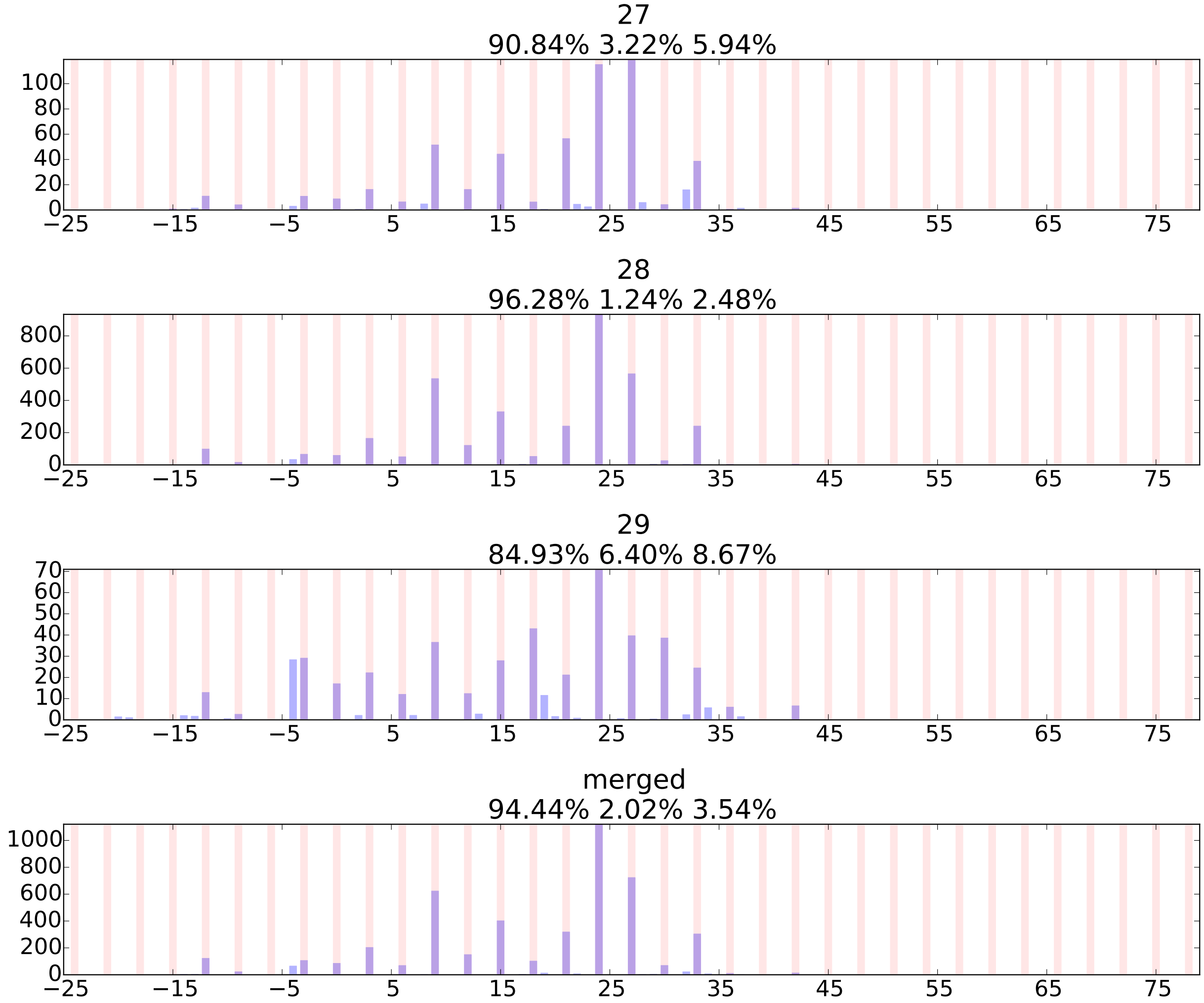
Profiles with different lengths can be aligned by observing that the start of the 3-nt periodicity is read length specific. We observe from the meta-profiles that the 3-nt peridicity for reads with length l starts at $-l+16$ (Figure 4.1, darker lines are where 3-nt periodicities start). Therefore the amount of shift between profile of length l_1 and profile of length l_2 is $-l_1 + 16 - (-l_2 + 16) = l_2 - l_1$. In our model, to align profiles with different lengths, $P_{true}(28)$ is used as the anchor, therefore $P_{true}(l)$ can be aligned to $P_{true}(28)$ by shifting $k_l = l - 28$ to the right. We denote by $P_{true}^{k_l}$ and $\varepsilon_t^{k_l}(l)$ the shifted version of the original vectors.

The starts of the 3-nt periodicity also indicate the locations of the majority of the ribosome read 5' boundaries when ribosomes start translating. They thus give the most probable A-site offsets for different read lengths. Although these offsets themselves cannot entirely capture the various distances between the A-site and the read boundaries, they serve as a good starting point for explaining the major digestion pattern of a given read length.

Putting everything together, the observed read locations $P_{obs}(l)$ of length l are assumed to be



(a) before deblur



(a) after deblur

Figure 4.4: One example of the ribosome profiles for different read lengths before and after debluring on transcript YOR302W. Read length and frame distribution are noted above each profile. In-frame (frame 0) loci are marked with light vertical lines, and read pileups are marked with dark vertical bars. The clear position signal is assumed to be consistent across read lengths, but slight shifts and deviations are allowed.

generated from the hidden P_{true} signal as:

$$P_{obs}(l) = \overbrace{\underbrace{\left(P_{true}^{k_l} - \varepsilon_t^{k_l}(l)\right)}_{P_{true}(l)} * b(l)}^{\widehat{P}_{obs}(l)} + \varepsilon_o(l) \ ,$$

where $\varepsilon_o(l)$ is the deviation of the modeled profile $\widehat{P}_{obs}(l)$ from the observed profile $P_{obs}(l)$. In short, the hidden consensus P_{true} is shifted with an additive difference $\varepsilon_t(l)$, convolved with a blur vector $b(l)$ to get the modeled profile $\widehat{P}_{obs}(l)$, and the difference between the observed profile

$P_{obs}(l)$ and the modeled profile is then measured with an additive error $\varepsilon_o(l)$. The parameters k_l, $\varepsilon_t(l)$, $\varepsilon_o(l)$, $b(l)$ must be optimized to find the hidden P_{true}. We explained above the rationale of choosing k_l, and we describe how other parameters are optimized in the following sections.

4.3.3 Deblurring ribosome profiles — a least square optimization

Our goal is to use the blurred observed profiles $P_{obs}(l)$ to deconvolve the clear ribosome position signal P_{true} of a transcript. Such clear signals should be consistent across all read lengths, and should be a good estimate of the observed ribosome distribution after applying the blurring effect. The consensus clear position signal P_{true} and the deviation between the consensus and the length-specific ribosome signal ($\varepsilon_t(l)$) are adjusted to minimize two terms: the difference between the observed profile and the modeled profile and the difference between the consensus and the length-specific ribosome signal. Specifically:

$$\min_{\boldsymbol{P_{true}}, \boldsymbol{\varepsilon_t(l)}} \sum_l \alpha(l) \left[\|P_{obs}(l) - \widehat{P}_{obs}(l)\|_2^2 + \|P_{true}^{k_l} - P_{true}(l)\|_2^2 \right] , \tag{4.1}$$

where $\alpha(l)$ is the total read count with length l for the tested transcript. Intuitively, if some read length is more abundant, the true position signal recovered from that read length should be weighted more.

Using $P_{true}(l) = P_{true}^{k_l} - \varepsilon_t^{k_l}(l)$ and $\widehat{P}_{obs}(l) = b(l) * \left(P_{true}^{k_l} - \varepsilon_t^{k_l}(l)\right)$, we rewrite (4.1) to be:

$$\min_{\boldsymbol{P_{true}}, \boldsymbol{\varepsilon_t(l)}} \sum_l \alpha(l) \left[\|P_{obs}(l) - b(l) * \left(\boldsymbol{P_{true}^{k_l}} - \boldsymbol{\varepsilon_t^{k_l}(l)}\right)\|_2^2 + \|\boldsymbol{\varepsilon_t^{k_l}(l)}\|_2^2 \right] . \tag{4.2}$$

If the blur vectors $b(l)$ are known, we can use an EM-like framework to find the least square solution:

M-step: We fix P_{true} and adjust $\varepsilon_t(l)$ to optimize the total least square problem in (4.2), where ε_t for each l can be optimized separately:

$$\min_{\boldsymbol{\varepsilon_t(l)}} \|b(l) * \boldsymbol{\varepsilon_t^{k_l}(l)} - \left(b(l) * P_{true}^{k_l} - P_{obs}(l)\right)\|_2^2 + \|\boldsymbol{\varepsilon_t^{k_l}(l)}\|_2^2 , \tag{4.3}$$

where bold indicates the variables we are optimizing. The optimal $\varepsilon_t(l)$ is found via a least square solver with Ridge regression [54] (damp $= 1$).

E-step: We fix $\widehat{P}_{obs}(l)$ and $P_{true}(l)$ as estimated from the M-step, and adjust the consensus P_{true} to minimize the objective in (4.1). The expected P_{true} is therefore the weighted average of all $P_{true}(l)$. After the M-step, the new estimation of $P_{true}(l)$ is $P_{true}^{k_l} - \varepsilon_t^{k_l}(l)$, so the weighted average of $P_{true}(l)$ is:

$$P'_{true} = \sum_l \frac{\alpha(l)(P_{true}^{k_l} - \varepsilon_t^{k_l}(l))}{\sum_{l'} \alpha(l')} = P_{true} - \sum_l \frac{\alpha(l)\varepsilon_t^{k_l}(l)}{\sum_{l'} \alpha(l')} . \tag{4.4}$$

We set all negative entries of P'_{true} to be zero, and renormalize P'_{true} so that it sums to 1. This constrains P_{true} to remain valid and in practice appears to have a minor effect on the shape of P_{true}.

We repeat the EM-like procedure until the change of the objective in (4.2) compared to the objective value from the previous step is smaller than 0.01.

We initially set P_{true} to be the in-frame values of the observed read pileups with length 28:

$$P_{init}[i] = \begin{cases} P_{obs}(28)[i] & \text{if } i \text{ is a multiple of } 3, \\ 0 & \text{otherwise.} \end{cases} \tag{4.5}$$

$P_{obs}(28)$ is used as the initial consensus because 28 is the typical ribosome footprint size for yeast. For such size, the real physical ribosome footprint boundaries should be most likely to overlap with the read ends. This is because an imperfect digestion for reads with length 28 has to be caused by a simultaneous over-digestion from one end and an under-digestion from the other end, which is likely to be relatively rare. Therefore, the observed read pileups with read length 28 should be the most clear and the closest to the ground truth position signal. Indeed, these profiles show the strongest frame skewness and the most visible 3-nt periodicity (Figure 4.1).

4.3.4 Estimating blur vectors from meta-profiles

The deblur process depends on a known set of blur vectors ($b(l)$) — a crucial element to model the imperfect digestion in ribosome reads. These vectors describe the probability of relocating ribosomes to transfer the clear ribosome position signal to the observed read pileups. Since these pileups are read 5' end pileups (see below), essentially the blur vectors adjust a true footprint boundary to the observed read boundary. They therefore indicate the probability of the amount of under/over digestion from the 5' end, and capture the read-length-specific digestion patterns.

These blur vectors can be estimated directly from the ribosome reads via meta-profiles. Meta-profiles are widely used to reveal the positional patterns of ribosome profiles [38, 60, 67, 79, 80, 98, 150]. They do so by summing read pileups from all transcripts for each position. The blur vectors can be estimated from these meta-profiles because convolution satisfies the distributive property.

To generate the meta-profiles, we group reads by lengths, and accumulate the positions of the 5' ends relative to the start codon for all transcripts. We then include the first 350 locations away from the start codon in the meta-profiles. We only use transcripts with length > 350 to reduce convolution boundary effect. Also, to avoid the outlier points biasing the shape of the blur vector, we exclude locations in the meta-profiles with the top 1.65% highest read counts. This threshold is chosen by assuming the top 5% of in-frame reads (1/3 of total reads) are outliers.

To estimate the blur vectors, we use an EM-like procedure similar to the earlier deblur optimization. In this procedure, the observed transcript profiles are replace by the meta-profiles, and the blur vectors are adjustable variables. The procedure is exactly the same as described in the previous section, except that the blur vector is first estimated prior to the M-step:

$$\min_{\boldsymbol{b(l)}} \| M^{k_l}_{true} * \boldsymbol{b(l)} - M_{obs}(l) \|_2^2 \ ,$$

where the "M" variables are the meta-profiles, and we replace P_{true} by M_{true} and $P_{obs}(l)$ by $M_{obs}(l)$ in (4.3) – (4.5). The blur vector size, which limits the diffusion range of the position signal, is set to 31. This way the signal can be diffused, either to the left or to the right, as far as approximately half the size of a ribosome. A non-negative least square solver (`scipy.optimize.nnls`) is used to find the best $b(l)$. All blur vectors with different read lengths are optimized separately.

4.3.5 Estimating the A-site profile

We merge the length-specific true profiles to get an overall ribosome position signal for a given transcript — the A-site profile. It is the weighted sum of all the length-specific true profiles, shifted to the right by 15:

$$C_{true}[i] = \sum_{l} \alpha(l) P_{true}(l)[i + k_l - 15] \; .$$

The shift is needed since the true profiles are estimated from the reads' 5' ends, and $P_{true}(28)$ is the anchor to align profiles with different length. We shift by 15 since it is the major A-site offset of reads with length 28, and it is the A-site offset under perfect digestion.

4.3.6 Codon decoding time estimation

To investigate the influence of tRNA abundance and wobble pairing on translation speed, we estimate codon decoding time using the procedure in [40]. The in-frame (frame-0) deblurred read counts are used as the input ribosome count for each codon position. Such counts are normalized by the average ribosome count for each transcript, as is done in [98, 178]. Following [40], these normalized counts are grouped by codon types to form codon count distributions, with the exclusion of the first and last 20 codon positions of each transcript and positions with ribosome counts less than 1. Each codon distribution is fit with a log normal distribution. The skewness of the log normal distribution is used as an estimate of the codon decoding time, as it has been shown to be informative for estimating the elongation speed from ribosome profiling data among various species [40].

4.3.7 Read alignment and data preprocessing

We test the deblur method on ribosome profiling data from *Saccharomyces cerevisiae*, where ambiguous mapping is not ubiquitous. We use ribosome reads from a yeast study, where the data is of high quality and of high sequencing depth (GSM1335348) [8].

Reads were first aligned to the yeast noncoding RNA reference, which includes rRNA, tRNA, snoRNA, etc., to remove noncoding contaminants. The remaining reads are then mapped to the yeast transcriptome. The yeast noncoding RNA reference and the transcriptome reference are downloaded from the Saccharomyces Genome Database [51]. Alignments are performed with STAR [44] with parameters `--clip3pAdapterSeq CTGTAGGCACCATCAAT --outFilterMismatchNmax 1`, which automatically 'softclips' the unaligned adapter sequences and any unaligned bases at the 5' end of the reads, allowing at most 1 mismatch. Only uniquely mapped reads, about 83% of the non-contaminated reads, are used to generate the observed profiles $P_{obs}(l)$.

The observed profile of a given length is included for a transcript in the deblur process if more than 50% of the in-frame loci have non-zero ribosome counts. Here, we define ‘in-frame’ as the frame with the highest total read count. Only transcripts with at least two observed profiles from different read lengths are tested for deblur.

Such filtering results in 1966 transcripts with high ribosome coverage. This transcript set size agrees with the size of highly expressed transcript set: 2108 transcripts share an estimated expression level > 100 transcript per million (TPM) (expression are estimated using Salmon [131] with the RNA-seq data from the same experiment (GSM1335347) [8]).

4.4 Results

4.4.1 Ribosome profiles are well explained by the blur model

We test whether the estimated blur vectors can truthfully characterize the observed read locations on the meta-profiles. Here, different blur vectors are convolved with the initial guess of the meta consensus — the in-frame values of $M_{obs}(28)$. Our modeled meta-profiles agree well with the observed meta-profiles. The blur process results in a good correlation and a small deviation between the modeled meta-profiles and the observed meta-profiles across all read lengths (Figure 4.5).

To test whether a single blur vector is sufficient to model all profiles for a given length, we train the blur vector on subsets of locations and on subsets of transcripts, and we get similar results compared to training the blur vector on the entire set (Figure 4.5). This indicates that the blur vectors are transcript and location independent.

In addition, allowing the length-specific clear position signals to be slightly deviated from the consensus further improves our model fitting. Instead of enforcing an identical consensus across all read lengths, these deviations result in both better modeled meta-profiles (Figure 4.6), and better modeled transcript profiles (Figure 4.7): the inconsistencies between the observed profiles and the modeled profiles are reduced by an average of 66% compared to not allowing such deviations.

4.4.2 Consistent read-length-specific profiles

To test how shifting and deblurring affect the consistencies among profiles with different read lengths, we compare read-length-specific profiles with the in-frame values of the observed profiles with length 28 (P_{init}, Eq. (4.5)). We choose P_{init} for comparison because it is the original data in which we have the most confidence. We use the Pearson correlation coefficient as a measurement of the consistency between the read-length-specific profile and P_{init}.

Two factors jointly improve the consistencies of ribosome profiles among different lengths: the deblur process and allowing a length-specific shift and deviation from the consensus. Initially, none of the raw observed profiles ($P_{obs}(l)$) correlate well with the in-frame values of

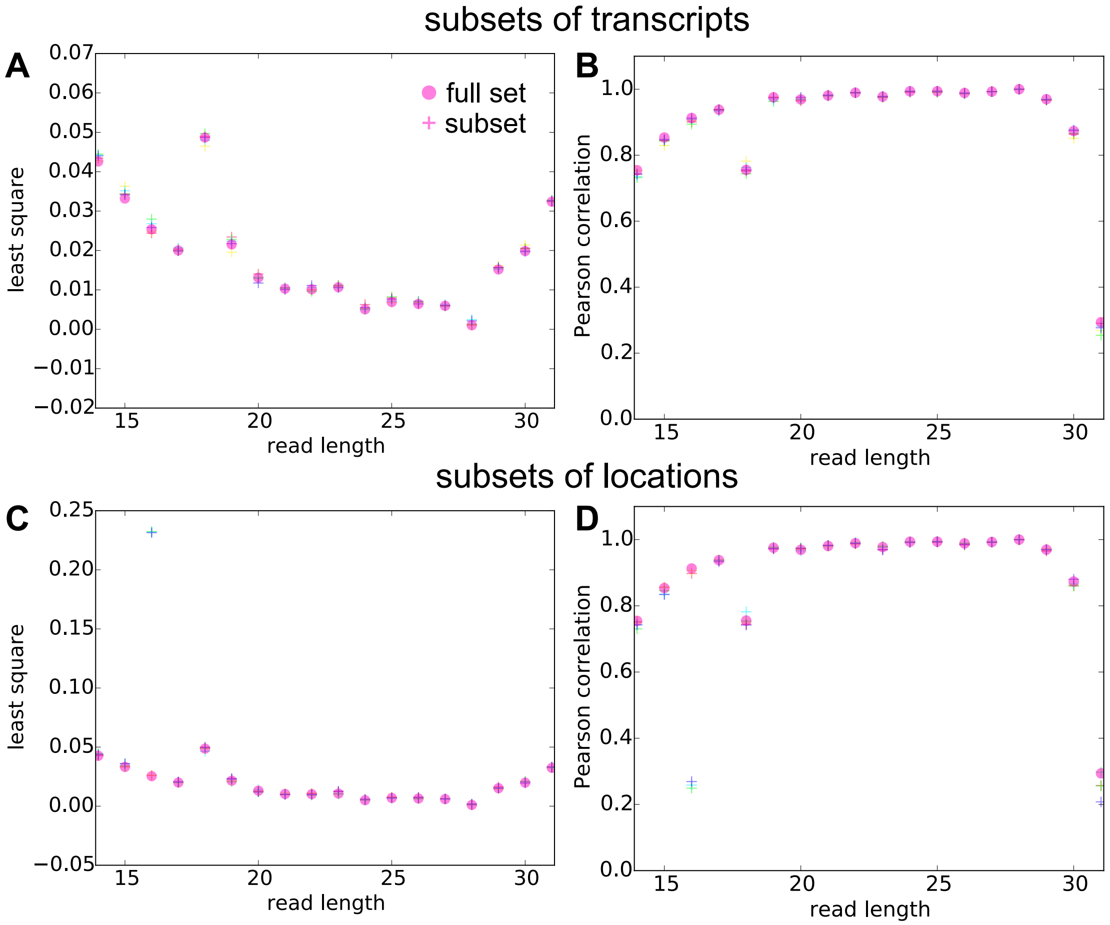


Figure 4.5: The agreement between the observed meta-profiles and the modeled meta-profiles with blur vectors trained on subsamples of data. (A, B) The transcripts are randomly divided into five groups with equal size, and the blur vectors are trained on the subset of transcripts. (C, D) The profile locations are randomly divided into five groups with equal sizes, and the blur vectors are trained on the subset of locations. The least square and the Pearson correlation are between the the observed meta-profiles ($M_{obs}(l)$) and the modeled meta-profiles ($\widehat{M}_{obs}(l) = M_{init} * b(l)$) for each read lengths l. Results are qualitatively similar regardless of whether the blur vectors are trained on a subsample of the data. Read length 18 is modeled slightly worse primarily due to low coverage and significant outliers.

$P_{obs}(28)$ (Figure 4.8A). However, the correlations are improved if the observed profiles are properly shifted and aligned to $P_{obs}(28)$ (Figure 4.8B). The correlations can be further increased by applying the deblur process to recover the length-specific clear profiles ($P_{true}(l)$) (Figure 4.8C). Lastly, compared to the initial guess of the consensus (P_{init}), at the end of the deblur process, the final consensus estimation (P_{true}) correlate better with the length-specific clear profiles (Figure 4.8D).

Overall, the correlation between $P_{true}(l)$ and P_{true} for most lengths is close to 1. Since P_{true} is the centroid of $P_{true}(l)$, the good correlation between the two indicates that the deblurred profiles are consistent across different read lengths.

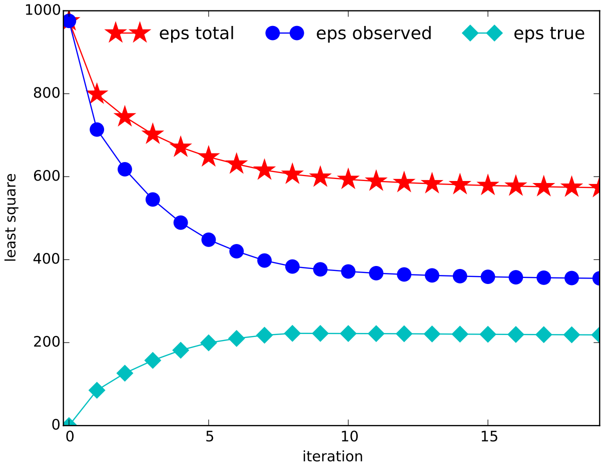


Figure 4.6: The change of the objective function in equation 4.1 for deblurring the meta-profiles. The circle line is the overall inconsistency between the modeled meta-profile ($\widehat{M_{obs}}(l)$) and the observed meta-profile ($M_{obs}(l)$), the diamond line is the overall inconsistency between the consensus clear signal ($M^{k_l}_{true}$) and the read-length-specific deblurred signal ($M_{true}(l)$), and the star line is the sum of the two. The overall inconsistency between the modeled meta-profiles and the observed meta-profiles successfully goes down during the optimization, and the observed profiles are better modeled by sacrificing the inconsistencies of clear signals across different lengths.

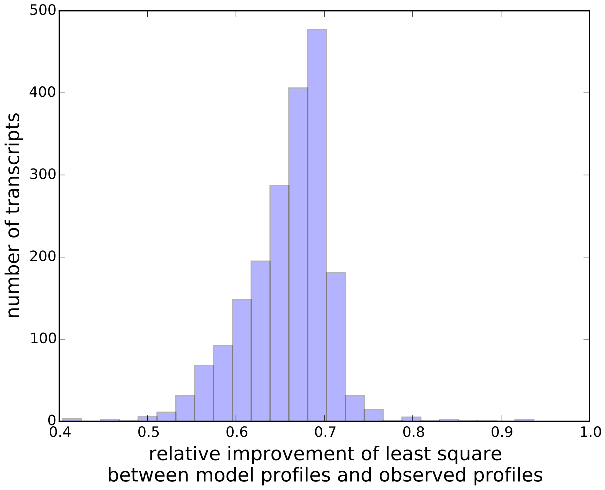


Figure 4.7: Histogram of the relative improvement of the overall inconsistency between the observed profiles P_{obs} and the modeled profiles $\widehat{P}_{obs}$ after the deblur procedure. The relative improvement is defined as the change of the inconsistency between P_{obs} and $\widehat{P}_{obs}$, over the initial inconsistency between the two.

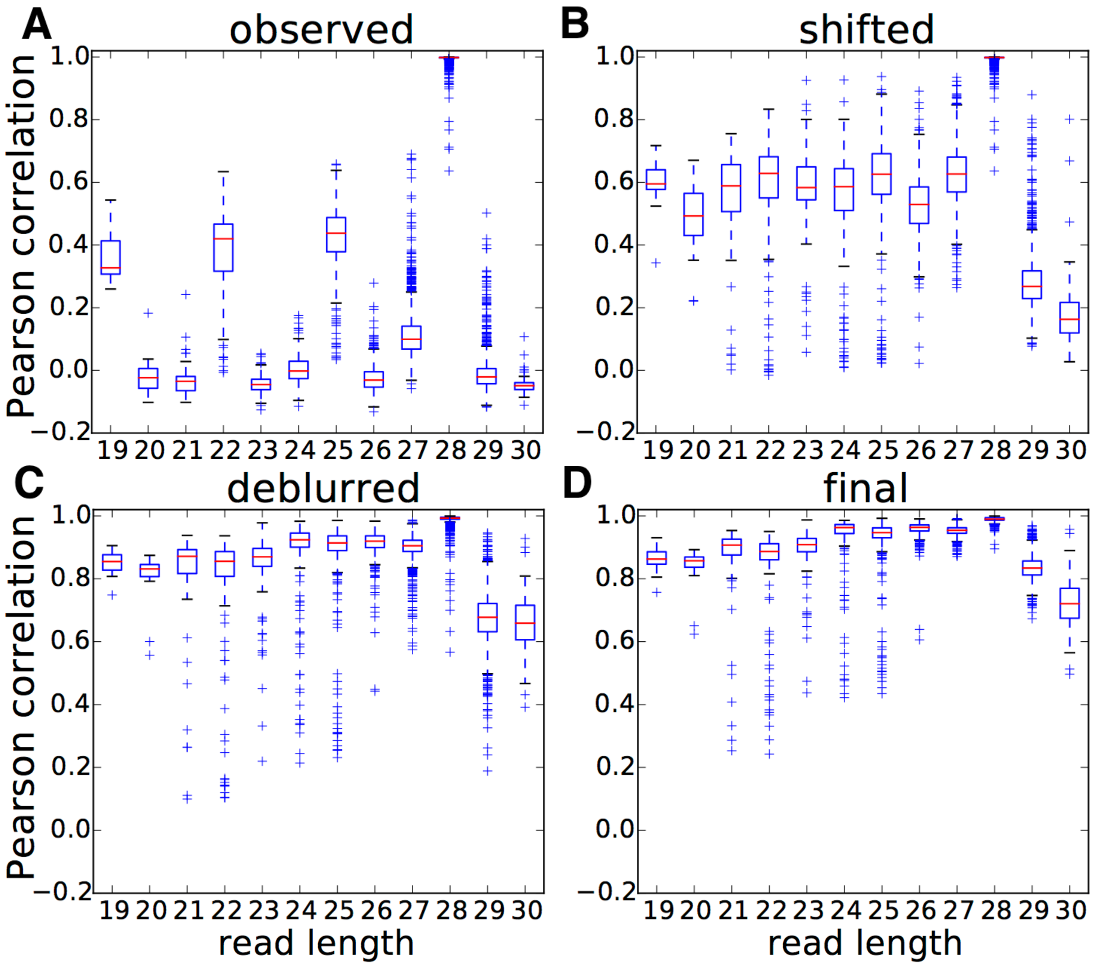


Figure 4.8: Effect of shifting and deblurring on profile consistencies across different read lengths. $P_{obs}(l)$ is the observed profile of length l, P_{init} is the in-frame values of $P_{obs}(28)$, which is also the initial guess of the true profile, k_l is the shift applied to a profile, $P_{true}(l)$ is the length-specific deblurred profile, and P_{true} is the consensus of $P_{true}(l)$s. Box plots of the Pearson correlation are between (A) $P_{obs}(l)$ and P_{init}, (B) $P_{obs}^{k_l}(l)$ and P_{init}, (C) $P_{true}^{k_l}(l)$ and P_{init}, and (D) $P_{true}^{k_l}(l)$ and P_{true}. The improvement of the Pearson correlation between the read-length-specific profiles and the true profiles is the combinational effect of the right amount of shifts and the success of deblurring.

4.4.3 Improved frame skewness

Our deblur process improves the frame skewness of the recovered A-site profiles, even if it does not explicitly optimize or force frame skewness. The in-frame position is the reading frame where reads are preferentially distributed within a codon. It is usually frame 0 for the A-sites of ribosome footprints with the absence of frameshifts. It is desirable for the recovered ribosome A-site profiles to be highly skewed towards frame 0. This is because ribosomes move in units of codons, so ribosome profiles should have 3-nt periodicity. Also, such profiles should be mainly concentrated on frame 0, since frameshifts are rare. Indeed, after deblur, the in-frame skewness does improve from an average of 71% to 92% (Mann-Whitney U test $p < 3 \times 10^{-308}$; Figure 4.9). This indicates that the deblur process produces ribosome profiles with less noise. It also enables more reads to be used in downstream analysis. For instance, if only the in-frame reads are used to represent the codon-level ribosome counts, the deblur process will allow on average 20% more reads to be used.

4.4.4 Deblur process produces sub-codon resolution profiles

The deblur procedure does not assume that the recovered A-site profiles are all from a fixed frame, thus it keeps the sub-codon resolution of the A-site profiles, and allows detection of potential programmed frameshifts. To test whether the deblur process can recover profiles with frameshifts, we synthetically generate frameshifts as follows: We first choose a random frame-0 location as the frameshift point in a transcript, we then shift all reads with a start location after such point to the right. This is to simulate an insertion in the transcript to induce a frameshift. The recovered A-site profiles should have a high skewness towards frame 0 before the frameshift point, and a high skewness towards frame 1 after the frameshift point. One example of our constructed frameshift events is shown in Figure 4.10.

The deblur process successfully maintained and improved the skewness of frame 0 before the frameshift point and the skewness of frame 1 after the frameshift point (Figure 4.11). Therefore, combining profiles with lengths other than 28 during the deblur process results in a recovery of a clear frameshifted A-site profile, regardless of the incorrect initial guess. To sum up, frameshift detection is an important task, but current frameshift detection method [121] suffers from high false positive rates. Our deblur process recovers ribosome profiles with a clear frame preference, which will promote the development of a better frameshift detection.

4.4.5 Wobble pairing codons translate slower than Watson-Crick pairing

The tRNA abundance was expected to be negatively correlated with the codon decoding time (CDT) [39, 40, 59, 98], and such correlation is strengthened using our deblurred profiles. After

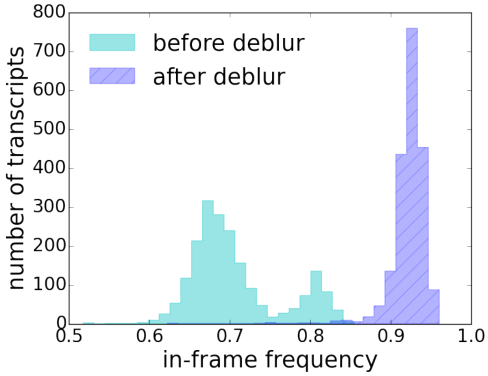


Figure 4.9: Histograms of in-frame (frame 0) portion of reads before and after deblur. The recovered A-site profiles have a higher in-frame skewness compared to the original profiles.

deblur, 85% of the codon distributions have a smaller variance, indicating the deblur process successfully removes noise from the observed read pileups. From these distributions, the estimated CDT is the skewness of a lognormal fit [40] (details in Methods). Such estimated CDT is compared with the tRNA Adaptation Index (tAI) [45] — proxy for the tRNA concentration. The deblur process strengthens the Spearman correlation between the tAI and the estimated CDT from -0.21 ($p = 0.1$) to -0.46 ($p = 1 \times 10^{-4}$). This provides stronger evidence that tRNA abundance play a role in elongation speed. Similarly, the raw frequency of codon usage also negatively correlates with the estimated CDT (Spearman correlation -0.5, $p = 3.7 \times 10^{-5}$), indicating frequent codons are translated faster than rare codons.

Wobble pairing could also affect the elongation speed. Since there are usually fewer tRNA types than codon types, some of the codons that encode the same amino acid must be decoded by the same tRNA. Wobble pairing allows a tRNA to recognize more than one codon. Within these synonymous codons, the determinant of the codon decoding speed is the efficiency of the tRNA recognizing the corresponding codon. According to the wobble hypothesis [36], the last two bases of the tRNA anticodon form Watson-Crick base pairs and bond strongly to the first two bases of the codon. However, the anticodon's first base can form a wobble pair: The base G can either Watson-Crick pair with C, or wobble pair with U; the base I (inosine, edited from A) can also both wobble pair with C and U, but I:U pairing has a less favorable geometry [156]; the base U can Watson-Crick pair with A, and wobble pair with G. It has been hypothesized that wobble paired codons tend to be translated slower than their synonymous Watson-Crick paired codons, since wobble pairs are more likely to be rejected before peptidyl transfer, causing the tRNA selection cycle to be repeated [162].

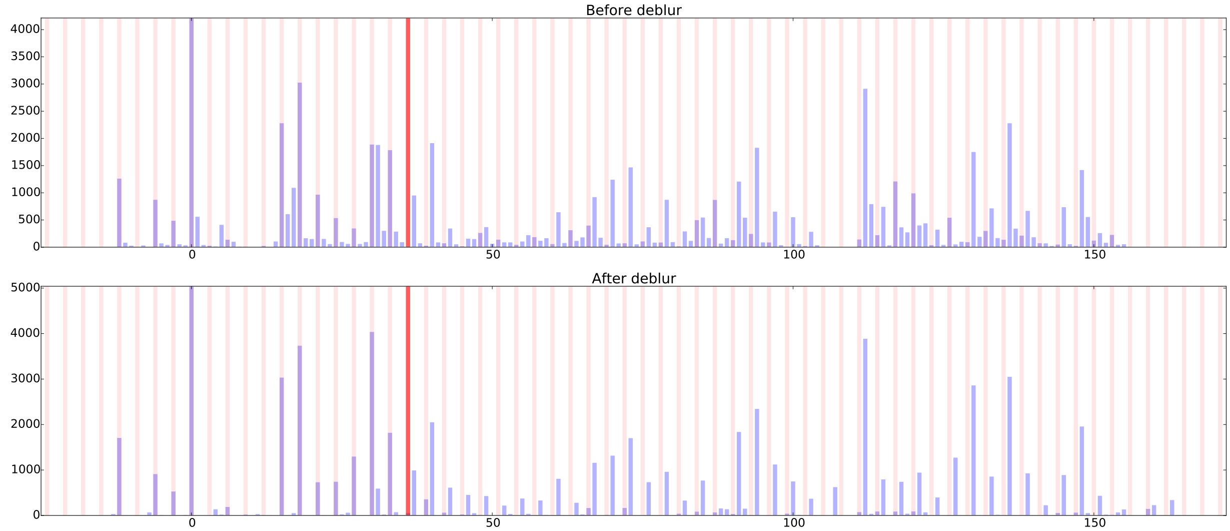


Figure 4.10: An example of the ribosome profiles with artificially programmed frameshift on transcript YDL061C before and after deblur. Frame-0 loci are marked with pink vertical lines and the frameshift point is marked with a darker vertical line. The high frame-0 skewness before the frameshift point and the high frame-1 skewness after the frameshift point are well maintained and strengthened by the deblur process.

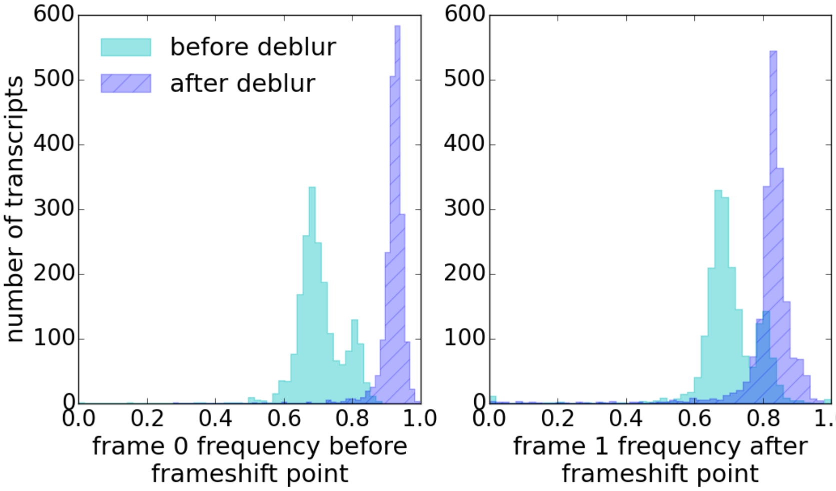


Figure 4.11: Histograms of the frame-0 portion of the ribosome profiles before the frameshift point and the frame-1 portion of the ribosome profiles after the frameshift point. The deblur process strengthens the frame skewness while keeping the estimated ribosome positions to be in the correct frame.

We investigated how wobble pairing influences CDT in yeast. We focus on pairs of codons that are translated by the same tRNA, so that the influence of tRNA concentration on the elongation speed is controlled. In this case, the codon pair shares the first two bases, and differs in the third base. We compared the estimated decoding time between the codon pairs, and find that the wobble pairing codons indeed are estimated to often translate slower than the Watson-Crick pairing codons (Figure 4.12).

We expect the decoding time difference between two wobble paired codons to be smaller than the difference between a wobble pair codon and a Watson-Crick pair codon, if the wobble-paired tRNA is truly more likely to leave the ribosome without successful peptidyl transfer [162]. For the three codon pairs being compared, we would therefore expect the time difference between I:C and I:U to be smaller than the time difference between G:C and G:U, and between U:A and U:G. To control for the absolute level of the translation time, we use the relative decoding time difference between a synonymous codon pair. It is defined as: $\Delta t = (t_{\text{wobble}} - t_{\text{Watson-Crick}})/t_{\text{Watson-Crick}}$, where t_x is the estimated decoding time for a codon with either wobble pairing or Watson-Crick pairing.

Using the profiles from the deblur process, the decoding time difference is inline with the above expectation. The decoding time difference between a wobble paired codon and a Watson-Crick paired codon is indeed visibly larger than the decoding time difference between two wobble pair codons (Figure 4.12). Although such a trend was first seen in metazoans [156], it was not observed for most wobble paired codons in yeast [13, 59, 127]. It is also less obvious when CDTs are estimated from the original ribosome profiles (Figure 4.12). This indicates that the uncorrected ribosome profiles obscure true ribosome A-site positions. Together, the CDT estimated from the deblurred profiles strengthen the conclusion that wobble pairing slows translation. These results also suggest that wobble pairing can be used as a mechanism to regulate elongation speed.

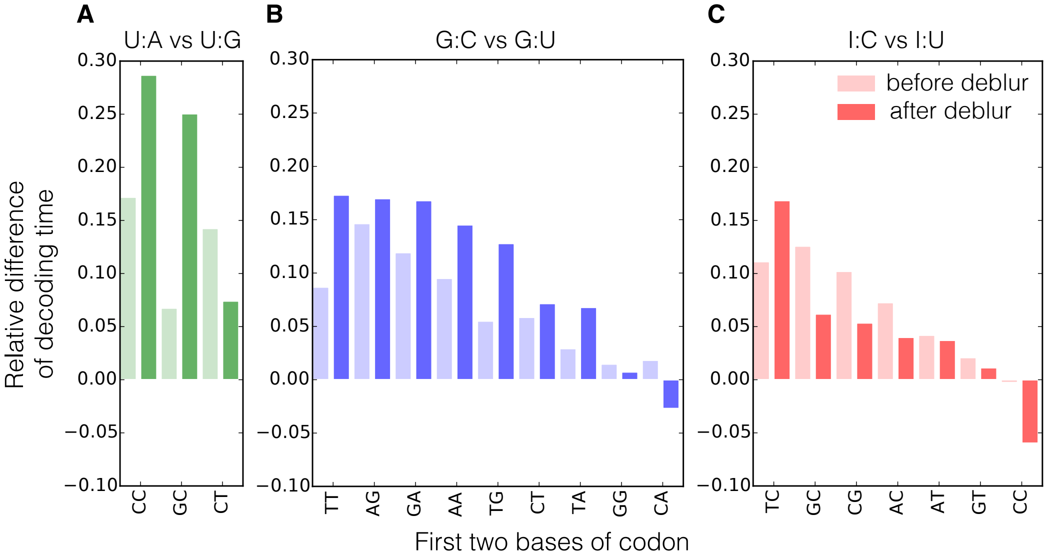


Figure 4.12: Relative differences of the CDT between pairs of codons that are decoded by the same tRNA. Lighter colors are time differences estimated from original ribosome profiles, and darker colors are time differences estimated from deblurred profiles. The anticodon : codon pairings are: (A) U:A (Watson-Crick) vs. U:G (wobble), (B) G:C (Watson-Crick) vs. G:U (wobble), (C) I:C (stronger wobble) vs. I:U (weaker wobble). The average relative time difference estimated from the deblurred profiles is: 0.2 between U:A and U:G, 0.1 between G:U and G:C, and 0.04 between I:U and I:C; the average relative time difference estimated from the original profiles is 0.12 between U:A and U:G, 0.07 between G:U and G:C, and 0.07 between I:U and I:C.

4.5 Discussion

Estimating ribosome A-site positions from ribosome profiling data is a challenging necessary step in quantifying codon-specific translation speed and ribosome pausing. There are controversial conclusions about whether the tRNA level plays an important role in CDT. Different analysis pipelines performed on different experiments show that the estimated CDT sometimes strongly correlates with the codon usage [59], sometimes weakly correlates with tAI [98], and sometimes does not correlate with the codon optimality [13] or tRNA level [135]. Different estimates of CDT alone produce different correlations between the estimated decoding time and the tRNA level among different species [39, 40]. The fact that there is evidence both for and against the correlation between tRNA levels and CDT indicates that a better codon decoding time analysis pipeline is needed.

We here by no means try to touch all aspects of the elongation time estimation, nor do we try to emphasize or diminish the impact of tRNA level on the translation dynamics. We focus on recovering the A-site positions from the ribosome profiling data, the first step of any quantitative analysis on codon rate or pausing strength. Rather than applying simple heuristics that cannot fully explain the observed read pileup patterns, we borrow an idea from signal processing to recover the A-site positions from ribo-seq data. Our deblur formulation is inspired by the problem of blind deconvolution in image processing [104], where both the original clear signal and the blur kernel (blur vector) are unknown. We make advantage of the strong asymmetry between the dimensionality of the blur vector and the ribo-seq positional signal, and formulate the clear signal recovery problem with a total least square optimization. We show via several lines of intrinsic and extrinsic evidence that our deblur method provides better estimates of A-site profiles, leading new insights on translation dynamics. Source code for the deblur method and the analysis can be found at: `http://www.cs.cmu.edu/~ckingsf/software/riboasitedeblur/`.

Chapter 5

Genome-wide analysis on modified ribosome profiling shows no preferred locations for ribosome collisions in yeast

This chapter presents the first genome-wide study of ribosome collision with a full experimental and computational pipeline. Ribosome collisions are extreme local ribosome slow-downs when two ribosomes bump into each other during translation. Understanding the patterns of ribosome collision is important for studying the mechanism of translational speed regulation. This chapter provides a detailed description of the analysis pipeline, and shows several lines of evidence to support random occurrence of ribosome collision. This was a joint work with Pieter Spealman, who executed the modified ribosome profiling experiments, Joel McManus, and Carl Kingsford, who both contributed to the design of the analysis.

5.1 Introduction

Ribosome stalling might play an important role in translational regulation during protein synthesis. Programmed ribosome stalling may modulate the amount of proteins being produced or assist proper protein foldings during translation. Although anecdotal examples are often exhibited to showcase ribosome stalling [69, 79, 80], methods that make full use of ribosome profiles and quantify the intensity of ribosome stalling are just starting to emerge. Recently, one statistical test was designed to analyze amino acid enrichments in ribosome stalling across different species [147], and a pause score was developed to measure the ribosome footprint enrichment of codon pairs [178]. We here study the extreme ribosome stalling where two ribosomes bump into each other during translation. We designed a protocol to captures ribosome collisions experimentally, which can provide a complimentary view of ribosome stallings from regular ribosome profiling experiments.

An overview of the ribosome profiling protocol is described in section 1.2.2. One of the key steps in the ribo-seq protocol is size selection to remove contaminants and select sequences bound by a single ribosome. In this chapter, we analyze ribo-seq data from a modified ribosome profiling protocol, in which the size selection step has been replaced with a step that selects regions consistent with two ribosomes being adjacently bound. This modified protocol allows us to investigate ribosome collision events experimentally. With this data set, we get both standard ribosome footprints and the so-called *doublet* footprints. We combined this data set we generated with our collaborators with two existing data sets that also contain doublets [69]. This allows us to begin to draw general conclusions about ribosome collisions.

Although the coverages of doublet libraries are low due to contaminants, all of our analysis are consistent with random occurrences of ribosome collision. Nonetheless we observe a consistent depletion of ribosome collisions within 50 bases away from the start codon. We further study stalling-induced collision sites (SICS) with a statistical framework to identify ribosome stalling sites from standard ribo-seq data coupled with enriched ribosome collisions. We notice that ribosome profiles often do not support co-occurrence of ribosome collision and stalling. Among identified SICS, we observe SICS locations are unique to experimental conditions. Further, we do not observe a differential pattern of codon usage within SICS compared to regular ribosome stalling sites, nor do we observe SICS to be preferentially located upstream of known RNA binding motifs. We provide for the first time a genome-wide ribosome collision study in yeast, and all of our analysis support the hypothesis that ribosome collisions occurs by chance.

Table 5.1: Terms and definitions used in ribosome collision analysis

Term	**Definition**
RPF	RPF stands for ribosome profiling footprint.
singlet	singlets are ribosome footprint reads from libraries with a typical ribosome footprint size.
doublet	doublets are collided ribosome reads from libraries with a size selection around 60, i.e. the size of 2 ribosome footprints back to back.
singlet/doublet load	A singlet or doublet load of a transcript is the sum of ribosome footprint counts from the specific library of that transcript.
transcript collision rate	transcript collision rate is computed as: doublet load / (singlet load + doublet load).
Stalling Induced Collision Site (SICS)	A SICS is a singlet peak with enriched upstream doublets.
Stalling Supported Collision Site (SSCS)	A SSCS is a doublet peak with enriched downstream singlets.
nonSICS	A nonSICS is a singlet peak without enrichment of upstream doublets.

5.2 Methods

5.2.1 Terms and definitions

We use several terms throughout this chapter and we have collected their definition here in Table 5.1 for easy reference.

5.2.2 Collecting singlet and doublet footprints

Ribosome profiling was performed with minor alterations as described in [155]. Briefly, two samples were generated, each having a different final concentration of cycloheximide before the cells were frozen. One sample had no cycloheximde added, while the other had 10x cycloheximide (1 mg/ml of EtOH) added. Cells were then vacuum filtered away from the media and resuspended in polysome lysis buffer and frozen, dropwise, into N2(l). Cells were disrupted by rounds of manual vortexing using acid-washed glass beads in conjunction with ice water in between each round. Cell lysate was digested with 80U of RNase I for 50 minutes at room temperature before digestion was stopped by adding $10\mu l$ of SUPERase-In. The digests were then loaded onto a 1M sucrose cushion and centrifuged at 4°C for 4 hours at 70,000G. Acid-phenol-chloroform extraction was performed with the purified product used for library construction. mRNA samples were enriched in poly-A transcripts using Oligo(dT) and then fragmented by incubating in alkaline fragmentation buffer. For mRNA 27–40nt size fractions were recovered, while sub-22nt fragments, 27–34nt, and 60–80nt size fractions were recovered for RPF samples. This range of RPF fragment sizes is notable as it is from this step that we generated both the monosome and doublet read populations. RNA was dephosphorylated using T4 PNK without ATP, followed by ligation of a universal miRNA linker using truncated T4 RNA ligase 2. For the RPF libraries, a subtractive hybridization step was included to remove ribosomal RNA sequences from circularized cDNA. Circularized cDNA templates were amplified by 12 cycles of PCR using Phusion-polymerase with primers incorporating barcoded library sequences.

5.2.3 Data preprocessing

Raw read trimming and barcode collapsing

Raw read sequences share the following format:

SEQ–CLONELINKER–BARCODE–ILLUMINALINKER

The barcode is to tag duplicate reads caused by sequencing bias. If two reads share both the exact footprint sequence and barcode, they are both products of PCR bias, and should not be counted twice. Therefore ribosome footprint sequences and the barcode sequences are first extracted with cutadapt [113], then reads that share the same footprint sequence and barcode are collapse, leaving only one record per footprint sequence. There are three steps in preprocessing the raw reads:

First, ribosome footprints are extracted by trimming the 3' adapter as the cloning linker sequence (CLONELINKER = CTGTAGGCACCATCAAT). Second, barcodes are extracted by simultaneously trimming the 5' adapter as the cloning linker sequence, and 3' adapter as the

Iluumina linker sequence (ILLUMINALINKER=AGATCGGAAGAG). Third, barcodes are collapsed as follows: both ribosome footprints and barcodes are streamed in, and only footprints with a barcode not seen so far are streamed out.

Aligning reads

Pre-trimmed reads from the previous step are first aligned to the yeast noncoding RNA reference, which includes rRNA, tRNA, snoRNA, etc., to remove noncoding contaminants. The remaining reads are then mapped to the yeast transcriptome. The yeast noncoding RNA reference and the transcriptome reference are downloaded from the Saccharomyces Genome Database [51]. Alignments are performed with STAR [44] with parameters `--outFilterMismatchNmax 1`, which automatically 'softclips' the first few unaligned bases and any unaligned bases at the 5' end of the reads, allowing at most 1 mismatch.

Generating ribosome profiles

Singlet A-site profiles are computed via uniquely mapped reads from singlet libraries. We use RiboDeblur [170], a method pipeline that automatically selects the appropriate range of read length and A-site offsets, to process these reads and output ribosome A-site profiles for each transcript.

Doublet read-start profiles are computed via uniquely mapped reads from doublet libraries. Only reads with length between 57 and 62 are included to estimate read start pileups at each nucleotide location. These reads have clear 3-nt periodicity for both our data sets and Guydosh and Greens' [69], which is evidence that they might represent real ribosome footprints.

Codon-position level profiles for both singlets and doublets are computed from the nucleotide-level profiles described above. Footprint counts from all 3 frames are summed together to get the codon-position level counts.

High coverage filtering

A transcript is included for downstream analysis if more than 50% of the codon-loci in its singlet profile have ribosome counts greater than 1.

5.2.4 Identifying co-occurrence of ribosome collision and ribosome stalling

We check co-occurrence of ribosome collision and ribosome stalling by identifying stalling induced collision sites (SICS) and stalling supported collision sites (SSCS). These two types of sites are two views of overlapping between ribosome collision and ribosome stalling. SICS is in the view of singlet profiles, where singlet peaks are coupled with enrichment of upstream doublets. SSCS is in the view of doublet profiles, where doublet peaks are coupled with enrichment of downstream singlets. The concepts of SICS and SSCS are illustrated in figure 5.1.

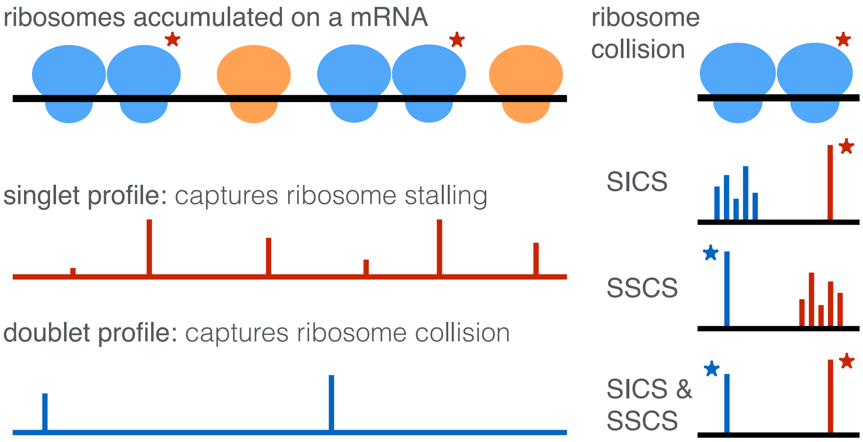

ribosomes accumulated on a mRNA
ribosome collision
singlet profile: captures ribosome stalling
doublet profile: captures ribosome collision
SICS
SSCS
SICS & SSCS

Figure 5.1: Illustration of simultaneous ribosome stalling and ribosome collision on ribosome profiles. Stars indicate peaks in profiles, singlet profiles are marked in red, and doublet profiles are marked in blue. SICS (stalling induced collision sites) are singlet peaks with enriched upstream doublets; SSCS (stalling supported collision sites) are doublet peaks with enriched downstream singlets; SICS & SSCS are doublet peaks with downstream singlet peaks.

Identifying stalling-induced collision sites (SICS)

From singlet profiles, stalling induced collision sites are singlet peaks with upstream doublet enrichment. They represent extreme ribosome stallings that lead to ribosome collision. Calling SICS are carried out in the following two steps:

Step I: Identifying potential ribosome stalling sites. Let $\boldsymbol{C_t}$ be the ribosome count vector of transcript t. We treat such a count vector as a distribution and compute its median (median($\boldsymbol{C_t}$)) and standard deviation (std($\boldsymbol{C_t}$)). A footprint count on a codon location i (c_{ti}) is identified as a potential stalling site if $c_{ti} > \text{median}(\boldsymbol{C_t}) + 2 \times \text{std}(\boldsymbol{C_t})$.

Step II: Identifying doublet enrichment. A stalling induced collision site is a potential ribosome stalling site with more upstream doublet count than expected by chance. The upstream doublet count is defined as the sum of doublet counts within a window of 13 – 17 codon positions upstream of the current stalling site.

The reason for such a window size is two fold: First, the typical doublet size is 58nt long [69]. Assuming that doublet footprints with such lengths are digested perfectly, then the A-site of the collision-causing ribosome should be 45nt away from the doublet read start, i.e.15 codon positions away. However doublet footprints are not always perfectly digested. Therefore, secondly, we allow a 2-codon window both before and after this 15th codon position, so that all potential ribosome collisions (captured by the imperfectly digested doublet reads) are included.

The enrichment of the upstream doublet count is compared with an empirical background doublet count. Since the foreground upstream doublet count is the sum of doublet counts within a 5-codon window upstream of the potential stalling sites, to generate a background doublet counts for a specific transcript, 5 codon locations are randomly selected to compute the doublet sum, and such sum is computed for 10,000 times. The significance of the upstream doublet count enrichment is measured by the empirical p-value: the fraction of times where the tested doublet count is smaller or equal to the background doublet count. The tested doublet count is called significant if p-value ≤ 0.1. To sum up, a stalling induced collision site (SICS) is a potential stalling site with significant enrichment of upstream doublets.

Identifying stalling supported collision sites (SSCS)

From doublet profiles, stalling supported collision sites are doublet peak loci with enriched singlets downstream. They represent high-confident ribosome collisions that are supported by ribosome stallings observed from singlets. In SSCS calling, both peak calling and enrichment testing are identical to the procedure in SSIC calling. SSCS are thus identified symmetrically to SICS, but in the view of doublet profiles. First, doublet peaks are identified from doublet profiles, and second, the sum of singlets within 13 – 17 codons downstream of a doublet peak is compared with an empirical background of singlet sums.

As described above, stalling induced collision sites are singlet peak loci with enriched upstream doublets. The underlying assumption is that ribosome collisions (doublet enrichment), are induced by the observed stallings (singlet peaks). Similarly, the rationale for the name SSCS is: Doublet peaks are potential non-random ribosome collisions, since they are loci accumulated with more doublets. These sites are thus more likely to be real programmed ribosome collisions that do not simply occur by chance. Theoretically, ribosome collisions should be caused by local slow-down of ribosome movements, and an enrichment of singlet downstream of these doublet peaks should be expected. In other words, downstream singlet enrichments, i.e. evidence of stallings, are *supporting* the observed doublet peaks.

In summary, SICS are identified on singlet profiles, where collisions are doublet enrichment, and stallings are singlet peaks; SSCS are identified on doublet profiles, where collisions are doublet peaks, and stallings are singlet enrichment. Both SICS and SSCS are used to study the relationship between ribosome stalling (measured by singlets) and ribosome collision (measured by doublets).

5.2.5 Codon usage analysis

We test whether SICS have unique codon composition by comparing SICS codon usage with the background and the codon usage of regular stalling sites without enrichment of upstream doublets — we call these sites nonSICS. Specifically, for a set of tested peaks (e.g. SICS), both a background and a foreground codon usage are computed. The background codon usage is the codon occurrences among all included transcripts, and the foreground codon usage is the codon occurrences within the peak set. SICS foreground codon usage is compared with both SICS background codon usage and nonSICS foreground codon usage. Chi-square test is used to quantify the similarity between two codon usage distributions. Since chi-square is sensitive to sample size, all codon usages are first adjusted to be on the same level as the codon usage of SICS, such that the sum of codon counts for each group of codon usage is the same as the number of identified SICS. If the tested codon usage vector is C_t (e.g. from SICS foreground), and the compared codon usage vector is C_c (e.g. from nonSICS foreground), then the chi-square is computed as: $\chi^2 = \sum_{i \in N} (C_t[i] - C_c[i])^2 / C_c[i]$, where N is the set of non-zero codons in C_c. A bigger chi-square indicates a larger deviance between the two compared distributions.

Individual codons are then examined for over- or under-represented in SICS. The ratio of the foreground and background codon usage of SICS are computed for any non-zero codons in the background. A codon is considered over-represented if such a ratio is > 2, and under-represented if such a ratio is $< 1/2$. These codon usage ratio after log transformation is plotted for all tested data sets.

5.2.6 RNA binding sites location analysis

Ribosome collisions might be caused by downstream RNA binding proteins (RBP) blocking the way of smooth translation, so we examine whether SICS are preferentially located upstream of RBPs. We compare the distance distribution of SICS and their closest downstream RBP binding sites with several choices of background distances. A smaller distance should be expected if SICS are preferentially located upstream of RBPs.

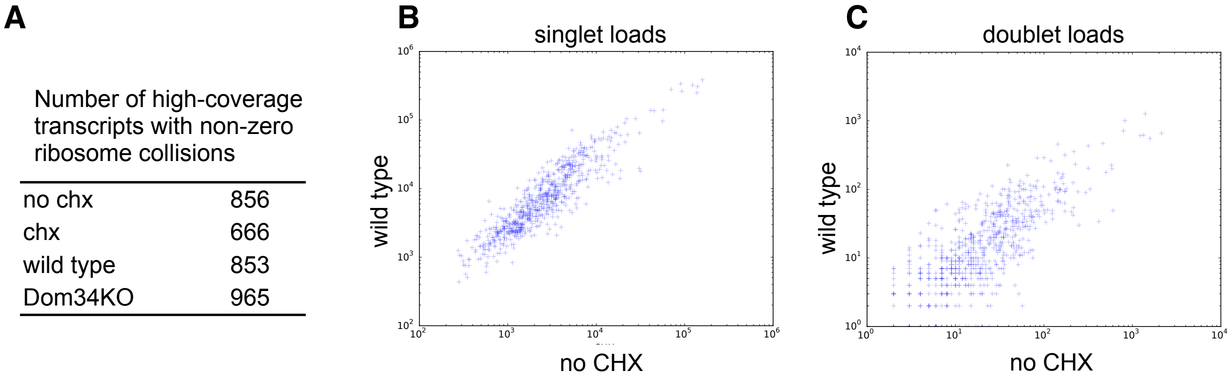
To get the distance distribution, first, thirteen RBP motifs are collected from a systematic motif finding paper in yeast [142]. Next, all motifs on the entire transcriptome are scanned using Biopython [33]. For motif calling, the background is based on a built-in 0^{th} order Markov model from Biopython, and the PSSM score is set with `threshold_balanced(1000)`, so the ratio between the false-negative rate and the false-positive rate is around 1000. The distance is then computed between a peak location in a ribosome profile and the closest downstream motifs from all RBP sites on a transcript. Lastly, such a closest distance distribution for SICS and their downstream motifs is collected and compared with the closest distances between SICS and their upstream motifs, nonSICS and their downstream motifs, and nonSICS and their upstream motifs.

5.2.7 Measuring reproducibility of SICS between samples

We test whether SICS are reproducible between different samples. Two samples are compared at a time: a tested sample and a compared sample. We match each SICS in the tested sample with the closest SICS in the compared sample, and we get a distance distribution of matching peaks. We then compare these matching distances with those computed on nonSICS and a random peak set. A smaller matching distance is expected if two SICS sets are reproducible.

For a given transcript, let S_t be the set of SICS in the tested sample, and S_c be the set of SICS in the compared sample. For each SICS in the tested sample s_t, the distance between s_t and each s_c in S_c is computed, and the closest distance is recorded. A distribution of closest distances for S_t compared to S_c is collected this way for all transcripts.

Such a closest distance distribution is compared with two other distributions: First, the closest distance distribution computed on SICS is compared with that on nonSICS. The closest distance is computed between the set of nonSICS in the tested sample N_t and the set of nonSICS in the compared sample N_c. Second, the distance distribution of two SICS sets is compared with that of two randomly sampled 'peak' sets. To generate random 'peaks' of a given transcript, if there are n_t SICS in the tested sample and n_c SICS in the compared sample, then two sets of codon locations are randomly selected with size n_t and n_c, and the two sets are called R_t and R_c



Number of high-coverage transcripts with non-zero ribosome collisions

no chx	856
chx	666
wild type	853
Dom34KO	965

Figure 5.2: Correlation of singlet and doublet loads across different conditions. (A) number of highly covered transcripts for different samples. (B) scatter plot of singlet loads for all highly-covered transcripts between our sample without cycloheximide treatment and Guydosh and Green's wild type sample. (C) scatter plot of doublet loads for all highly-covered transcripts between no CHX sample and wild type sample.

respectively. The closest distances are computed between R_t and R_c, and a distance distribution is collected for all random sets of all transcripts.

5.3 Results

We studied ribosome collision both globally and individually. Globally, we observe doublet loads to be highly correlated with singlet loads on each transcript, and doublets are depleted near the start codon. Individually, we observe ribosome collisions do not often overlap with ribosome stallings. Further, co-occurrence of ribosome collision and ribosome stalling do not seem to be different from regular stallings in terms of both codon usage and location preferences to RBP sites. Lastly, co-occurrence of ribosome collision and ribosome stalling do not seem to be reproducible across different tested conditions. All of the observations above support the hypothesis that ribosome collisions occur randomly.

5.3.1 Ribosome collisions are rare and proportional to ribosome loads

We generate both doublet and singlet libraries with modified ribosome profiling to quantify ribosome collisions. Two samples are prepared, one treated with cycloheximide and one without, and good correlation of both singlet loads and doublet loads are observed between the two samples. Only transcripts with high coverage (see Sec 5.2.3) are included to reduce the effect of noise. This results in about 700 – 900 transcripts to be included in each condition (Fig 5.2A). The singlet loads (see def. in Sec 5.2.1) from the sample without cycloheximide treatment are almost perfectly correlated with the cycloheximide treated sample (Pearson $r \simeq 1$, Spearman $r = 0.99$). The doublets loads (see def. in Sec. 5.2.1) from the two conditions also correlate strongly (Pearson $r = 0.98$, Spearman $r = 0.73$). Consistent with previous studies [174], these results show that cycloheximide does not affect ribosome loads at the gene level. They also show

Table 5.2: Pearson r between singlet loads and doublet loads.

no chx	0.94
chx	0.91
wild type	0.90
Dom34KO	0.86

Table 5.3: Average collision rate (%) on all tested data sets.

no chx	0.90
chx	0.11
wild type	0.27
Dom34KO	0.66

good reproducibility of ribosome loads among biological replicates we generated.

To validate our results, we also used another published ribosome profiling data set with available doublet libraries [69]. The original purpose of that study was to understand the effect of Dom34 in living cells, so the Dom34 knock out strand (Dom34KO) is compared against the wild type strand using ribosome profiling. Ribosome loads between samples from this study are also well reproduced: Singlet loads from Dom34KO correlate near perfectly with wild type (Pearson $r = 0.99$, Spearman $r = 0.99$), and doublet loads from the two strands also correlate almost perfectly (Pearson $r = 0.97$, Spearman $r = 0.90$). Since the correlation level from our lab is comparable with Guydosh and Green's [69], this provides another line of evidence that successful experiments should lead to good reproducibility of ribosome loads within labs.

Ribosome loads are also reproducible across labs. We compare our sample without cycloheximide treatment with the wild type sample from Guydosh and Green [69], since these two samples are processed with a similar experimental pipeline. Singlets between the two libraries are highly correlated (Fig 5.2B, Pearson $r = 0.94$, Spearman $r = 0.90$), and doublets between the two are strongly correlated (Fig 5.2C, Pearson $r = 0.81$, Spearman $r = 0.74$). Together, these results show that our libraries are of good quality.

All tested samples consistently show that doublet loads are directly proportional to singlet loads: All libraries show a strong correlation between singlet loads and doublet loads (Table 5.2). That is, highly expressed transcripts with more abundant ribosome loads also tend to have more ribosome collisions. This is consistent with random occurrence of ribosome collisions: If ribosomes collide randomly, we would expect more ribosome collisions on transcripts with more ribosomes. Although the good correlation between singlet loads and doublet loads alone cannot indicate that ribosome collisions are random.

Nonetheless ribosome collision rates (see def. in Sec. 5.2.1) for all tested transcripts are very low across all conditions. Most transcripts have less than one doublet read observed over every 100 ribosome footprint reads (Table 5.3). This shows that doublet libraries have very low coverage compared to its companion singlet libraries. Therefore, the measurements of ribosome collision tend to have high variance across samples and are less reliable than quantification of singlet loads. In summary, our observation support that ribosome collisions are rare and are often likely to happen randomly.

5.3.2 Ribosome collisions are depleted near start codons

Ribosome collisions are depleted near the start codon for all tested conditions. Such a trend is observed by comparing the cumulative distribution of both singlet counts and doublet counts

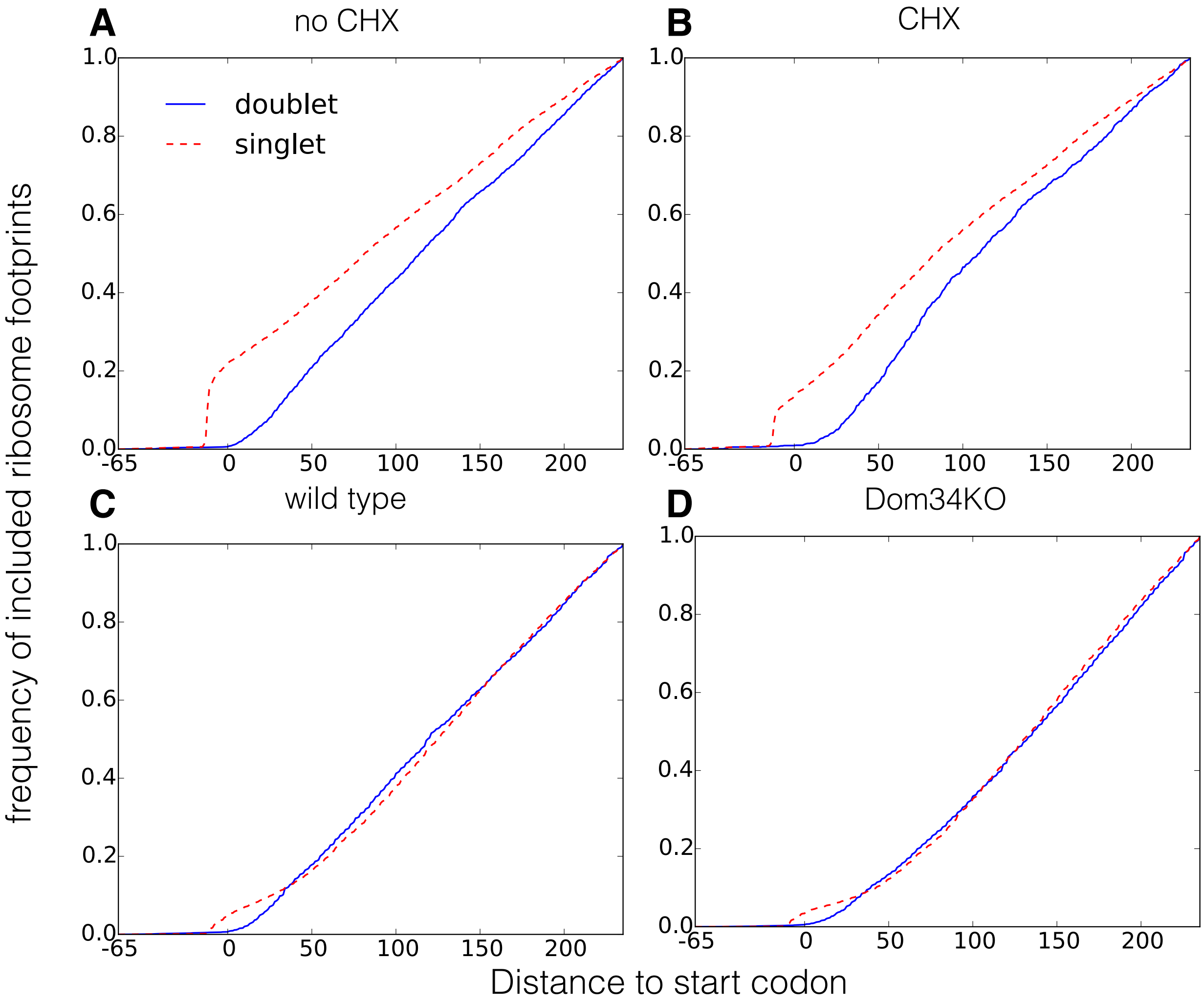


Figure 5.3: Cumulative distribution of ribosome footprint counts over transcript locations near start codon for different samples: (A) no CHX. (B) CHX. (C) wild type (D) Dom34KO. Cumulative distributions are computed based on the first 300 bases from all transcripts.

across all tested samples. As shown in Figure 5.3, singlet CDFs have a sharp increase near the 12^{th} positions before the start codon. This indicates an accumulation of ribosome footprints at the start codon (Here, read start positions are used to plot the CDFs, and the footprint start position will be at base 12 before the start codon while the ribosome P-site is at the start codon). However, the doublet CDFs increase gradually over the transcript positions. In fact, the CDFs are convex near the start codon, indicating a slow accumulation of doublets. Such a depletion of doublet near start codon is more apparent considering doublets are longer than singlets. While the doublet read start is at the start codon, the leading A-site of the two collided ribosomes are about 47 bases away from the read start (as explained in section 5.2.4), and the leading A-site is the study of interest since it is the exact location where ribosome collides. Therefore, the gradual increase of doublet CDFs at the start codon indicates a doublet depletion for the first 47 bases near the start codon.

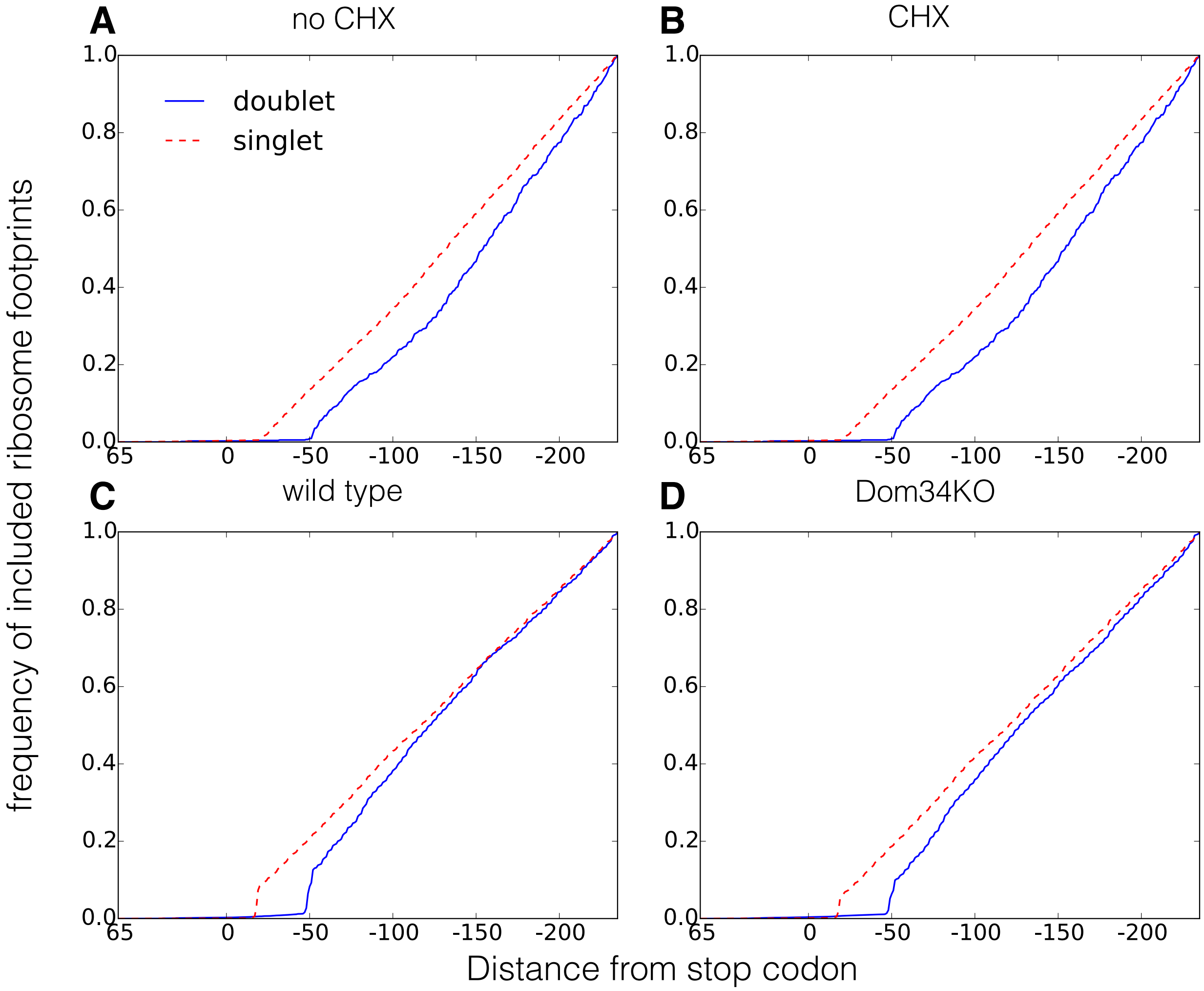


Figure 5.4: Cumulative distribution of ribosome footprint counts over transcript locations near stop codon for different samples: (A) no CHX. (B) CHX. (C) wild type (D) Dom34KO. Cumulative distributions are computed based on the last 300 bases from all transcripts.

Symmetrically, doublets stop accumulating about 50 bases before the stop codon, while singlets stop accumulating about 15 bases before the stop codon (Fig 5.4). This is consistent with the assumption above: When a ribosome is terminating on the transcript, its A-site is sitting on the stop codon, while the doublet footprint start is 50 bases away from the leading A-site, and the singlet footprint start is 15 bases away. However, unlike CDFs near the start codon, doublet CDFs and singlet CDFs near the stop codon shares a similar shape: There is an accumulation of ribosome pileups at the stop codon, but both CDFs increase gradually over the transcript locations before the stop codon. This indicates that there is no depletion of doublets near the stop codon.

Although all tested data set show a consistent depletion of ribosome collisions near the start codon, each lab has a unique singlet profile pattern. For example, the two singlet libraries generated by us show a clear enrichment near the start codon (Fig 5.3 A, B) but no enrichment near

Table 5.4: Overlaps between ribosome collision and ribosome stalling. Here SICS& SSCS are doublet peaks with downstream singlet peaks, and SICS& SSCS /SSCS are the percentage of doublet peaks with downstream singlet peaks among all doublet peaks.

	SICS	SICS/ singlet peaks	SSCS	SSCS/ doublet peaks	SICS& SSCS	SICS& SSCS /SSCS
no chx	582	4.73%	69	11.79%	34	49%
chx	332	2.57%	1	2.63%	1	100%
wild type	889	5.68%	139	13.61%	93	67%
Dom34KO	1410	7.68%	358	13.71%	265	74%

the stop codon (Fig 5.4 A, B). On the other hand, the two singlet libraries generated by Guydosh and Green [69] show a clear enrichment near the stop codon (Fig 5.4 C, D) but no enrichment near the start codon (Fig 5.3 C, D). These lab-specific features indicate a limited reproducibility of ribosome profiles, and further experiments are needed to validate the observation of doublet depletions near the start codon.

5.3.3 Ribosome collisions do not often overlap with ribosome stallings

We further study individual ribosome collision events on each transcript. Specifically, we test whether ribosome collisions often co-occur with ribosome stallings. Ribosome stalling is identified as an enrichment of singlet reads, and ribosome collision is identified as an enrichment of doublet reads. The co-occurrence of ribosome stalling and ribosome collision are detected from two angles. From the view of singlets, stalling induced collision sites (SICS) are detected as singlet peaks (stalling) with upstream enrichment of doublets. From the view of doublets, stalling supported collision sites (SSCS) are detected as doublet peaks (collision) with downstream enrichment of singlets (Figure 5.1). We developed a statistical framework to identify enrichment of footprint reads in both SICS and SSCS. Instead of detecting coupled footprint peaks in both singlet and doublet libraries, we choose to detect footprint peaks in one library coupled with footprint enrichment in another. This is because imperfect digestion in ribosome profiling might blur the locations of peaks in ribosome profiles, which might impede simultaneous peak detections in both libraries. More details of the SICS and SSCS calling is described in section 5.2.4.

Admittedly, both SICS and SSCS calling depend on their definitions. For footprint enrichment testing, a *distance* needs to be chosen upstream or downstream of the tested peak, a *window size* needs to be chosen to summarize the footprint counts within such a window, and a *p-value cutoff* needs to be chosen to call significance. We notice under a lenient p-value cutoff (0.1), the sets of SICS with different distances given a fixed window size are very similar with each other. Similarly, the sets of SICS with different window size given a fixed distance are also largely shared. Therefore, our results will likely to hold under alternative choices of distance and window size. We thus define SICS and SSCS with first principles, as described in section 5.2.4.

Although stalling events are not rare, only a small portion of singlet peaks are SICS (Table 5.4). Among all of the tested data sets, Dom34KO library has the highest percentage of

Table 5.5: Chi-square statistics between pairs of codon usage distributions

	SICS foreground vs. background	nonSICS foreground vs. background	foreground SICS vs. nonSICS	background SICS vs. nonSICS
no chx	206*	151*	56	1
chx	166*	82*	67	1
wild type	250*	224*	78	1
Dom34KO	387*	323*	72	1

* p-value < 0.05.

SICS among singlet peaks, which is still less than 8%. This shows that singlet peaks are rarely enriched with upstream doublets. Similarly, only a small portion of doublet peaks are SSCS: all tested libraries have about 10% doublet peaks as SSCS. Compared with other data sets, our cycloheximide treated sample have very few identified SICS and SSCS. This is likely due to severe contamination, which lead to a low coverage of the doublet library. Analogous to SICS, doublet peaks are also rarely enriched with downstream doublets. Nevertheless, the majority of identified SSCS have downstream singlet peaks (Table 5.4): most tested data sets have more that 50% SSCS with downstream singlet peaks. This indicates that the two angles of detecting co-occurrence of ribosome collision and ribosome stalling — from the view of singlet profiles and from the view of doublet profiles — are often consistent with each other. In short, we do not often observe ribosome collision overlapping with ribosome stalling.

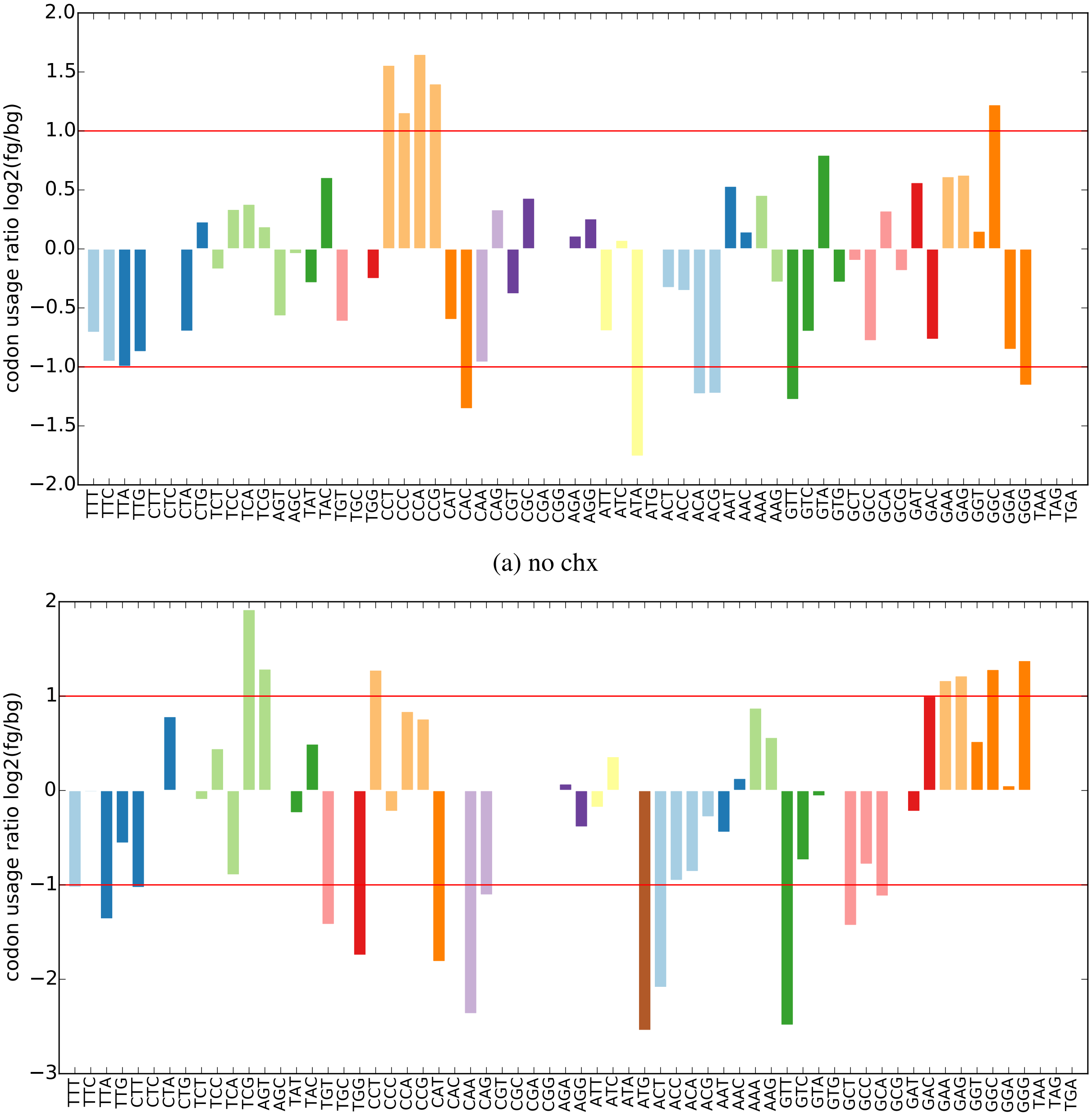
5.3.4 SICS shares similar codon usage with nonSICS

SICS do not share a unique codon usage compared to regular stalling sites without enrichment of upstream doublets (nonSICS). Codon usage in SICS are compared with the background codon usage and codon usage in nonSICS. Chi-square is used to quantify the similarity between two codon usage distributions after controlling for sample size (more details in section 5.2.5). Although bigger than two backgrounds computed from different sets of transcripts, the Chi-square statistics of codon usage between SICS and nonSICS are much smaller than that between SICS and its background. In fact, both SICS and nonSICS have significantly different codon usage compared to the background (Table 5.5). Further, such an observation is consistent across all tested data sets. This shows that the codon usage of SICS are not significantly different from regular stalling sites.

However, codon usage preferences seem to be unique to each data set. To test whether specific codons are over- or under-represented in SICS, the ratio between the foreground and the background codon frequencies of SICS are computed. These ratio patterns are very different among different samples. For example, in our sample without cycloheximide treatment, all proline codons in SICS are observed more than twice as frequently compared to background, while ATA (Isoleucine) is observed less than half of the times. On the other hand, SICS in Guydosh and Green's data are enriched with Arginine codons and depleted with Leucine codons (Figure 5.6). Along with the previous observations, this shows that codon usage in SICS are unique to experimental conditions, and such a codon preference are likely due to common trends in ribosome stalling sites.

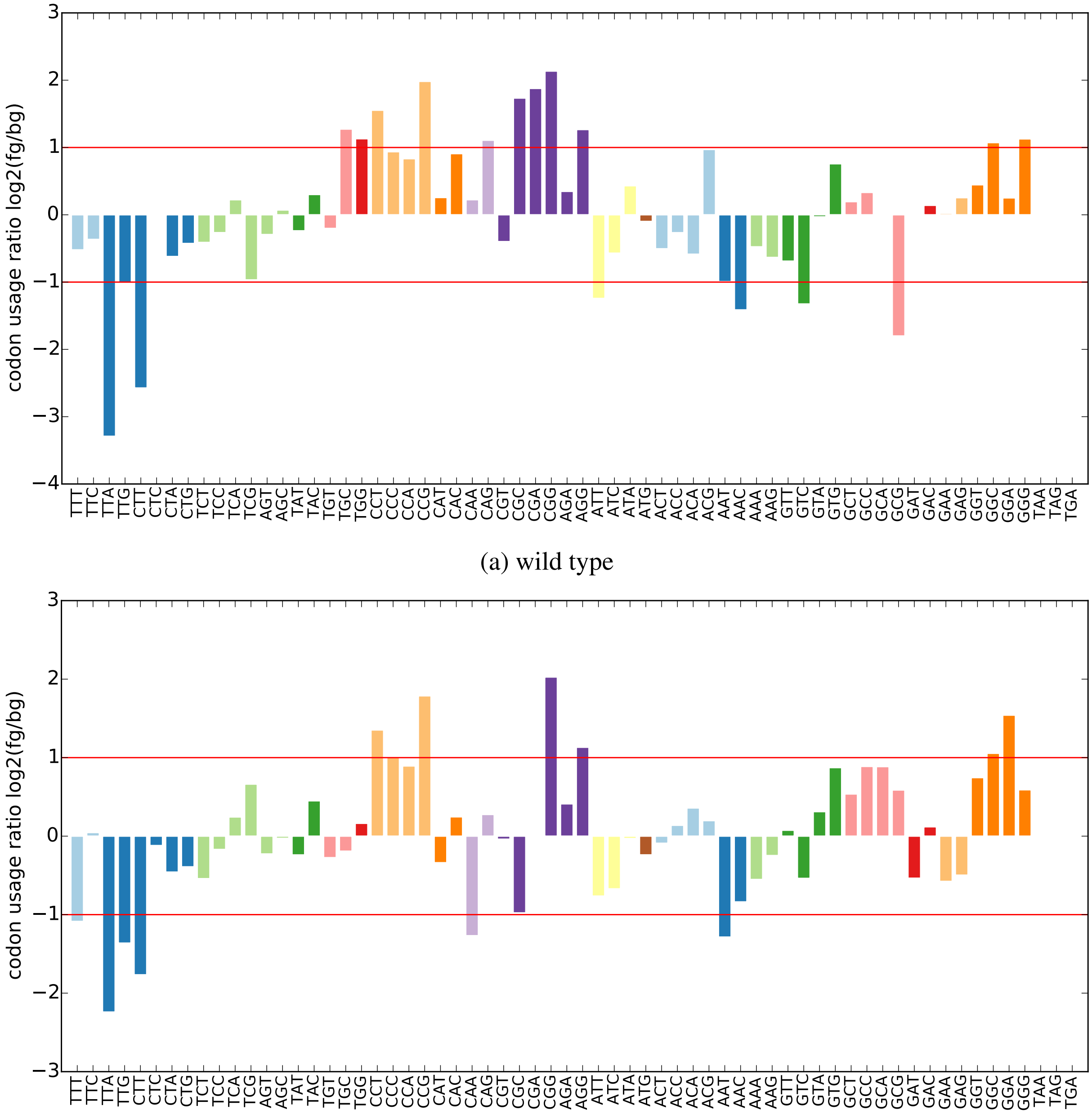
5.3.5 SICS do not preferentially locate upstream of RNA binding sites

One potential reason for ribosome collision is RNA binding proteins (RBP) blocking the way of ribosome movements. However SICS do not seem to be more likely to appear upstream of a RBP site. To test such a preferential location of SICS, all RBP sites on the transcriptome are iden-



(a) no chx

(b) chx



(a) wild type

(b) Dom34 Knock Out

Figure 5.6: Ratio of foreground and background codon frequencies of SICS

tified, and the closest downstream and upstream motifs to both SICS and nonSICS are recorded (more details in section 5.2.6). The distance distribution of SICS and its closest downstream motif is not significantly different from the distance distribution of SICS and its closest upstream motif according to the Kolmogorov-Smirnov test, nor is it different from the distance distribution between nonSICS and its closest downstream motif (Figure 5.7). Further, there does not seem to be dominant RBP motifs from downstreams of SICS (Figure 5.8). Such results again consistently hold across all tested data sets. This provides evidence that SICS do not preferentially

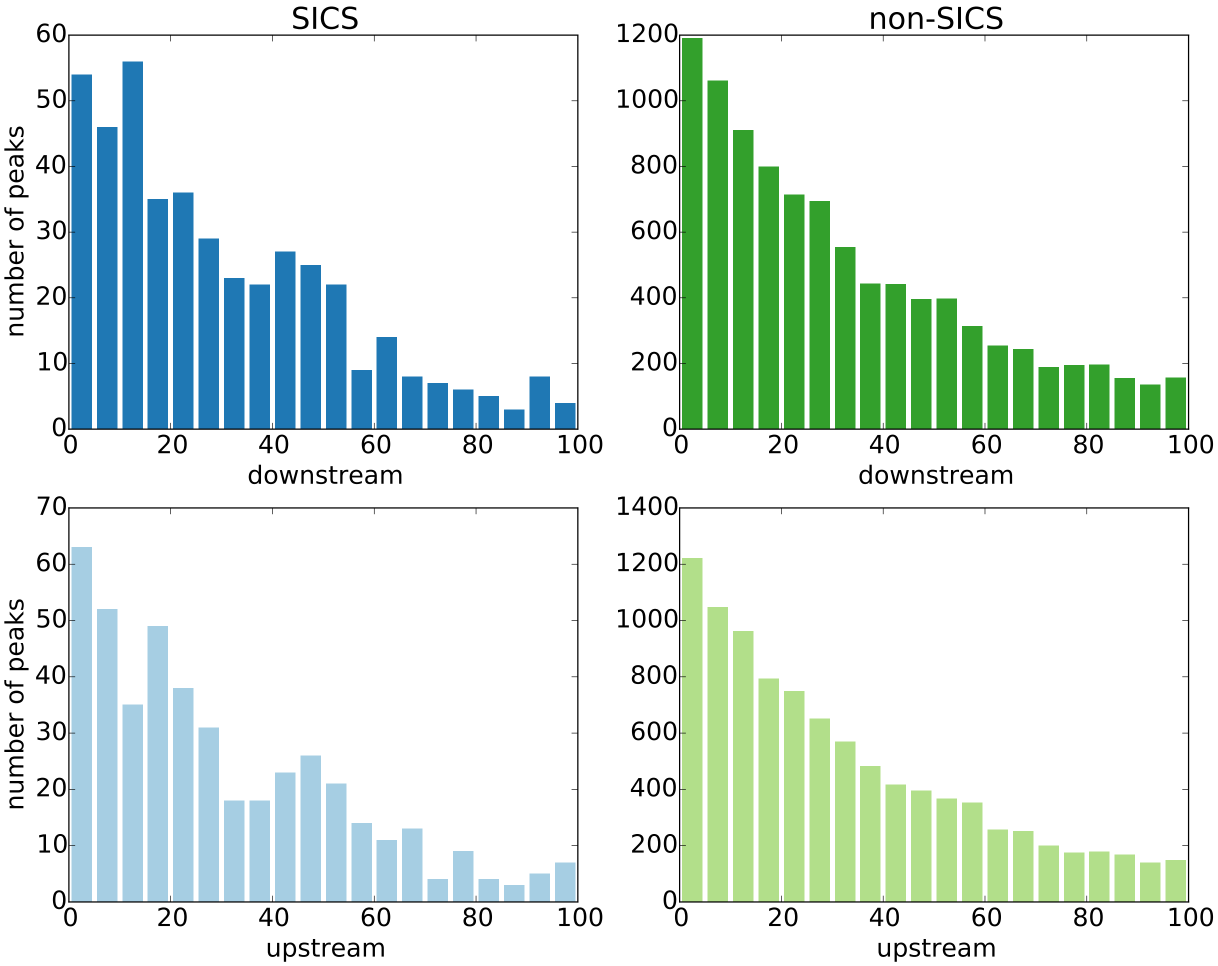


Figure 5.7: Distance distribution of SICS/nonSICS to its closest RBP motif. Distributions are computed from the no chx sample. The distribution of SICS and downstream motifs is not significantly different from the distribution of SICS and upstream motifs (KS statistics $= 0.03, p = 0.99$), nor is it different from the distribution of nonSICS and downstream motifs (KS statistics $= 0.04, p = 0.33$). Other samples share qualitatively similar results.

locate upstream of RBP sites.

5.3.6 SICS locations are unique to experimental conditions

We test whether SICS detected from one condition can also be observed in another condition. Again, we compare our no-cycloheximide sample with Guydosh and Green's wild type sample. For each of our SICS, the closest distance to SICS in wild type is computed. The closest distances of nonSICS between the two samples are also computed (more details in section 5.2.7). The distances of SICS are significantly larger than the distances of nonSICS (Figure 5.9, Mann-Whitney U test $p = 3.01 \times 10^{-100}$). If two SICS between two samples are considered *matched* or reproducible when they are within 10 codons away, only 13% SICS in no-cycloheximide

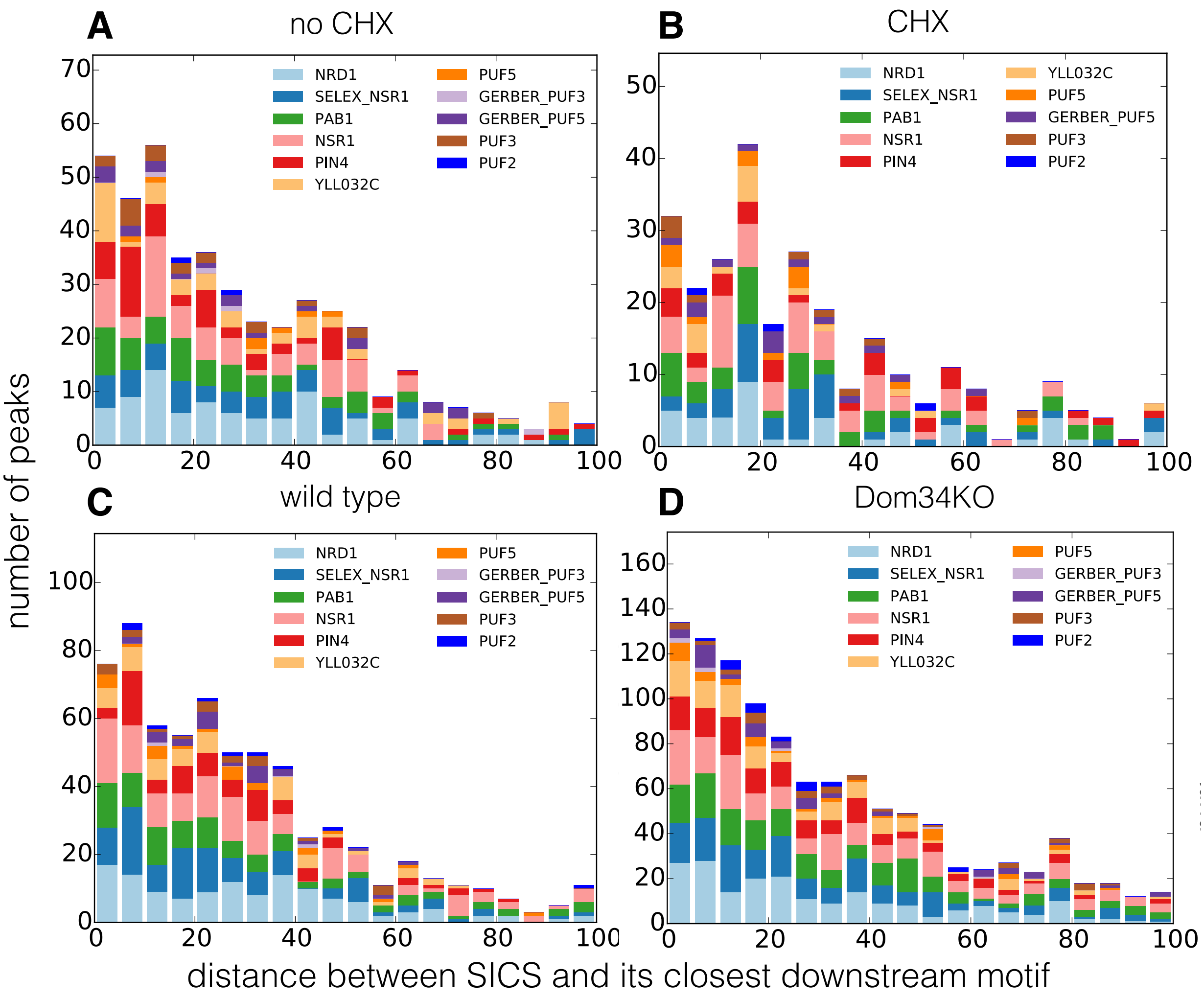


Figure 5.8: Distance distribution of SICS to its closest downstream RBP motif. All tested data sets are shown, different motifs are marked with different colors.

have a matched SICS in wild type. On the other hand, 62% nonSICS in no-cycloheximde have a matched nonSICS in wild type. Other pair-wise sample comparisons all have qualitatively similar results. This indicates that SICS are unique to experimental conditions.

The longer distances between the closest two SICS might simply due to a smaller sample size: if peak positions are randomly drawn, a shorter distance is expected if more points are included. As mentioned above, SICS only count for about 5% of all singlet peaks, so nonSICS have a much bigger set size compare to SICS. To test whether the longer distance is indeed caused by two sparser set of points, we randomly generate two sets of peak locations and compute the closest distances between the two (more details in section 5.2.7). The distance distribution of the random peaks are qualitatively the same as the distance distribution of SICS (KS statistics $= 0.06, p = 0.57$). Further, increasing the sample size indeed makes the distance distribution skew towards a shorter value (Figure 5.9). This shows that the distance distribution of SICS and nonSICS are indistinguishable from the distance distribution of randomly drawn peak locations, thus supporting that random occurrences of ribosome collisions.

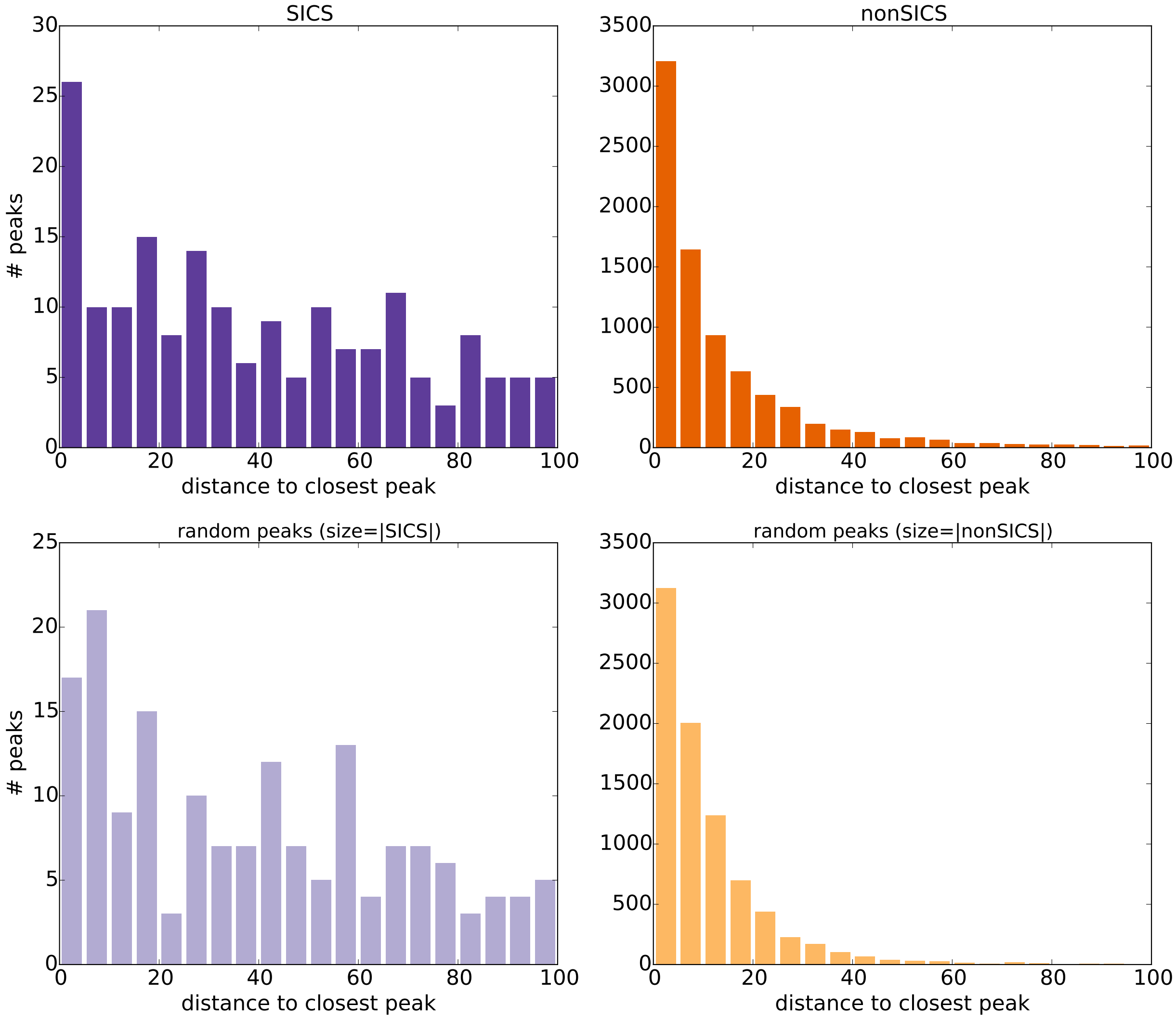


Figure 5.9: Matching SICS of wild type to no chx sample.

It is desirable to test whether SICS are reproducible between biological replicates. However, we do not have ribo-seq samples from biological replicates. Our no-cycloheximide sample shares a similar experimental protocol as Guydosh and Green's wild type, and these two data sets are the closest we can find as biological replicates. However, the two samples are from different yeast strains, and the experiments are conducted by different labs, both of which might indeed lead to unique patterns of ribosome collision.

5.4 Discussion

We developed an experimental and computational pipeline to systematically capture ribosome collisions in yeast. This is the first genome-wide analysis to quantify extreme local ribosome slow downs. From several lines of evidence, our analysis consistently suggests that ribosome collisions might occur randomly.

Interestingly, we do observe a global trend of doublet depletion near the start codon among all tested data sets: Cumulative distribution of doublet counts shows a near-zero coverage for the first 50 bases of all transcripts (Figure 5.3). One potential explanation is that the engaging initiation factors during the initiation process will block some space on the mRNA [69], thus disabling ribosome collisions near the start codon. It is also possible that since initiation might be much slower than elongation [67, 79, 168], ribosomes will be naturally spaced out so that ribosome collisions are avoided at the beginning of a transcript. But such a spacing effect is reduced as ribosomes move down the transcript, and ribosomes that enter the transcript later start to catch up and bump into ribosomes in front of them.

Unfortunately, reads from doublet libraries can be rarely mapped to the transcriptome. For singlet libraries, since real singlet footprints share a strikingly different size signature from contaminants, contaminant sequences such as ribosomal RNAs can be effectively removed via a physical size selection procedure. This results in about 50% of raw reads as true ribosome footprints (under the assumption that reads that can be mapped to the transcriptome are 'true ribosome footprints'). On the contrary, doublet footprints share a similar size as contaminant sequences, thus contaminants are the major components of doublet libraries. Table 5.6 shows that more than 90% of raw reads from doublet libraries can be mapped to ribosomal RNAs, tRNAs, and other non-coding RNAs. Besides these mappable contaminants, there are often a small portion of reads (less than 5%) that cannot be mapped to any non-coding RNAs, mRNAs, or the genome.

The high portion of contaminants in doublet libraries is problematic. For singlet libraries, contaminants are harmless since the true signal, i.e., the percentage of true ribosome footprints, is stronger or comparable to contaminants, therefore ribosome footprints are likely to be more accurately quantified. However, for the same amount of raw cell material, doublet libraries often have less than 1% potential true doublet footprints. Even noise from non-mappable contaminants is stronger than the doublet footprints signal. This might lead to a false conclusion that ribosome collisions are rare during translation, while it might very well be that ribosome collisions are prevalent on many mRNAs, but the current library preparation protocol cannot effectively capture them. Moreover, these low coverage doublet libraries can also impede an accurate observation and estimation of ribosome collision events. This is a clear call for a better experimental protocol for better observing extreme local ribosome slow-downs.

From a statistical framework for identifying co-occurrence of ribosome collision and ribosome stalling from two angles, we do not observe much overlapping between ribosome collision and ribosome stalling. This supports the hypothesis that ribosome collisions happen by chance. It is possible that ribosome collisions are indeed often coupled with ribosome stalling, but the current doublet libraries do not have enough sequencing depth to support that. It is also possible that abundant ribosome collisions will mask out severe ribosome stalling, thus ribosome collisions and ribosome stallings will happen individually instead of simultaneously. Either hypothesis requires a better designed doublet profiling technique to reliably identify ribosome collisions.

Different experimental conditions also exhibit different codon usage preferences. Additionally, codon usage of SICS is very similar to codon usage of regular stalling sites without ribosome collisions. This is also inline with random occurrence of ribosome collisions. However, individual codons alone might not be sufficient to characterize ribosome collision patterns. A

Table 5.6: Read mapping statistics for all tested data sets. Contaminants refer to mappable contaminants from rRNA, tRNA, and other non-coding RNAs. Unmappable contaminants refer to reads that cannot be mapped to the genome.

Singlets	Raw	Contaminants	Transcriptome	Genome
no chx	11,703,032	5,209,671 (44%)	6,308,169 (54%)	11,556,933 (99%)
chx	11,252,157	4,644,867 (41%)	6,422,036 (57%)	11,099,571 (99%)
wild type	34,710,064	16,809,904 (48%)	16,357,454 (47%)	33,038,058 (95%)
Dom34KO	29,584,871	13,520,796 (46%)	15,258,457 (52%)	28,550,877 (97%)
Doublets	Raw	Contaminants	Transcriptome	Genome
no chx	6,253,737	5,927,522 (95%)	127,794 (2%)	6,129,541 (98%)
chx	3,337,230	3,226,506 (97%)	29,033 (0.9%)	3,289,686 (99%)
wild type	69,277,151	63,188,638 (91%)	686,060 (1%)	63,935,303 (92%)
Dom34KO	89,062,886	82,403,395 (93%)	448,064 (0.5%)	82,834,707 (93%)

recent study shows that interactions between adjacent codon pairs might strongly affect translational efficiency [56], thus codon pair might be the right unit to study sequence patterns in ribosome collision. Yet the very rare co-occurrence between ribosome stalling and ribosome collision hinders such analysis. In fact, almost all codons from all tested data sets have less than 50 co-occurrences of ribosome collision and ribosome stalling. The current doublet data sets are therefore too sparse to support codon pair analysis.

We do not observe collision sites to be preferentially located upstream of RNA binding motifs. Admittedly, RNA binding protein blocking the way of ribosomes during translation is only one potential reason for ribosome slow-down. Other factors might also contribute to translational speed regulation. One alternative case is a complicated secondary mRNA structure, such as a hairpin, might substantially slow down ribosome movements. In addition, only 13 existing RBP motifs are tested. RBP sites might be hard to identify solely based on sequence contents due to the lack of a strong consensus motif. Nevertheless, our result at least shows that SICS do not preferentially locate upstream of 13 known RBP motifs, agreeing with the hypothesis that ribosome collisions might happen randomly.

Lastly, we do not observe consistent occurrences of SICS across different experimental conditions. This again supports that ribosome collisions tend to arise randomly. It would be interesting to test whether SICS can be reproduced between biological replicates, and lacking reproducibility between replicates will provide a stronger evidence for ribosome collisions being random. However, such a test is currently not available, and it should be a necessary next step for validating the reproducibility of ribosome collision sites.

Most of our analysis relies on the identification of stalling induced collision sites, which detects simultaneous enrichments of singlets coupled with upstream doublets. However, footprint accumulations can be biased by experimental protocols. For example, the reversible binding nature of cycloheximide is known to redistribute footprints on a transcript after translation being halted [76]. Anecdotal examples also show that RNA-seq shares a somewhat correlated coverage patterns with ribo-seq [115], indicating that ribosome profiling might share a similar sequencing bias from RNA-seq technology. These biases might distort the true pictures of ribosome pile-ups. True ribosome footprint accumulation might be missed, while false footprint peaks might be identified. Unfortunately, currently, there is no software or algorithm developed to systematically correct for sequencing biases for ribo-seq reads. Without correctly handling biases introduced during the experiments, we can only get a noisy observation of ribosome pileup patterns. Along with the low coverage of doublet libraries, it might be very difficult to confidently draw out real patterns of programmed ribosome slow-downs.

In summary, we provide a preliminary study on ribosome collision quantification. Our study consistently supports the hypothesis that ribosome collisions are likely to be random. However, to further validate our conclusion, or alternatively extract meaningful programmed ribosome stalling patterns, a better experimental protocol needs to be developed to effectively capture ribosome collisions.

Chapter 6

Conclusion and future work

In this thesis, we developed various statistical tests and mathematical solutions to help study gene regulation mechanisms from high-throughput sequencing data like chromosome conformation capture data and ribosome profiling data. Pre-transcriptionally, we designed a statistical framework to study whether functionally related elements also tend to be spatially close. Our test is crucial in understanding the role of chromatin's spatial arrangement in transcriptional regulation. Post-transcriptionally, we developed several methods and analysis to study key factors that regulate translational speed. Understanding the translational dynamics is important because it can help us understand the mechanism of protein production, and further the engineering of protein synthesis design.

6.1 Conclusion and limitations

We first study the relationship between chromatin spatial arrangement and its functionality in chapter 2. Specifically, we propose a statistical framework to identify spatially compact sets that are functionally related. We construct 3C interaction graphs directly from raw 3C interaction data, and we apply novel topological features for a chromatin spatial closeness test. Our approach therefore avoids the computation of a three-dimensional chromatin embedding that is often time consuming. In our statistical framework, our resampling scheme successfully controls for chromosome-specific patterns, thus outputs robust and unbiased p-value estimations. Our approach can accurately recognize spatially compact regions across multiple chromosomes, but misses cases where the spatially closed region is mainly located on a single chromosome. This is because of only inter-chromosomal interactions are included for constructing the 3C graph.

We then study translational dynamic regulations from ribosome profiling data. Ribosome profiling is a powerful technique to study translational dynamics by capturing ribosome locations during translation. It provides a picture of where ribosomes are over time. However, several challenges need to be addressed to convert raw ribo-seq reads to ribosome locations. In this thesis, we tackle two of the main challenges in ribo-seq read processing. Particularly, we presented an automatic pipeline in Chapter 3 to resolve multi-mappings in ribo-seq reads and output isoform-level ribosome profile estimation. This is the first systematic approach to handle multi-mappings caused by alternative splicing, which is shown to be the main cause for multi-mappings in mammalian ribo-seq data. By reasoning that transcript abundance partially determines the observed ribosome pileups, we assign ribo-seq reads to candidate locations with the guidance of transcript abundance estimation. Our approach accurately recovers the true ribosome profiles from synthetic ribo-seq data, and therefore improves overall transcript-level total ribosome load estimations. One of the remaining challenge in such a read-assignment scheme is to simultaneously model ribosome movements and estimate translation rate while assigning ribo-seq reads. The limitation of our method is that we only consider transcript abundance when resolving multi-mapping reads. We do not make any assumptions about ribosome motion while assigning ribo-seq reads.

We further tackle another challenge in converting ribo-seq reads to ribosome locations — estimating ribosome active sites from ribo-seq reads. In chapter 4, we show that imperfect digestion during the ribosome profiling experiment complicates the observed ribosome pileups. Further, imperfect digestions are very common among ribo-seq reads. We propose a signal processing approach based on blind deconvolution to recover the unknown clear ribosome positions from the blurred observations. Our estimated profiles are consistent across different read lengths, have a sharper 3-nt periodicity — a key feature for a clean ribosome pileup signal, and retain sub-codon resolutions for off-frame events detection. From these better estimated profiles, we compute the codon decoding time, and show that both tRNA abundance and wobble pairing affect translation speed — a conclusion that is inline with our prior understanding about translation, but was not observable from the original ribosome profiles. This shows that our method successfully unveils biologically meaningful insights from the noisy ribo-seq data. The limitation of our method is that we assume that each read length shares a universal digestion pattern, while other factors such as sequencing bias and read positions might also alter such a pattern.

Finally, in chapter 5, we provide the first genome-wide analysis on ribosome-collisions in yeast. This is the first attempt to study extreme local ribosome slow-downs during translation with the help of a modified ribosome profiling experiment that captures ribosome collisions. Our preliminary analysis suggests that ribosome collisions might be random and rare, and the sequencing depth in our experiments is not sufficient to capture strong features in programmed ribosome stalling.

6.2 Future work

For the problem of using topological properties of 3C graph to infer spatial proximity, incorporating 3C interactions within a chromosome into spatial compactness test still remains to be challenging. Normalization schemes are yet to be developed to embed these data into 3C graph without diluting the spatial closeness estimation from 3C interactions among different chromosomes.

For the problem of resolving multi-mappings in ribo-seq data, our method can be strengthened by incorporating position-wise ribosome occupancy probabilities into the framework. However, currently there is no universally recognized approaches to model ribosome movements, yet the choice of such a model will largely affect the results of assigning ribo-seq reads. Therefore, a valid model of ribosome occupancy probability will surely improve the estimation of ribosome profiles.

For the problem of recovering ribosome A-sites from ribo-seq reads, it would be beneficial to extend our approach to take into account other factors that might influence ribo-seq digestion patterns, such as sequencing bias and positional preferences. It would also be of great use to the community if de-convolving multi-mapping reads and recovering ribosome active site can be combined into a unified pipeline. A new objective function that captures both intents needs to be developed.

For the problem of quantifying ribosome collisions, one of the challenges in capturing ribosome collision is that the size of two collided ribosome back to back is similar to the size of typical contaminants in ribo-seq reads. Unlike the classical ribosome profiling, where contaminants can be easily removed by a physical size selection step, here, contaminants need to be targeted by special agents for removal. However, effectively removing these contaminants will substantially strengthen the amplification of the true signals during PCR. Further, if ribosome collisions are indeed rare, then the signal-to-noise ratio for detecting ribosome collisions is expected to be low, thus accurately quantifying ribosome collisions will be difficult. To amplify the signal-to-noise ratio, inducing ribosome stalling might be a necessary step in cell preparation. Therefore, any biological conclusions on programmed ribosome stalling determinants hinge on a better experiment design to effectively capture ribosome collisions, along with the reproducibility of the discoveries across multiple biological replicates.

Together, this thesis provides several methods to prepare two types of high-throughput sequencing data to study gene regulatory mechanisms. These methods are the necessary first step to explore mechanism of genome structure and translational control.

Bibliography

[1] Human tRNA gene copy number. 2009. `http://gtrnadb.ucsc.edu/Hsapi/Hsapi-summary-codon.html` Accessed: 2015-02-18. 3.1

[2] Eukaryotic tRNA fasta. 2009. `http://gtrnadb.ucsc.edu/download/tRNAs/eukaryotic-tRNAs.fa.gz` Accessed: 2015-02-18. 3.1

[3] GENCODE Human protein-coding transcript fasta file (version 18). 2013. `ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_18/gencode.v18.pc_transcripts.fa.gz` Accessed: 2015-02-18. 3.1

[4] GENCODE Human gene annotation GTF file (version 18). 2013. `ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_human/release_18/gencode.v18.annotation.gtf.gz` Accessed: 2015-02-18. 3.1

[5] GENCODE Mouse protein-coding transcript fasta file (version M4). 2014. `ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_mouse/release_M4/gencode.vM4.pc_transcripts.fa.gz` Accessed: 2015-02-18. 3.1

[6] GENCODE Mouse gene annotation GTF file (version M4). 2014. `ftp://ftp.sanger.ac.uk/pub/gencode/Gencode_mouse/release_M4/gencode.vM4.annotation.gtf.gz` Accessed: 2015-02-18. 3.1

[7] Florian Aeschimann, Jieyi Xiong, Andreas Arnold, Christoph Dieterich, and Helge Großhans. Transcriptome-wide measurement of ribosomal occupancy by ribosome profiling. *Methods*, 85:75–89, 2015. 1.2.2

[8] F. W. Albert, D. Muzzey, J. S. Weissman, and L. Kruglyak. Genetic influences on translation in yeast. *PLoS Genet.*, 10(10):e1004692, 2014. 1.2.1, 4.1.1, 4.1.2, 4.2, 4.3.7

[9] Dmitry E Andreev, Patrick BF O'Connor, Ciara Fahey, Elaine M Kenny, Ilya M Terenin, Sergey E Dmitriev, Paul Cormican, Derek W Morris, Ivan N Shatsky, and Pavel V Baranov. Translation of 5' leaders is pervasive in genes resistant to eiF2 repression. *Elife*, 4: e03971, 2015. 1.2.1

[10] Dmitry E Andreev, Patrick BF OConnor, Alexander V Zhdanov, Ruslan I Dmitriev, Ivan N Shatsky, Dmitri B Papkovsky, and Pavel V Baranov. Oxygen and glucose deprivation induces widespread alterations in mrna translation within 20 minutes. *Genome biology*, 16(1):1, 2015. 1.2.1

[11] Carolina Arias, Ben Weisburd, Noam Stern-Ginossar, Alexandre Mercier, Alexis S Madrid, Priya Bellare, Meghan Holdorf, Jonathan S Weissman, and Don Ganem. KSHV

2.0: a comprehensive annotation of the Kaposi's sarcoma-associated herpesvirus genome using next-generation sequencing reveals novel genomic and functional features. *PLoS Pathog*, 10(1):e1003847, 2014. 1.2.1

[12] C. G. Artieri and H. B. Fraser. Evolution at two levels of gene expression in yeast. *Genome Res.*, 24(3):411–421, 2014. 1.2.1, 4.1.1

[13] C. G. Artieri and H. B. Fraser. Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation. *Genome Res.*, 24(12):2011–2021, 2014. 1.2.1, 1.2.2, 1.1, 1.2.4, 3.3.3, 4.1.1, 4.4.5, 4.5

[14] Rohan Balakrishnan, Kenji Oman, Shinichiro Shoji, Ralf Bundschuh, and Kurt Fredrick. The conserved GTPase LepA contributes mainly to translation initiation in *Escherichia coli*. *Nucleic acids research*, page gku1098, 2014. 1.2.1

[15] Alexander Bartholomäus, Cristian Del Campo, and Zoya Ignatova. Mapping the non-standardized biases of ribosome profiling. *Biological chemistry*, 397(1):23–35, 2016. 1.2.2

[16] Davide Baù et al. The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nat. Struct. Mol. Biol.*, 18(1):107–114, 2010. 2.1, 2.1, 2.5.1

[17] Ariel A. Bazzini, Miler T. Lee, and Antonio J. Giraldez. Ribosome Profiling Shows That miR-430 Reduces Translation Before Causing mRNA Decay in Zebrafish. *Science*, 336 (6078):233–237, 2012. 1.2.1

[18] Ariel A Bazzini, Timothy G Johnstone, Romain Christiano, Sebastian D Mackowiak, Benedikt Obermayer, Elizabeth S Fleming, Charles E Vejnar, Miler T Lee, Nikolaus Rajewsky, Tobias C Walther, et al. Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *The EMBO journal*, page e201488411, 2014. 1.2.1, 1.1, 1.2.3

[19] Becker, Annemarie H and Oh, Eugene and Weissman, Jonathan S and Kramer, Günter and Bukau, Bernd. Selective ribosome profiling as a tool for studying the interaction of chaperones and targeting factors with nascent polypeptide chains and ribosomes. *Nature Protocols*, 8(11):2212–2239, 2013. 1.2.1

[20] Ben-Elazar, Shay and Yakhini, Zohar and Yanai, Itai. Spatial localization of co-regulated genes exceeds genomic gene clustering in the *Saccharomyces cerevisiae* genome. *Nucleic Acids Res.*, 41(4):2191–2201, 2013. 1.1.1, 2.1, 2.2

[21] Botao Liu and Yan Han and Shu-Bing Qian. Cotranslational Response to Proteotoxic Stress by Elongation Pausing of Ribosomes. *Molecular Cell*, 49(3):453–463, 2013. 1.2.1

[22] G. A. Brar et al. High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science*, 335(6068):552–557, 2012. 1.2.1, 4.1.1

[23] Gloria A Brar and Jonathan S Weissman. Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nature Reviews Molecular Cell Biology*, 2015. 1.2.1

[24] Lorenzo Calviello, Neelanjan Mukherjee, Emanuel Wyler, Henrik Zauber, Antje Hirsekorn, Matthias Selbach, Markus Landthaler, Benedikt Obermayer, and Uwe Ohler.

Detecting actively translated open reading frames in ribosome profiling data. *Nature methods*, 2015. 1.1, 1.2.3

[25] CGAL. CGAL, Computational Geometry Algorithms Library, 2012. 2.6.1

[26] Patricia P. Chan and Todd M. Lowe. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res.*, 37(Database issue):D93–D97, 2009. 3.1

[27] Guo-Liang Chew, Andrea Pauli, John L Rinn, Aviv Regev, Alexander F Schier, and Eivind Valen. Ribosome profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding RNAs. *Development*, 140(13):2828–2834, 2013. 1.2.1, 1.1, 1.2.3

[28] Dominique Chu and Tobias von der Haar. The architecture of eukaryotic translation. *Nucleic Acids Research*, 40(20):10098–10106, 2012. 1.2.4

[29] Dominique Chu, Nicolae Zabet, and Tobias von der Haar. A novel and versatile computational tool to model translation. *Bioinformatics*, 28(2):292–293, 2012. 1.2.4

[30] Sang Y Chun, Caitlin M Rodriguez, Peter K Todd, and Ryan E Mills. SPECtre: a spectral coherence-based classifier of actively translated transcripts from ribosome profiling sequence data. *bioRxiv*, page 034777, 2015. 1.1, 1.2.3

[31] Betty Y Chung, Thomas J Hardcastle, Joshua D Jones, Nerea Irigoyen, Andrew E Firth, David C Baulcombe, and Ian Brierley. The use of duplex-specific nuclease in ribosome profiling and a user-friendly software package for ribo-seq data analysis. *RNA*, 21(10): 1731–1745, 2015. 1.2.1, 1.2.2, 1.1, 1.2.3

[32] Luca Ciandrini, Ian Stansfield, and M Carmen Romano. Ribosome traffic on mRNAs maps to Gene Ontology: genome-wide quantification of translation initiation rates and polysome size regulation. *PLoS Computational Biology*, 9(1):e1002866, 2013. 1.2.4

[33] Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009. 5.2.6

[34] Axel Cournac, Hervé Marie-Nelly, Martial Marbouty, Romain Koszul, and Julien Mozziconacci. Normalization of a chromosomal contact map. *BMC Genomics*, 13(1):436, 2012. 2.5.1

[35] J. Crappe et al. PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration. *Nucleic Acids Res.*, 2014. 3.1

[36] F. H. Crick. Codon–anticodon pairing: the wobble hypothesis. *J. Mol. Biol.*, 19(2):548–555, 1966. 4.4.5

[37] Zhiming Dai and Xianhua Dai. Nuclear colocalization of transcription factor target genes strengthens coregulation in yeast. *Nucleic Acids Res.*, 40(1):27–36, 2012. 2.1, 2.2

[38] A. Dana and T. Tuller. Determinants of translation elongation speed and ribosomal profiling biases in mouse embryonic stem cells. *PLoS Comput. Biol.*, 8(11):e1002755, 2012. 1.2.4, 3.1, 4.3.4

[39] A. Dana and T. Tuller. The effect of tRNA levels on decoding times of mRNA codons.

Nucleic Acids Res., 42(14):9171–9181, 2014. 1.2.4, 3.1, 4.4.5, 4.5

[40] A. Dana and T. Tuller. Properties and determinants of codon decoding time distributions. *BMC Genomics*, 15 Suppl 6:S13, 2014. 1.2.4, 4.1.3, 4.2, 4.3.6, 4.4.5, 4.5

[41] de Wit, Elzo and de Laat, Wouter. A decade of 3C technologies: insights into nuclear organization. *Genes & Development*, 26(1):11–24, 2012. 1.1.1, 2.1

[42] Alon Diament and Tamir Tuller. Estimation of ribosome profiling performance and reproducibility at various levels of resolution. *Biology direct*, 11(1):1, 2016. 1.2.4

[43] Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 2012. 2.1

[44] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29 (1):15–21, 2013. 3.3.2, 3.3.3, 4.3.7, 5.2.3

[45] M. dos Reis, R. Savva, and L. Wernisch. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.*, 32(17):5036–5044, 2004. 3.4.1, 4.4.5

[46] Josée Dostie, Todd A. Richmond, Ramy A. Arnaout, Rebecca R. Selzer, William L. Lee, Tracey A. Honan, Eric D. Rubio, Anton Krumm, Justin Lamb, Chad Nusbaum, Roland D. Green, and Job Dekker. Chromosome conformation capture carbon copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Res.*, 16(10):1299–1309, 2006. 2.1

[47] Zhijun Duan et al. A three-dimensional model of the yeast genome. *Nature*, 465(7296): 363–367, 2010. 2.1, 2.1, 2.1, 2.2, a, 2.5.1, 2.5.2, 2.5.3, 2.6.1, 2.6.5, 2.2, 2.6.6

[48] Geet Duggal, Rob Patro, Emre Sefer, Hao Wang, Darya Filippova, Samir Khuller, and Carl Kingsford. Resolving spatial inconsistencies in chromosome conformation measurements. *Alg. Mol. Biol.*, 8(1):8, 2013. 2.1

[49] J. G. Dunn, C. K. Foo, N. G. Belletier, E. R. Gavis, and J. S. Weissman. Ribosome profiling reveals pervasive and regulated stop codon readthrough in *Drosophila* melanogaster. *Elife*, 2:e01179, 2013. 1.2.1, 4.1.1

[50] Sara Elgamal, Assaf Katz, Steven J Hersch, David Newsom, Peter White, William Wiley Navarre, and Michael Ibba. EF-P dependent pauses integrate proximal and distal signals during translation. *PLoS Genet.*, 10(8):e1004553, 2014. 1.2.1

[51] S. R. Engel and J. M. Cherry. The new modern era of yeast genomics: community sequencing and the resulting annotation of multiple *Saccharomyces cerevisiae* strains at the Saccharomyces Genome Database. *Database (Oxford)*, 2013:bat012, 2013. 4.3.7, 5.2.3

[52] Alexander P Fields, Edwin H Rodriguez, Marko Jovanovic, Noam Stern-Ginossar, Brian J Haas, Philipp Mertins, Raktima Raychowdhury, Nir Hacohen, Steven A Carr, Nicholas T Ingolia, et al. A regression-based analysis of ribosome-profiling data reveals a conserved complexity to mammalian translation. *Molecular cell*, 60(5):816–827, 2015. 1.2.1, 1.1, 1.2.3

[53] Flicek, Paul and Ahmed et al. Ensembl 2013. *Nucleic Acids Research*, 41(D1):D48–D55,

2013. 3.1

[54] D. C. Fong and M. Saunders. LSMR: An iterative algorithm for sparse least-squares problems. *SIAM Journal on Scientific Computing*, 33(5):2950–2971, 2011. 4.3.3

[55] Geoff Fudenberg, Gad Getz, Matthew Meyerson, and Leonid A. Mirny. High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. *Nat. Biotechnol.*, 29(12):1109–1113, 2011. 2.1

[56] Caitlin E Gamble, Christina E Brule, Kimberly M Dean, Stanley Fields, and Elizabeth J Grayhack. Adjacent codons act in concert to modulate translation efficiency in yeast. *Cell*, 2016. 5.4

[57] X. Gao, J. Wan, B. Liu, M. Ma, B. Shen, and S. B. Qian. Quantitative profiling of initiating ribosomes *in vivo*. *Nat. Methods*, 12(2):147–153, 2015. 1.2.1, 4.1.1

[58] Xiangwei Gao et al. Quantitative profiling of initiating ribosomes *in vivo*. *Nat. Methods*, 2014. 3.1

[59] J. Gardin, R. Yeasmin, A. Yurovsky, Y. Cai, S. Skiena, and B. Futcher. Measurement of average decoding rates of the 61 sense codons *in vivo*. *Elife*, 3:e03735, 2014. 1.2.1, 1.2.4, 4.1.1, 4.2, 4.4.5, 4.4.5, 4.5

[60] M. V. Gerashchenko and V. N. Gladyshev. Translation inhibitors cause abnormalities in ribosome profiling experiments. *Nucleic Acids Res.*, 42(17):e134, 2014. 1.2.1, 1.2.2, 4.3.4

[61] M. V. Gerashchenko, A. V. Lobanov, and V. N. Gladyshev. Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress. *Proc. Natl. Acad. Sci. U.S.A.*, 109(43):17394–17399, 2012. 1.2.1, 3.1, 4.1.1

[62] A. V. Goldberg. Finding a maximum density subgraph. Technical report, CSD-84-171, Berkeley, CA, USA, 1984. d

[63] R. E. Gomory and T. C. Hu. Multi-terminal network flows. *J. Soc. Ind. Appl. Math.*, 9(4): pp. 551–570, 1961. c

[64] Christian Gonzalez, Jennifer S Sims, Nicholas Hornstein, Angeliki Mela, Franklin Garcia, Liang Lei, David A Gass, Benjamin Amendolara, Jeffrey N Bruce, Peter Canoll, et al. Ribosome profiling reveals a cell-type-specific translational landscape in brain tumors. *The Journal of Neuroscience*, 34(33):10924–10936, 2014. 1.2.1

[65] Thomas E Gorochowski, Zoya Ignatova, Roel AL Bovenberg, and Johannes A Roubos. Trade-offs between tRNA abundance and mRNA secondary structure support smoothing of translation elongation rate. *Nucleic acids research*, page gkv199, 2015. 1.2.4

[66] Alexey A Gritsenko, Marc Hulsman, Marcel JT Reinders, and Dick de Ridder. Unbiased quantitative models of protein translation derived from ribosome profiling data. *PLoS Comput Biol*, 11(8):e1004336, 2015. 1.1, 1.2.4

[67] H. Guo, N. T. Ingolia, J. S. Weissman, and D. P. Bartel. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, 466(7308):835–840, 2010. 1.2.1, 3.1a, 3.1, 3.1, 3.4.1, 3.1, 3.5.1, 3.5.2, 3.5.3, 3.6, 4.1.1, 4.3.4, 5.4

[68] Mitchell Guttman, Pamela Russell, Nicholas T Ingolia, Jonathan S Weissman, and Eric S

Lander. Ribosome profiling provides evidence that large noncoding rnas do not encode proteins. *Cell*, 154(1):240–251, 2013. 1.1, 1.2.3

[69] N. R. Guydosh and R. Green. Dom34 rescues ribosomes in 3' untranslated regions. *Cell*, 156(5):950–962, 2014. 1.2.1, 4.1.1, 4.1.3, 5.1, 5.2.3, 5.2.4, 5.3.1, 5.3.2, 5.4

[70] Rembrandt JF Haft, David H Keating, Tyler Schwaegler, Michael S Schwalbach, Jeffrey Vinokur, Mary Tremaine, Jason M Peters, Matthew V Kotlajich, Edward L Pohlmann, Irene M Ong, et al. Correcting direct effects of ethanol on translation and transcription machinery confers ethanol tolerance in bacteria. *Proceedings of the National Academy of Sciences*, 111(25):E2576–E2585, 2014. 1.2.1

[71] Gert-Jan Hendriks, Dimos Gaidatzis, Florian Aeschimann, and Helge Großhans. Extensive oscillatory gene expression during *C. elegans* larval development. *Molecular cell*, 53 (3):380–392, 2014. 1.2.1

[72] Michael T. Howard, Bradley A. Carlson, Christine B. Anderson, and Dolph L. Hatfield. Translational redefinition of UGA codons is regulated by selenium availability. *Journal of Biological Chemistry*, 2013. 1.2.1

[73] Andrew C Hsieh et al. The translational landscape of mTOR signalling steers cancer initiation and metastasis. *Nature*, 485(7396):55–61, 2012. 1.2.1, 3.1

[74] Ming Hu, Ke Deng, Siddarth Selvaraj, Zhaohui Qin, Bing Ren, and Jun S Liu. HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics*, 28(23):3131–3133, 2012. 2.5.1

[75] Tao Huang, Sibao Wan, Zhongping Xu, Yufang Zheng, Kai-Yan Feng, Hai-Peng Li, Xiangyin Kong, and Yu-Dong Cai. Analysis and prediction of translation rate based on sequence and functional features of the mRNA. *PLoS ONE*, 6(1):e16036, 2011. 1.1, 1.2.4

[76] Jeffrey A Hussmann, Stephanie Patchett, Arlen Johnson, Sara Sawyer, and William H Press. Understanding biases in ribosome profiling experiments reveals signatures of translation dynamics in yeast. *PLoS Genet*, 11(12):e1005732, 2015. 1.2.2, 1.1, 1.2.4, 5.4

[77] Maxim Imakaev, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R Lajoie, Job Dekker, and Leonid A Mirny. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature Methods*, 9(10):999–1003, 2012. 2.5.1

[78] N. T. Ingolia. Ribosome profiling: new views of translation, from single codons to genome scale. *Nat. Rev. Genet.*, 15(3):205–213, 2014. 1.2.1, 1.2.2, 3.1, 3.1, 3.3.2, 3.6, 4.1.1, 4.1.3

[79] N. T. Ingolia, S. Ghaemmaghami, J. R. Newman, and J. S. Weissman. Genome-wide analysis *in vivo* of translation with nucleotide resolution using ribosome profiling. *Science*, 324(5924):218–223, 2009. 1.2.1, 1.2.4, 3.1, 3.1, 3.4.1, 3.6, 4.1.1, 4.1.2, 4.1.3, 4.3.4, 5.1, 5.4

[80] N. T. Ingolia, L. F. Lareau, and J. S. Weissman. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, 147(4): 789–802, 2011. 1.2.1, 3.1b, 3.1, 3.3.3, 3.1, 3.5.3, 3.6, 4.1.1, 4.1.3, 4.3.4, 5.1

[81] Nicholas T Ingolia. Ribosome footprint profiling of translation throughout the genome.

Cell, 165(1):22–33, 2016. 1.2.1, 1.2.2, 1.2.2

[82] Nicholas T Ingolia, Gloria A Brar, Silvia Rouskin, Anna M McGeachy, and Jonathan S Weissman. Genome-Wide Annotation and Quantitation of Translation by Ribosome Profiling. *Current Protocols in Molecular Biology*, pages 4–18, 2013. 1.2.2

[83] Nicholas T Ingolia, Gloria A Brar, Noam Stern-Ginossar, Michael S Harris, Gaëlle JS Talhouarne, Sarah E Jackson, Mark R Wills, and Jonathan S Weissman. Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. *Cell reports*, 8 (5):1365–1379, 2014. 1.2.1, 1.1, 1.2.3

[84] Nicholas T Ingolia et al. The ribosome profiling strategy for monitoring translation *in vivo* by deep sequencing of ribosome-protected mRNA fragments. *Nat Protoc*, 7(8):1534–1550, 2012. 1.2.1, 1.2.2, 3.1

[85] Nerea Irigoyen, Andrew E Firth, Joshua D Jones, Betty Y-W Chung, Stuart G Siddell, and Ian Brierley. High-resolution analysis of coronavirus gene expression by rna sequencing and ribosome profiling. *PLoS Pathog*, 12(2):e1005473, 2016. 1.2.1

[86] Calvin H Jan, Christopher C Williams, and Jonathan S Weissman. Principles of er cotranslational translocation revealed by proximity-specific ribosome profiling. *Science*, 346(6210):1257521, 2014. 1.2.1

[87] Yujin Jeong, Ji-Nu Kim, Min Woo Kim, Giselda Bucca, Suhyung Cho, Yeo Joon Yoon, Byung-Gee Kim, Jung-Hye Roe, Sun Chang Kim, Colin P Smith, et al. The dynamic transcriptional and translational landscape of the model antibiotic producer Streptomyces coelicolor A3 (2). *Nature communications*, 7, 2016. 1.2.1

[88] Zhe Ji, Ruisheng Song, Aviv Regev, and Kevin Struhl. Many lncRNAs, 5UTRs, and pseudogenes are translated and some are likely to express functional proteins. *Elife*, 4: e08890, 2015. 1.2.1, 1.1, 1.2.3

[89] Hui Jiang and Wing Hung Wong. Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, 25(8):1026–1032, 2009. 3.1

[90] Timothy G Johnstone, Ariel A Bazzini, and Antonio J Giraldez. Upstream ORFs are prevalent translational repressors in vertebrates. *The EMBO journal*, page e201592759, 2016. 1.2.1

[91] Piyada Juntawong, Thomas Girke, Jérémie Bazin, and Julia Bailey-Serres. Translational dynamics revealed by genome-wide profiling of ribosome footprints in arabidopsis. *Proceedings of the National Academy of Sciences*, 111(1):E203–E212, 2014. 1.2.1

[92] Reza Kalhor et al. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat. Biotechnol.*, 30(1):90–98, 2012. 2.1

[93] Krishna Kannan, Pinal Kanabar, David Schryer, Tanja Florin, Eugene Oh, Neil Bahroos, Tanel Tenson, Jonathan S Weissman, and Alexander S Mankin. The general mode of translation inhibition by macrolide antibiotics. *Proceedings of the National Academy of Sciences*, 111(45):15958–15963, 2014. 1.2.1

[94] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at UCSC. *Genome*

Research, 12(6):996–1006, 2002. 3.5

[95] Anthony Khong, Jennifer M Bonderoff, Ruth V Spriggs, Erik Tammpere, Craig H Kerr, Thomas J Jackson, Anne E Willis, and Eric Jan. Temporal regulation of distinct internal ribosome entry sites of the dicistroviridae cricket paralysis virus. *Viruses*, 8(1):25, 2016. 1.2.1

[96] Kai Kruse, Sven Sewitz, and M. Madan Babu. A complex network framework for unbiased statistical analyses of DNA–DNA contact maps. *Nucleic Acids Res.*, 41(2):701–710, 2013. 2.1, 2.2, 2.4.2, 2.5.1, 2.6.3, 2.6.4, 2.6.5

[97] S. Kuersten et al. Translation regulation gets its 'omics' moment. *Wiley Interdiscip Rev RNA*, 4(6):617–630, 2013. 1.2.1, 3.1

[98] L. F. Lareau, D. H. Hite, G. J. Hogan, and P. O. Brown. Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. *Elife*, 3: e01257, 2014. 1.2.1, 1.2.4, 4.1.1, 4.1.3, 4.3.4, 4.3.6, 4.4.5, 4.5

[99] Liana F Lareau et al. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements. *Nature*, 446(7138):926–929, 2007. 3.1

[100] Miler T Lee, Ashley R Bonneau, Carter M Takacs, Ariel A Bazzini, Kate R DiVito, Elizabeth S Fleming, and Antonio J Giraldez. Nanog, Pou5f1 and SoxB1 activate zygotic gene expression during the maternal-to-zygotic transition. *Nature*, 503(7476):360–364, 2013. 1.2.1

[101] S. Lee, B. Liu, S. Lee, S. X. Huang, B. Shen, and S. B. Qian. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U.S.A.*, 109(37):E2424–2432, 2012. 1.2.1, 4.1.1

[102] Rachel Legendre, Agnès Baudin-Baillieu, Isabelle Hatin, and Olivier Namy. RiboTools: a Galaxy toolbox for qualitative ribosome profiling analysis. *Bioinformatics*, 31(15):2586–2588, 2015. 1.1, 1.2.3

[103] Lei Lei, Junpeng Shi, Jian Chen, Mei Zhang, Silong Sun, Shaojun Xie, Xiaojie Li, Biao Zeng, Lizeng Peng, Andrew Hauck, et al. Ribosome profiling reveals dynamic translational landscape in maize seedlings under drought stress. *The Plant Journal*, 84(6): 1206–1218, 2015. 1.2.1

[104] Anat Levin, Yair Weiss, Fredo Durand, and William T Freeman. Understanding and evaluating blind deconvolution algorithms. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1964–1971. IEEE, 2009. 4.5

[105] G. W. Li, D. Burkhardt, C. Gross, and J. S. Weissman. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell*, 157(3): 624–635, 2014. 1.2.1, 4.1.1

[106] Gene-Wei Li, Eugene Oh, and Jonathan S Weissman. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature*, 484(7395):538–541, 2012. 1.2.1

[107] Erez Lieberman-Aiden, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R. Lajoie, Peter J. Sabo, Michael O. Dorschner,

Richard Sandstrom, Bradley Bernstein, M. A. Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A. Mirny, Eric S. Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009. 2.1

[108] Ming-Jung Liu, Szu-Hsien Wu, Jing-Fen Wu, Wen-Dar Lin, Yi-Chen Wu, Tsung-Ying Tsai, Huang-Lung Tsai, and Shu-Hsing Wu. Translational landscape of photomorphogenic arabidopsis. *The Plant Cell*, 25(10):3699–3710, 2013. 1.2.1

[109] Tzu-Yu Liu and Yun S. Song. Prediction of ribosome footprint profile shapes from transcript sequences. *Bioinformatics*, 32(12):i183–i191, 2016. 1.1, 1.2.3

[110] Xiaoqiu Liu, Huifeng Jiang, Zhenglong Gu, and Jeffrey W Roberts. High-resolution view of bacteriophage lambda gene expression by ribosome profiling. *Proceedings of the National Academy of Sciences*, 110(29):11928–11933, 2013. 1.2.1

[111] Fabricio Loayza-Puch, Jarno Drost, Koos Rooijers, Rui Lopes, Ran Elkon, and Reuven Agami. p53 induces transcriptional and translational programs to suppress cell proliferation and growth. *Genome biology*, 14(4):1, 2013. 1.2.1

[112] A. T. Martens, J. Taylor, and V. J. Hilser. Ribosome A and P sites revealed by length analysis of ribosome profiling data. *Nucleic Acids Res.*, 43(7):3680–3687, 2015. 4.1.1, 4.1.3

[113] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1):pp–10, 2011. 5.2.3

[114] Rachel Patton McCord, Ashley Nazario-Toole, Haoyue Zhang, Peter S. Chines, Ye Zhan, Michael R. Erdos, Francis S. Collins, Job Dekker, and Kan Cao. Correlated alterations in genome organization, histone methylation, and DNA-lamin A/C interactions in Hutchinson-Gilford Progeria syndrome. *Genome Res.*, 23(2):260–269, 2013. 2.1

[115] C. J. McManus, G. E. May, P. Spealman, and A. Shteyman. Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome Res.*, 24(3): 422–430, 2014. 1.2.1, 3.1, 4.1.1, 4.1.3, 5.4

[116] Amit Mehra and Vassily Hatzimanikatis. An algorithmic framework for genome-wide modeling and analysis of translation networks. *Biophysical Journal*, 90(4):1136–1146, 2006. 1.2.4

[117] Audrey M Michel and Pavel V Baranov. Ribosome profiling: a Hi-Def monitor for protein synthesis at the genome-wide scale. *Wiley Interdisciplinary Reviews: RNA*, 4(5):473–490, 2013. 1.2.1

[118] Audrey M. Michel, Gearoid Fox, Anmol M. Kiran, Christof De Bo, Patrick B. F. O'Connor, Stephen M. Heaphy, James P. A. Mullan, Claire A. Donohue, Desmond G. Higgins, and Pavel V. Baranov. GWIPS-viz: development of a ribo-seq genome browser. *Nucleic Acids Research*, 42(D1):D859–D864, 2014. 1.1, 1.2.3, 3.1

[119] Audrey M Michel, Anna M Ahern, Claire A Donohue, and Pavel V Baranov. GWIPS-viz as a tool for exploring ribosome profiling evidence supporting the synthesis of alternative proteoforms. *Proteomics*, 15(14):2410–2416, 2015. 1.1, 1.2.3

[120] Audrey M Michel, James PA Mullan, Vimalkumar Velayudhan, Patrick BF O'Connor, Claire A Donohue, and Pavel V Baranov. Ribogalaxy: A browser based platform for the alignment, analysis and visualization of ribosome profiling data. *RNA biology*, 13(3): 316–319, 2016. 1.1, 1.2.3

[121] Audrey M. Michel et al. Observation of dually decoded regions of the human genome using ribosome profiling data. *Genome Res.*, 22(11):2219–2229, 2012. 1.2.1, 1.1, 1.2.3, 3.1, 4.1.1, 4.1.3, 4.4.4

[122] Teemu P Miettinen and Mikael Björklund. Modified ribosome profiling reveals high abundance of ribosome protected mRNA fragments derived from 3' untranslated regions. *Nucleic acids research*, 43(2):1019–1034, 2015. 1.2.1

[123] Fuad Mohammad, Christopher J Woolstenhulme, Rachel Green, and Allen R Buskirk. Clarifying the translational pausing landscape in bacteria by ribosome profiling. *Cell reports*, 14(4):686–694, 2016. 1.2.1

[124] Mortazavi, Ali and Williams, Brian A and McCue, Kenneth and Schaeffer, Lorian and Wold, Barbara. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7):621–628, 2008. 1.2, 3.1

[125] Kenji Nakahigashi, Yuki Takai, Yuh Shiwa, Mei Wada, Masayuki Honma, Hirofumi Yoshikawa, Masaru Tomita, Akio Kanai, and Hirotada Mori. Effect of codon adaptation on codon-level and gene-level translation efficiency *in vivo*. *BMC genomics*, 15(1):1, 2014. 1.2.1, 1.2.4

[126] D. D. Nedialkova and S. A. Leidel. Optimization of Codon Translation Rates via tRNA Modifications Maintains Proteome Integrity. *Cell*, 161(7):1606–1618, 2015. 1.2.1

[127] P. O'Connor, D. Andreev, and P. Baranov. Surveying the relative impact of mRNA features on local ribosome profiling read density in 28 datasets. *bioRxiv*, page 018762, 2015. 1.1, 1.2.3, 4.4.5

[128] Eugene Oh, Annemarie H. Becker, Arzu Sandikci, Damon Huber, Rachna Chaba, Felix Gloge, Robert J. Nichols, Athanasios Typas, Carol A. Gross, Gnter Kramer, Jonathan S. Weissman, and Bernd Bukau. Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor *in vivo*. *Cell*, 147(6):1295–1308, 2011. 1.2.1, 3.1

[129] Adam B. Olshen et al. Assessing gene-level translational control from ribosome profiling. *Bioinformatics*, 29(23):2995–3002, 2013. 1.1, 1.2.3, 3.1

[130] Pachter, L. Models for transcript quantification from RNA-Seq. *arXiv preprint arXiv:1104.3889*, 2011. 3.1

[131] R. Patro, G. Duggal, and C. Kingsford. Salmon: Accurate, Versatile and Ultrafast Quantification from RNA-seq Data using Lightweight-Alignment. *bioRxiv*, page 021592, 2015. 4.3.7

[132] R. Patro et al. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat. Biotechnol.*, 32(5):462–464, 2014. 3.3.2, 3.4.1, 3.4.2

[133] Andrea Pauli, Megan L Norris, Eivind Valen, Guo-Liang Chew, James A Gagnon, Steven Zimmerman, Andrew Mitchell, Jiao Ma, Julien Dubrulle, Deepak Reyon, et al. Tod-

dler: an embryonic signal that promotes cell movement via apelin receptors. *Science*, 343 (6172):1248636, 2014. 1.2.1, 1.1, 1.2.3

[134] Jonas Paulsen, Tonje G. Lien, Geir Kjetil Sandve, Lars Holden, Ornulf Borgan, Ingrid K. Glad, and Eivind Hovig. Handling realistic assumptions in hypothesis testing of 3D co-localization of genomic elements. *Nucleic Acids Res.*, 2013. 2.6.3, 2.6.5

[135] C. Pop, S. Rouskin, N. T. Ingolia, L. Han, E. M. Phizicky, J. S. Weissman, and D. Koller. Causal signals between codon bias, mRNA structure, and the efficiency of translation and elongation. *Mol. Syst. Biol.*, 10:770, 2014. 1.2.1, 1.1, 1.2.4, 4.1.1, 4.2, 4.5

[136] A Popa, K Lebrigand, A Paquet, N Nottet, K Robbe-Sermesant, R Waldmann, and P Barbry. RiboProfiling: a Bioconductor package for standard Ribo-seq pipeline processing. *F1000Research*, 5(1309), 2016. 1.1, 1.2.3

[137] Julien Racle, Jan Overney, and Vassily Hatzimanikatis. A computational framework for the design of optimal protein synthesis. *Biotechnology and Bioengineering*, 109(8):2127–2133, 2012. 1.2.4

[138] Anil Raj, Sidney H Wang, Heejung Shim, Arbel Harpak, Yang I Li, Brett Engelmann, Matthew Stephens, Yoav Gilad, and Jonathan K Pritchard. Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *eLife*, 5:e13328, 2016. 1.1, 1.2.3

[139] David W Reid, Qiang Chen, Angeline S-L Tay, Shirish Shenolikar, and Christopher V Nicchitta. The unfolded protein response triggers selective mrna release from the endoplasmic reticulum. *Cell*, 158(6):1362–1374, 2014. 1.2.1

[140] Reut Shalgi and Jessica A. Hurt and Irina Krykbaeva and Mikko Taipale and Susan Lindquist and Christopher B. Burge. Widespread Regulation of Translation by Elongation Pausing in Heat Shock . *Molecular Cell* , 49(3):439–452, 2013. 1.2.1

[141] S. Reuveni et al. Genome-scale analysis of translation elongation with a ribosome flow model. *PLoS Comput. Biol.*, 7(9):e1002127, 2011. 3.4.1, 3.4.1, 3.5.1

[142] Daniel P Riordan, Daniel Herschlag, and Patrick O Brown. Identification of RNA recognition elements in the *Saccharomyces cerevisiae* transcriptome. *Nucleic acids research*, 39(4):1501–1509, 2011. 5.2.6

[143] A. Roberts and L. Pachter. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nat. Methods*, 10(1):71–73, 2013. 3.3.2, 3.7.4

[144] Koos Rooijers, Fabricio Loayza-Puch, Leo G Nijtmans, and Reuven Agami. Ribosome profiling reveals features of normal and disease-associated mitochondrial translation. *Nat Commun*, 4:2886, 2013. 1.2.1, 3.1

[145] Claudia A Rubio, Benjamin Weisburd, Matthew Holderfield, Carolina Arias, Eric Fang, Joseph L DeRisi, and Abdallah Fanidi. Transcriptome-wide characterization of the eIF4A signature highlights plasticity in translation regulation. *Genome biology*, 15(10):1, 2014. 1.2.1

[146] Andrzej J Rutkowski, Florian Erhard, Anne L'Hernault, Thomas Bonfert, Markus Schilhabel, Colin Crump, Philip Rosenstiel, Stacey Efstathiou, Ralf Zimmer, Caroline C Friedel,

et al. Widespread disruption of host transcription termination in HSV-1 infection. *Nature communications*, 6, 2015. 1.2.1

[147] R. Sabi and T. Tuller. A comparative genomics study on the effect of individual amino acids on ribosome stalling. *BMC Genomics*, 16 Suppl 10:S5, 2015. 4.1.3, 5.1

[148] Neelam Dabas Sen, Fujun Zhou, Nicholas T Ingolia, and Alan G Hinnebusch. Genome-wide analysis of translational efficiency reveals distinct but overlapping functions of yeast DEAD-box RNA helicases Ded1 and eIF4A. *Genome research*, 25(8):1196–1205, 2015. 1.2.1

[149] Tom Sexton et al. Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell*, 148(3):458–472, 2012. 2.1

[150] P. Shah, Y. Ding, M. Niemczyk, G. Kudla, and J. B. Plotkin. Rate-limiting steps in yeast protein translation. *Cell*, 153(7):1589–1601, 2013. 1.1, 1.2.4, 1.2.4, 3.4.1, 4.3.4

[151] Marieke Simonis, Petra Klous, Erik Splinter, Yuri Moshkin, Rob Willemsen, Elzo De Wit, Bas Van Steensel, and Wouter De Laat. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nature Genetics*, 38(11):1348–1354, 2006. 2.1

[152] B. Sipos, G. Slodkowicz, T. Massingham, and N. Goldman. Realistic simulations reveal extensive sample-specificity of RNA-seq biases. *arXiv preprint arXiv:1308.3172*, 2013. 3.4.2

[153] Marlena Siwiak and Piotr Zielenkiewicz. A comprehensive, quantitative, and genome-wide model of translation. *PLoS Comput Biol*, 6(7):e1000865, 2010. 1.1, 1.2.4

[154] Jenna E Smith, Juan R Alvarez-Dominguez, Nicholas Kline, Nathan J Huynh, Sarah Geisler, Wenqian Hu, Jeff Coller, and Kristian E Baker. Translation of small open reading frames within unannotated RNA transcripts in *Saccharomyces cerevisiae*. *Cell reports*, 7 (6):1858–1866, 2014. 1.2.1

[155] Pieter Spealman, Hao Wang, Gemma May, Carl Kingsford, and C Joel McManus. Exploring ribosome positioning on translating transcripts with ribosome profiling. *Post-Transcriptional Gene Regulation*, pages 71–97, 2016. 1.2.2, 5.2.2

[156] M. Stadler and A. Fire. Wobble base-pairing slows *in vivo* translation elongation in metazoans. *RNA*, 17(12):2063–2073, 2011. 1.2.1, 4.1.1, 4.1.3, 4.2, 4.4.5, 4.4.5

[157] M. Stadler, K. Artiles, J. Pak, and A. Fire. Contributions of mRNA abundance, ribosome loading, and post- or peri-translational effects to temporal repression of *C. elegans* heterochronic miRNA targets. *Genome Res.*, 22(12):2418–2426, 2012. 1.2.1, 4.1.1

[158] Noam Stern-Ginossar, Ben Weisburd, Annette Michalski, Vu Thuy Khanh Le, Marco Y. Hein, Sheng-Xiong Huang, Ming Ma, Ben Shen, Shu-Bing Qian, Hartmut Hengel, Matthias Mann, Nicholas T. Ingolia, and Jonathan S. Weissman. Decoding human cytomegalovirus. *Science*, 338(6110):1088–1093, 2012. 1.2.1

[159] T. Sterne-Weiler et al. Frac-seq reveals isoform-specific recruitment to polyribosomes. *Genome Res.*, 23(10):1615–1623, 2013. 3.1

[160] Craig R. Stumpf, Melissa V. Moreno, Adam B. Olshen, Barry S. Taylor, and Davide Rug-

gero. The translational landscape of the mammalian cell cycle. *Molecular Cell*, 52(4): 574–582, 2013. 1.2.1, 3.1

[161] H. Tanizawa et al. Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nuc. Acids Res.*, 38(22):8164–8177, 2010. 2.1, 2.1, 2.1

[162] D. Tarrant and T. von der Haar. Synonymous codons, ribosome speed, and eukaryotic gene expression regulation. *Cell. Mol. Life Sci.*, 71(21):4195–4206, 2014. 4.2, 4.4.5, 4.4.5

[163] Carson C Thoreen et al. A unifying model for mTORC1-mediated regulation of mRNA translation. *Nature*, 485(7396):109–113, 2012. 1.2.1, 3.1, 3.1

[164] Harianto Tjong, Ke Gong, Lin Chen, and Frank Alber. Physical tethering and volume exclusion determine higher-order genome organization in budding yeast. *Genome Res.*, 22(7):1295–1305, 2012. 1

[165] C. Trapnell et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, 28(5):511–515, 2010. 3.3.2, 3.7.4

[166] Mark A Umbarger et al. The three-dimensional architecture of a bacterial genome and its alteration by genetic perturbation. *Mol. Cell*, 44(2):252–264, 2011. 2.1, 2.1

[167] P. P. Vaidyanathan, B. Zinshteyn, M. K. Thompson, and W. V. Gilbert. Protein kinase A regulates gene-specific translational adaptation in differentiating yeast. *RNA*, 20(6): 912–922, 2014. 1.2.1, 4.1.1

[168] Wan, Ji and Qian, Shu-Bing. TISdb: a database for alternative translation initiation in mammalian cells. *Nucleic Acids Research*, 2013. 5.4

[169] Hao Wang, Geet Duggal, Rob Patro, Michelle Girvan, Sridhar Hannenhalli, and Carl Kingsford. Topological properties of chromosome conformation graphs reflect spatial proximities within chromatin. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, pages 306–315, 2013. 2

[170] Hao Wang, Joel McManus, and Carl Kingsford. Accurate recovery of ribosome positions reveals slow translation of wobble-pairing codons in yeast. In *Research in Computational Molecular Biology*, pages 37–52. Springer, 2016. 4, 5.2.3

[171] Hao Wang, Joel McManus, and Carl Kingsford. Isoform-level ribosome occupancy estimation guided by transcript abundance with Ribomap. *Bioinformatics*, 32(12):1880–1882, 2016. 3

[172] Jianbin Wang, Julianne Garrey, and Richard E Davis. Transcription in pronuclei and one- to four-cell embryos drives early development in a nematode. *Current Biology*, 24(2): 124–133, 2014. 1.2.1

[173] Jing Wang, William Rennie, Chaochun Liu, Charles S Carmack, Karine Prévost, Marie-Pier Caron, Eric Massé, Ye Ding, and Joseph T Wade. Identification of bacterial sRNA regulatory targets using ribosome profiling. *Nucleic acids research*, 43(21):10308–10320, 2015. 1.2.1

[174] David E. Weinberg, Premal Shah, Stephen W. Eichhorn, Jeffrey A. Hussmann, Joshua B.

Plotkin, and David P. Bartel. Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. *Cell Reports*, 14(7): 1787 – 1799, 2016. 1.2.1, 1.2.2, 1.2.4, 5.3.1

[175] Arun P Wiita, Etay Ziv, Paul J Wiita, Anatoly Urisman, Olivier Julien, Alma L Burlingame, Jonathan S Weissman, and James A Wells. Global cellular response to chemotherapy-induced apoptosis. *Elife*, 2:e01236, 2013. 1.2.1

[176] Christopher C Williams, Calvin H Jan, and Jonathan S Weissman. Targeting and plasticity of mitochondrial proteins revealed by proximity-specific ribosome profiling. *Science*, 346 (6210):748–751, 2014. 1.2.1

[177] Daniela M. Witten and William Stafford Noble. On the assessment of statistical significance of three-dimensional colocalization of sets of genomic elements. *Nucleic Acids Res.*, 40(9):3849–3855, 2012. 2.1, 2.2, a, 2.4.2, 2.5.1, 2.6.1, 2.6.3, 2.5, 2.6.6

[178] C. J. Woolstenhulme, N. R. Guydosh, R. Green, and A. R. Buskirk. High-precision analysis of translational pausing by ribosome profiling in bacteria lacking EFP. *Cell Rep*, 11 (1):13–21, 2015. 1.2.1, 4.1.1, 4.1.3, 4.3.6, 5.1

[179] Zhengtao Xiao, Qin Zou, Yu Liu, and Xuerui Yang. Genome-wide assessment of differential translations with ribosome profiling data. *Nature communications*, 7, 2016. 1.1, 1.2.3

[180] Shang-Qian Xie, Peng Nie, Yan Wang, Hongwei Wang, Hongyu Li, Zhilong Yang, Yizhi Liu, Jian Ren, and Zhi Xie. RPFdb: a database for genome wide information of translated mrna generated from ribosome profiling. *Nucleic acids research*, page gkv972, 2015. 1.1, 1.2.3

[181] Eitan Yaffe and Amos Tanay. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.*, 43(11): 1059–1065, 2011. 2.5.1

[182] Jian-Rong Yang, Xiaoshu Chen, and Jianzhi Zhang. Codon-by-codon modulation of translational speed and accuracy via mRNA folding. *PLoS Biol.*, 12(7):e1001910, 2014. 1.2.4

[183] Zhilong Yang, Shuai Cao, Craig A Martens, Stephen F Porcella, Zhi Xie, Ming Ma, Ben Shen, and Bernard Moss. Deciphering poxvirus gene expression by RNA sequencing and ribosome profiling. *Journal of virology*, 89(13):6874–6886, 2015. 1.2.1

[184] Tao You, George M Coghill, and Alistair J. P. Brown. A quantitative model for mRNA translation in *Saccharomyces cerevisiae*. *Yeast*, 27(10):785–800, 2010. 1.2.4

[185] David J Young, Nicholas R Guydosh, Fan Zhang, Alan G Hinnebusch, and Rachel Green. Rli1/ABCE1 recycles terminating ribosomes and controls translation reinitiation in 3' UTRs *in vivo*. *Cell*, 162(4):872–884, 2015. 1.2.1

[186] Huayan Zhao, Shiyou Lü, Ruixi Li, Tao Chen, Huoming Zhang, Peng Cui, Feng Ding, Pei Liu, Guangchao Wang, Yiji Xia, et al. The Arabidopsis gene DIG6 encodes a large 60S subunit nuclear export GTPase 1 that is involved in ribosome biogenesis and affects multiple auxin-regulated development processes. *Journal of experimental botany*, 66(21): 6863–6875, 2015. 1.2.1

[187] Yi Zhong, Theofanis Karaletsos, Philipp Drewe, Vipin Thankam T Sreedharan, David Kuo, Kamini Singh, Hans-Guido Wendel, and Gunnar Rätsch. RiboDiff: Detecting Changes of Translation Efficiency from Ribosome Footprints. *bioRxiv*, page 017111, 2016. 1.1, 1.2.3

[188] Christophe Zimmer and Emmanuelle Fabre. Principles of chromosomal organization: lessons from yeast. *J. Cell Biol.*, 192(5):723–733, 2011. 2.6.5

[189] Boris Zinshteyn and Wendy V Gilbert. Loss of a conserved trna anticodon modification perturbs cellular signaling. *PLoS Genet*, 9(8):e1003675, 2013. 1.2.1

[190] Reimo Zoschke and Alice Barkan. Genome-wide analysis of thylakoid-bound ribosomes in maize reveals principles of cotranslational targeting to the thylakoid membrane. *Proceedings of the National Academy of Sciences*, 112(13):E1678–E1687, 2015. 1.2.1

[191] Hermioni Zouridis and Vassily Hatzimanikatis. Effects of codon distributions and tRNA competition on protein translation. *Biophysical Journal*, 95(3):1018–1033, 2008. 1.2.4

[192] A. Zupanic, C. Meplan, S. N. Grellscheid, J. C. Mathers, T. B. Kirkwood, J. E. Hesketh, and D. P. Shanley. Detecting translational regulation by change point analysis of ribosome profiling data sets. *RNA*, 20(10):1507–1518, 2014. 1.1, 1.2.3, 3.3.3, 4.2