2017

# High Throughput Detection of Pseudouridine: Caveats, Conundrums, and a Case for Open Science

Maryam Zaringhalam


# HIGH-THROUGHPUT DETECTION OF PSEUDOURIDINE: CAVEATS, CONUNDRUMS, AND A CASE FOR OPEN SCIENCE

A Thesis Presented to the Faculty of

The Rockefeller University

in Partial Fulfillment of the Requirements for

the degree of Doctor of Philosophy

by

Maryam Zaringhalam

June 2017

# HIGH-THROUGHPUT DETECTION OF PSEUDOURIDINE: CAVEATS, CONUNDRUMS, AND A CASE FOR OPEN SCIENCE

Maryam Zaringhalam, Ph.D.

The Rockefeller University 2017

The isomerization of uridine to pseudouridine (Ψ), known as pseudouridylation, is the most abundant post-transcriptional modification of stable RNAs. Due to technical limitations in pseudouridine detection methods, studies on pseudouridylation have historically focused on ribosomal RNAs, transfer RNAs, and spliceosomal small nuclear RNAs, where Ψs play a critical role in RNA biogenesis and function. For decades, Ψ research was confined to this small subset of cellular RNAs ,owing to limitations in methods for Ψ detection. Interest in this modification was reinvigorated, however, with reports that Ψ is conditionally induced in different environmental contexts and that pseudouridylation of certain codons recoded amino acid incorporation. Pseudouridine has thus revealed itself as a dynamic modification capable of fine-tuning RNA function.

In this thesis, I describe how I attempted to develop a high-throughput technique to identify novel sites of pseudouridylation throughout the whole transcriptome. By identifying what transcripts are subject to pseudouridylation, I hoped to better understand Ψ's functional role. While pursuing this work, a series of deep sequencing methods — Pseudo-seq, Ψ-seq, PSI-seq, and CeU-seq — were published that mapped Ψ positions across the entire transcriptome with single nucleotide resolution. Collectively, these methods greatly expanded the catalogue of pseudouridylated transcripts and revealed conditionally-dependent sites of pseudouridylation in response to cellular stress. With

four techniques available, I undertook a critical analysis of their results, uncovering a comparatively small subset of robustly detectible putative Ψ sites. This analysis underscored the merits and limitations of each approach.

Having identified areas for improvement in the available Ψ-detection approaches, I adapted Ψ-seq to profile sites of pseudouridylation in the protozoan parasite *Trypanosoma brucei*. My efforts at transcriptome-wide Ψ-detection, however, were undercut by an inability to experimentally replicate Ψ-seq.

As much as this thesis documents an endeavor to better understand the functional role of pseudouridylation, it also documents systematic and thorough experimental failure. In so doing, the work detailed in this thesis highlights a need within the sciences to foster increased transparency and reproducibility.

*For Mom, Dad, and Nad.*

**ACKNOWLEDGMENTS**

I would first like to thank my advisor, Nina Papavasiliou, not only for her mentorship and preternatural optimism and encouragement, but also for teaching me to openly embrace ignorance and failure as pillars of the scientific endeavor. Her support and encouragement of even my wackiest side projects have guided me to step off the beaten path and out into the wide world of science policy and advocacy.

Thank you also to George Cross, for taking me under your wing and trekking to the Big Apple from your home in the woods to guide me through the wild world of trypanosome biology. I thank my committee members, Sandy Simon, Fred Cross, and Tom Meier, who have given me great feedback and guidance through the years, along with Jayne Raper who has kindly given up her Thursday afternoon to serve as the external examiner of my thesis defense.

I am so grateful to my collaborators — Yi-Tao Yu, Guowei Wu, Mark Helm, Katharina Schmid, and Tricia Serio — many of whom have been kind enough to put me up in their labs to learn from their expertise.

I owe so much moral and scientific support to my labmates. I am so grateful for our closeness and camaraderie. A special thanks to Monica Mugnier and Danae Schulz for guiding me through five scooped projects with sound scientific advice and wine to boot; to Jason Pinger, Violeta Rayon, and Linda Molla for joining the lab alongside me and sticking it out through thick and through thin; to Eric Fritz, Claire Hamilton, and Rebecca Delker for showing me lay of the lab land.

Thank you mom and dad and family for listening to my rambling about the latest in a series of promising findings or failed experiments. To Alice Lu, my fellow scientist

in arms, thank you for spontaneously travelling the world with me and for all our late-night cupcake-filled study sessions. To my roommates and best friends Tim and Chelsey, who have seen me through bed bugs and burglaries, I thank you for making sure my home away from lab was well-stocked with beer, TV marathons, and some semblance of sanity.

And finally, I would like to thank all of the people and groups who have taught me that science not shared with the public is science not done. Thank you to the good people at Neuwrite for workshopping my often rambling, long-winded attempts at writing; to all of the guests who ever participated in the crazy experiment that was ArtLab; to Devon Collins and Avital Percher for letting me join in on the Science Soapbox team; to the organizers of NerdNite, Story Collider, Drunk Science, and The Empiricist League for taking a chance on me and letting me speak to your audiences; to the BioBus and BioBase for being a magical lab away from labs; to Jeanne Garbarino and the Outreach Lab at Rockefeller for opening so many doors for others and for myself.

I would not be the scientist I am today without all of you.

**TABLE OF CONTENTS**

## LIST OF FIGURES

## LIST OF TABLES

# CHAPTER 1. Introduction[1]

> *The central dogma, enunciated by Crick in 1958 and the keystone of molecular biology ever since, is likely to prove a considerable oversimplification.*
>
> — Anonymous, 1970

A central question in biology is how life's great diversity and complexity results from a genetic alphabet composed of a mere four letters: adenine (A), cytosine (C), guanine (G), and thymine (T). Conceived in 1958 by Francis Crick, the central dogma of biology proposes a neat linear flow of genetic information from one gene to one protein — that DNA is transcribed into RNA that is then translated into protein. While the central dogma is certainly a workable model, it has long been considered a crude oversimplification that fails to recognize the plastic processes that occur beyond what is strictly encoded in the genomic sequence. In particular, RNA is subject to a whole host of modifications — from splicing to transcript-content modification — as it relays genomic information to the cellular machinery. The expansive catalog of transcriptional modifications highlights that RNA is no mere mediator of "hard-coded" genetic content, but instead plays a vital and dynamic role in cellular function.

## 1.1 A primer on post-transcriptional modifications

RNA is subject to over 100 types of chemically distinct post-transcriptional modifications that span all three phylogenetic domains — Archaea, Bacteria, and Eukarya [79]. RNA

[1] Portions of this chapter were published in [131].

modifications were first identified in the 1950s during the early days of RNA biology, underscoring their ubiquity in the transcriptome [26,28]. Over the last six decades, modifications have been identified in a range of RNA species where they play a pivotal role in refining RNA structure and function.

Transfer RNAs (tRNAs) are the most highly modified transcripts, with an average of 17% of their total nucleotide content subject to post-transcriptional modification [57]. Although RNA modifications are not required for tRNAs to adopt their famed cloverleaf shape, modifications allow tRNAs to adopt subtly different conformations as needed. For instance, while dihydrouridine adds conformational flexibility where present, pseudouridine adds rigidity. Three-dimensional nucleotide maps of *Escherichia coli* and *Saccharomyces cerevisiae* ribosomal RNAs (rRNAs) have also revealed that the bulk of modifications (~95% and 60%, respectively) occur in regions important for translation, such as the A, P, and E sites of tRNA- and mRNA-binding [29]. In messenger RNAs (mRNAs), 2′-*O*-methylated ribonucleotides, such as $N^6$,2′-*O*-dimethyladenosine ($m^6Am$), are often found in the 5′ untranslated region (UTR) and mark the beginning of transcripts [68]. Deamination modifications in mRNAs, which convert adenosine to inosine or cytidine to uracil, can also diversify the coding sequence of target transcripts or alter their stability when directed to the 3′ UTR [13,23,98,102].

Our increasing knowledge of the location of RNA modifications, like the ones listed above, has deepened appreciation for the wide-ranging roles they play in fine-tuning molecular function. Once thought to be constitutive, some chemical modifications, such as ribose methylation, have been found to be reversible, while others, such as pseudouridylation, can be induced in response to changes in environment. The dynamic

changes in RNA modification states are reminiscent of the DNA epigenome, and have thus led to coinage of the terms “RNA epigenome” or the “epitranscriptome.”

### 1.2 Pseudouridine: the fifth ribonucleoside

The most abundant of the post-transcriptional modifications, pseudouridine (Ψ) was the first to be discovered and is often referred to as “the fifth ribonucleoside” [26,28,79]. Ψ is the C5-glycoside isomer of uridine that results when the N1-C1′ bond linking the uracil base to the ribose sugar is broken. The base is then rotated 180º around the N3-C6 axis and a non-canonical C5-C1′ glycosidic bond is formed (Figure 1.1) [21].

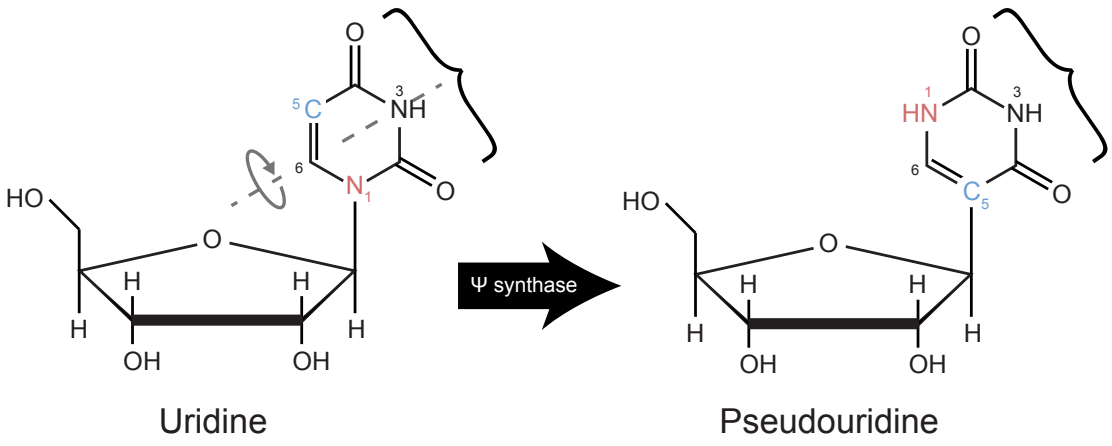


**Figure 1.1. Isomerization of uridine to pseudouridine.**
Pseudouridylation begins with the breakage of the N1-C1′ bond followed by a 180° base rotation around the N3-C6 axis. The resulting Ψ contains an additional hydrogen bond donor (red) and a C5-C1′ base-sugar linkage (blue).

Ψ’s designation as a fifth ribonucleoside is fitting given its unique physiochemical properties with respect to its U isomer. Following isomerization, the Watson-Crick edge of uridine remains unchanged, allowing for Ψ-A base pairing. Important to note, the

resulting Ψ has an additional hydrogen bond donor at the N1 position. In an RNA chain, Ψ's ability to coordinate a structural water molecule via its N1H group confers added rigidity to RNA structure by increasing base stacking and adding extra hydrogen bonds between the base and its phosphate backbone. Additionally, N1H-mediated water coordination has been reported to increase Ψ/A base-pairing stability compared to the U/A pair [90]. Ψ's additional hydrogen bond donor has also been thought to contribute to novel base pairing interactions in Ψ-containing RNA [21,96]. In fact, recent structural studies demonstrate that the ribosome can accommodate non-canonical codon-anticodon base pairing mediated by a pseudouridylated sense codon, the functional import of which is discussed later in this chapter [37].

**1.3 Site-specific pseudouridylation is catalyzed by two distinct mechanisms**

Site-specific pseudouridylation is catalyzed by pseudouridine synthases (PUSs) through one of two distinct mechanisms: a protein-only (stand-alone) mechanism and a box H/ACA snoRNP-catalyzed (guide-dependent) mechanism [112]. Stand-alone pseudouridylation is catalyzed by a single PUS that recognizes its particular substrate, either through a specific consensus motif or secondary structure [14,16,78,110].

On the other hand, RNA-dependent pseudouridylation is mediated by an RNA-protein (RNP) complex, consisting of four core proteins — Nhp2p, Gar1p, Nop10p, and the Ψ-synthase Cbf5 (Nap57/dyskerin in mammals) — assembled on a box H/ACA small nucleolar RNA (snoRNA) scaffold. Each H/ACA snoRNA folds into a conserved hairpin-hinge-hairpin-tail structure (Figure 1.2). Each hairpin contains a unique single-stranded internal loop — the pseudouridylation pocket — that is complementary to a specific sequence in a substrate RNA, flanking 3-10 nucleotides on either side of a

particular target uridine [41,91]. The substrate RNA base pairs with the Ψ pocket, positioning the target uridine at the base of the upper stem of the hairpin where Cbf5 then site-specifically catalyzes pseudouridylation. Notably, the pseudouridylation pocket's short guide sequence is split by a hairpin structure that is variable in length, making it difficult to computationally predict a particular H/ACA snoRNA's target RNA for pseudouridylation. While several stand-alone PUSs are not required for cell viability, Cbf5 deficiency is lethal. High-throughput sequencing techniques have identified a growing set of snoRNAs with unknown target sites, suggesting there is still much of the Ψ landscape left to be charted [22,54,71,105,128].

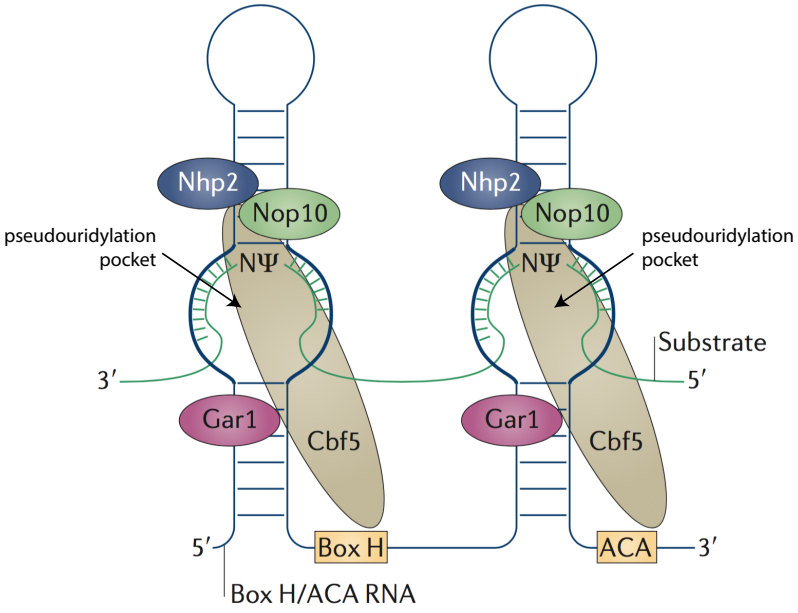


**Figure 1.2. Schematic of eukaryotic box H/ACA snoRNP complex.**
H/ACA snoRNA forms a hairpin-hinge-hairpin tail structure, which coordinates four core proteins: Nhp2, Nop10, Gar1, and Cbf5. The guide sequence in the pseudouridylation pocket base pairs with the complementary substrate RNA, directing the site-specific isomerization of the target U by the Ψ-synthase Cbf5 (figure courtesy of Yi-Tao Yu).

### 1.4 Hints at the biological significance of pseudouridylation

Pseudouridine's distinct structural properties make it unsurprising that Ψs are well-known to cluster in evolutionarily conserved and functionally important regions of stable noncoding RNAs (ncRNAs). Over the years, appreciation for the significant role pseudouridylation plays in RNA function has grown. Ψ's functional relevance has been well-documented in rRNAs, where pseudouridylation is required for ribosome biogenesis and translational fidelity and efficiency, and in small nuclear RNAs (snRNAs), where specific Ψ residues have been identified as necessary for proper pre-mRNA splicing [12,56,70,127,130]. Furthermore, many Ψs in rRNAs and snRNAs are conserved across species, occurring at identical or near-identical sites [29,126].

#### 1.4.1 Pseudouridylation is conditionally induced in different cellular contexts

Once thought to be a constitutive modification, pseudouridylation has been found to be inducible in response to cellular stress and differentiation, suggesting pseudouridylation may provide a dynamic regulatory mechanism for RNA function [11,88,125].

Following heat shock and nutrient deprivation, two novel Ψs were identified in yeast U2 spliceosomal snRNA: Ψ56 and Ψ93 [125]. While Ψ56 conversion is catalyzed by the stand-alone PUS Pus7, Ψ93 is targeted by the H/ACA snoRNP complex guided by snR81 [125]. Notably, both inducible Ψs are flanked by sequences that deviate from the canonical motifs recognized by Pus7 and snR81. For instance, the Ψ pocket of snR81 — known to modify Ψ42 in U2 snRNA and Ψ1051 in 25S rRNA — pairs with two mismatches to the sequence flanking Ψ93. This finding is contradictory to previously identified constitutive RNA-dependent targets of pseudouridylation, which pair with perfect sequence complementarity (i.e. without mismatches) to their corresponding guide

snoRNAs. Imperfect sequence complementarity between the guide and substrate RNAs is therefore a likely hallmark of conditionally inducible Ψ targets. Importantly, Ψ93 interferes with pre-mRNA splicing, suggesting a role in altering gene regulation in response to nutrient deprivation.

Developmentally dependent Ψ28 in U6 spliceosomal snRNA, which is guided by the stand-alone PUS Pus1, has been found to initiate a filamentous growth program in yeast, which is triggered by, for instance, nitrogen- or glucose-starved environments or exposure to fusel alcohols [11]. Ψ28 is not present during log-phase growth and is not induced by other standard stress conditions, such as heat-shock, indicating that this alternate site of pseudouridylation is induced by filamentation-specific environmental stressors. Like Ψ93 in U2 snRNA, U6-Ψ28 affects pre-mRNA splicing, this time reducing the splicing efficiency of suboptimal introns. Altered splicing to target transcripts may therefore activate mRNAs necessary for filamentous growth, or inactivate those that inhibit such a growth program.

### 1.4.2 Pseudouridylation alters amino acid decoding

Ψ has long been known to play a role in translation of mRNAs. rRNA pseudouridylation is essential for translation fidelity, and pseudouridylated anticodons have been shown to alter ribosomal decoding in echinodermal mitochondrial RNA [115]. However, because the possibility of mRNA pseudouridylation had never been closely studied, the effect of pseudouridylation in protein-coding transcripts remained unknown. While studies on pseudouridylation had traditionally focused on its role in tRNAs, snRNAs, and rRNAs, largely due to their abundance, there was no reason to assume that pseudouridylation substrates should be restricted to this class of noncoding RNAs.

Given that the structure of H/ACA snoRNAs is so well-conserved, guide RNAs can theoretically be engineered to target pseudouridylation to any RNA of interest by modifying the guide sequence in the Ψ pocket [53]. In a proof of principle experiment, the Yu group at University of Rochester Medical Center engineered guide RNAs derived from the naturally occurring yeast H/ACA snoRNA *SNR81* to target pseudouridylation to mRNA to investigate the effect of Ψ in protein-coding transcripts. Ψ was artificially targeted to a premature stop codon within a reporter mRNA to monitor translation termination efficiency. Interestingly, introducing Ψ into each of the known stop codons (UAA, UAG, UGA) suppresses translation termination by directing the incorporation of biochemically and structurally similar amino acids. Specifically, ΨAA and ΨAG code for serine and threonine, while ΨGA codes for tyrosine and phenylalanine [61]. Further studies have confirmed similar nonsense-to-sense codon conversion in bacteria, suggesting that Ψ-mediated recoding is conserved in prokaryotes and eukaryotes [37].

Ψ's recoding potential is strengthened by structural studies that demonstrate the ribosome can accommodate non-canonical codon-anticodon base pairing mediated by a pseudouridylated sense codon [37]. The crystal structure was resolved for ΨAG pairing with the $tRNA^{Ser}$ anticodon stem loop AGI. The decoding center's unexpected plasticity suggests that Ψ may similarly recode sense codons, thereby expanding the genetic code and generating protein diversity beyond what is encoded in genomic DNA. This finding is all the more intriguing given the possibility of condition-dependent pseudouridylation events in coding regions in response to changes in environment. In fact, ΨUU, which is derived from the phenylalanine-encoding UUU codon, has been found to code for cysteine and tyrosine (Yu, personal communication). While some groups have

theoretically predicted what other pseudouridylated sense codons could be coding for, they still agree that more experimental data are needed before more accurate predictions can be made [96].

*In vitro*-transcribed mRNAs in which every U residue is pseudouridylated have also been found to exhibit enhanced stability and translation efficiency when delivered *in vivo* [62]. Important to note, however, is that fully pseudouridylated mRNAs synthesized for this study were translated into functional proteins (i.e. GFP, lacZ, and luciferase). While the protein products were not sequenced to determine if Ψ facilitated alternate amino acid incorporation, the likelihood that a functional protein would result from multiple codon recoding events is low. Consequently, the number or density of Ψs within a particular protein-coding transcript could perhaps play a role in Ψ-mediated recoding.

### 1.5 Methods of pseudouridine detection

Pseudouridine was first identified as an unknown ribonucleoside in 1951 by subjecting calf liver RNA isolates to ion-exchange chromatography [26]. Because Ψ is mass-silent with respect to U, rather labor-intensive chromatographic techniques continued to be the prevailing method for Ψ detection. These methods took advantage of the effect of Ψ's additional hydrogen bond donor on migration. As the field advanced, a combination of RNase digestion, radiolabeling, and chromatography-based methods produced the first pseudouridine maps in tRNAs and rRNAs [47,52,113]. Notably, these approaches required large amounts of purified RNA as a starting material, and were thus limited to studying only highly abundant RNA species.

In 1993, a method was developed by Bakin and Ofengand taking advantage of the carbodiimide CMC (*N*-Cyclohexyl-*N*′-(2-morpholinoethyl)carbodiimide metho-*p*-

toluenesulfonate) to label Ψ residues [7]. Under physiological conditions, CMC acylates guanosine (G) at the N1 position and uracil at the N3 position (Figure 1.3A). Notably, isomerization to Ψ creates an additional CMC conjugation site, so CMC acylates Ψ residues at the N1 and N3 positions. CMC adducts are susceptible to alkaline hydrolysis (pH=10.4), except in Ψ where CMC remains specifically and irreversibly bound at the N3 position. Traditionally, the method has been coupled to primer-extension assays to map sites of pseudouridylation, as Ψ-CMC adducts result in reverse transcriptional (RT) arrest one base downstream of a Ψ site (Figure 1.3B). Ψ-CMC is thus detectable as a distinct stop, whereas without conjugation to CMC, Ψs are indistinguishable from U by the reverse transcriptional machinery. Since the CMC/RT approach was introduced, it has become the primary means of Ψ detection.

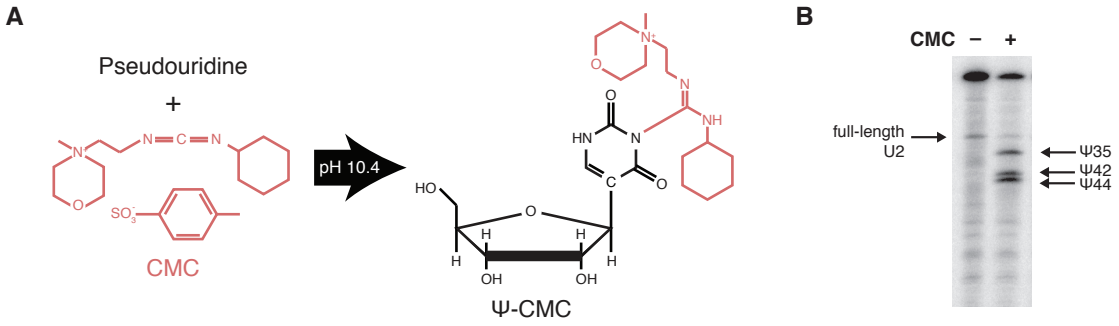


**Figure 1.3. CMC specifically labels Ψ and causes RT arrest one base 3' to Ψ.**
(A) CMC specifically labels pseudouridine. Following alkaline hydrolysis, CMC (red) remains bound to the N3 position of Ψ. (B) Reverse transcription using a primer specific to U2 snRNA maps Ψ-CMC-mediated RT arrest sites (right lane) when compared to mock-treated control (left lane).

While the CMC/RT approach is not quantitative, CMC derivatization has been coupled to matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS) to quantify the relative abundance of derivatized Ψ [97]. When CMC is conjugated to its Ψ target, it can be detected as a distinct mass shift of 252 Da. A key limitation of this method, which is discussed more in the next chapter, is its reliance on uniform derivatization of CMC to its targets for accurate quantitation.

In more recent years, limitations inherent in CMC derivatization have incentivized the development of CMC-independent techniques. In particular, site-specific RNase H cleavage of a candidate Ψ site was combined with splinted ligation, ribonuclease digestion, and thin layer chromatography to identify hypothesized sites of pseudouridylation [75]. The method, termed **S**ite-specific **C**leavage **A**nd **R**adioactive-labeling followed by **L**igation-assisted **E**xtraction and **T**hin-layer chromatography (SCARLET), has the added benefit of quantitatively detecting the extent to which a particular Ψ is modified. In addition, mass spectrometry techniques have been developed to exploit Ψ's unique physiochemical features independent of CMC conjugation. More specifically, Ψ's noncanonical C–C glycosidic bond yields a unique fragmentation pathway following collision-induced dissociation (CID), the products of which can be detected by liquid chromatography tandem mass spectrometry (LC-MS/MS) [1].

Site-specific Ψ-mapping for each of the methods summarized above, however, requires prior knowledge of the Ψ-containing sequence of interest, preventing an unbiased detection approach. In addition, with the exception of SCARLET, the current methods have been developed to detect Ψ in relatively abundant RNAs, ruling out detection of Ψ in more lowly expressed transcripts, such as mRNAs.

### 1.6 Statement of the problem

The advent of high-throughput RNA sequencing and the development of increasingly sophisticated bioinformatic methods to analyze the resulting data have led to the creation of techniques to specifically map RNA modifications across the transcriptome. For instance, utilizing an $m^6A$-specific antibody to immunocapture modified transcripts has allowed for transcriptome-wide localization of $m^6A$, while analyzing specific RT-arrest and nucleotide misincorporation profiles has led to the global identification of $N^1$-methyladenosine ($m^1A$) residues [32,49]. Modification maps have allowed for the generation of testable hypotheses to continue probing the functional relevance of the modification in question.

In contrast to the growing body of work pointing to the biological functions of pseudouridylation, further inquiry was limited by the available methods for site-specific Ψ detection. For instance, despite pseudouridine's recoding potential, pseudouridylation of native mRNA transcripts had never been observed. Elucidating the role of pseudouridylation in naturally occurring RNAs would therefore require the development of a high-throughput, unbiased, and sensitive approach to identify Ψs. As a result, I set out to develop a deep-sequencing approach for Ψ detection, outlined in Chapter 2, adapting CMC derivatization to a high-throughput format.

During my pilot Ψ-profiling experiments, three CMC-based approaches to transcriptome-wide detection were published, with a fourth technique released shortly thereafter. Collectively, these methods — called Pseudo-seq, Ψ-seq, PSI-seq, and CeU-seq, in order of publication — catalogued thousands of novel sites of pseudouridylation across a number of species and in a range of environmental contexts. The availability of

four independent yet interrelated methods provided a unique opportunity for a critical, cross-method comparison of their respective results. I therefore undertook such a comparative analysis, which revealed previously undiscussed shortcomings of each approach, detailed in Chapter 3. I then applied lessons learned from the caveats I uncovered to improve the now available Ψ-detection approaches.

The original aim of mapping *where* Ψ sites are was to understand *what* Ψ sites do. In other words, Ψ-detection approaches were developed to better understand the role pseudouridylation plays in biological systems. As a result, I chose to apply my improvements to characterize Ψ profiles at two life cycle stages in the digenetic protozoan parasite *Trypanosoma brucei.* In so doing, I hoped to begin to unravel the role differential pseudouridylation might play in cellular differentiation. However, my preliminary experiments in this system revealed unanticipated concerns surrounding robust, reproducible high-throughput CMC-based Ψ detection, discussed in Chapter 4.

Beyond pseudouridylation, an underlying theme in this thesis is the importance of well-documented experimental failure. Therefore, I have attempted to rigorously investigate and characterize potential sources of my failure to implement high-throughput Ψ-detection. I hope the work undertaken in this thesis might set an example for how to transparently and productively discuss caveats and experimental limitations of the scientific practice more broadly.

## CHAPTER 2. Developing a high-throughput approach for Ψ detection

To gain a better understanding of Ψ and its potential role in modulating cellular function, we must first know what subset of transcripts are targeted for pseudouridylation and where in those transcripts Ψ occurs.

Since its introduction in 1993, CMC derivatization and subsequent alkaline hydrolysis coupled to primer extension has become the primary means of Ψ detection [7]. Despite its popularity, the technique comes with two primary limitations:

1. Primer design requires prior knowledge of the Ψ-containing sequence of interest, precluding unbiased discovery of pseudouridine residues.
2. CMC-dependent pseudouridine mapping was developed to detect Ψ residues in relatively abundant RNA species (i.e. rRNAs, tRNAs, snRNAs), where uridine is highly isomerized to pseudouridine. For instance, the majority of Ψ residues in *Schizosaccharomyces pombe* are isomerized from U to Ψ at an efficiency of 85% or higher [114]. Therefore, low efficiency pseudouridylation events and Ψs in lowly abundant transcripts, like mRNAs, are unlikely to be detected using the traditional, low throughput CMC-based approach.

Both limitations can be circumvented with the advent of next-generation sequencing technologies, which can interrogate the entire transcriptome at high depth for sites of Ψ-CMC-mediated reverse transcriptional arrest. Coupling CMC conjugation with stranded RNA sequencing (RNA-seq) therefore allows for mapping of novel Ψ targets, and is the basis of the high-throughput sequencing approach laid out in this chapter.

### 2.1 CMC derivatization and alkaline hydrolysis optimization

CMC conjugation is not without its challenges. Specifically:

1. CMC does not conjugate to all G- and U-like residues with uniform efficiency. Thus, the presence of underivatized Ψ residues will lead to false negatives [33].
2. Likewise, alkaline cleavage of CMC adducts to non-Ψ residues occurs at incomplete efficiency, so failure to cleave CMC from G-like and U-like residues will result in false positives [33].

Conditions for CMC treatment and alkaline hydrolysis vary within the literature, specifically with respect to three variables: (1) alkaline solution pH, (2) incubation time, and (3) incubation temperature [7,33]. Before proceeding to a pilot study coupling CMC derivatization with RNA-seq, it was therefore essential to establish a standardized derivatization protocol to maximize CMC conjugation efficiency to pseudouridine, while minimizing RNA degradation resulting from alkaline hydrolysis.

To ensure optimal reaction conditions, a method of monitoring CMC derivatization and subsequent cleavage from non-Ψ residues was required. As a result, I collaborated with Dr. Mark Helm's group at Johannes Gutenberg-Universität Mainz, where they had synthesized a CMC derivative called *N*-cyclohexyl-*N*′-β-(4-propargylmorpholinum) ethylcarbodiimide or CMCyne. The compound importantly contained an alkyne group for Copper(I)-Catalyzed Azide-Alkyne Cycloaddition, the classic "click" chemistry reaction (Figure 2.1A). To track CMC adducts, I could then take advantage of a fluorescent azide, $atto_{488}$, which could be conjugated to CMCyne following derivatization to its target U- and G-like residues (Figure 2.1B). While $atto_{488}$ does not provide an absolute measurement of CMCyne conjugation, diminishment of a fluorescence signal following hydrolysis provided a relative gauge of cleavage efficiency.

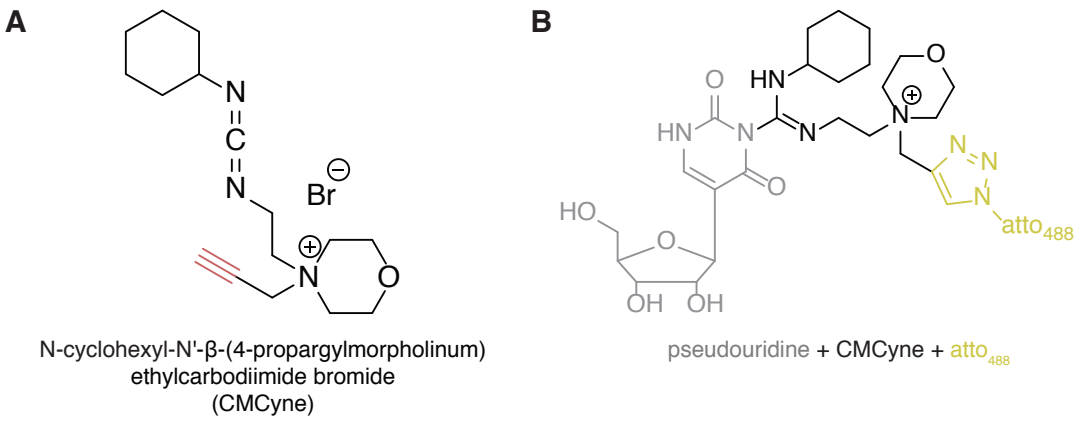


**Figure 2.1. Structures of CMCyne derivatives.**

(A) Structure of CMCyne with alkyne group in red. (B) Structure of CMCyne conjugated to pseudouridine (grey) and atto$_{488}$-azide (yellow) following azide-alkyne cycloaddition.

I first established that CMCyne could conjugate to U- and G-like residues like the commercially available CMC. *In vitro*-transcribed tRNA tyrosine (IVT tRNA$^{Tyr}$) containing only unmodified ribonucleotides was therefore derivatized with CMCyne followed by atto$_{488}$ conjugation. The product of the reaction was visualized on a 15% SDS-PAGE gel, which confirmed CMCyne had conjugated to U and G residues in IVT tRNA$^{Tyr}$. Next, I tested a number of conditions for alkaline hydrolysis with 50 mM $(NH_4)_2CO_3$ by varying the pH, temperature, and time of the reaction, which are detailed in Table 2.1.

**Table 2.1. Detailed conditions for alkaline hydrolysis optimization.**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **pH** | 10.5 | 11.0 | 11.5 | 10.5 | 11.0 | 11.5 | 10.5 | 11.0 | 11.5 | 10.5 | 11.0 | 11.5 |
| **Temperature** | 37˚C | | | | | | 42˚C | | | | | |
| **Time** | 2.5 hours | | | | | | 3.0 hours | | | | | |

The schematic for alkaline hydrolysis optimization is detailed in Figure 2.2A. Following CMCyne conjugation to IVT $tRNA^{Tyr}$, samples were subjected to alkaline hydrolysis under all of the above conditions. Treatment with $H_2O$ in place of alkaline solution was used as a control. Because IVT $tRNA^{Tyr}$ contains no pseudouridine residues, all CMCyne should be cleaved from U and G residues under efficient alkaline hydrolysis. As a result, $atto_{488}$-azide would have no available substrate for conjugation. I could therefore monitor CMCyne cleavage efficiency as a loss of fluorescence signal. I also monitored RNA degradation under alkaline conditions by subjecting my samples to a GelRed stain following fluorescence imaging on a 15% SDS-PAGE gel (Figure 2.2B).

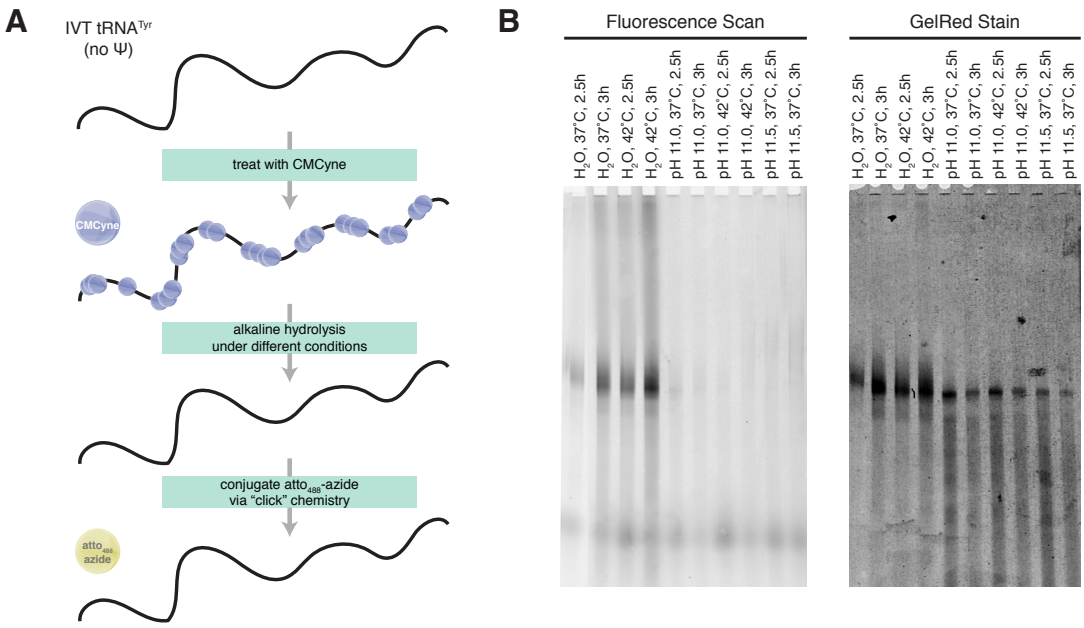


**Figure 2.2. Optimizing conditions for efficient alkaline hydrolysis.**
(A) Schematic of perfectly efficient alkaline hydrolysis reaction in which all CMCyne is cleaved from non-Ψ residues leaving $atto_{488}$ without a substrate. (B) Following CMCyne derivatization, samples were subjected to alkaline hydrolysis under a range of conditions. RNA was visualized on a 15% SDS-PAGE gel, scanned for fluorescence (left), and then stained with GelRed (right). Samples treated at pH 10.5 are not shown.

Based on the results detailed in Figure 2.2B, I settled on treatment with 50 mM $(NH_4)_2CO_3$ at pH 11.0 at 37˚C for 2.5 hours as sufficiently efficient. Under these conditions, fluorescence from conjugated $atto_{488}$ substantially diminished with respect to the $H_2O$ control, while RNA degradation remained minimal.

Finally, I confirmed that under these conditions, I could distinguish a uridine from a pseudouridine. To do so, I took advantage of two short oligoribonucleotides (~30 nt long) synthesized by the Helm group; both were identical except one contained a single U residue, while the other contained a Ψ in its place (Figure 2.3A). Each oligoribonucleotide was subjected to optimized CMCyne treatment. Following fluorescent labeling with $atto_{488}$-azide, the reaction conditions were found to be sufficient for distinguishing a Ψ from a U within the oligonucleotide (Figure 2.3B).

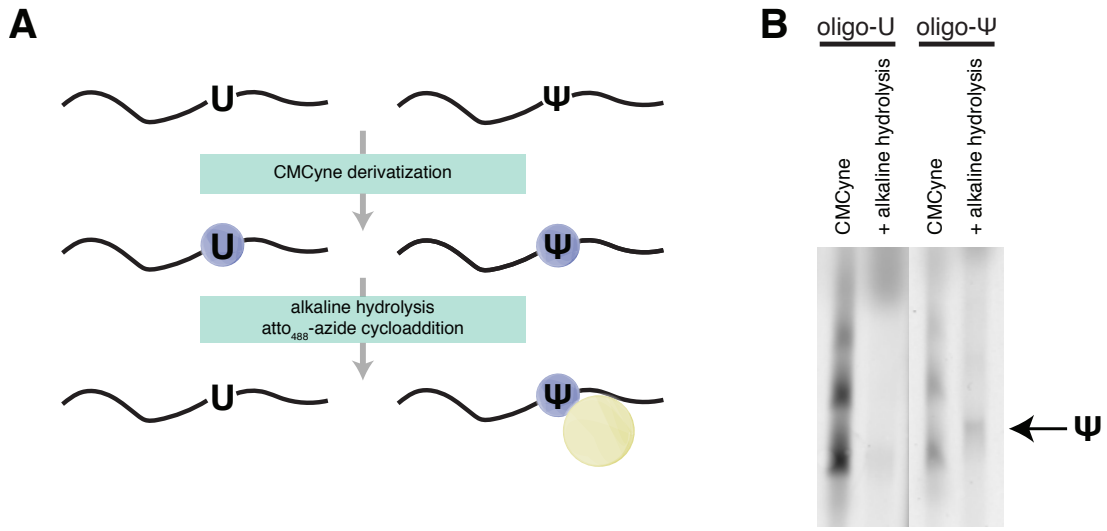


**Figure 2.3. Optimized CMCyne-alkali treatment can distinguish Ψ from U.**
(A) Schematic of CMCyne-alkali treatment with $atto_{488}$ conjugation. (B) Derivatization of the Ψ-containing oligonucleotide results in a fluorescent signal (indicated by arrow), which is absent in its U-containing counterpart.

### 2.2 Pilot high-through Ψ-detection experiment with *S. cerevisiae r*RNA

Having established an optimized protocol for CMCyne conjugation and subsequent cleavage, I began a pilot experiment to map sites of pseudouridylation using RNA sequencing. CMCyne would have been the ideal compound for derivatization because its alkyne group allowed for conjugation with a biotin-azide, which would in turn allow for enrichment of Ψ-CMCyne-containing transcripts prior to sequencing. However, the compound was available only in limited quantity. As a result, I chose to move forward with commercially available CMC by Sigma for this experiment.

*S. cerevisiae* (budding yeast) was initially used as a model organism, given the relative low complexity of the yeast genome, which has the additional benefit of being very well annotated. I chose to pilot a high-throughput Ψ-detection method on yeast ribosomal RNAs. In addition to the high abundance of rRNA species (~95% of the transcriptome), sites of pseudouridylation within yeast rRNAs, as well as snRNAs and tRNAs, have been well-characterized and corroborated by a number of independent studies conducted during log phase and stationary phase growth [6,8,103]. Several of these studies have also attributed specific PUS or H/ACA snoRNA activity to a particular Ψ [30,41]. A successful high-throughput Ψ-mapping method, which I will call CMC-seq, would robustly detect these known sites of pseudouridylation with a low false positive rate, validating the success of my approach.

#### 2.2.1 Library preparation method for pilot CMC-seq experiment

To prepare libraries for CMC-seq, total RNA was extracted from yeast cells grown to log phase, and treated in duplicate with CMC followed by alkaline hydrolysis; mock-treated (i.e. without CMC) samples were processed in parallel. Mock-treated libraries ensured

that premature RT termination sites were due specifically to Ψ-CMC and not, for instance, natural stops due to RNA secondary structure. Standard library preparation protocols next call for fragmentation of RNA to a uniform length prior to first strand synthesis to ensure uniform coverage across the transcriptome (Figure 2.4). However, because alkaline hydrolysis leads to fragmentation, this step was skipped to avoid generating sequencing reads that were too small to be reliably mapped back to the genome. To confirm that hydrolysis resulted in uniform fragment length, I ran samples before and after treatment on the Agilent 2100 Bioanalyzer system (Figure 2.5).

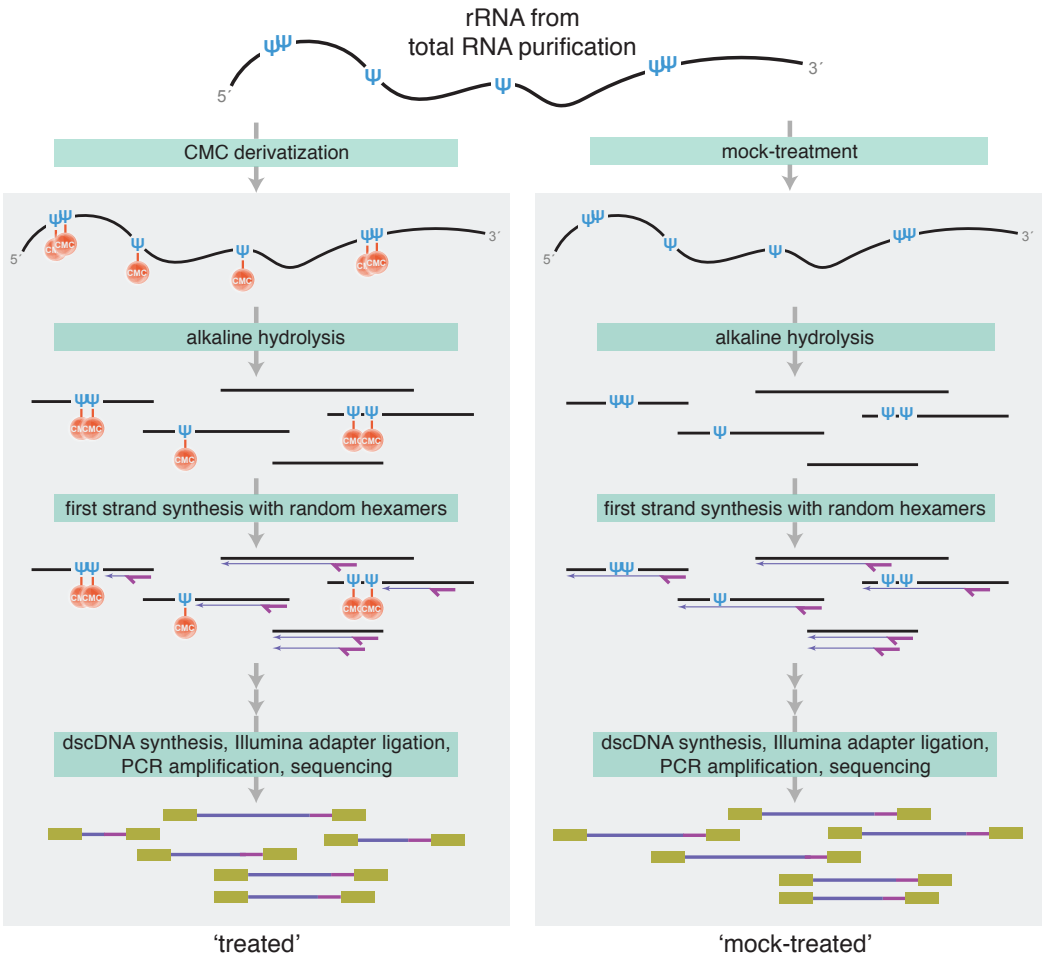


**Figure 2.4. Schematic of CMC-seq library preparation for total RNA.**

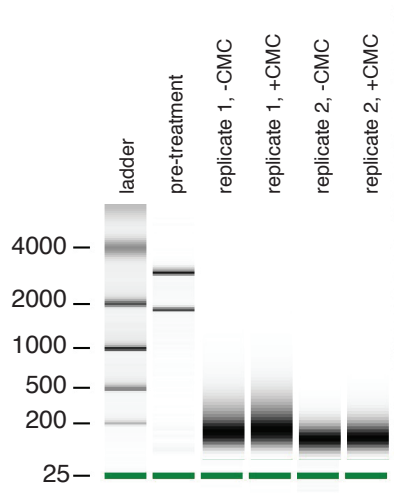


**Figure 2.5. Alkaline hydrolysis is sufficient for uniform RNA fragmentation.**

Stranded CMC-seq libraries were then prepared by reverse transcription with random hexameric primers, followed by second strand synthesis to generate double-stranded complementary DNA (dscDNA). Illumina adapters were then ligated onto dscDNA after end repair and A-tailing, followed by PCR amplification. The resulting libraries were finally sequenced with the Illumina HiSeq 2000 in 100bp single-end mode.

### 2.2.2 Initial strategy for data analysis: comparing read coverage in treated versus mock-treated CMC-seq libraries

The initial strategy for bioinformatic analysis was to first map RNA-seq reads back to a single rDNA repeat within the rDNA locus of the yeast genome (SacCer3) and calculate read coverage at each position. Per base coverage was normalized with DESeq to account for differences in library size [3]. Because Ψ-CMC adducts mediate premature RT arrest, CMC-treated libraries will exhibit an increase in truncated reads around sites of pseudouridylation, which translates into valleys in read coverage around the bases flanking a putative Ψ. By comparison, mock-treated libraries will contain a higher

number of reads in pseudouridylated regions of transcripts. I therefore calculated what I called the 'CMC-stat' for each position along rRNA (Equation 2.1). I defined CMC-stat as the $\log_2$-transformed ratio of the median number of reads covering a given position in mock-treated versus treated libraries.

**Equation 2.1.** $$\text{CMC-stat} = \log_2\left(\frac{\text{reads at position}^{-\text{CMC}}}{\text{reads at position}^{+\text{CMC}}}\right)$$

The CMC-stat was then plotted for each position along the length of a given rRNA transcript (Figure 2.6). As predicted, known sites of rRNA pseudouridylation correspond well with peaks in CMC-stat values for 18S and 25S rRNA as a result of more reads covering mock-treated transcripts compared to their treated counterparts. However, peaks were not observed for the shorter 5.8S and 5S rRNAs, which is likely attributed to a size selection step following dscDNA synthesis that removed fragments of less than 100 nucleotides in length. In fact, DESeq-normalized read coverage (corrected for transcript length) of both 5.8S and 5S rRNA is significantly lower than that of 18S and 25S rRNA (Figure 2.7). Because both transcripts are less than 150 nucleotides long, reads that spanned the length of the transcripts were enriched following dscDNA clean-up. In other words, short dscDNA fragments resulting from Ψ-CMC-mediated reverse transcriptional arrest were removed prior to sequencing, particularly given that Ψ is located roughly in the middle of both 5.8S and 5S rRNA.

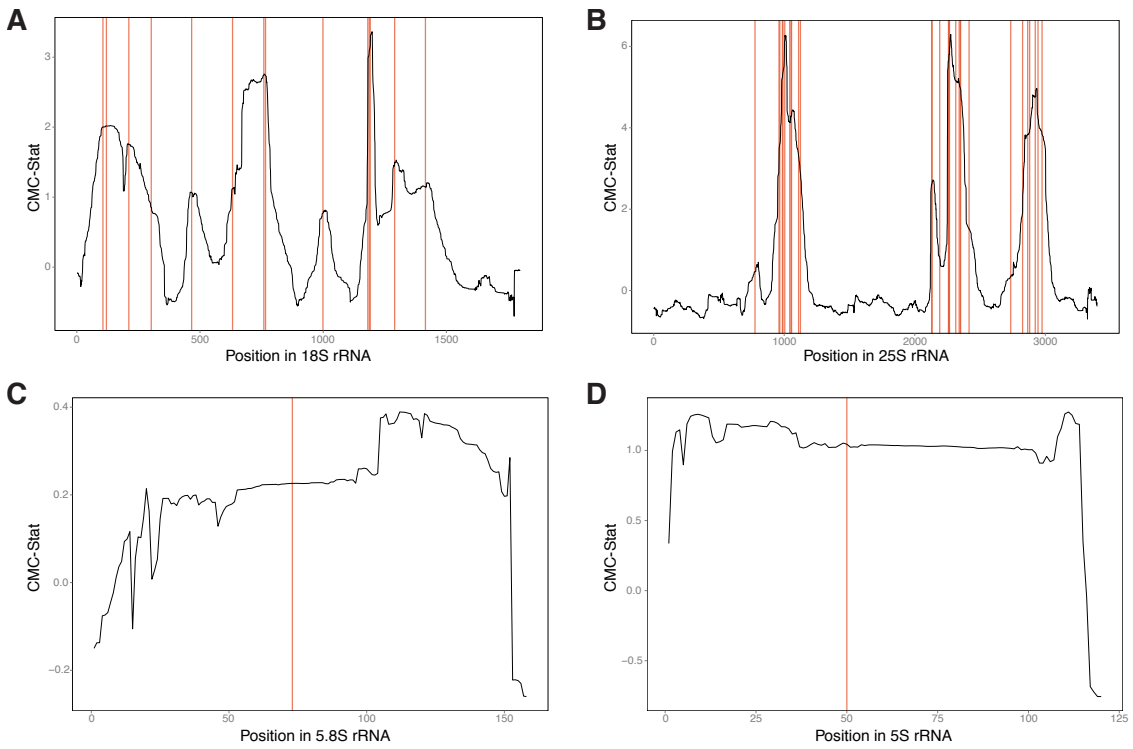


**Figure 2.6. CMC-stat plots show broad peaks around clusters of rRNA Ψ sites.**

CMC-stat values were plotted for each position along (A) 18S rRNA, (B) 25S rRNA, (C) 5.8S rRNA, and (D) 5S rRNA transcripts. Known Ψs are indicated with red vertical lines.

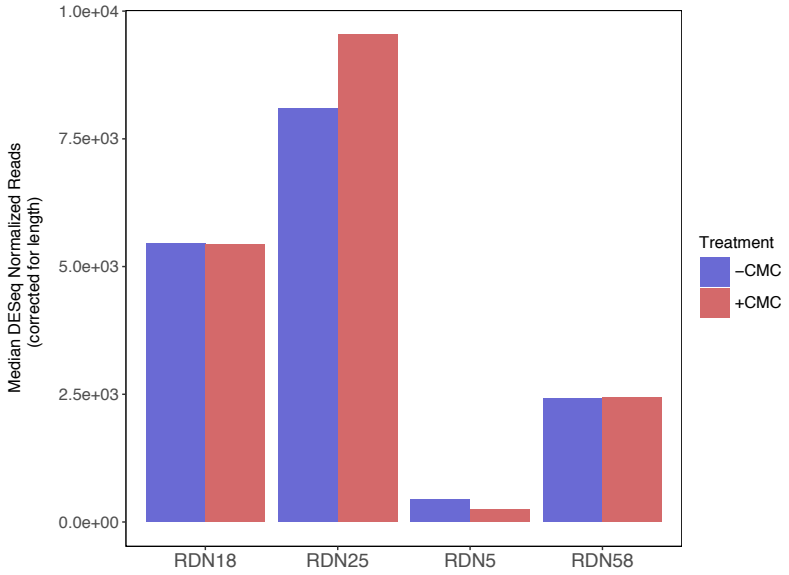


**Figure 2.7. DESeq normalized reads corrected for length of rRNA transcripts.**

While CMC-stat plots were able to qualitatively identify Ψs as peaks in the CMC-stat metric, this approach could not computationally pinpoint putative Ψs at single nucleotide resolution, which is essential to detecting novel targets of pseudouridylation across the entire transcriptome. If the peaks were of a uniform shape, I could interrogate the transcriptome for CMC-stat peaks of that same shape to narrow down a set of putative Ψ sites, which could then be validated using low-throughput methods. However, peak shape was quite variable, so clear Ψ-calling cutoffs could not be concretely defined, even based on known sites of pseudouridylation. rRNA represents the best case scenario for Ψ detection, with its high levels of expression guaranteeing high coverage and its highly isomerized Ψ residues facilitating a strong Ψ-CMC RT stop signal. Because CMC-stat analysis was difficult to interpret for rRNA, I concluded that it would be poorly suited for a transcriptome-wide analysis. As a result, I tested an alternative approach to Ψ mapping, detailed in the next section.

### 2.2.3 Alternative data analysis strategy: Analyzing nucleotide misincorporation during cDNA synthesis to identify Ψs

Post-transcriptionally modified nucleotides have been known to alter reverse transcriptase processivity [123]. Alterations to the Watson-Crick face of a nucleotide can impose a roadblock to complementary nucleotide incorporation by the reverse transcriptional machinery. As a result, the polymerase stalls at the modified site, which in turn can lead to nucleotide misincorporation. By analyzing RNA-seq data, several groups have identified characteristic nucleotide misincorporation signatures that distinguish a particular RNA modification from random sequencing errors [49,50,104].

Because CMC derivatizes to the Watson-Crick edge of Ψ, and because Ψ-CMC adducts have been found to cause a "stuttered" stop in RT at the modification site (as opposed to one base 3′ of Ψ), I decided to take a closer look at alternative nucleotide incorporation frequencies [6]. Given yeast contain over 100 rDNA repeats, intra-genomic DNA polymorphisms naturally exist among the many gene copies, which must be distinguished from modification-driven mismatched nucleotide incorporation [58,122]. I therefore filtered out known polymorphisms from my analysis. In addition, mispriming events introduced by random hexamers during first strand synthesis and a drop in base calling quality result in higher sequencing error rates at the ends of reads. As a result, I trimmed two bases off the ends of each read. Importantly, rRNA positions exhibit excellent read coverage; however, if this method were applied to the entire transcriptome, only positions with sufficient coverage should be analyzed to ensure a modification-mediated mismatch can be distinguished from sequencing errors or mis-mapping events.

Following these filtering steps, I calculated the ratio of non-reference nucleotide incorporation, or the 'mismatch rate' (MR), for each position (Figure 2.8A). I then calculated the $\log_2$-transformed ratio of treated versus mock-treated mismatch rates to compare whether mismatch frequencies at modified nucleobases were indeed higher (Figure 2.8B). Analysis was next narrowed to a subset of rRNA positions with a median mismatch rate of greater than 1.5% in treated samples (Figure 2.8A, purple points) and with over four times the median MR in treated versus mock-treated samples (Figure 2.8B, maroon and green points). The MR cutoff was based on an assumed base-calling error rate of ~1% using the Illumina sequencing platform [77].

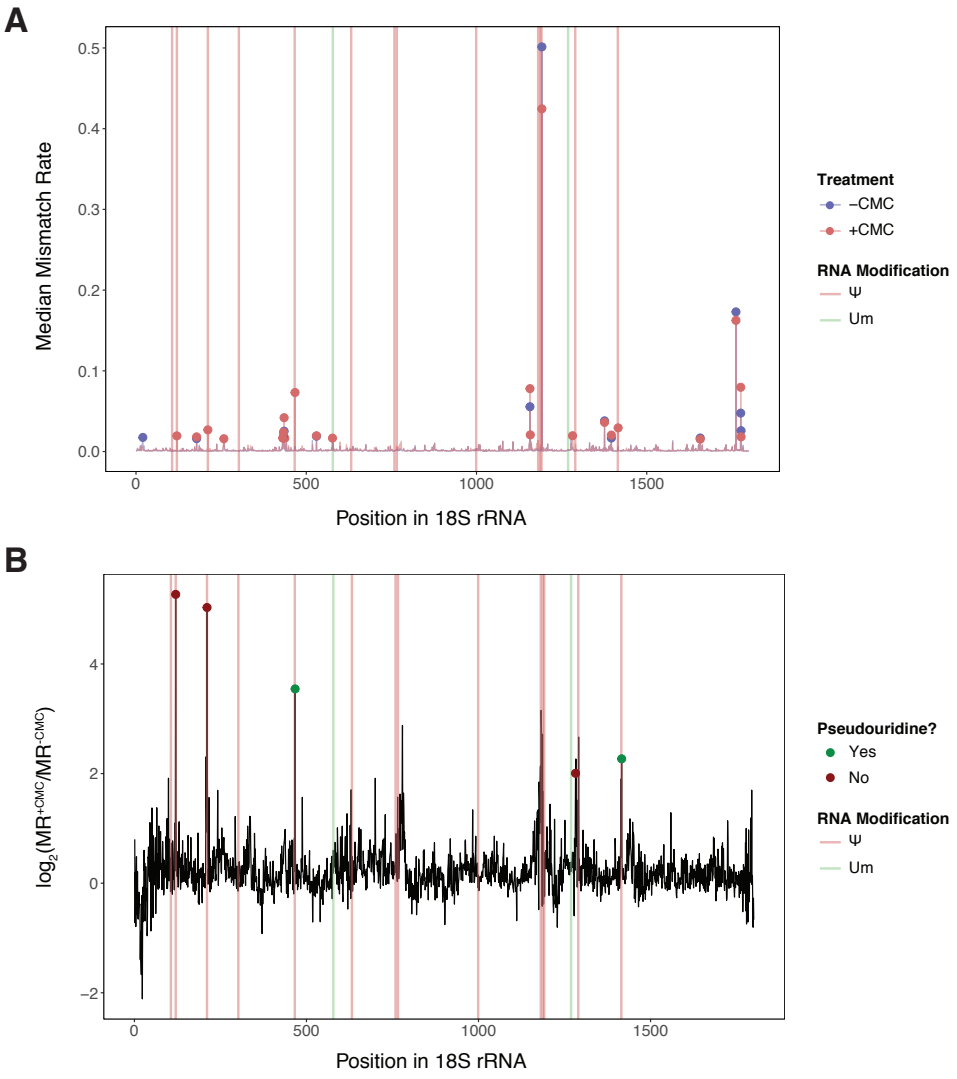


**Figure 2.8. Non-reference nucleotide incorporation rates for rRNA.**

(A) The median mismatch rate was plotted for each position. MRs of greater than 1.5% are indicated as points. (B) The $\log_2$-transformed ratio of treated versus mock-treated median MRs was plotted for each position. Sites with a greater than four-fold higher treated mismatch rate of at least 1.5% are indicated as points. Green points indicate a true Ψ site. Known RNA modification sites are indicated by vertical lines.

With this mismatch analysis, three of 46 total pseudouridines (6.52%) were successfully detected across all four rRNAs, while an additional 12 called sites were false

positives. If I included an additional filter to require a 'U' as the reference position, the number of false positives would decrease to three. Clearly, however, employing mismatch rate as a Ψ-detecting metric is not sufficiently adequate for identifying the bulk of known Ψ sites. Still, I chose to take a closer look at the alternate nucleotide incorporation profiles by base of the three true Ψ hits from this mismatch analysis (Figure 2.9, yellow highlight). All three Ψ positions show a clear bias towards sequencing of C with varying levels of A, which is in line with observations made by other groups, and with closer inspection of all Ψ residues in my CMC-seq data set [104].

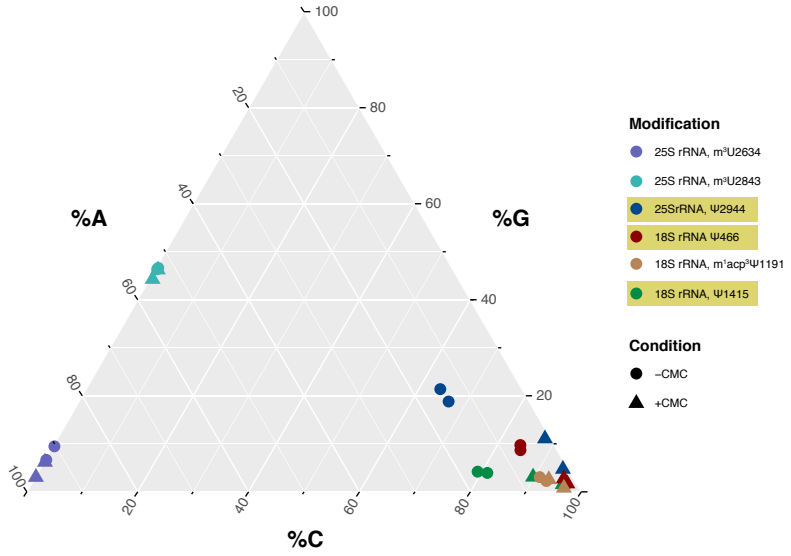


**Figure 2.9. Nonreference nucleotide incorporation profiles at U positions of interest.** Nonreference nucleotide incorporation frequencies for each treated (triangle) and mock-treated (circle) replicate were plotted for true Ψ sites identified during mismatch analysis (highlighted in yellow). Three additional points of interest were identified and plotted.

In addition, I analyzed the mismatch profiles of three modified sites that had high mismatch rates in both treated and mock-treated samples: $N^1$-methyl-$N^3$-(3-amino-3-

carboxypropyl) pseudouridine ($m^1acp^3\Psi$) at position 1191 in 18S rRNA and 2′-*O*-methyluridines (Um) at positions 2634 and 2843 in 25S rRNA. Interestingly, CMC cannot derivatize to $m^1acp^3\Psi$ because both derivatization points on the Ψ base are otherwise occupied by chemical groups. In fact, this hypermodified nucleotide has been found to block reverse transcription on its own, likely owing to the $acp^3$ group at the N3 position of Ψ [81]. $m^1acp^3\Psi$ shows a similar bias towards C(/A) as its non-hypermodified counterpart. Um, on the other hand, shows a bias towards incorporation of A with varying levels of G. To date, available methods for Um mapping either take advantage of 2′-*O*-methylated residues resistance to (1) alkaline hydrolysis, (2) RNase digestion, or (3) 2′-OMe-specific reverse transcriptional stalling under limiting dNTP concentrations [63,80,129]. Recently, the latter strategy was adapted to a high-throughput format in a method called 2OMe-seq [55]. However, none of these methods has identified nucleotide misincorporation at Um as an additional validation strategy for putative Um sites.

Although mismatch analysis is insufficient for robust *de novo* Ψ identification, I concluded that adding a filter for a mismatch incorporation profile of C(/A) at a putative Ψ can serve as an additional bioinformatic layer to increase confidence in that site.

While analysis of my preliminary CMC-seq experiments showed promise for *de novo* mapping of Ψ sites, my efforts were interrupted with the concurrent publication of three similar CMC-based methods for high throughput Ψ-detection. These methods, as well as a critical analysis of their results, are detailed in the next chapter.

## CHAPTER 3. A comparative analysis of high-throughput Ψ-detection methods[2]

With interest surrounding pseudouridylation growing — particularly around its potential role in stabilizing or recoding mRNAs — several groups had independently begun work developing a high-throughput Ψ-mapping approach similar to CMC-seq. As a result, within the month of October 2014 alone, three methods were published — called Pseudo-seq, Ψ-seq, and PSI-seq, in order of publication — that coupled CMC derivatization with RNA-seq to map the pseudouridine landscape in yeast and human RNAs [19,76,107]. Shortly thereafter, a fourth group published a similar technique called CeU-seq that employed a CMC derivative with an azide group for click chemistry, similar to the principal behind CMCyne [69]. Collectively, these papers revealed hundreds of novel Ψs that were found throughout the transcriptome. Intriguingly, pseudouridines were found in mRNA and ncRNA transcripts for the first time. What follows in this section is an outline of these methods, comparing and contrasting their approaches, and a comparative analysis of their results, which revealed areas of improvement for high-throughput Ψ detection.

### 3.1 Strategies underlying four published high-throughput Ψ-detection methods

Given all four techniques rely on CMC derivatization and subsequent deep sequencing, little technical difference exists between the library preparation protocols for each (Figure 3.1). All began by treating poly(A)-selected RNA with CMC, followed by alkaline hydrolysis to selectively label Ψ residues. Following treatment, an adapter was ligated to the 3′ end of RNAs and transcripts were reverse transcribed, with truncated cDNA

[2] Portions of this chapter were published in [131].

products resulting from Ψ-CMC-induced RT arrest. Depending on the method used, either a 3′ adapter was ligated to the resulting cDNAs or RT products were circularized for subsequent PCR amplification and deep sequencing. As a control, libraries were also prepared from mock-treated samples processed in parallel.

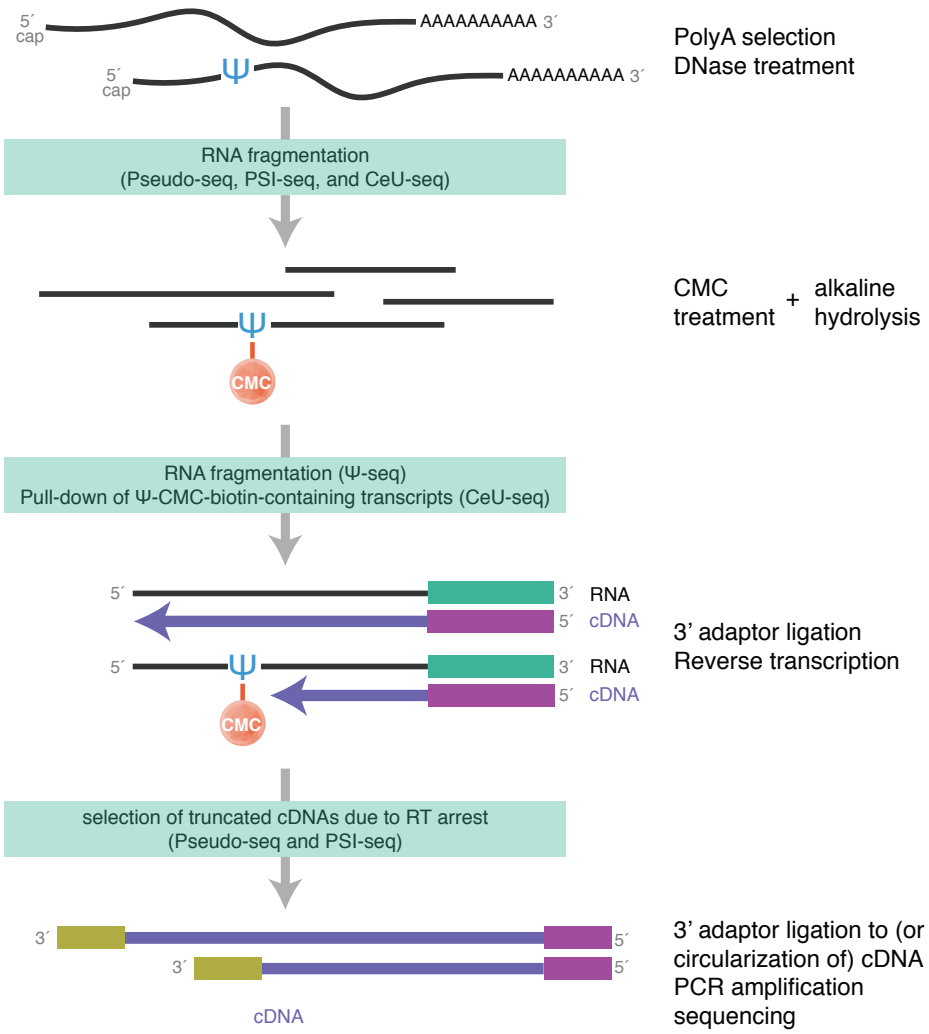


**Figure 3.1. Four methods of transcriptome-wide identification of Ψ residues are based on the same CMC-derivatization principles.**

Generalized library preparation procedure for Ψ-detection methods. Method-specific details are highlighted in green boxes.

While all the aforementioned methods follow the general outline detailed above, a few notable exceptions exist, particularly in how each method enriches for Ψ-containing transcripts (Figure 3.1, green boxes). Most notably, CeU-seq chemically enriches for Ψ-CMC-containing transcripts, as its full name, $N_3$-**C**MC-**e**nriched pseudo**u**ridine sequencing implies. A CMC-azide derivative was utilized for CMC-treatment, which allows for biotin conjugation with click chemistry following derivatization and subsequent hydrolysis. Ψ-CMC-biotin-containing transcripts were then pulled down with streptavidin beads, increasing the method's sensitivity with the benefit of approximately 15-20-fold enrichment of pseudouridylated RNAs [69].

The production of truncated reverse transcriptional products due to Ψ-CMC is central to all four methods and poses unique challenges for bioinformatic detection. Whereas my initial CMC-seq analyses centered on analysis of read coverage, each of these methods has developed similarly derived bioinformatic approaches to identify Ψ-CMC-mediated reverse transcriptional stops to chart the pseudouridine landscape. Importantly, *reverse transcriptional stops* correspond to *sequencing read starts*. Instead of focusing on overall read coverage for a given RNA position, as I had in my initial data analysis strategy, each method computationally identified an increase in CMC-treated reads beginning one position 3′ to a putative Ψ with respect to the mock-treated control (Figure 3.2A). Sites not immediately preceded by a U were removed from analysis.

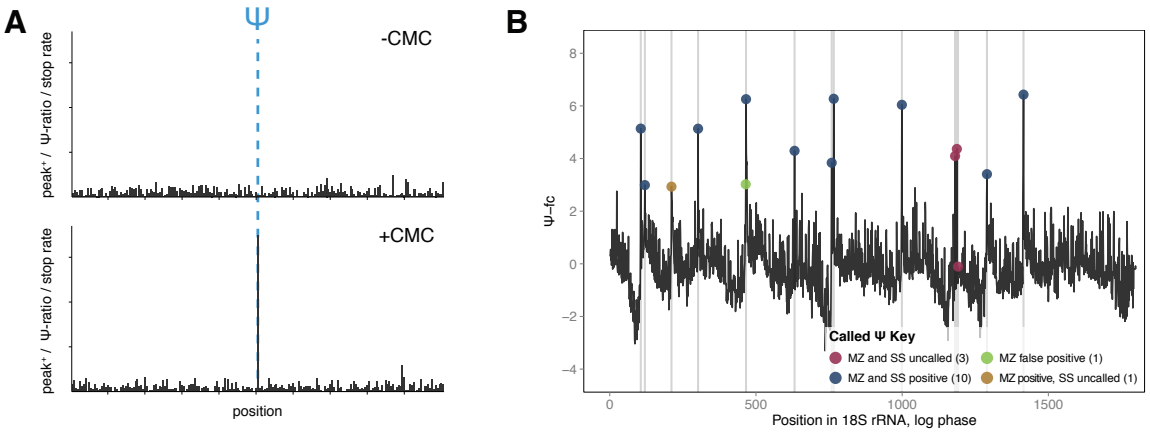


**Figure 3.2. Ψ-CMC adducts correspond to peaks in read starts.**

(A) Sample output of Pseudo-seq, Ψ-seq, and CeU-seq Ψ-detection metrics. (B) Ψ-fc plot for 18S rRNA generated with data from Schwartz, *et al*. analyzed using my own custom scripts. Grey lines indicate known rRNA Ψs. Similarities and differences between Ψs called by my analysis (MZ) and that of Schwartz *et al.* (SS) are marked by colored points.

PSI-seq utilized a regression analysis comparing reads initiating at a given position between treated and mock-treated libraries [76]. Pseudo-seq, Ψ-seq, and CeU-seq did not rely on such a statistical approach. Rather, they computationally identified peaks in the number of reads initiating at a particular U-adjacent site. Ψ-seq and CeU-seq calculated the ratio of reads (the 'Ψ-ratio' and the 'stop rate,' respectively) beginning at each mapped position to the total number of reads covering that position (Equation 3.1) [69,107]. The treated and mock-treated ratios were then compared to call putative Ψ sites, requiring the treated ratio, the $\log_2$-transformed ratio difference (Ψ-fc), and the number of reads initiating at that position exceed a particular cutoff. Using RNA-seq data from Schwartz *et al*., I replicated their bioinformatic analysis and results in rRNA using custom scripts (Figure 3.2B). CeU-seq also relied on 'CMC sensitivity' — which was

adapted from related work profiling RNA secondary structure using DMS-mediated RT stops — as an additional measure of the difference in stop reads at a particular site [31].

**Equation 3.1.**
$$\Psi\text{-ratio} = \text{stop rate} = \left( \frac{\text{reads beginning at position}}{\text{total reads at position}} \right)$$

Pseudo-seq utilized a metric similar to the Ψ-ratio/stop rate, calculated with 150-nucleotide windows (Equation 3.2, WS) centered on a U site. The number of reads beginning 1 base 3′ of the central U (Equation 3.2, URS) and the total number of reads initiating at any other position within the window (Equation 3.2, WRS) were determined for treated and mock-treated libraries to calculate the 'peak$^{+}$' (Equation 3.2) [19]. Peak$^{+}$ values above a specified cutoff and exceeding a minimal number of supporting reads were used to call putative Ψs, requiring reproducibility over a given number of replicates.

**Equation 3.2.**
$$\text{peak}^{+} = \text{WS} \times \left( \frac{\text{URS}^{+\text{CMC}} - \text{URS}^{-\text{CMC}}}{\text{WRS}^{+\text{CMC}} - \text{WRS}^{-\text{CMC}}} \right)$$

With the four published methods now outlined above, two notable differences can be drawn between their approaches and the one I employed during CMC-seq library preparation. First, I used random hexamers for priming during first strand synthesis, which has been found to bias read coverage towards certain 13-mer sequences [48]. In contrast, the methods outlined in this section prime from 3′ ligated adapters for more uniform coverage of the transcriptome. Second, each method prepared libraries so the 3′ ends of cDNA made during first strand synthesis would correspond to a read start site. Ψ-

seq did so using paired-end sequencing to sequence both ends of a particular dscDNA. The remaining three methods employed intramolecular ligation in lieu of 5′ adapter ligation to circularize cDNA following first strand synthesis; the site of RT termination was thus exactly 3′ of the sequencing primer. Because I did not circularize cDNA and employed 100bp single-end sequencing, I would only sequence the RT stop site if the dscDNA fragment were less than 100 nucleotides long. Thus, my CMC-seq libraries are insufficient for detecting an enrichment in read starts at a putative Ψ-CMC site. Indeed, I attempted a similar approach on my CMC-seq samples — computationally detecting an enrichment in read stops in place of starts — and could only identify one Ψ in 18S rRNA (Figure 3.3).

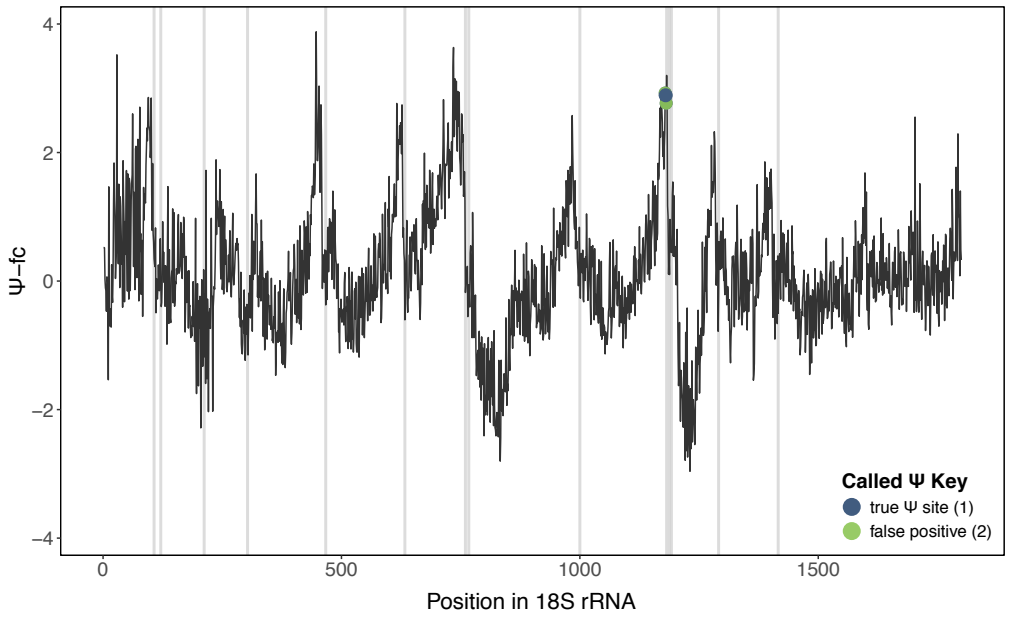


**Figure 3.3. Ψ-seq analysis on CMC-seq libraries detects only one Ψ 18S rRNA.**

### 3.2 Key results from Pseudo-seq, Ψ-seq, PSI-seq, and CeU-seq

Pseudo-seq, Ψ-seq, PSI-seq, and CeU-seq were all performed on a number of cell types and growth conditions, revealing a tremendous amount of diversity and complexity in the pseudouridylation landscape. But before the methods were applied to a transcriptome-wide analysis, known sites of rRNA pseudouridylation were first used to calibrate each method's respective Ψ-detecting metrics to balance the sensitivity and specificity of each approach. This analysis also confirmed CMC's specificity for Ψ derivatization, which was a concern of mine during CMC treatment optimization. Each method filters for hits that correspond to a 'U' in the transcriptome, which excludes analysis of G residues that may still be conjugated to CMC due to incomplete hydrolysis. Nevertheless, according to the Ψ-seq study conducted with log phase yeast, all predicted Ψ sites were either preceded by a U residue or adjacent to a called Ψ as a result of 'stuttered' RT arrest [6,107]. Additionally, the CeU-seq study demonstrated high specificity of $N_3$-CMC to Ψ, with no cross-reactivity to U or the G-like inosine [69].

Having established the appropriate computational cutoffs for Ψ-detecting metrics, Ψ maps were detailed for budding yeast, human cells (HEK293, HEK293T, HeLa, and fibroblasts), and mouse brain and liver cells. In addition to detecting known sites of pseudouridylation in tRNAs, snRNAs, and rRNAs, Ψs were found for the first time in a range of functionally relevant noncoding RNAs and mRNAs [19,69,76,107]. A subset of these newly identified Ψs were attributed to a specific PUS or Ψ-guiding snoRNA through a series of systematic knockdown/knockout experiments. Combined, genetic perturbation experiments and computational analyses linked approximately 20-50% of

putative Ψs to guide RNA or PUS activity, depending on which Ψ-detection method was used and which PUSs and snoRNAs were further investigated.

While matching predicted sites of pseudouridylation to PUS or guide RNA activity indirectly validated a subset of Ψ candidates, Li *et al*. went one step further, directly validating four of their hits from CeU-seq. A quantitative, CMC-independent Ψ-detection method called SCARLET (detailed in Chapter 1.5) was utilized to verify that the aforementioned three previously unknown Ψ sites detected in human rRNA were modified to greater than 90% [69]. Even more intriguingly, SCARLET was applied to demonstrate U519 in *EEF1A1* mRNA was indeed pseudouridylated to approximately 56%, providing the first documented experimental evidence of mRNA pseudouridylation.

Each method also identified a conditionally dependent set of mRNA pseudouridylation sites. In particular, hundreds of stress-dependent pseudouridylation events were identified in yeast by Ψ-seq (265 Ψs) and PSI-seq (314 Ψs) analysis of cells following heat shock [19,107]. 60% of Ψ-seq hits perfectly corresponded to the conserved Pus7 recognition motif and became undetectable in the *pus7* strain, suggesting Pus7 plays a major role in orchestrating heat-shock-specific pseudouridylation. Notably, Pus7 had previously been implicated in the inducible modification of U2 snRNA at Ψ56 following heat shock and nutrient deprivation [125]. Stress-induced Ψs were also found in human cells; CeU-seq profiled sites following heat-shock (464 Ψs) and $H_2O_2$ treatment (477 Ψs), while Pseudo-seq profiled sites in serum-starved versus serum-fed HeLa cells [19,69]. CeU-seq profiling was additionally performed on mouse cells derived from liver and brain tissue. 1,741 and 1,543 Ψ sites were identified in brain and liver mRNAs, respectively; however, only 54 of those sites were shared between the two cell types.

Remarkably, pseudouridylated transcripts were strongly enriched for tissue-specific function. For example, Ψ-containing mRNAs from the brain encoded proteins involved in nervous system development and signal transduction.

### 3.3 Comparative analyses of approaches reveal opportunities for improvement

The existence of four independent CMC-based deep sequencing approaches for Ψ detection afforded a unique opportunity for critical comparison of their respective results to determine the robustness of each approach. Because each was applied to a diverse set of cell types and growth conditions, I was careful to compare Ψ maps provided only for transcripts isolated from the same cell line grown under similar conditions. Consequently, an in-depth analysis was restricted only to yeast cells grown in log phase, though I did also compare human-derived Ψ maps. The resulting comparative analysis revealed a subset of high-confidence Ψ sites, independently detected by multiple methods, and underscored opportunities to improve the available Ψ-detection approaches.

#### 3.3.1 Comparing pseudouridylation candidates in budding yeast

Pseudo-seq, Ψ-seq, and PSI-seq all profiled pseudouridylation events in yeast undergoing log phase growth ($OD_{600} \approx 1.0$, Pseudo-seq; midlog phase hits were used for Ψ-seq, with log phase defined as $OD_{600} = 2$, though midlog $OD_{600}$ was undefined; $OD_{600} = 0.6$-$0.8$, PSI-seq), which became the focus of my comparisons. Analysis was further restricted to include only Ψ candidates in coding DNA sequences, as UTRs were not analyzed in Ψ-seq. Because PSI-seq aligned reads to an earlier genome assembly (SacCer2 versus SacCer3), however, site-specific events could only be compared between Pseudo-seq and Ψ-seq. Nevertheless, I was able to interrogate the three methods to uncover a subset of genes with independently called putative Ψs (Figure 3.4, left panel).

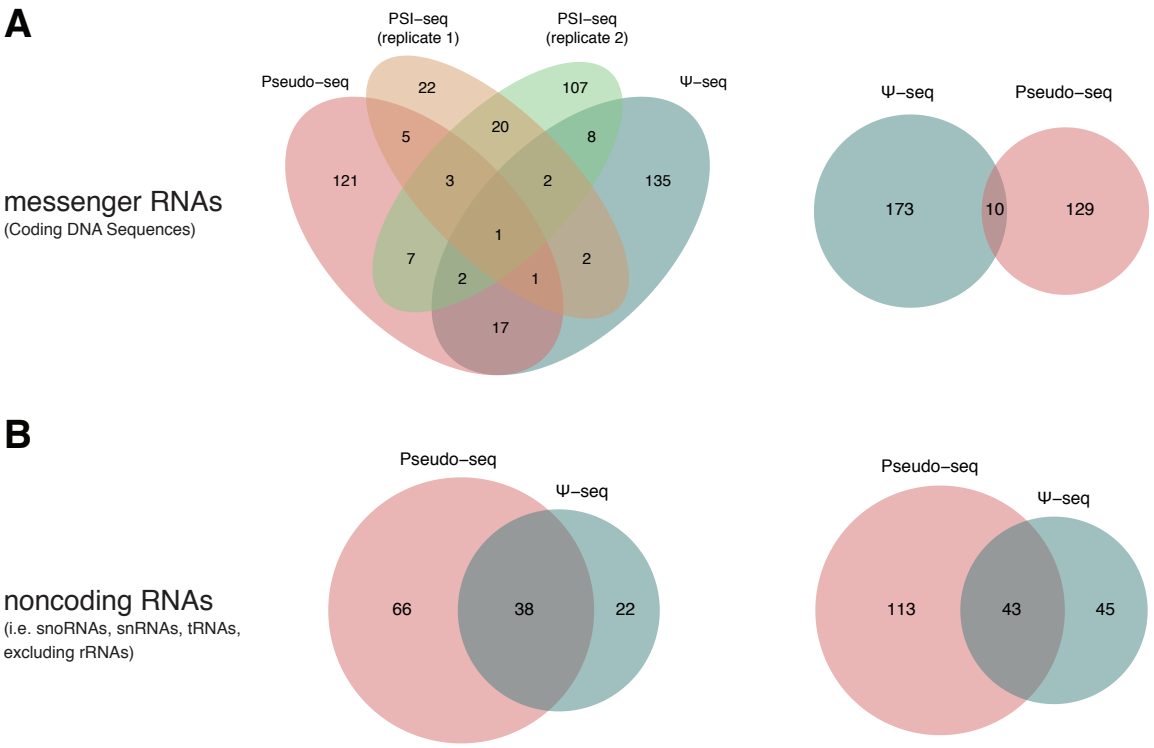


**Figure 3.4. Comparative analysis of candidate pseudouridylation targets in *S. cerevisiae* during log phase growth.**

(A) Putatively pseudouridylated coding DNA regions detected by Pseudo-seq, Ψ-seq, and two replicates of PSI-seq (left) and site-specific Ψ sites detected by Pseudo-seq and Ψ-seq (right) were compared to identify overlapping hits. (B) The same analysis was performed for noncoding transcripts (left) and specific ncRNA-internal Ψ sites (right) identified by Pseudo-seq and Ψ-seq.

In total, pseudouridylation was detected within the CDSs of 402 unique genes. Of those genes, however, only *RPL11a* (a 60S ribosomal subunit protein) was consistently found to contain a CDS-internal Ψ at position 68. On closer inspection, Ψ239 was detected in *TEF1* (a translation elongation factor) by both Pseudo-seq and PSI-seq, and at the same position in *TEF2* by Pseudo-seq and Ψ-seq. Because *TEF1* and *TEF2* are

paralogous genes that resulted from gene duplication, it is likely that one or both transcripts are pseudouridylated. Importantly, each of these detection techniques is inherently biased towards detecting sites in more abundant transcripts. Indeed, *RPL11a* and *TEF1*/*TEF2* are both within the top 30 most highly expressed genes in the yeast genome, which may account for their reproducible detection by independent methods [7,121]. Both Pseudo-seq and PSI-seq cite Pus1 dependency for *RPL11a* Ψ68, and all three studies cite Pus4 dependency for *TEF1*/*TEF2*. Furthermore, using the low-throughput CMC-Ψ/RT approach, Lovejoy *et al*. identified *RPL11a* Ψ68 in the related yeast *Saccharomyces mikitae* and *TEF1* Ψ239 in both *S. mikitae* and *S. pombe* [76]. The evident evolutionary conservation of these modifications further points to the potential biological relevance of these particular pseudouridylation events.

Site-specific pseudouridine candidates identified by Pseudo-seq and Ψ-seq were next analyzed. Of the 21 overlapping putatively pseudouridylated CDSs, 10 predicted Ψ positions in 10 genes exactly overlapped (Figure 3.4, right panel, Table 3.1). The mean distance between the remaining Ψ sites within overlapping CDSs was approximately 740, ruling out the possibility that non-overlapping sites were the result of stuttered CMC-Ψ-mediated RT termination. In both studies, five of the ten Ψ sites were also found to be dependent on activity from the same PUS (either Pus1 or Pus4). While it would be reasonable to assume that a high Ψ-ratio or peak$^{+}$ value would increase relative confidence in a given Ψ site, the Ψs belonging to this overlapping set did not necessarily have the highest Ψ-detection metrics. In fact, Ψ239 in *TEF1*/*TEF2* just barely passed the cutoff requirements for Ψ-seq (Table 3.1). Still, given the approximately 2.5 million U residues in yeast coding sequences, an overlap of 10 independently called pseudouridines

is highly significant ($P = 1.12 \times 10^{-8}$ by the hypergeometric test), increasing confidence that this set contains true targets of pseudouridylation. It is worth noting, however, that while all Ψ members of this set were also detected by Pseudo-seq under post-diauxic cell growth, only one (Ψ1916 in *KAR2*) was detected by Ψ-seq following heat shock.

**Table 3.1. CDS-internal Ψ candidates detected by Pseudo-seq and Ψ-seq.**

| | | | Ψ-seq metrics | | Pseudo-seq metrics |
|---|---|---|---|---|---|
| **Coordinate** | **Gene** | **Position in Gene** | **Ψ-ratio** | **Ψ-fc** | **peak$^+$** |
| chr10:383242 | KAR2 | 1916 | not available | not available | 2.82 |
| chr16:126070 | YPL225W | 65 | not available | not available | 1.15 |
| chr4:331025 | BDF2 | 2 | 0.66 | 4.25 | 13.90 |
| chr3:51028 | GLK1 | 191 | 0.29 | 4.31 | 7.03 |
| chr10:314164 | MPM1 | 709 | 0.24 | 4.66 | 2.02 |
| chr2:477909 | TEF2 (TEF1) | 239 | 0.11 | 3.17 | 6.25 |
| chr1:32596 | GDH3 | 1030 | 0.23 | 5.41 | 3.52 |
| chr16:731681 | RPL11A | 68 | 0.23 | 3.5 | 11.97 |
| chr8:499441 | RPN10 | 383 | 0.17 | 3.74 | 5.96 |
| chr7:623051 | YGR067C | 1736 | 0.14 | 3.78 | 2.68 |

The minuscule percentage of overlapping pseudouridylated CDSs (~0.5%) and specific Ψ positions (~3.2%) does nevertheless highlight the limitations of the high-throughput detection of pseudouridylation events. Specifically, because high coverage at each surveyed position is essential to robust Ψ detection, the output of each method is highly dependent on sequencing depth, which likely varied between each group. All the methods outlined above also favor specificity over sensitivity, which necessitates rather conservative cutoffs for Ψ detection. As a result, the reported Ψs are likely a small sampling of several true pseudouridylation events missed by each method. Additionally, the efficiency of native mRNA pseudouridylation has not been concretely established and may be highly variable [61]. Karijolich *et al*. noted low isomerization efficiency (~7-

10%) when artificially targeting mRNA pseudouridylation, while the one experimentally verified native Ψ target identified by CeU-seq was pseudouridylated to a much higher extent (~56%) [61,69]. High variance in the efficiency of naturally occurring pseudouridylation events coupled with stringent Ψ-detection cutoffs therefore introduces yet another challenge to reproducible Ψ mapping.

A core finding of all four Ψ-detection methods was the conditional inducibility of pseudouridylation, which further complicates Ψ profiling. Changes in the Ψ landscape in response to large environmental perturbations were investigated; however, the robustness of particular pseudouridylation events to smaller environmental fluctuations was not examined. For instance, small differences in $CO_2$ levels in the incubators of different laboratory spaces may produce different Ψ landscapes. The difference in Ψs identified by these different methods may then be a reflection of biological fluctuations in pseudouridylation in even slightly different environmental contexts. Furthermore, all of the above methods query pseudouridylation events in populations of cells, aggregating cells that likely differ, for instance, in cell cycle stage or microenvironment. These distinct subpopulations may likewise differ with respect to pseudouridylation substrates. Population averaging effects may thus be an additional contributor to variance. We may speculate, then, that the Ψs identified by multiple methods are more frequently pseudouridylated under a broader spectrum of environments, suggesting they play some core role in mRNA structure or function, at least under logarithmic cell growth.

With the above challenges in mind, I turned my attention to analyzing the set of pseudouridines detected in noncoding transcripts in log-phase yeast, excluding rRNAs. Because PSI-seq did not detail Ψs in this subset of transcripts, I compared only the

outputs from Pseudo-seq and Ψ-seq. Here, the percentage of pseudouridylated transcripts (~30%) and specific Ψ sites (~20%) independently detected by each method was markedly greater and highly statistically significant ($P = 6.25 \times 10^{-9}$ by the hypergeometric test) (Figure 3.4B). This overlap is well in line with the generally higher expression of ncRNA species with respect to their protein-coding counterparts. Important to note as well, these ncRNA transcripts include snRNAs and tRNAs — long established targets of site-specific pseudouridylation. Moreover, Ψ is known to be essential for proper structure and function of these classes of RNAs, necessitating constitutive modification of specific uridine residues. Combined, higher expression and functional importance thus facilitate reproducible Ψ detection by multiple methods.

### 3.3.2 Comparing pseudouridylation candidates in human cells

Pseudo-seq detailed the Ψ landscape for epithelia-derived HeLa cells, Ψ-seq for a combination of embryonic kidney-derived HEK293 cells and fibroblasts, and CeU-seq for HEK293T cells. While the comparative analysis undertaken above would suggest that these three methods are not directly comparable, I still wondered if I might determine to what extent pseudouridylation was conserved across all transcripts in these different human cell types. Importantly, CeU-seq pre-enriches for Ψ-CMC-containing transcripts by up to 20-fold to increase the method's sensitivity to low-abundance transcripts, which accounts for the large difference in the reported number of hits with respect to Pseudo-seq and Ψ-seq. Pseudo-seq-analyzed HeLa cells shared no putative Ψs with HEK293T cells or the combination of HEK293 cells and fibroblasts, aside from the previously mentioned Ψ5160 and Ψ5590 in the lncRNA *MALAT1*. The lack of commonly predicted Ψs between these cell lines derived from different tissues is in line with the low overlap

in Ψ sites detected by CeU-seq in mouse brain and liver cells [69]. On the other hand, HEK293/fibroblast cells and HEK293T cells shared 47 putative Ψs out of the 396 and 2,084 called sites in Ψ-seq and CeU-seq, respectively. Rather interestingly, nearly 90% of those overlapping positions were detected in mRNAs, distributed primarily in the 3′ UTR and CDS regions. Once again, the magnitude of each method's respective Ψ-detecting metrics (Ψ-ratio, Ψ-fc, and stop rate difference) does not necessarily correlate with their inclusion in this overlapping set.

### 3.3.3 Conclusions from comparative analysis of cross-method putative Ψ sites

Of the many pseudouridylation events that have been collectively identified by the available Ψ-detection methods, a small subset has been identified by more than one study, further increasing confidence in the Ψ-detecting power of these techniques with the necessary caveats detailed above. Nevertheless, the motivation behind developing such Ψ-detection methods is to elucidate the functional role of this modification. Having further established Ψ's ubiquity by cataloguing a remarkable number of putatively modified sites, it is imperative to next narrow down the list to a set of promising, robustly modified and detectable candidates to interrogate experimentally through, for instance, site-specific Ψ knockout experiments. While comparing the outputs of each respective method has filtered the set of putative Ψs for yeast grown to log phase (and to a lesser extent for HEK293/HEK293T/fibroblast cells), to perform all four methods for every cell type and growth environment of interest is impractical. Consequently, each method could benefit from additional parameters that measure the extent to which a given uridine is isomerized, particularly because a high Ψ-detection metric from any one method does not

guarantee detection by an independent technique. I propose one such parameter in the following section.

### 3.4 Towards quantitative Ψ profiling: a case for molecular barcoding

Of the four methods for high-throughput Ψ detection, Ψ-seq was the only to demonstrate its ability to quantitatively capture the relative level of pseudouridylation by comparing the Ψ-ratios at a particular position across two or more samples. This quantitative power was demonstrated in a synthetic spike-in experiment that mixed different ratios of oligoribonucleotides that either contained a Ψ at a specific site or not [107]. Importantly, however, Ψ-seq was unable to measure absolute levels of pseudouridylation within a given sample, perhaps reflecting incomplete CMC derivatization to Ψ residues or Ψ-CMC readthrough events. For instance, ribosomal pseudouridines are considered to be constitutively modified at near 100% efficiency. Recent work in *S. pombe* has experimentally demonstrated that the majority of pseudouridylated residues in rRNA are indeed highly modified (>85% isomerization) [114]. I was therefore curious to examine the variation in Ψ-ratios across ribosomal pseudouridines identified by Ψ-seq, given this method's special focus on quantitative measurement (Figure 3.5A).

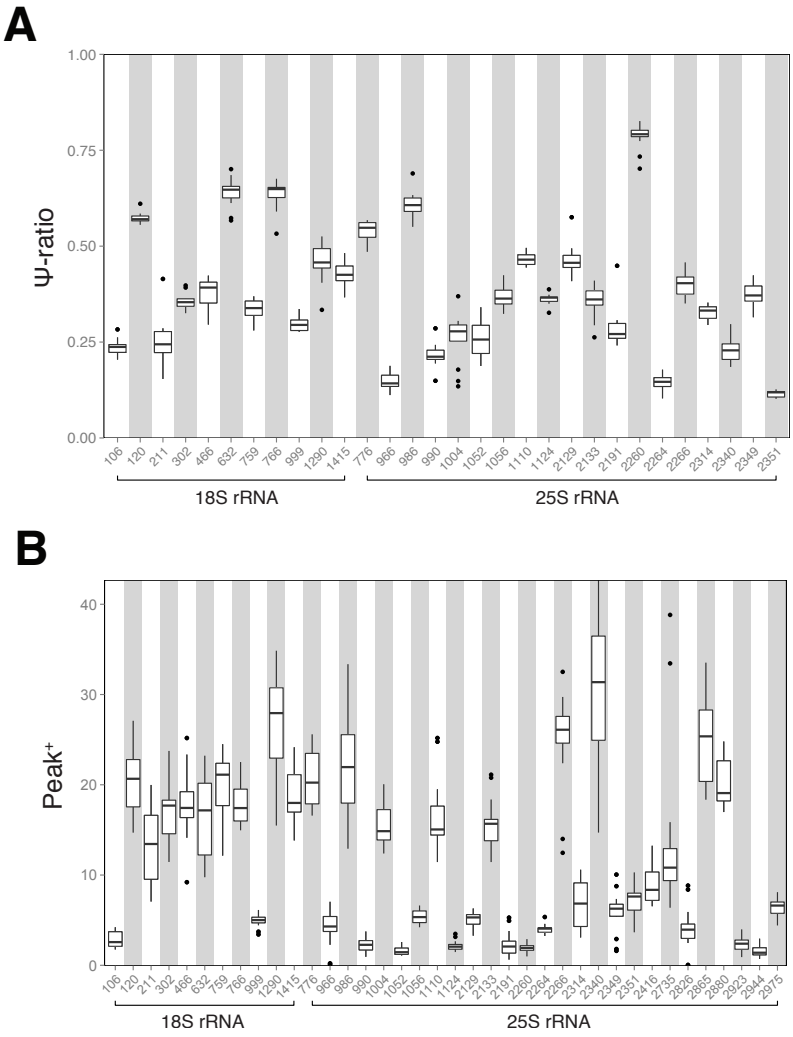


**Figure 3.5. Ψ-detecting metrics are unable to provide absolute quantitation of pseudouridlyation levels.**

(A) Ψ-ratios of known sites of pseudouridylation in 18S and 25S rRNAs detected by Ψ-seq were plotted to assess the variance at each position and across all positions. (B) The same analysis was applied to the peak$^{+}$ Ψ-detection metric used in Pseudo-seq.

Ψ-ratios are remarkably reproducible among replicates at a given position, which well supports Ψ-seq's ability to quantitatively compare Ψ levels between samples. However, the relatively uniform level of rRNA pseudouridylation is not reflected in the variable distribution of Ψ-ratios across all rRNA positions. A similar trend can be seen in

the peak$^{+}$ Ψ-detection metric utilized by Pseudo-seq, though peak$^{+}$ values exhibit a higher degree of variability at each position (Figure 3.5B). Variation at a given Ψ residue reflects the variability intrinsic to RNA-seq library preparations using the CMC/RT approach, which requires multiple steps that likewise introduce multiple opportunities for inconsistency in the hands of different operators. Variation across all known rRNA Ψ residues, however, may be the result of chemical limitations inherent in CMC's ability to uniformly derivatize to pseudouridine, which may be due, for instance, to restrictions imposed by RNA secondary structure or interference from surrounding RNA modifications. While studies have been undertaken to optimize CMC derivatization efficiency, substrate preferences for CMC derivatization, if any, have not been characterized and published [33].

An alternative explanation to intrasample variability in Ψ-detection metrics lies in the high sequencing depth required for each method outlined in this chapter. Increasing sequencing coverage captures more rare cDNA fragments resulting from lowly expressed transcripts; however, increased depth also results in sequencing redundant PCR amplification products more frequently. This trade-off is particularly important given that each of the Ψ-detection techniques identify putative Ψs by *an enrichment in identical reads initiating at the same position*. Importantly, single-end sequencing produces Ψ-CMC-derived reads that are indistinguishable from PCR duplicates (assuming no mismatches). Discarding duplicates therefore interferes with Ψ-detecting power, as multiple Ψ-CMC-initiating reads are collapsed into one (Figure 3.6). Requiring several replicates for confident Ψ detection does mitigate the possibility of false positive Ψ calls due to PCR duplicates. Still, it is difficult to determine the true proportion of reads

initiating at a position due to Ψ-CMC, which could more accurately reflect the level of pseudouridylation at that position. Notably, Ψ-seq performed paired-end sequencing, which improves read mapping resolution by sequencing both the 5′ and 3′ ends, to the extent that cDNA fragment length is sufficiently diverse. Redundant reads can therefore be collapsed with reduced loss of sequencing information (Figure 3.6).

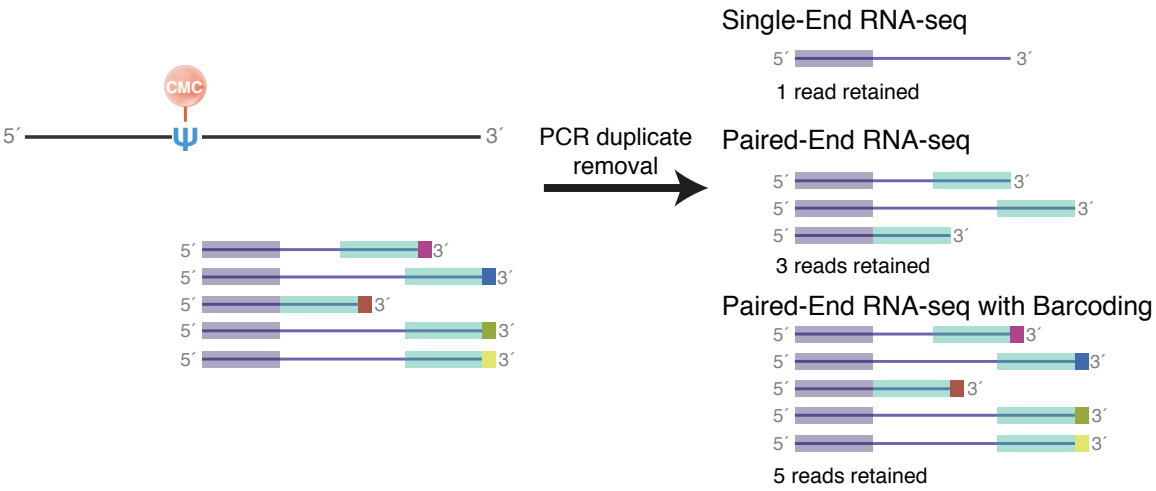


**Figure 3.6. Molecular barcoding improves quantitation of unique reads initiating from Ψ-CMC adducts.**

Schematics of PCR duplicate removal of Ψ-CMC-derived reads using single-end RNA-seq (purple box), paired-end RNA-seq (purple and green boxes), and paired end RNA-seq coupled with molecular barcoding (purple, green, and multi-colored boxes) are depicted. Depending on the sequencing mode used, more reads may be retained.

The requisite sequencing depth may still be ensured while conserving reads resulting from identical Ψ-CMC-derived cDNA fragments (as opposed to identical PCR duplicates). Coupling molecular barcoding with RNA-seq has been shown to more accurately and reproducibly quantify the absolute number of cDNA fragments in a given

sample [40,109]. I therefore proposed similarly incorporating short randomized DNA sequences, through end ligation or reverse transcription, prior to PCR amplification to uniquely identify cDNA fragments. The specific length of the barcodes is dictated by the size and complexity of the transcriptome under investigation [20,39]. Replicate clones due to PCR amplification are identified as reads with matching barcodes and sequences that map to the same location. These reads are then collapsed into one, allowing for single-copy resolution. Most importantly, identical reads initiating from the same position due to Ψ-CMC may be distinguished by their unique barcodes, providing absolute quantitation of the number of reads initiating at and covering a given position (Figure 3.6). In other words, the number of unique barcodes, rather than the number of reads, would be used to count and calculate the relevant Ψ-detection metrics. While molecular barcodes cannot completely overcome limitations due to inefficient CMC conjugation, barcoding provides a more quantitative approach that could facilitate a more direct comparison of pseudouridylation levels across putative Ψ sites by comparing absolute proportions of reads. Furthermore, the extent to which given positions are pseudouridylated provides additional information to discern which particular Ψ residues are most promising for further functional investigation.

In the next chapter, I employed an approach coupling Ψ-seq with molecular barcoding to attempt to identify novel sites of pseudouridylation using the African trypanosome, *Trypanosoma brucei*, as a model organism. I also examined the feasibility of incorporating nonreference nucleotide incorporation patterns as an additional point of validation for putative Ψ sites.

## CHAPTER 4. Improving high-throughput Ψ detection in *Trypanosoma brucei*

The purpose of developing an unbiased, high-throughput approach for Ψ detection was ultimately to expand our understanding of the biological role of pseudouridylation beyond a relatively small subset of cellular RNAs. Each of the published methods catalogued hundreds to thousands of novel pseudouridylation sites, a very small subset of which have been identified by more than one study. However, beyond a GO term analysis of putative Ψ-containing transcripts, these studies did little to experimentally ascertain a physiological role for pseudouridylation to a given transcript. For instance, does Ψ alter the half-life of a particular transcript in response to stress that is necessary to facilitate a biological stress response?

My comparative analysis of the currently available Ψ-detection approaches detailed in the previous chapter shed insights into opportunities for improvement that I sought to apply while simultaneously probing the physiological function of this ubiquitous modification. In this chapter I discuss my work coupling a molecular barcoding scheme with the Ψ-seq method laid out by Schwartz *et al*. in order to detail the pseudouridylation profiles of *Trypanosoma brucei* at two distinct points of the parasite's life cycle.

### 4.1 Utilizing the *T. brucei* life cycle as a model system to investigate the functional consequences of differential pseudouridylation

The conditional inducibility of pseudouridylation has been demonstrated under a number of different conditions of environmental stress and nutrient deprivation by both low-throughput and high-throughput means of Ψ detection [19,69,76,88,107,124,125]. In addition, pseudouridine has been implicated in initiating a cellular differentiation

program in yeast, and distinct tissue-specific Ψ profiles have been detailed in mice [11,69]. The growing catalog of inducible Ψ sites has thus strengthened the hypothesis that pseudouridylation fine-tunes gene expression in response to changing environmental conditions, adding a post-transcriptional layer to gene regulation by, for instance, altering pre-mRNA splicing or recoding specific mRNA codons. In fact, Schwartz *et al.* more closely examined the possibility that Ψ may contribute to enhanced transcript stability by examining the abundance of mRNAs containing heat-shock-induced Ψs [107]. A large subset of these induced Ψs were attributed to Pus7 activity; so mRNA expression levels were compared in wild-type versus Δ*pus7* strains under normal and heat-stressed conditions. While Ψ-containing transcripts were expressed at comparable levels between wild-type and Pus7-deficient cells under normal growth conditions, these transcripts were expressed at ~25% higher levels following heat shock in wild-type cells [107].

As a result, I chose to investigate the biological implications of pseudouridylation at two points during the *T. brucei* life cycle, which is an ideal model system for reasons detailed in this section.

### 4.1.1 *T. brucei* differentiation requires adaptation to different host environments

African trypanosomes are unicellular protozoan parasites that are the causative agents of African trypanosomiasis — African sleeping sickness in humans and *nagana* in their vertebrate zoonotic counterparts — affecting sub-Saharan Africa. Within the lab, the most commonly studied and well-characterized species of trypanosome is *T. brucei brucei*, which is the focus of the experiments detailed in this chapter.

*T. brucei* is transmitted between its mammalian hosts through the tsetse vector (*Glossina* sp.). Because the parasite lives extracellularly throughout the entirety of its life

cycle, it must adapt to a variety of host-specific environmental conditions differing in pH, nutrient availability, and temperature that cue a particular set of diverse biological changes (Figure 4.1). Following a bloodmeal, the parasite establishes in the tsetse midgut — a relatively cool environment (27˚C) with low pH and a variety of harsh proteases — and differentiates into its proliferative, asexual procyclic form (PF) [43,83]. From there, procyclic forms migrate to the fly's salivary glands where they attach and differentiate into the epimastigote form, which is capable of undergoing meiosis allowing for diversification through genetic recombination [44]. Eventually, epimastigotes develop into non-proliferative metacyclic form parasites capable of infecting a mammalian host through the fly's next bloodmeal [83]. On bite, metacyclic trypanosomes migrate from the skin into the glucose-rich bloodstream, where they differentiate to the aptly named bloodstream form (BSF). Early during infection, BSF parasites adopt the long 'slender form,' dividing rapidly both in the bloodstream and within the extravascular space (with its own distinct metabolic requirements) until they reach high density and develop into the non-dividing 'stumpy form' [100,116]. Interestingly, stumpy BSF parasites are primed for transfer into the tsetse midgut, as they exhibit increased resistance to acidic and proteolytic stress [60,93].

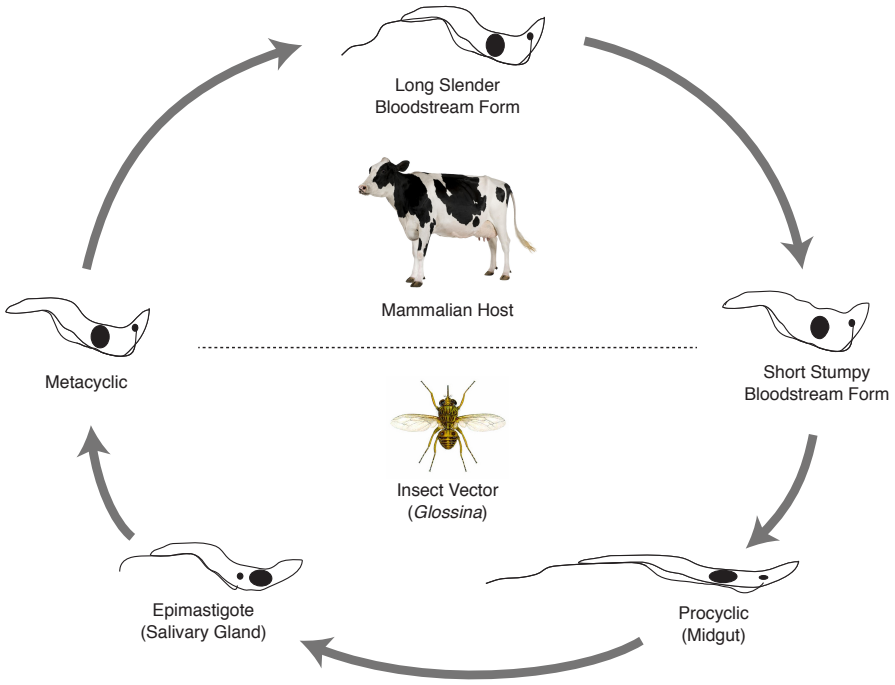


**Figure 4.1. The life cycle of *T. brucei*.**

The biological feat required to rapidly adapt to and differentiate in diverse host environments is well-reflected in this particular passage from Gibson *et al.*:

> *"The trypanosomes migrate anteriorly from this stronghold [in the midgut] on a tortuous journey to the paired salivary glands* [43]*."*

The life cycle of *T. brucei* therefore provides an excellent biological context for host-dependent inducible pseudouridylation to further probe the role of Ψ in adaptation and differentiation.

### 4.1.2 Gene expression in *T. brucei* remains mysterious and occurs largely at the post-transcriptional level

Highly programmed differentiation is made all the more impressive considering the organization — or lack thereof — of the *T. brucei* genome. Specifically, protein-coding genes are arranged in polycistronic transcription units (PTUs) encoding ten to hundreds of mRNAs (Figure 4.2). Genes oriented in the same direction are co-transcribed and rapidly processed through coordinated *trans*-splicing, endonucleolytic cleavage, and polyadenylation at a fixed distance from the splice signal, resulting in mature mRNA transcripts [84,117]. Notably, *trans*-splicing requires joining of a small, capped spliced leader (SL) RNA to the 5′ end of a pre-mRNA [72]. The SL notably contains Ψ at position –12 with respect to the splice acceptor site. Pseudouridylation at this residue is catalyzed by the RNA-dependent Ψ synthase Cbf5 guided by the spliced leader-associated (SLA1) RNA, an H/ACA snoRNA [72]. In contrast to yeast and mammalian mRNAs surveyed for pseudouridylation, all mature mRNAs contain at least a single Ψ.

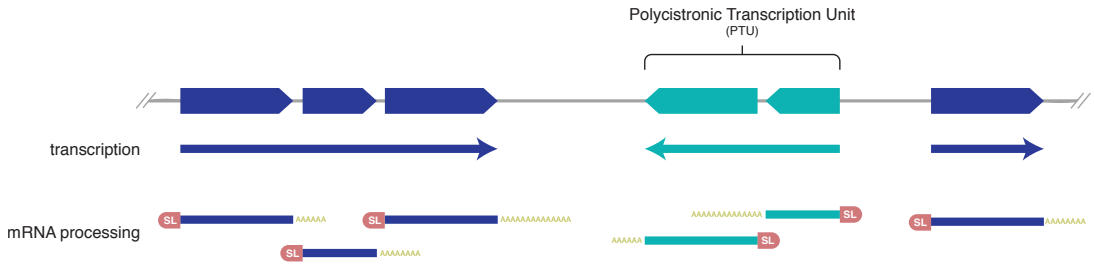


**Figure 4.2. Schematic of *T. brucei* and mRNA post-transcriptional processing.**

Though PTUs resemble bacterial operons, genes within a given polycistron appear to lack functional clustering, with the exception of the alternating α/β-tubulin gene

array [18]. In addition, while pol I promoters (responsible for rRNA and life cycle-specific surface protein transcription) and pol III promoters (responsible for tRNA and small nuclear U RNA transcription) have been identified, pol II promoters responsible for the bulk of mRNA transcription remain elusive. In fact, only one gene-specific pol II promoter has been identified for transcription of the small spliced-leader RNA [27,46]. A growing body of work has pointed to histone marks to delineate the boundaries of transcription, and has suggested that histone modifications may regulate polycistronic transcription [101,106,111]. For instance, while pursuing the Ψ-profiling efforts detailed in this thesis, I, in collaboration with Danae Schulz and Hee-sook Kim, found that the histone variant H3 (H3.V) and a kinetoplastid-specific DNA base modification known as base J (ß-D-glucosyl-hydroxymethyluracil), together regulate transcription termination [106]. In the kinetoplastid *Leishmania major*, depletion of base J alone has been shown to be sufficient to result in transcriptional readthrough at sites of transcription termination [101,118]. Still, transcriptional control at the individual mRNA level appears impossible given the lack of promoters or other *cis*-regulatory elements specific to a single PTU-internal gene [25]. Control of mRNA levels therefore occurs primarily at the post-transcriptional level.

Individual gene regulation is achieved by modulating mRNA stability, translation efficiency, and protein stability. Interestingly, a transcriptome-wide study analyzing the kinetics of trypanosome mRNA decay reported regulated decay of developmentally regulated mRNAs, switching their decay patterns during differentiation [35]. The specific mechanism of this life-cycle dependent "switch" is largely unknown, though the *T. brucei* genome encodes a wealth of RNA-binding proteins (RBPs) with unknown RNA targets

[18]. Several groups have therefore undertaken the task of identifying conserved motifs that regulate developmental gene expression. For instance, a specific 3′ UTR sequence motif (UAUUUUUU) has been found to be highly conserved in procyclic-enriched transcripts, and makes up the core of a 26mer element negatively regulating expression in the bloodstream form [85]. However, developmentally regulated expression of these motif-containing genes was unaffected by the RNA silencing machinery, leaving open the question of how exactly these transcripts are stage-specifically stabilized or destabilized. This is in line with the finding that differentiation progresses normally in Argonaute1-deficient cells [59].

Gene regulation has also been observed through alternative *trans*-splicing of the spliced leader RNA, which can affect gene regulation by altering the open reading frame. Because polyadenylation occurs at a fixed distance from the splice acceptor site, alternative splicing can also affect polyadenylation site choice, which can impact RNA half-life by changing the 3′ UTR [84,120]. Spliced leader trapping experiments illuminated over 2,500 alternative splicing events, several of which appeared to be developmentally regulated [92].

The examples of gene regulation summarized above are by no means comprehensive. Rather, they serve to illustrate the gaps in our understanding of gene regulation at the post-transcriptional level, and areas in which investigation into the developmental Ψ landscape may bridge some of those gaps.

### 4.1.3 Pseudouridylation has been documented and studied in *T. brucei*

Having laid out some of the intriguing quirks surrounding trypanosome biology and differentiation, it merits mentioning that Ψ has been relatively well studied in *T. brucei*,

which is helpful for studying functional consequences of pseudouridylation in this model organism. As in other eukaryotic cells, pseudouridine formation is catalyzed by a guide RNA-dependent mechanism through the activity of Cbf5 guided by H/ACA snoRNAs, and likely through the activity of a number of putative Ψ synthases that bear homology to known stand-alone PUSs [10]. *T. brucei* H/ACA snoRNAs notably differ from those of other eukaryotic cells in that they are composed of only a single hairpin, in contrast to the more common double hairpin structure [73]. Recent small RNA-seq experiments have identified 83 H/ACA snoRNAs capable of guiding pseudouridylation and 79 C/D snoRNAs, which guide 2′-*O*-methylation [24,89].

Sites of pseudouridylation have also been documented within the *T. brucei* genome. Aside from the aforementioned Ψ-containing SL transcript, Ψs have been mapped within ribosomal RNAs using the traditional low-throughput CMC/RT approach [71]. In addition, a recent study published by Chikne *et al.* during the writing of this thesis catalogued additional sites of ribosomal pseudouridylation in BSF and PF trypanosomes [24]. In total, 68 of 75 known Ψs were identified across the two life cycle stages, with 62 of those sites mapping to a specific H/ACA snoRNA guide. Despite having a similar genome size as budding yeast, trypanosomes have over 20 more ribosomal pseudouridylation events. A possible explanation for higher Ψ content with respect to yeast ribosomes is that more ribosomal Ψs may allow the trypanosome to rapidly adapt to fluctuating temperatures as it cycles between hosts while preserving the structural integrity of the ribosome, or possibly modulating ribosomal function [10]. Notably, Chikne *et al*. reported that the level of isomerization from U in 21 Ψs was increased 1.3- to 2.7-fold in BSF ribosomes by comparing Ψ-fc values at a given position

between BSF and PF parasites. A corresponding upregulation of 43 H/ACA snoRNAs was also observed in the bloodstream form, though they did not guide pseudouridylation to the "hyper" modified BSF Ψ sites.

Life cycle-specific differential levels of pseudouridylation in the trypanosome ribosome requires closer investigation than what was conducted by Chikne *et al.* Comparison of Ψ-fc values to ascertain relative Ψ stoichiometries between any two samples depends on a qualitatively "high" sequencing depth; the exact depth, however, has not been experimentally confirmed in biological samples using low-throughput quantitative assays like SCARLET [75,107]. Modest differences in Ψ-fc values between BSF and PF Ψ sites therefore warrant closer investigation, particularly as the source of their "hyper" modification remains unknown. Furthermore, differential levels of pseudouridylation in the ribosomes of BSF compared to PF trypanosomes implies that at least some ribosomal Ψ residues are modified at low efficiency. For instance, if a Ψ residue is modified at a level 2.7-fold greater in BSF versus PF trypanosomes, it cannot be isomerized at a level greater than 37% in the procyclic form. This implication runs counter to the prevailing observation that ribosomal pseudouridylation occurs at a near-uniform high efficiency of >85% in yeast and human ribosomes [69,114]. The finding thus warrants absolute quantification of uridine isomerization at differentially modified Ψ sites using SCARLET or mass spectrometry, which was notably not performed by Chikne *et al*. Still, eight of the 21 differentially pseudouridylated sites were predicted to impact rRNA structure, which may contribute to fine-tuning ribosome function in a given host environment. Perturbation of these sites in order to determine whether changing Ψ levels is necessary for transitioning through the life cycle remains an open question.

### 4.1.4 Statement of the problem

Taken together, the adaptation required to differentiate in a new host environment and the lack of gene-specific transcriptional regulation make *T. brucei* an ideal model system for probing the functional relevance of Ψ. Certainly, developmentally regulated pseudouridylation events are an intriguing possibility for modulating RNA function. I therefore selected bloodstream form and procyclic form trypanosomes for Ψ profiling due to the ease with which they are cultured in the laboratory. I could then ask whether pseudouridylation events were developmentally regulated based on differential Ψ profiles. If BSF- and PF-specific Ψ profiles did indeed differ, I could ask how developmentally regulated Ψs contribute to differentiation. There are, of course many possibilities, which include, but are by no means limited to:

1. Ψ-containing transcripts may exhibit increased stability, which could contribute to their upregulation in one life cycle stage compared to the other.
2. Ψ-containing transcripts may affect translational efficiency. Specifically, pseudouridylation to mRNA codons, tRNA anticodons, or rRNA may mediate amino acid recoding resulting in altered protein products, or may enhance or impede translational efficiency.
3. Differential pseudouridylation of spliceosomal U RNAs may alter pre-mRNA splicing, which could, for instance, alter RNA half-life by changing the site of polyadenylation, changing the open reading frame, or decreasing the efficiency of mature RNA processing.

### 4.2 Experimental design with molecular barcodes

The additional quantitation afforded by coupling molecular barcoding to Ψ-seq could serve to more robustly differentiate levels of pseudouridylation between bloodstream form and procyclic form trypanosomes beyond comparison of Ψ-fc values. Before proceeding, I had to develop and test a molecular barcoding strategy experimentally suited for Ψ-seq library preparation with trypanosome RNA, and develop a bioinformatic pipeline for subsequent deduplication and analysis. The results of my pilot experiments and deduplication analysis are detailed in this section.

#### 4.2.1 Molecular barcode design

Library preparation proceeded exactly as detailed by Ψ-seq (Figure 3.1), except the 3′ adapters ligated onto RNA fragments were specially designed to contain a molecular barcode (Figure 4.3A). Given *T. brucei*'s relatively small genome size (26 Mb), a randomized 6mer would be sufficient to serve as a barcode sequence to distinguish up to 4,096 ($4^6$) unique reads initiating from a given position. Regardless of the barcode sequence, every adapter contained a common 5′ 4mer ligation linker and a 3′ 20mer sequence selected from the *Cyprinus carpio* (carp) genome not found in *T. brucei* for priming during first strand synthesis (FSS). As an additional consideration, nucleotide diversity is essential during initial rounds of sequencing for accurate cluster coordinate identification by the Illumina HiSeq 2000/2500 [65]. Under paired-end sequencing, though, the common 20mer priming sequence will act as the start of the 'left-hand' read, so the initial 20 cycles of sequencing will be identical, impeding effective cluster calling (schematic of paired end reads depicted in Figure 4.3C). To increase sequence diversity, I therefore designed FSS primers complementary to the common priming sequence that

either contained two, one, or no randomized nucleotides at to the 5′ end (Figure 4.3B). The three classes of primers were mixed in equal proportions during first strand synthesis to phase the common priming sequence during sequencing. To further increase diversity, I had also initially designed three adapters (A, B, and C), which differed only in the common priming sequence selected from carp DNA, to generate libraries for different replicates, which were then pooled and run on a single lane.

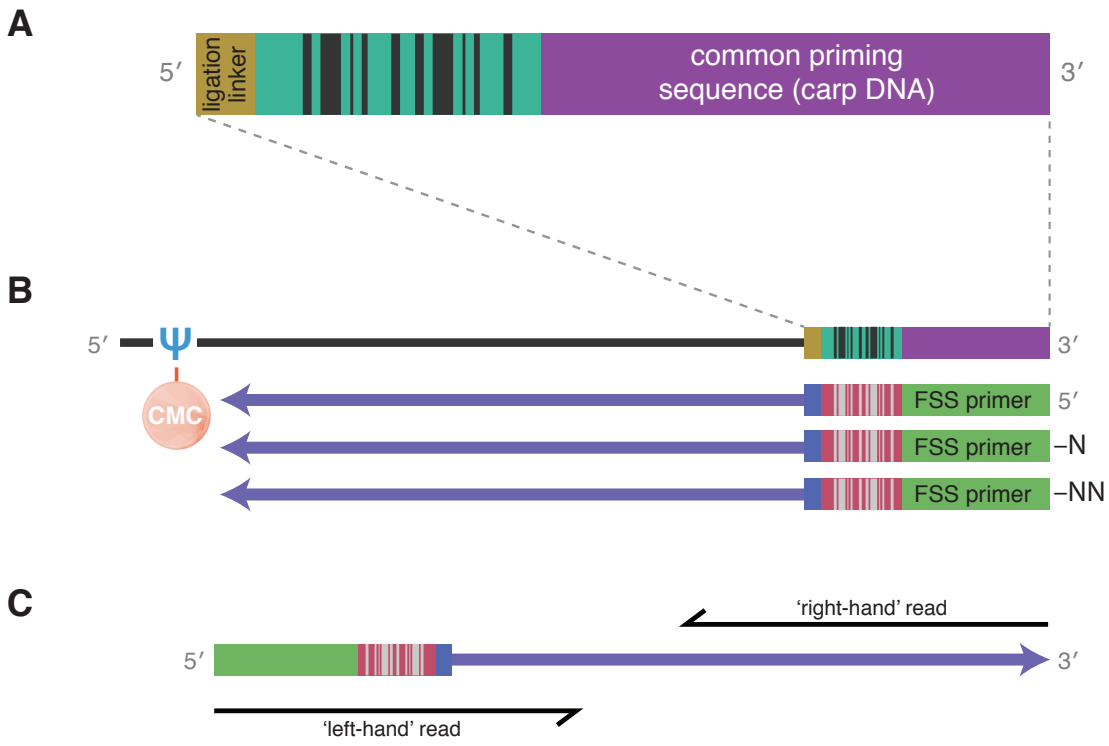


**Figure 4.3. 3′ adapter design with molecular barcodes.**
(A) Schematic of 3′ adapters with randomized 6N molecular barcode. (B) First strand synthesis with priming by one of three possible 0N/1N/2N FSS primers ensures phasing of the common priming sequence. (C) The resulting cDNA will have the adapter at its 5′ end, read as the ‘left-hand’ read, and the Ψ-CMC-mediated RT arrest site at its 3′ end, read as the ‘right-hand’ read start.

### 4.2.2 Pilot experiment reveals barcode diversity is essential for deduplication

In a pilot experiment to test the feasibility of my barcoding scheme, I isolated poly(A)-enriched RNA from BSF and PF trypanosomes, and prepared barcoded Ψ-seq libraries for CMC-treated and mock-treated (input) samples in duplicate. Following 50bp paired-end sequencing, I designed a custom tool to computationally subset 'left-hand' reads that contained the common priming sequence, along with the partnered 'right-hand' reads, and extracted the downstream barcode sequence (Figure 4.3C). Because adapter trimming from left-hand reads resulted in reads too short to be mapped, only right-hand reads were used for alignment, first to the rDNA locus using bowtie2. Reads unmapped to rDNA were then mapped to the whole genome. The results for BSF samples from this initial experiment are summarized in Table 4.1.

**Table 4.1. Summary of results for BSF samples prepared using Ψ-seq with molecular barcoding scheme.**

| Sample | Adapter | Reads with Barcode | Barcode Reads of High Quality | Mapped to rDNA | Mapped to Genome |
|---|---|---|---|---|---|
| input-BSF-1 | C | 79.89% | 99.56% | 6.98% | 59.04% |
| input-BSF-2 | A | 65.86% | 99.08% | 3.94% | 29.24% |
| treated-BSF-1 | A | 73.45% | 99.48% | 4.38% | 57.51% |
| treated-BSF-2 | B | 13.31% | 99.37% | 4.75% | 52.81% |

On average, about 60% of sequenced reads contained barcoded adapter sequence, and of those, over 99% were determined to be of high quality by trim_galore (Babraham Bioinformatics). Libraries prepared with adapter B, however, contained substantially fewer barcoded reads, perhaps due to inefficient ligation compared to adapters A and C.

Nevertheless, reads resulting from libraries prepared with all three adapters mapped to the genome, so I next confirmed that a 6N barcode was sufficient for deduplicating genomic reads. To do so, I calculated the number of reads initiating at a given position in the genome (Figure 4.4). Indeed, all but six positions were covered by less than 4,096 reads initiating at a particular position, indicating that the barcode length was sufficient, assuming that each unique read was in turn marked by a unique barcode sequence.

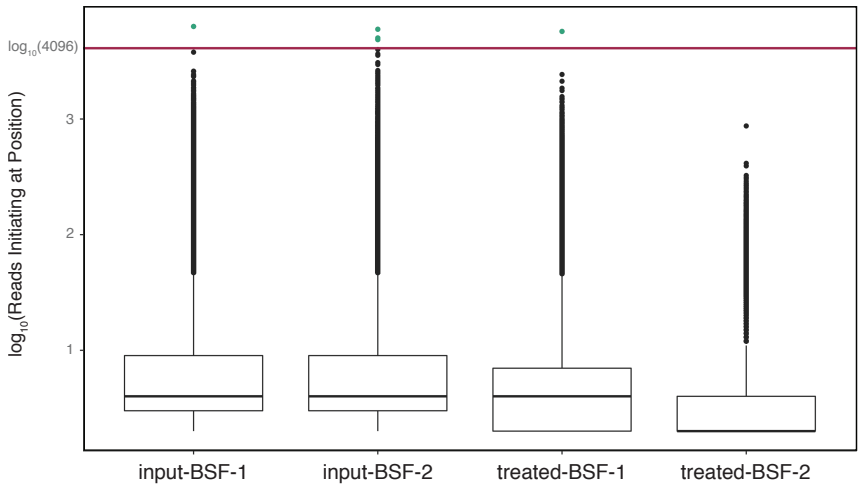


**Figure 4.4. Barcode length is sufficient for deduplication of reads mapping to the *T. brucei* genome.**
$\log_{10}$-transformed reads mapping to every position within the *T. brucei* genome were plotted for each BSF Ψ-seq library. $\log_{10}(4{,}096)$ is indicated by the red horizontal line.

Having established that 4,096 unique barcodes is sufficient for deduplication, I next proceeded to remove reads that appeared to result from PCR duplicates. To do so, I computationally eliminated reads that initiated at the same position and contained the same barcode sequence (Figure 4.5). Depending on the adapter used, between 18-95% of reads were eliminated following deduplication. I was concerned by the range in the

proportion of discarded reads because it implied that library complexity varied widely across samples. However, because the variability in reads discarded following deduplication was largely due to the 3′ adapter used during library preparation, I chose to investigate the diversity of specific 6mer barcode sequences (Figure 4.6). To do so, I plotted the number of reads with a given barcode against all 4,096 barcodes. While all possible barcodes were represented within the total population of reads, adapters A and B contained a marked overrepresentation of a specific 6mer sequence. Adapter C also contained bias towards certain 6mers, but to a far less pronounced degree, which accounts for the higher proportion of reads retained following deduplication.

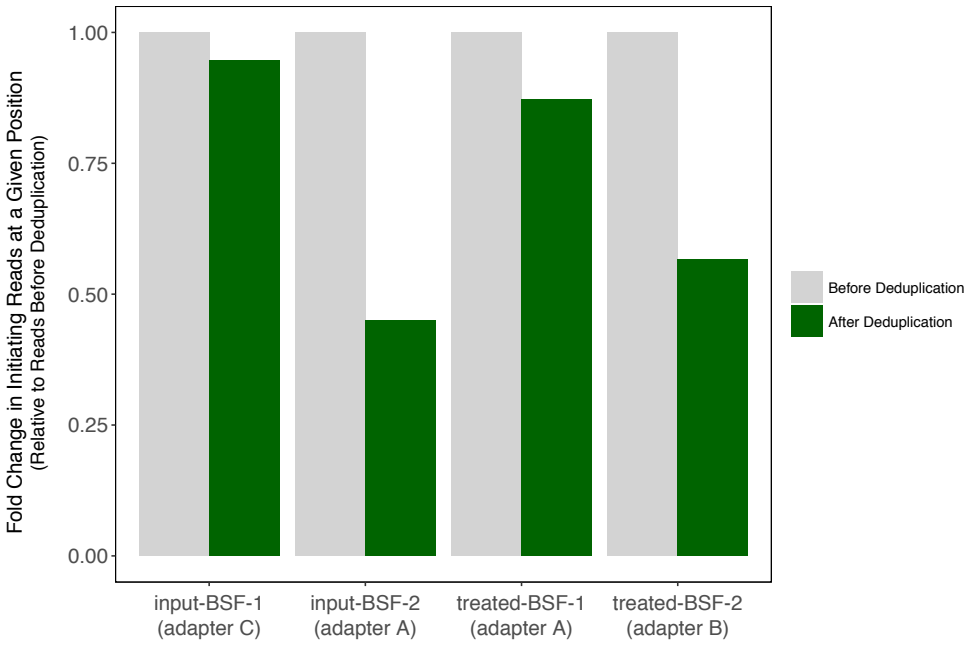


**Figure 4.5. A wide range of reads is discarded following deduplication.**
The fold change in reads before and after deduplication was plotted and normalized with respect to reads before deduplication for each BSF Ψ-seq library.

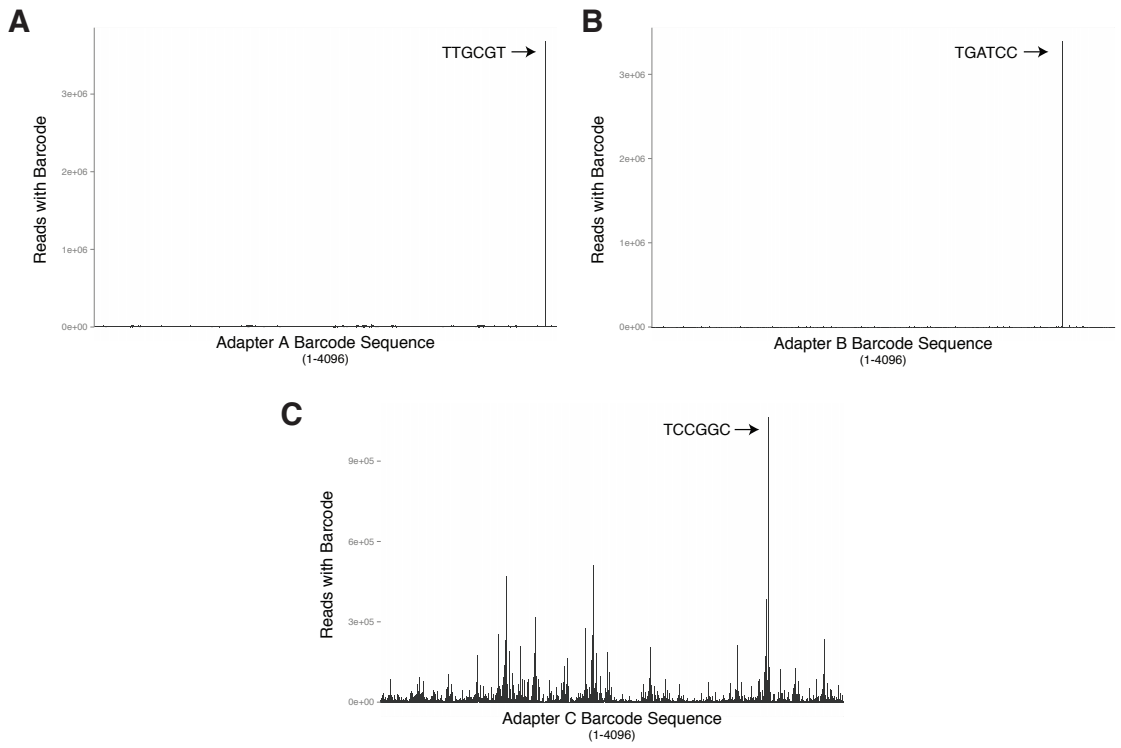


**Figure 4.6. Adapters display bias towards certain 6mer barcode sequences.**
The distribution of barcode 6mers was graphed for adapters (A) A, (B) B, and (C) C. The most highly overrepresented 6mer is indicated with an arrow.

I contacted Integrated DNA Technologies (IDT), who reported a known (though unpublished) bias in the automated oligonucleotide synthesis protocol used to synthesize these adapters. They therefore resynthesized adapter A with a new manual protocol to ensure all 6mers were equally represented in the oligonucleotide pool. The newly optimized adapter was utilized for the experiments carried out in the next section.

### 4.2.3 Optimized adapters eliminate 6mer bias, allowing effective deduplication

The IDT-optimized adapter A was utilized to prepare Ψ-seq libraries from poly(A)-enriched RNA samples from BSF and PF trypanosomes in triplicate. Libraries were

sequenced and reads were mapped to the *T. brucei* genome as before. The results of this experiment are summarized in Table 4.2. Notably, a far greater percentage of reads were barcoded (~88%), and of these reads, a high percentage mapped to the genome. In addition, the distribution of 6mers was far more uniform than in the previous experiment (Figure 4.7A), which allowed for the conservation of a higher percentage of reads following deduplication (Figure 4.7B). Satisfied that IDT adapter optimization accounted for the biases observed in my initial pilot experiment, I moved on to a deeper Ψ-seq-like analysis of the resulting data, detailed in the next section.

**Table 4.2. Summary of results for BSF samples prepared using IDT-optimized barcoded adapters with Ψ-seq.**

| Sample | Reads with Barcode | Barcode Reads of High Quality | Mapped to rDNA | Mapped to Genome |
|---|---|---|---|---|
| input-BSF-1 | 91.98% | 99.83% | 6.13% | 83.521% |
| input-BSF-2 | 87.37% | 99.82% | 5.83% | 84.208% |
| input-BSF-3 | 89.76% | 99.75% | 5.50% | 82.693% |
| treated-BSF-1 | 88.23% | 99.73% | 4.18% | 72.916% |
| treated-BSF-2 | 89.68% | 99.78% | 4.84% | 80.369% |
| treated-BSF-3 | 84.33% | 99.76% | 4.93% | 78.077% |

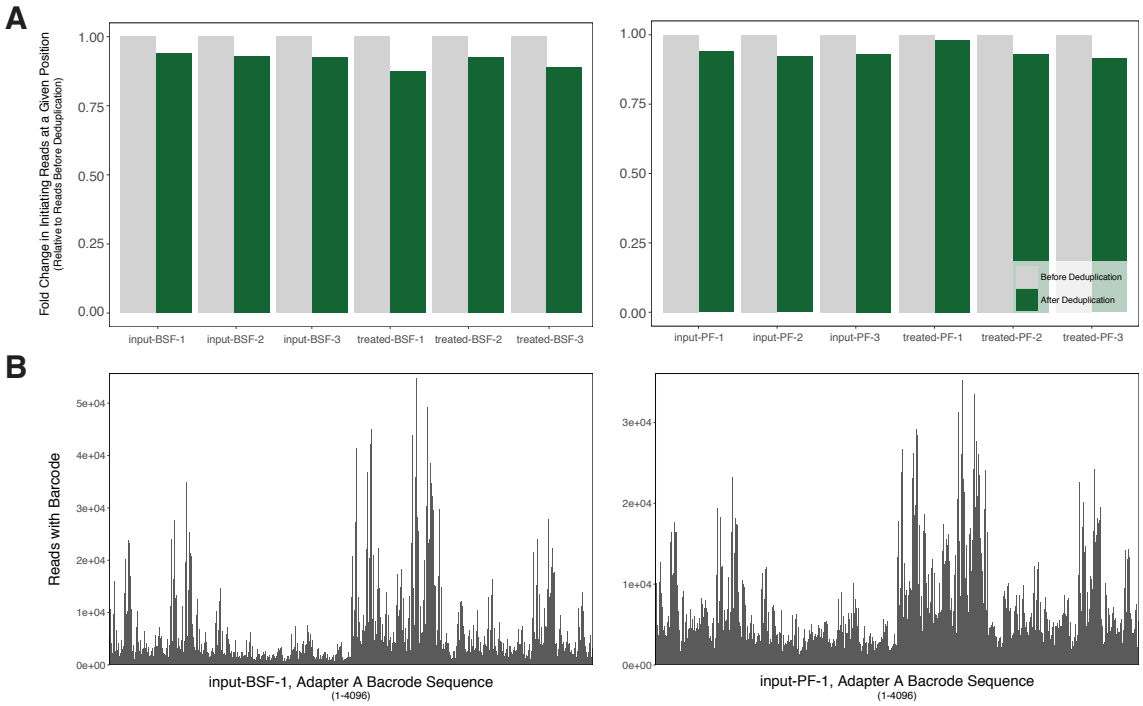


**Figure 4.7. IDT-optimized barcode adapter diversity better suited for deduplication.** (A) A higher proportion of reads are retained following deduplication of reads mapped to the whole genome for BSF (left) and PF (right) Ψ-seq libraries. (B) The 6mer barcode distribution for a representative BSF (left) and PF (right) library was plotted.

## 4.3 Ψ-seq results reveal deeper problems with high-throughput Ψ-detection

### 4.3.1 Detection of known sites of rRNA pseudouridylation

I first analyzed reads mapping to the rDNA locus to determine whether I could detect known ribosomal Ψ sites detailed in [71] and [24] using the bioinformatic pipeline I built to replicate Schwartz *et al.*'s Ψ-seq results (Chapter 3.1, Figure 3.2). In order to get a more accurate sense of my false positive rate, I did not filter out hits that did not correspond to a U in the reference genome. Because each position in the rDNA locus was covered by over 4,096 initiating reads, I performed analysis on alignment files prior to

deduplication to avoid collapsing reads that were not truly PCR duplicates. Unfortunately, I failed to detect any putative Ψ sites — false positives or true positives — in either BSF or PF trypanosomes. In addition, relative Ψ-fc peaks did not correspond to known Ψs based on Ψ-fc plots for each rRNA, a representative example of which is plotted in Figure 4.8.

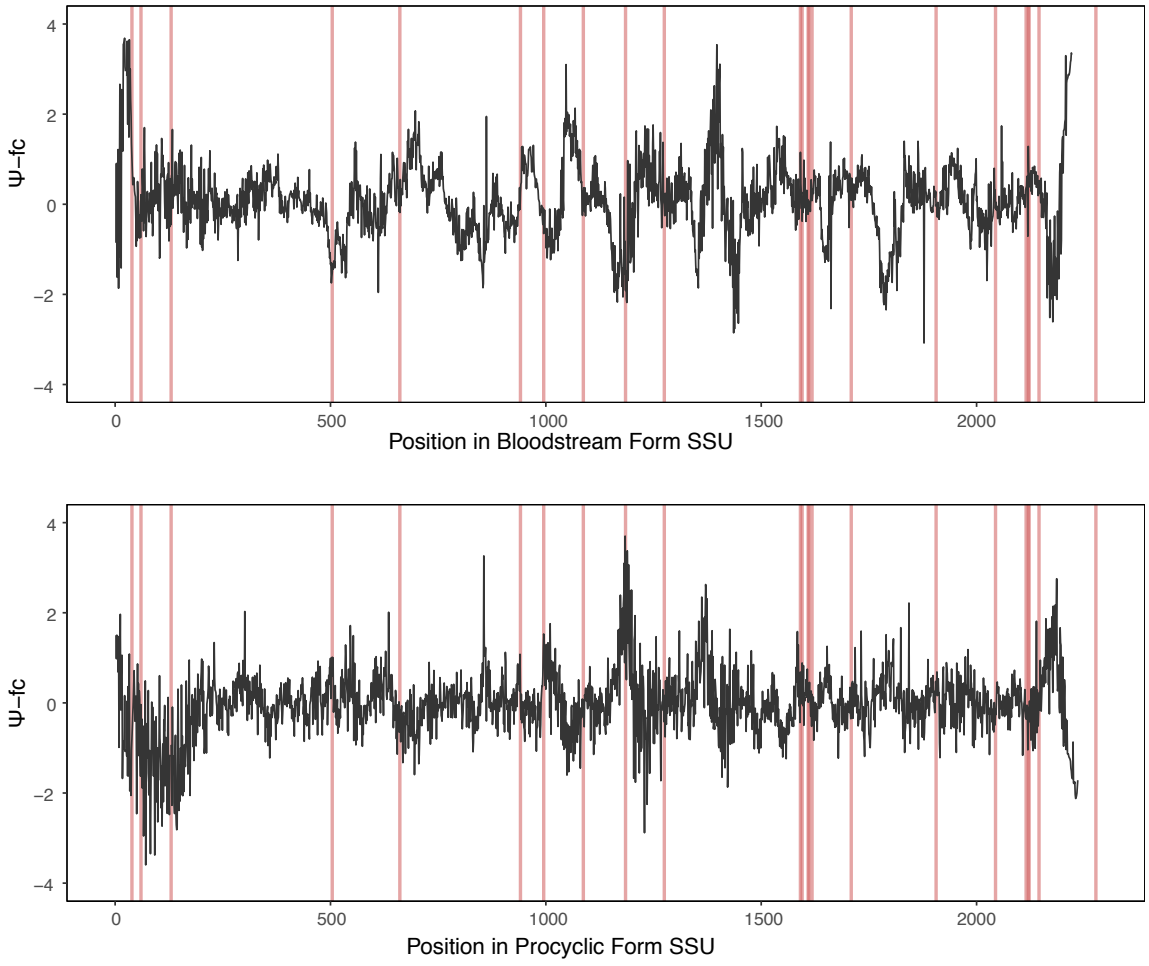


**Figure 4.8. Known rRNA Ψs were not detected in poly(A)-enriched Ψ-seq libraries.** Ψ-fc values were plotted for the small ribosomal subunit (SSU) RNA in BSF (top) and PF (bottom) trypanosomes. Known Ψ sites are marked by vertical red lines.

Furthermore, each of the Ψ-detecting metrics utilized by Ψ-seq showed no discriminatory power between true Ψ sites and false positives when applied to my data set (Figure 4.9). The area under the Receiver Operating Characteristic (ROC) curves (AUCs) for both the treated Ψ-ratio and the Ψ-fc were far lower than those calculated using the published yeast Ψ-seq data set (AUC=0.544 versus AUC=0.951 and AUC=0.442 versus AUC=0.985, respectively). I could therefore conclude that even if I had used different cutoff values for each Ψ-detection metric, I still would not have been able to accurately detect Ψ sites.

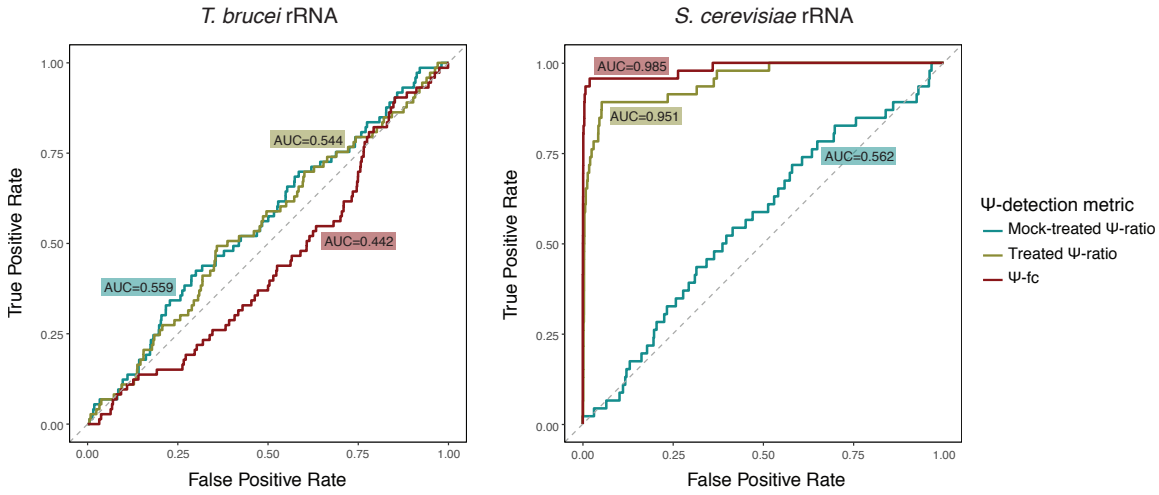


**Figure 4.9. ROC curves demonstrate no discriminatory power in Ψ-detection metrics in Ψ-seq libraries prepared with *T. brucei* poly(A)-enriched RNA.** Receiver Operating Characteristic (ROC) curves for different Ψ-calling metrics were calculated for my trypanosome rRNA and for Schwartz *et al.*'s yeast rRNA. The line of no discrimination is plotted as a dashed grey line.

While parsing through the data compiled by Schwartz *et al*. during my comparative analysis, I did notice that analysis of poly(A)-enriched Ψ-seq libraries failed

to detect several known sites of rRNA pseudouridylation. I therefore prepared libraries from total RNA isolated only from PF parasites to again test whether I could detect known Ψs. Ψ-seq analysis of this data, which is not shared here, once again returned neither false nor true positive rRNA hits and produced ROC curves that indicated virtually no discriminatory power in Ψ-detection metrics.

Because I was unable to identify Ψs by Ψ-seq analysis, I next sought to determine whether CMC treatment was successful during library preparation. While I could not directly assess whether CMC had successfully conjugated to Ψ targets and hydrolyzed from non-Ψ residues, I could calculate the CMC-stat for each rRNA position (described in Chapter 2.2.2, Equation 2.1) to see if coverage decreased around Ψs in treated versus untreated libraries. If CMC treatment were successful, I would have observed strong peaks in CMC-stat values around known Ψ sites, similar to those observed in Figure 2.6. Peaks that lined up with trypanosome rRNA Ψ sites, however, were not readily distinguishable from peaks occurring at non-Ψ positions (Figure 4.10). I also entertained the possibility that the documented Ψ sites were not accurate, so I compared rRNA CMC-stat profiles for PF libraries prepared from total RNA and from poly(A)-enriched RNA. If I could see a consistent CMC-stat pattern, with peaks lining up across the rDNA locus, then perhaps the reported Ψs were incorrect. However, no such pattern was observed, leading me to conclude that CMC treatment was likely unsuccessful.

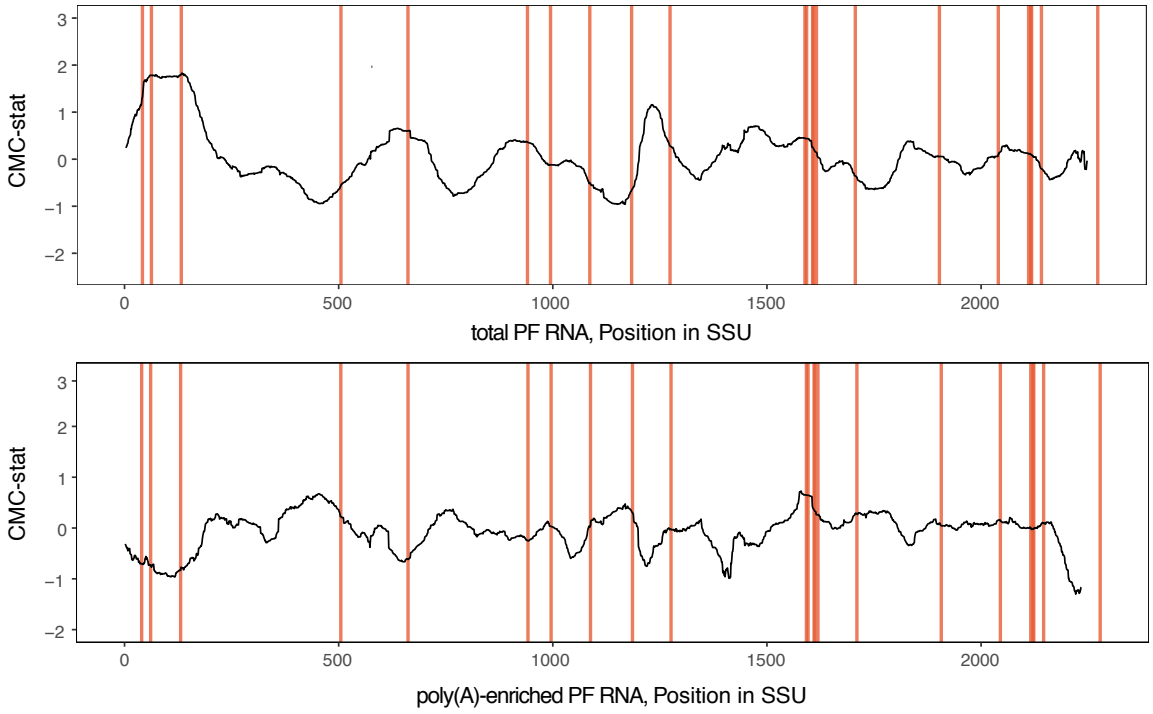


**Figure 4.10. CMC-stat analysis of trypanosome Ψ-seq libraries at the SSU locus.** CMC-stat was calculated for PF Ψ-seq libraries prepared from total RNA (top) and poly(A)-enriched RNA (bottom). Known Ψs are indicated as red vertical lines.

The lack of experimental reproducibility — in contrast to that of the readily reproduced Ψ-seq bioinformatic pipeline — raises concerns around the tractability of high-throughput Ψ-detection methods. Nevertheless, I decided to proceed and apply Ψ-seq analysis to reads mapped to the whole genome, stopping at a number of analysis "checkpoints" to better characterize the experimental source of failure to detect Ψ sites in rRNA, and uncover any other caveats that might apply to Ψ-seq detection. Of course, because CMC conjugation to Ψ — or hydrolysis from non-Ψ residues — likely did not work, any quirks found in my Ψ-seq libraries may not apply to those prepared with successful, Ψ-specific CMC conjugation.

### 4.3.2 Detection of Ψs in the whole transcriptome

In this section, I filter candidate Ψ sites through multiple checkpoints, each of which illuminates a caveat or flaw, either in my libraries specifically or in the published Ψ-detection methods as a whole.

#### 4.3.2.1 Putative Ψ sites are not exclusively called at reference U positions

Ψ-seq-specified cutoffs were used to call putative Ψ sites with reads aligned to the whole genome. Because no false positives were called with these particular thresholds applied to the rDNA locus, I was curious to know whether putative Ψs would be called at positions that did not correspond to a reference U. Without filtering, 156 and 114 sites were called before and after deduplication in BSF cells (73.1% hit retention), while 23 and 10 sites were called before and after deduplication in PF cells (43.5% hit retention). I then analyzed the breakdown of reference nucleotides at each called position and found no significant enrichment for called sites at a U, as I would expect following Ψ-specific CMC treatment (Figure 4.11). Inefficient CMC hydrolysis from G-like residues — including inosine resulting from A-to-I RNA editing — would have resulted in hits at positions corresponding to a reference A, G, or U. However, positions corresponding to a reference C were also called, which notably does not contain a CMC conjugation site so could not be the result of inefficient CMC hydrolysis. These positions were likely not called due to reverse transcriptional arrest mediated by secondary structure, as the Ψ-ratio would be the same in both CMC-treated and mock-treated libraries. The called sites may thus be the result of computational noise — sites that happened to pass the defined Ψ-detection metrics — or the consequence of technical shortcomings inherent in the Ψ-seq library preparation protocol.

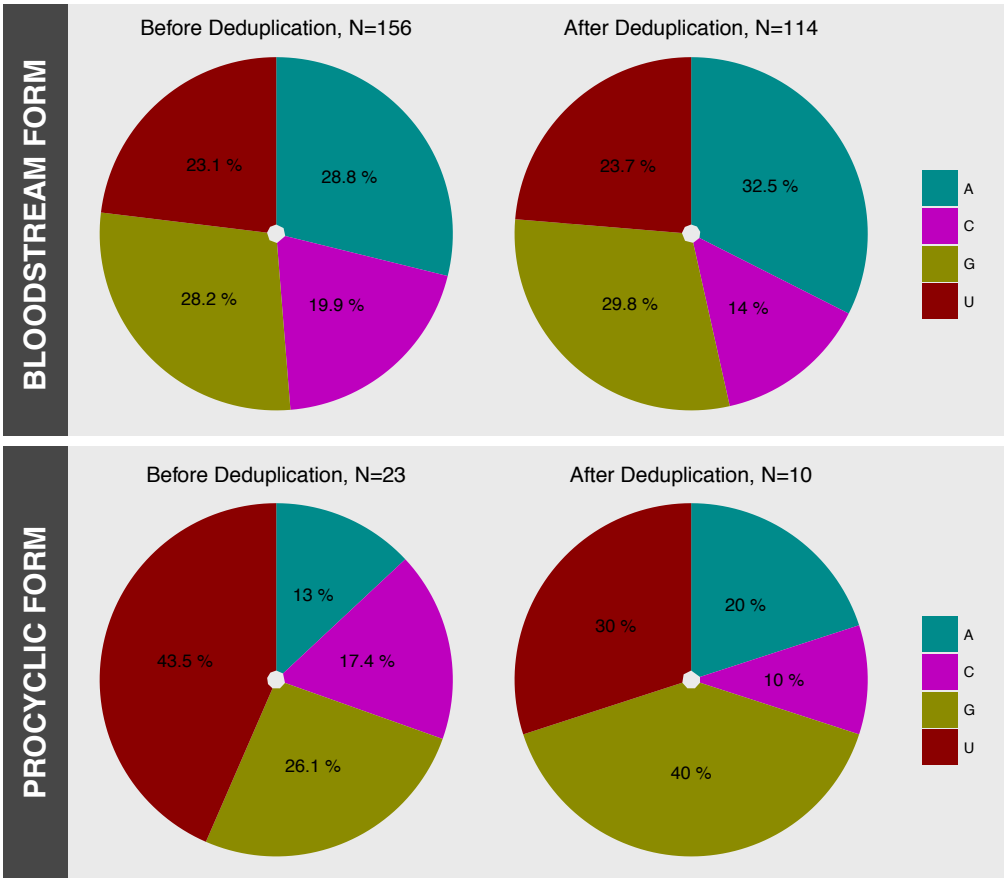


**Figure 4.11. Reference nucleotide breakdown of called Ψ sites from whole-genome deduplicated reads.**

### 4.3.2.2 Deduplication results in newly called Ψ sites following filtering for U

I next investigated the effect of deduplication on called Ψ sites. Specifically, were new sites called following deduplication that were otherwise obscured due to PCR duplication events? I therefore filtered for called sites occurring at a reference U and compared sets of putative Ψs in BSF and PF samples to determine whether deduplicated hits were contained in the set of called sites prior to deduplication (Figure 4.12). While all deduplicated Ψ hits fell within the set of PF hits called prior to deduplication, four new Ψ sites were called in the deduplicated set for BSF trypanosomes. These additional sites

might have been called due to elimination of background due to PCR duplication events. This result should still be taken with a grain of salt given my initial low confidence that called Ψ sites in these problematic libraries are true sites of pseudouridylation.

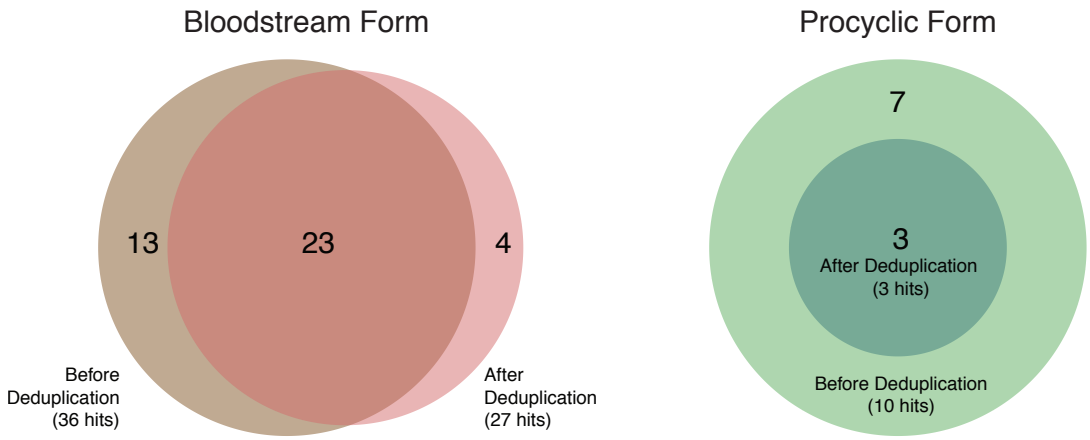


**Figure 4.12. Overlap in called Ψ sites before and after deduplication.**

#### 4.3.2.3 Ψ-calling is sensitive to sequencing depth

I also observed nine times as many putative Ψ sites in BSF versus PF trypanosomes following deduplication, which was intriguing given the increased levels of uridine isomerization reported in BSF versus PF rRNA [24]. To make any meaningful direct comparisons between the two life cycle stages requires equal coverage. BSF libraries, however, were sequenced at three times the depth. I therefore sampled one-third of BSF mapped reads and performed Ψ-seq analysis on this fraction, which returned four putative Ψ sites (filtered by reference nucleotide). The fact that different Ψ sites may be called following deduplication or downsampling of reads indicates just how sensitive Ψ-detection is to sequencing depth. Regardless of depth or deduplication, however, BSF and PF putative Ψ sites did not overlap (Figure 4.13).

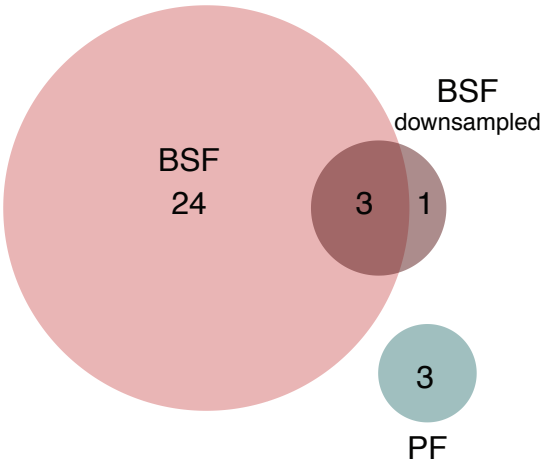


**Figure 4.13. All called Ψ sites are life-cycle stage specific.**

### 4.3.4 Mismatch analysis further filters putative Ψ hits

As a final checkpoint, I applied the mismatch filters described in Chapter 2.2.3 on the final set of Ψ sites curated following U-filtering, deduplication, and appropriate downsampling. Of the four BSF and three PF Ψ sites, three passed the appropriate mismatch rate filters — that is, a greater than four-fold higher CMC-treated mismatch rate (with respect to mock-treated samples) that is itself greater than 1.5%. Of those remaining three sites, only one in the BSF candidate Ψ set — U2030 in the gene Tb927.11.6440 — exhibited the C(/A) nonreference nucleotide incorporation profile characteristic of Ψ (Figure 4.14). The sequence surrounding this remaining position (GTG**T**TCA), however, did not map to any known H/ACA snoRNAs or guide-RNA independent Ψ-synthases, which did not increase confidence in this putative site.

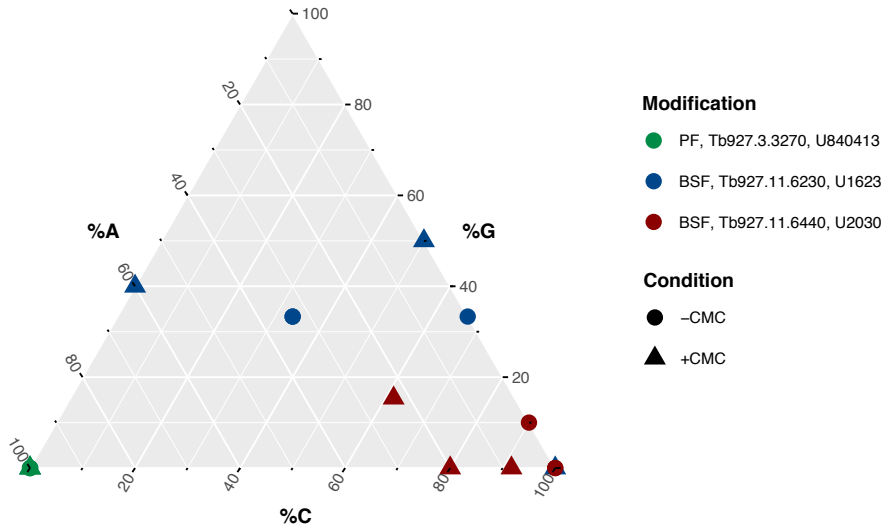


**Figure 4.14. Nonreference nucleotide incorporation profiles for putative Ψ sites passing mismatch rate filters.**

I analyzed the expression difference in Tb927.11.6440, which is annotated only as "hypothetical protein," between BSF and PF cells using DESeq, and found the gene is expressed at a two-fold higher level ($p_{adj} = 6.14 \times 10^{-21}$) in BSF. The expression difference is in line with data from a comparative RNA-seq analysis of BSF and PF transcript expression, which was conducted in parallel with comparative ribosome profiling that found increased ribosome occupancy on the PF transcript [119]. Counterintuitively, the half-life of this transcript is increased in BSF versus PF trypanosomes, as indicated in another transcriptome-wide analysis of life-cycle specific mRNA decay [35]. Considering how Ψ factors into the relationship between transcript stability and translation efficiency during differentiation is certainly interesting; however my low confidence in this particular Ψ site, combined with the translated protein's unknown function, makes pseudouridylation to this transcript a poor subject for further study.

Thus, from 184 candidate Ψ sites in the combined bloodstream form and procyclic form transcriptome, I computationally winnowed my way down to only one low-confidence Ψ-site in BSF cells that mapped to a protein of unknown function. The initial pool of hits was likely called either due to inefficiencies in CMC conjugation and subsequent alkaline hydrolysis, or because of computational noise that makes it impossible to meaningfully distinguish signal from noise.

### 4.4 Barcoded Ψ-seq method revisited: a post-thesis defense addendum

During the defense of this thesis, a committee member raised a potential flaw in the protocol used to generate Ψ-seq libraries with molecular barcodes. Following first strand synthesis, the Illumina TruSeq® Stranded mRNA kit or Total RNA kit was used for second strand synthesis and subsequent library preparation steps. With these kits, second strand dscDNA is synthesized by first digesting the original RNA template with RNase H to produce short fragments for priming and extension by DNA Polymerase I (Figure 4.15). The resulting dscDNA is then subjected to end repair, which digests away 5′ and 3′ overhangs. As a result, the sequence immediately downstream of a Ψ-CMC may be lost, so read starts may not correspond to sites of reverse transcriptional termination. In the context of putative Ψ detection, the exact position of Ψ-CMC-mediated RT arrest may be obfuscated when second strand synthesis priming does not occur at the exact end of the cDNA (Figure 4.15, depicted by maroon shaded box). Loss of sequence information at the site of RT termination may therefore have been the reason I could not accurately detect known sites of pseudouridylation in the *T. brucei* ribosome.

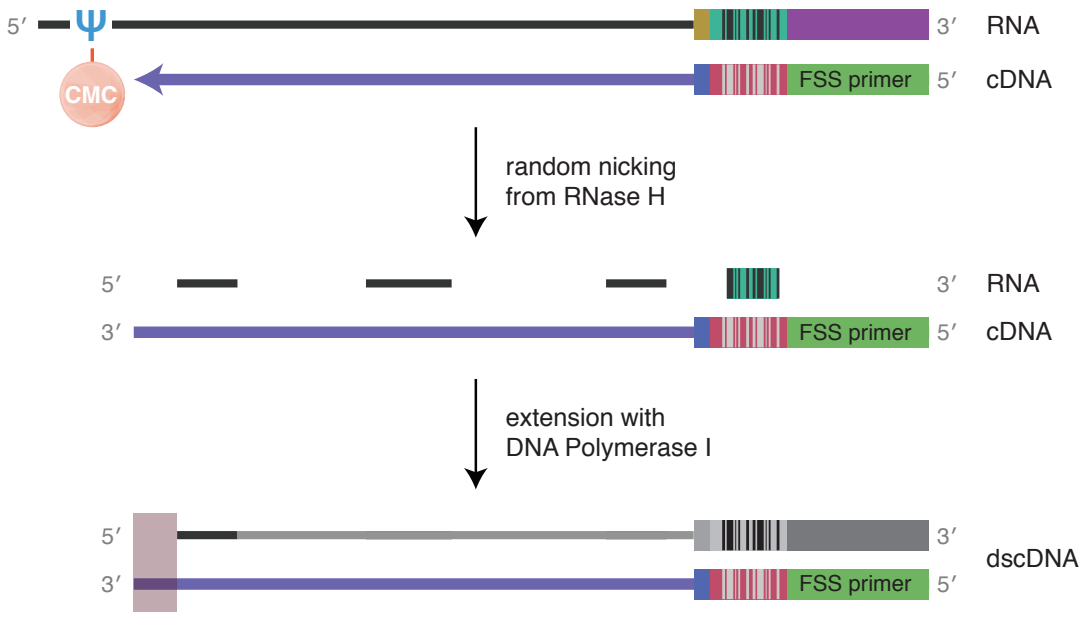


**Figure 4.15. Schematic of second strand synthesis with Illumina TruSeq® kit.** Illumina's TruSeq® library preparation kits perform second strand synthesis using RNase H to digest the original RNA template into short fragments to allow for priming and extension with DNA Polymerase I. An end repair reaction is then performed to digest away 5′ and 3′ overhangs. As a result, the sequence just downstream of a Ψ-CMC site may be removed prior to sequencing (depicted by maroon shaded box).

Following my thesis defense, I modified the Ψ-seq library preparation protocol to ensure that the sequence corresponding to RT termination was included in the final dscDNA product (Figure 4.16). CMC treatment was performed as before. Following first strand synthesis, a 3′ adapter containing a randomized 6-nucleotide barcode was ligated onto the resulting cDNA. The adapter contained a universal priming sequence, which was used to generate the second strand using the high fidelity Phusion polymerase. Second strand synthesis was then followed by end repair, A-tailing, and Illumina adapter ligation.

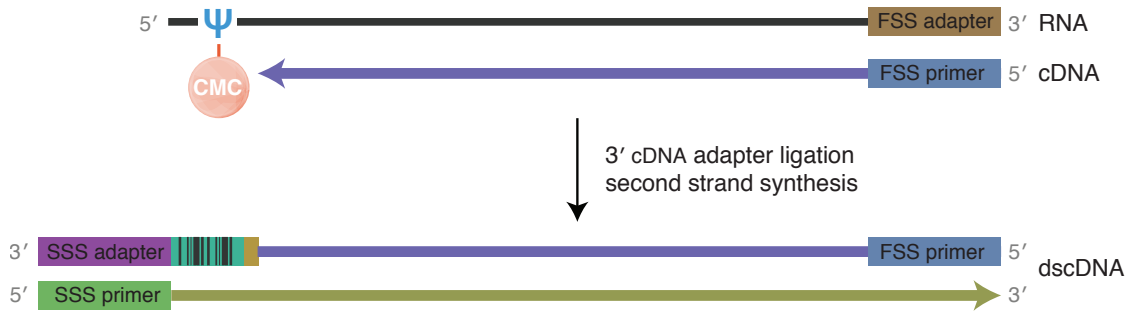


**Figure 4.16. Schematic of modified Ψ-seq library preparation.**
A first strand synthesis adapter (FSS adapter) without a barcode was ligated to the 3′ end of CMC-treated or mock-treated RNA. A primer (FSS primer) complementary to the adapter was used for first strand synthesis. A second strand synthesis adapter (SSS adapter) containing a randomized 6-nucleotide barcode was then ligated to the 3′ end of the resulting cDNA and second strand synthesis was performed with a primer (SSS primer) complementary to a universal priming sequence.

I applied this modified Ψ-seq library preparation protocol to ribosomal RNA isolated in duplicate from bloodstream form trypanosomes to determine whether I could now accurately detect known sites of pseudouridylation. Analysis of the resulting sequencing reads was performed as before, except the barcode was extracted from the 'right-hand' read (Figure 4.3C). Once again, however, each position in the rDNA locus was covered by over 4,096 initiating reads, so I did not deduplicate reads at the risk of collapsing reads that were not truly PCR duplicates. Utilizing this new Ψ-seq protocol, three Ψ sites were called, all three corresponding to known sites of pseudouridylation (Figure 4.17, Table 4.3).

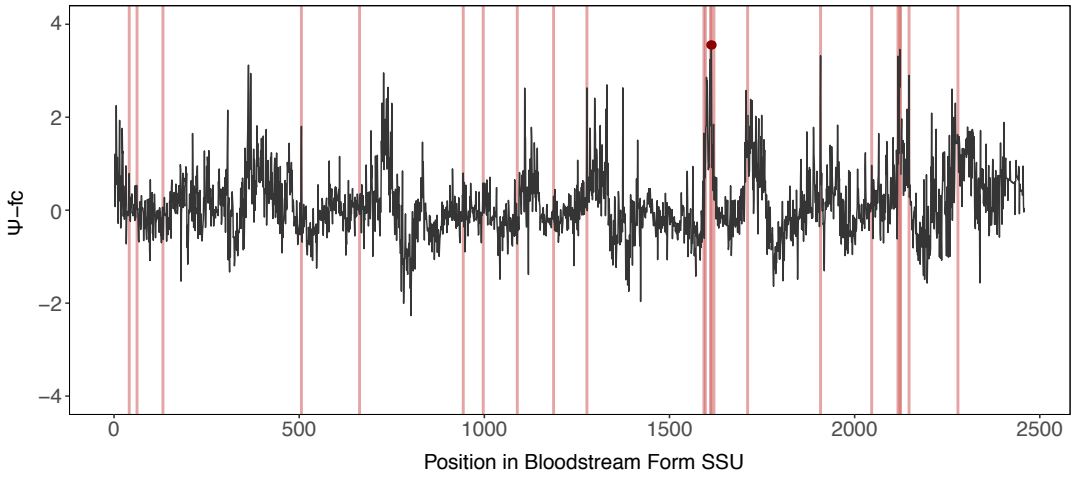


**Figure 4.17. Modified Ψ-seq protocol detects one known Ψ in the *T. brucei* SSU.** Ψ-fc values were plotted for the small ribosomal subunit (SSU) RNA in BSF trypanosomes. Known Ψ sites are marked by vertical red lines. Known Ψ called by Ψ-seq indicated by red point.

**Table 4.3. Putative Ψ sites called by Ψ-seq with modified library preparation.**

| Gene | Position | CMC-treated Ψ-ratio | Ψ-fc |
|---|---|---|---|
| LSU5′ | 935 | 0.127 | 3.96 |
| LSU3′ | 1377 | 0.154 | 3.73 |
| SSU | 1612 | 0.279 | 3.64 |

The results of this pilot experiment are encouraging, providing an explanation for the high rate at which putative Ψ sites were called at positions that corresponded to a non-U nucleotide in the reference genome, as detailed in Figure 4.11. These sites could be the products of second strand synthesis priming further from the exact 3′ end of the cDNA. As a result, they may not map exactly to the site of reverse transcriptional termination. Nevertheless, the possibility remains that these sites were called due to

computational noise, rather than corresponding to sites in the near vicinity of a putative pseudouridylation target.

Furthermore, only three of 75 previously detected Ψ sites were identified, which speaks to limitations inherent in the CMC-based Ψ detection approach. My inability to detect the remaining Ψ sites could be due to sequencing depth, though the average number of reads from each prepared library was ~ 1 x $10^7$, three-fold more than the number of reads sufficient to call known sites of pseudouridylation by Schwartz *et al.* [108]. Alternatively, the failure to detect more known Ψ sites could be due to CMC conjugation efficiency. If CMC does not efficiently conjugate to all U- and G-like residues, there is a higher likelihood of a high false negative rate, as observed here. Because no false positive sites were called — in other words, no sites that correspond to a 'G' in the reference genome — I am reasonably confident that the previous library preparation scheme, rather than inefficient alkaline hydrolysis, was the culprit behind the non-U reference Ψ calling detailed in Figure 4.11. The inefficiency of the method as adapted here, combined with the caveats to Ψ detection discussed throughout the body of this thesis, thus raises concerns that the available CMC-based Ψ-detection methods are not robust "plug and play" methods for *de novo* Ψ detection.

# CHAPTER 5. Discussion

*Ever tried. Ever failed. No matter. Try Again. Fail again.*

*Fail better.*

— Samuel Beckett

## 5.1 Reproducibility and reusability of high-throughput Ψ-detection methods

Since I began the work detailed in this thesis, the field of pseudouridylation has grown tremendously with reinvigorated interest in this intriguing modification. In that time, hundreds to thousands of Ψ sites have been catalogued depending on the species surveyed and the technique used [19,69,107]. Amidst the staggering number of novel putative sites, however, only three — two in human rRNA and one in mRNA — have been experimentally verified [69]. In addition, my comparative analysis of each method's results revealed only a small subset of Ψ sites that were reproducibly detected by two or more approaches [131].

The aim of developing these approaches was to generate testable hypotheses as to the functional role of pseudouridylation by identifying pseudouridylated transcripts of biological interest. The currently available methods as they stand are not robust enough to produce such informed hypotheses. While my comparative analysis did generate a confident set of 10 mRNA Ψ sites in yeast, groups generating Ψ maps for new species will not have the benefit of comparing their results to three other labs using three different-yet-related techniques. The bioinformatic pipeline detailed by Schwartz *et al*. in Ψ-seq, and later adapted for CeU-seq, is readily reproduced due to its intuitive derivation. On the contrary, the experimental pipeline for Ψ-seq library preparation and RNA

sequencing is not readily reused, based on my experiences in *T. brucei*. Following my attempts to combine Ψ-seq with molecular barcoding, I homed in on three major areas for improvement — apart from the absolute quantitation discussed in Chapter 3.4 — that likely affect the low reusability I experienced.

First and foremost, for the field to move forward, a common Ψ-detection method should be agreed upon, especially if users are to compare results with one another. CeU-seq is by far the most sensitive detection method, though the commercial availability of non-clickable CMC makes Ψ-seq a more practical option for wider use.

Second, CMC-based techniques would benefit from a quality control step to ensure efficient CMC derivatization and alkaline hydrolysis before proceeding to library preparation. For instance, a fluorescent biotin moiety could be conjugated to the CMC-azide utilized in CeU-seq to visualize CMC conjugated to treated transcripts. This is a very useful control step that I was developing together with the Helm group, but fell by the wayside due to limited availability of CMCyne and publication of the other Ψ-detection approaches. Of course, fluorescence could not confirm efficient hydrolysis from non-Ψ residues. Nevertheless, quantifying some minimum standardized level of fluorescence could create confidence that the conjugation step at least worked. To assess efficient hydrolysis following sequencing, I propose eliminating the filter removing putative Ψ sites at positions that do not correspond to a U in the reference genome to assess hits resulting from unhydrolyzed G-CMC and I-CMC.

Third, deeper sequencing leads to identification of more putative Ψ sites, as demonstrated during my attempts to profile Ψs in bloodstream form versus procyclic form trypanosomes. In order to accurately compare putative Ψ maps, coverage must

therefore be consistent. Consequently, I propose delineating a standardized range for sequencing depth for each method. With that said, a cutoff threshold cannot account for differential pseudouridylation events in response to slightly different environmental contexts intrinsic to different laboratories or population averaging effects. Still, the closer the field can move towards standardized experimental protocols, the more likely multiple users are to corroborate and build confidence around a given set of pseudouridylation events within the same biological system.

### 5.2 Need CMC-independent approaches for Ψ-detection

Limitations in CMC derivatization and hydrolysis efficiency, as demonstrated both in the literature and in my Ψ-seq results with trypanosome RNA, underscore the need for high throughput methods that circumvent CMC. Even with the modified Ψ-seq method utilized after the defense of this thesis, only a small sampling of known ribosomal Ψ sites were detected. The high false negative rate further highlights the limitations of CMC-based detection approaches, as they are highly reliant on CMC conjugation efficiency, which is well-known for being temperamental. One promising method involves real-time monitoring of the rate of nucleotide dissociation and incorporation for biological polymerases along a single RNA or DNA molecule. The technology, known as **S**ingle **M**olecule, **R**eal **T**ime (SMRT®) sequencing by PacBio, can identify certain nucleic acid modifications as pauses in the polymerization machinery when the modification interferes with base pairing at the Watson-Crick face. In the case of RNA sequencing, the SMRT platform directly monitors the kinetics of the reverse transcription reaction, removing the need to synthesize and amplify dscDNA. Accurate detection of methylated deoxyribonucleotides and β-D-glucosyl-hydroxymethyluracil (base J), and of the

modified ribonucleotide $N^6$-methyladenosine by SMRT sequencing have been published [38,42,123]. In addition, a patent has been filed citing successful detection of a number of other modified nucleotides including Ψ [64]. Despite the great potential of single molecule RNA sequencing, however, PacBio has focused on using the SMRT sequencing technology for DNA sequencing, particularly of long (~10,000nt long) DNA reads (personal communication).

Advances in mass spectrometry have also allowed for RNA modification mapping of multiple RNA species at one time [74]. However, high-throughput MS has been limited to profiling modifications in highly abundant, highly modified RNAs like tRNAs and rRNAs. Still, mass spectrometry methods may be used to validate putative pseudouridylation events in mRNAs as an orthogonal validation approach before designing experiments to investigate a given Ψ site's potential function. Additionally, antibodies have also been raised against Ψ, which could facilitate an immunocapture-based approach similar to $m^6$A-seq; however, these antibodies are not selective (Helm, Motorin, and Meier, personal communication) [32]. Finally, a lesser explored approach to Ψ detection exploits Ψ's (and $m^5$U's) increased resistance to hydrazinolysis compared to other pyrimidines [9]. Hydrazine treatment followed by aniline treatment cleaves the polynucleotide chain at non-Ψ pyrimidines, causing termination of RT one base 3′ to hydrazine-sensitive non-Ψ (i.e. U- and C-like) residues. Thus, this approach is the direct complement of the CMC/RT approach, though experimental caveats have been investigated to a far lesser degree.

### 5.3 Policy reforms to incentivize collaboration, corroboration, and revision

Science not published is often regarded as science not done. By this standard, a large fraction of replication studies or research that results in negative or non-confirmatory data has never come to pass because it is not widely accessible by the scientific community. Modern day mainstream scientific publishing, however, screens for articles based on perceived importance and impact, which often biases publication towards studies that report novel and positive results [36]. There is also an implied judgment on studies that fail to replicate or result in negative data — namely, that they were conducted by a set of sloppy scientists. Interpersonal politics and etiquette within the culture of science may thus further disincentivize entering the results of post-publication peer review into the public record. This mentality can also foster an unwillingness to transparently share data, lest inconsistencies or overlooked caveats be uncovered.

Nevertheless, science proceeds by collaboration, corroboration, and revision. Research of the sort undertaken in this thesis is therefore crucial to correct or improve upon the scientific record. Technological advances of the day — particularly the increased capacity of servers to efficiently store and share information — have allowed for policy innovations that incentivize transparency, replication, and publication of data of all sorts. In this section, I discuss some of these policies, which collectively form the basis of the open science movement, and how they relate to my own experiences comparing published data, attempting to replicate a published technique, and seeking an accessible platform for publication of the results.

### 5.3.1 Incentivizing transparency

To conduct my comparative analysis, I had to parse through data generated by four different groups. Raw sequencing reads were readily accessible, owing in large part to data-sharing requirements. Since 2008, the National Institutes of Health (NIH) has required that funded projects share relevant data within one year of publication [86]. U.S. government funding agencies like the National Science Foundation (NSF) and the Centers for Disease Control (CDC) have since followed suit, particularly under further incentive by a 2013 memo from the White House Office of Science and Technology Policy (OSTP) on increasing public access [51]. Similar policies have swept the field, implemented by governmental organizations outside the U.S. like CERN and charitable foundations like the Bill & Melinda Gates Foundation. The Transparency and Openness Promotion (TOP) Committee, organized by the Center for Open Science, *Science* Magazine, and the Berkeley Initiative for Transparency in Social Science, has also generated a series of guidelines for best practices in transparent publication, which is under review by over 500 journals [2,94].

Top-down mandates to submit data used in publication, however, ignore inconclusive data sets like the ones generated during my Ψ-seq experiments in *T. brucei*. It could well be that my data sets are just one in a sea of several failed Ψ-seq experiments that could collectively highlight compelling flaws in and caveats to Ψ detection. As a result, scientific fields as a whole have little opportunity to learn from collaborative analysis of negative data. Publication of this brand of data thus requires bottom-up cultural shifts that view negative data as an equal contribution to scientific practice [87].

Even with confirmatory data flooding repositories, there is still a need to integrate data-processing workflows so that they are likewise open for use and inspection. For instance, while I was able to easily access raw reads through their Gene Expression Omnibus (GEO) Accession Numbers, the scripts utilized to process the data were not openly accessible. I was therefore left to replicate the pipelines based on each technique's Methods section. In fact, I chose Ψ-seq's pipeline in large part because I could not replicate Pseudo-seq's computational approach even on the authors' own data.

Independent platforms, like the Open Science Framework (OSF), now exist to easily and transparently link workflow to data. For instance, the OSF allows researchers to store and selectively share every step of their work, allowing both data and pipelines used for analysis to be shared even before publication. Utilizing these platforms for next-generation sequencing studies can increase the robustness of data analysis, allowing research stakeholders to collaboratively, for instance, catch coding errors or debate the use of certain statistical tests. In so doing, data analysis pipelines may become more standardized so as to be reused by multiple groups, instead of existing as standalone computational methods associated with only one or a few publications that then need to be re-derived on an ad hoc basis. Adoption of frameworks like the OSF therefore has the potential to reduce competition and foster communal collaboration in an ongoing peer review process even after publication.

### 5.3.2 Incentivizing replication studies

Transparent data and analysis sharing are often posed as solutions to a broader conversation centered on reproducibility of results. A poll conducted by *Nature* and published in May 2016 reports that of 1,500 scientists surveyed, 90% believe there is a

slight to significant crisis around reproducibility [5]. In fact, most of the scientists surveyed had tried and failed to replicate either their own or someone else's experiments. The so-called "reproducibility crisis" originally gained mainstream attention in 2012 following the finding that many preclinical research studies could not be reproduced [15]. In 2015, reproducibility hit the media radar once again, eliciting coverage from the likes of *The Atlantic, Vox.com*, and *The New York Times*, following the completion of The Reproducibility Project: Psychology in August 2015 [95]. A collaboration of 270 researchers undertook the replication of 100 studies published in three psychology journals, where the data were openly accessed. The rationale was that potentially problematic practices — "selective reporting, selective analysis, and insufficient specification of the conditions necessary or sufficient to obtain the results" — might bias results so that they appear statistically significant, while an alternative, more objective third party analysis may show otherwise.

97% of the studies originally reported statistically significant results, while only about a third of the replications were able to corroborate those results. These findings sparked controversy as to the validity of the methodology of the replication studies in which each side essentially accused the other of not understanding statistics [4,45]. The debate itself, which took place in *Science*, demonstrated that research is often open to interpretation and that the more points of view we have addressing a given problem, the more opportunities for productive conversation around methodology and results.

While psychology is often dismissed as not a "hard science," the Reproducibility Project: Cancer Biology is currently under way, which may reveal similar obstacles to replication in a "harder" science. Researchers are attempting to independently replicate

37 high-impact experimental results in preclinical cancer biology studies published between 2010 and 2012 [34]. The project notably scaled back from replicating 50 studies due to budgetary constraints, which highlights a lack of support for replication studies even in a field like cancer biology where false leads can translate to tremendous time and resources wasted on clinical trials.

Next-generation sequencing (NGS) analysis, however, is rather inexpensive by comparison, only requiring a computer, coding skills, and time.[3] My comparative analysis of the replicability of putative Ψ sites using different NGS Ψ-detection techniques required only an investment in time. I was therefore curious to track reproducibility efforts undertaken with NGS sequencing data, apart from my own. As a crude first approximation, I searched for articles containing the phrase "next generation sequencing" with or without the term "reproducible" or "reproducibility" in the title or abstract on pubmed.gov. While articles containing "next generation sequencing" have gone up tremendously since the advent of NGS technologies, articles with "reproducible" or "reproducibility" have remained consistently low by comparison (Figure 5.1). As biology continues to generate big data, efforts to at the very least replicate analyses are crucial to ensuring that fields are pursuing true biological leads that result from robust data generation and analysis methods.

[3] As an aside, I could imagine undergraduate students replicating NGS analysis on openly accessed data as part of an advanced curriculum or thesis project, combining an educational opportunity with an inexpensive and valuable contribution to the scientific community.

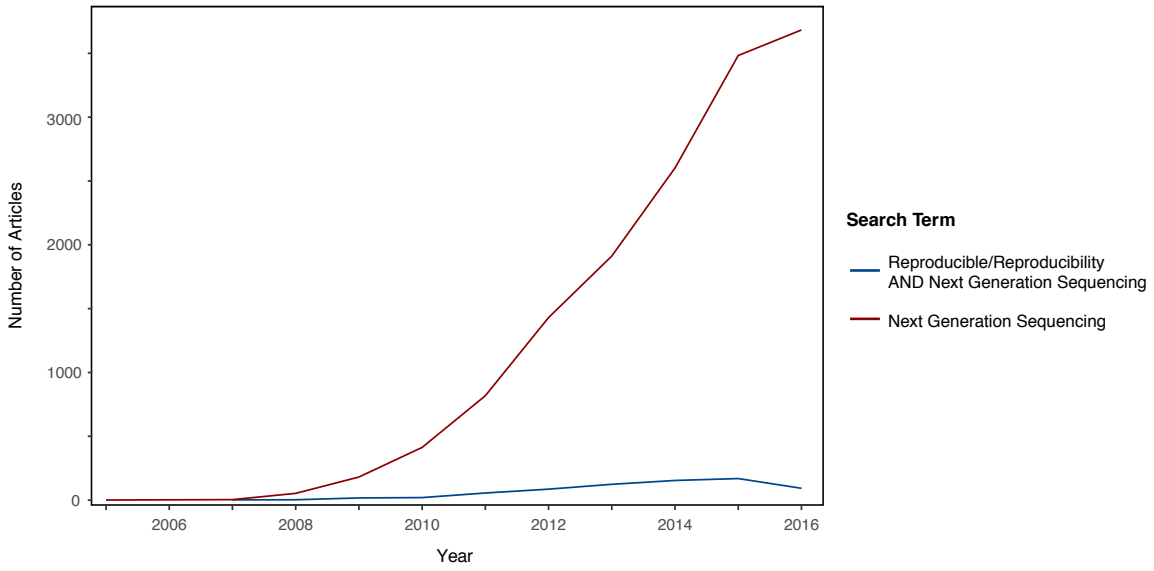


**Figure 5.1. Pubmed.gov search terms since the advent of NGS technologies.**

### 5.3.3 Incentivizing alternatives to traditional journal publication

Even without budgetary restrictions, there is little incentive to replicate results, particularly for early stage scientists like myself, who must innovate in order to move up the career ladder. We are more susceptible to the "publish or perish" mantra — the very same mantra that can pressure publication of irreproducible results in the first place — because a long list of publications on a resume is a conspicuous indicator of merit. In fact, my comparative analysis of high-throughput Ψ-detection methods was incentivized, in part, by a publication opportunity. Specifically, I was invited to write a "critical" review of the currently available methods as an objective fifth party. The results of my Ψ-seq experiments with *T. brucei*, however, do not readily fit into the traditional publishing structures, as they are neither innovative nor novel. Rather, they are inconclusive findings that could nevertheless benefit from peer review and open comment.

Fortunately, norms around publishing are evolving to fill the need for more transparent reporting and replication studies. Open-access journals like the *Journal of Negative Results in BioMedicine* (BioMed Central) and *F1000Research* provide a peer-reviewed platform for studies that can, for instance, prevent fellow researchers from pursuing false leads and generate collaborative discussions on how to improve existing techniques. In addition, preprint servers allow for direct uploading of complete scientific manuscripts not subjected to traditional peer review, which are accessible by the public.

Paper preprints were introduced in the 1960s, while electronic preprints were first introduced in 1991 with arXiv, founded by Paul Ginsparg for rapid communication of scientific findings in physics, mathematics, computer science, quantitative biology, quantitative finance, and statistics. ArXiv has become embedded in the culture of these fields, which enjoy submissions of over 100,000 papers to the server each year. Preprints are currently used minimally in the life sciences — where the preprint server of choice is bioRxiv — perhaps owing to a cultural reluctance to share results before they have gone through the quality check of peer review. Results in the biomedical sciences also have the potential to be commercialized, creating another obstacle to sharing proprietary findings or techniques. The onus of establishing quality is therefore left to the reader of the preprint, as opposed to a review board established by the journal. Proponents of preprints counter with the benefits of democratizing and de-anonymizing peer review through bioRxiv's comments section. Preprints have the added benefit of establishing priority in discoveries when publishing positive results, and many journals now accept manuscripts previously shared on a preprint servers [17].

### 5.5 Concluding remarks

The work presented in this thesis highlights not only limitations in high-throughput approaches to Ψ-detection, but also cultural limitations in the way that negative or non-confirmatory results can be broadly shared with the scientific community. As science continues to move more towards the generation and analysis of big data, there is great merit in detailed attention to techniques that are not readily reused and to results that are not replicated. In line with the tenets of the open science movement, I plan to assemble a condensed but thorough version of the work detailed in Chapter 4 for preprint publication. I will also deposit my data into PubMed Central and experiment with platforms like the Open Science Framework to share my workflow for comment and reuse. My transcriptome-wide hunt for uncharted sites of pseudouridylation has demonstrated that even work that results in experimental failure can be a valuable contribution to the practice and products of scientific inquiry.

## CHAPTER 6. A thesis condensed for nonscientists[4]

> *Knowledge is a big subject. Ignorance is bigger. And it is more interesting.*
>
> —Stuart Firestein, *Ignorance: How it Drives Science*

> *Ever tried. Ever failed. No matter. Try again. Fail again. Fail better.*
>
> —Samuel Beckett in *Failure: Why Science Is So Successful*

My first week in the lab, my boss plopped a book with the bold title *Ignorance: How it Drives Science*. And now, as I wrap up writing my dissertation, she has given me its sequel, *Failure: Why Science Is So Successful*. Preternatural optimist that she is, she did not gift these books out of pessimism or wry passive aggression. Rather, she believed they contained important lessons. Lessons that perfectly bookend my Ph.D. career.

My time in the lab began with ignorance — not the wide-eyed, first-year graduate student variety, but the rigorous brand that embraces an open question. A great conundrum in modern biology is how life's great diversity stems from four letters — A, C, G, and T — arranged in a near-infinite array to compose life's blueprint molecule: DNA. Now, consider that every cell in your body contains the exact same complement of DNA. Yet a heart cell looks and acts completely different from a brain cell which looks and acts completely different from a skin cell. So how did a heart cell, a brain cell, and a

[4] This chapter will be submitted to *Scientific American* blogs for publication. Underlined portions correspond to hyperlinks in the text.

skin cell arrive at such different biological fates when given the exact same set of molecular blueprints?

To deploy the blueprint's directions, instructions must first be transcribed to an intermediate molecule — the RNA — which then delivers them to the cellular machinery for execution. So understanding the dynamics of RNA, smack at the front lines of cellular activity, can help us understand how diversity emerges from the same DNA blueprint.

RNA is similarly composed of a four-letter alphabet: A, C, G, and U. That alphabet can be expanded upon with a library of over 100 chemical tweaks to fine-tune RNA function — a small M added to an A or a chemical S to a U. Of these alphabetical adornments, one stands out as the most ubiquitous: a subtle structural change in the genetic letter U to a pseudo-U, or pseudouridine (Ψ). Here, ignorance comes to play.

While Ψ was first discovered in the 1950s, we still don't know much about its precise biological function today, except that without Ψ, cells die. We do, however, have some clues — one that particularly piqued my interest. Introducing Ψs into a set of instructions that dictate how a protein is made changed the way those instructions were interpreted by the cell. Ψ unexpectedly recoded RNA's message beyond the mandates of the genetic code — a code considered fully cracked in the 1960s.

So in Ψ, I found a candidate for how diversity arises from DNA's hard-coded instructions. But that study was undertaken in an artificial system, which left open the question: where does Ψ naturally lie? By understanding *where* Ψs are, we might begin to uncover *what* exactly they do to affect how cells behave. When I wound my way to this question, we still had no methods to map Ψs beyond a few varieties of RNA. So, with the

power of next-generation sequencing technologies that first emerged to map the human genome, I went Ψ-hunting.

Meanwhile, the allure of Ψ had entered into the zeitgeist, calling researchers from around the world to endeavor on the same Ψ-charting quest. I was beat to the punch when four methods — three of which were released back-to-back-to-back — were published spotting Ψs in a whole host of RNAs. I decided to make the best of being quadruply beat to the punch and compared each group's Ψ maps, partly out of curiosity, but mostly because I was asked to review the techniques as an objective fifth party. All four methods were based on the same principle, so their results should overlap well with one another. But they did not. And here enters failure.

Of the hundreds to thousands of Ψs catalogued by each method, only a small fraction of sites were found by them all. I was genuinely surprised by the result. So I hunkered down and thought through a host of technical and biological caveats that were not detailed in the original publications. I then tried to apply one of those methods to map Ψs in African trypanosomes, the single-celled parasites that cause African sleeping sickness. But, try as I might, I could not get the method to work. And so, more failure.

Failure is the natural product of risk, and there's nothing riskier than the pursuit of ignorance — asking those big bold questions that probe the unknown. But while the practice of science is riddled with failures — from the banal failures of day-to-day life at the bench to the heroic, paradigm shifting failures that populate the book called *Failure* — many scientists are uncomfortable with the idea. We publish our innovations, the stories of how our ignorance led to success. Where the "publish or perish" mantra prevails, these stories are essential to making a name for ourselves and securing grant

money. So there is little incentive to replicate the work of others or report experimental failure. In fact, there is barely a medium to publish these sorts of efforts, which are relegated to the bottom of the file drawer.

But the scientific method hinges on self-correction, which requires transparent reporting of positive (or negative) data and corroboration (or contradiction) of previous experiments. And so I wanted to share my work, to open it up to comment, to transform my failure into something productive. If I couldn't get these Ψ mapping methods to work in my hands, that's a problem worth sharing because chances are, I'm not alone. This is how we avoid chasing false leads, how we improve our practices, how we move science forward. These tenets lie at the heart of the "open science" movement, which I have come to embrace (despite its New-Agey name) as I have ventured to share the failed fruits of my doctoral work.

Of course, open science is easier said than done. The increasing competitiveness of certain scientific fields has disincentivized transparency and collaboration. There is also a value judgment that comes with sharing experimental failure — a vulnerability that your peers will view your efforts as sloppy, rather than earnest and honest. So distributing negative or non-confirmatory data comes with an extra burden of proof.

Still, policy reforms and open science advocates are working to incentivize practices that foster open collaboration. Open-source software like the Open Science Framework now exist for collaborative sharing of data and data-processing workflows. Peer-reviewed publications like *F1000Research* are now accepting negative or non-confirmatory data of the sort I generated during my thesis. Preprint servers — which allow for direct uploading of complete manuscripts without formal peer review (but open

for comment) and have long been embraced by the physics community — are now gaining steam in the life sciences thanks to the work of advocacy groups like ASAPbio.

While I haven't uncovered any mysteries in the world of RNA biology, I have learned that science needs to fail better. I am now conducting further investigation into the source of my failures with the hopes of finding and publishing their root so it may be of use to all those Ψ chasers. Because in science, things often don't work out the way we think they should, and we are left with our ignorance. But the narratives we form around failure — transparently, openly, and together — can be just as valuable as those we form around success.

## CHAPTER 7. Materials and methods

### 7.1 Culture methods and strains

74-D694 (SY670) yeast cells, courtesy of the Serio laboratory at the University of Arizona, were used for CMC-seq pilot experiments. Yeast were grown to log phase ($OD_{600} \approx 1.0$) at 30˚C in YPD media (Sigma-Aldrich).

*T. brucei* bloodstream form cells used were cultured from the strain Lister 427 (antigenic type MITat1.2 clone 221a) in HMI-9 media at 37˚C. Procyclic form Lister 427 cells were grown in SDM-79 media at 30˚C.

### 7.2 CMCyne derivatization and "click" chemistry

15 pmol of *in vitro* transcribed $tRNA^{Tyr}$ were suspended either in 47 mM CMCyne in BEU buffer (7 M urea, 4 mM EDTA at pH 8.5, 50 mM bicine) or in BEU buffer alone for a 10 μL total reaction volume. Samples were incubated for 2 hours at 37˚C and were then subject to ethanol precipitation. CMCyne-treated or mock-treated samples were then resuspended in hydrolysis buffer (50 mM $(NH_4)_2CO_3$, 2 mM EDTA) at pH 10.5, 11.0, or 11.5, and incubated at either 37˚C or 42˚C for either 2.5 or 3 hours. Following hydrolysis, RNA was again ethanol precipitated and subjected to "click" chemistry with the fluorescent azide $atto_{488}$. $Atto_{488}$-azide was conjugated to $tRNA^{Tyr}$ in a 20 μL reaction with 2.5 mM TPTA, 5 mM sodium ascorbate, 0.5 mM $CuSO_4 \cdot 5H_2O$, and 0.05 mM $atto_{488}$-azide. The light-protected reaction proceeded at 21˚C for 2 hours, shaking at 350 rpm. RNA was precipitated a final time and resuspended in water before being run on a 15% SDS-PAGE gel. The gel was scanned first for fluorescence, then stained with GelRed and scanned for RNA integrity. For the 30mer Ψ-containing oligonucleotide (or

its U-containing counterpart), CMC treatment and atto488 conjugation proceeded as detailed above, except 30 pmol of the RNA oligonucleotide were used to start.

### 7.3 Generation of sequencing libraries

#### 7.3.1 CMC-seq library preparation

Yeast cells were harvested by centrifugation and total RNA was extracted using the MasterPure™ Yeast RNA Purification Kit by Epicentre, which included a DNase treatment step. For each replicate, 12 μg of total RNA was either suspended in 0.17 M CMC (Sigma 29469, ≥99.0%, now discontinued) in BEU buffer (treated) or in BEU buffer alone (mock-treated), and incubated for 2 hours at 37˚C. The reaction was stopped by ethanol precipitating RNA. The pellet was resuspended in 50 μL of hydrolysis buffer at pH 11.0 and the mixture was incubated for 2.5 hours at 37˚C. RNA was again precipitated. Libraries were prepared with the TruSeq® Stranded Total RNA kit (Illumina) without the RiboZero Deplete and Fragment RNA step in the protocol. The resulting dscDNA libraries were then diluted to 15 nM and pooled for sequencing on the Illumina HiSeq 2000 for single-end 100bp reads.

#### 7.3.2 Ψ-seq library preparation with molecular barcodes

Trypanosome cells were harvested by centrifugation and total RNA was extracted using RNA STAT-60 (Tel-Test, Inc.), according to the manufacturer's protocol. For poly(A)-enriched libraries, 300 μg of total RNA was used as starting material per replicate, followed by poly(A) enrichment using the μMACS mRNA Isolation Kit for total RNA. RNA was then DNase treated using RQ1 RNase-Free DNase (Promega) and cleaned with 2x RNAClean® XP beads (Agencourt), yielding approximately 1-3 μg of starting material

for CMC treatment. For libraries prepared from total RNA, 12 μg of RNA (following DNase digestion and clean-up) was used as starting material per replicate.

Prior to CMC derivatization, RNA was fragmented for 15 minutes with RNA fragmentation reagent and stop solution (Ambion), according to the manufacturer's specifications, and cleaned with 2.5x RNAClean® XP beads. CMC treatment and hydrolysis proceeded as in 7.3.1, except that RNA was cleaned with 3x RNAClean® XP beads following alkaline hydrolysis. In addition, CMC at ≥99.0% purity was discontinued, so I had to use CMC at ≥95% purity from Sigma.

RNA was then dephosphorylated with FastAP Thermosensitive Alkaline Phosphatase (Thermo Scientific), cleaned with 3x RNAClean® XP beads, followed by overnight 3′ adapter ligation with T4 RNA ligase (New England Biolabs) at 16˚C. After cleaning RNA once again with 3x RNAClean® XP beads, first strand synthesis primers with 5′ 2N, 1N, or 0N ends were mixed at equimolar concentrations and used for first strand synthesis with Superscript III Reverse Transcriptase (Invitrogen). I then proceeded to library preparation with the TruSeq® Stranded mRNA kit (Illumina), starting at the second strand synthesis step. The only deviation from the manufacturer's protocol was that dscDNA was cleaned with 3x AMPure XP beads (Agencourt) following second strand synthesis. The resulting libraries were diluted to 15 nM and pooled for sequencing on the Illumina HiSeq 2500 for 50bp paired-end reads with 10% spike-in of PhiX Sequencing Control (Illumina) to ensure accurate cluster calling.

### 7.3.3 Ψ-seq library preparation with molecular barcodes modified

Total RNA was extracted from bloodstream form trypanosome cells harvested as in 7.3.2. For each library, 12 μg of total RNA (following DNase digestion and clean-up) was used

as starting material per replicate. This time, however, RNA was not fragmented prior to CMC treatment and hydrolysis, which proceeded as in 7.3.2.

RNA was dephosphorylated with FastAP Thermosensitive Alkaline Phosphatase, cleaned with 2.5x RNAClean® XP beads. Overnight 3′ adapter ligation with T4 RNA ligase at 16˚C followed, this time with a short adapter that lacked a barcode. The RNA was cleaned once again with 2.5x RNAClean® XP beads. First strand synthesis was performed with Superscript III and first strand synthesis primers with 5′ 2N, 1N, or 0N ends mixed at a 4:2:1 molar ratio. The resulting cDNA was cleaned with 2.5x AMPure XP beads, followed by dephosphorylation with FastAP Thermosensitive Alkaline Phosphatase, and cleaned again with 2.5x AMPure XP beads. The sample was subjected to overnight 3′ adapter ligation with T4 RNA ligase at 16˚C with a DNA oligonucleotide that contained a randomized six-nucleotide barcode and a universal priming sequence. The oligonucleotides used for this experiment are preceded by 'mod Ψ-seq' in Table 7.1.

The reaction was cleaned with 2.5x AMPure XP beads and subjected to second strand synthesis with Phusion High-Fidelity DNA Polymerase (Thermo Fisher Scientific). Second strand synthesis primers with 5′ 2N, 1N, or 0N ends were mixed at a 4:2:1 molar ratio. To ensure a stranded cDNA library, I utilized an oligonucleotide mixture at a final concentration of 200 μM dATP, dCTP, and dGTP and 400 μM dUTP. The resulting dscDNA product was cleaned with 2.5x AMPure XP beads, followed by end repair using the NEBNext® End Repair Module. I then proceeded to library preparation with the TruSeq® Stranded mRNA kit (Illumina), starting at the A-tailing step. The resulting libraries were diluted to 10 nM and pooled for sequencing on the

Illumina HiSeq 2500 for 100bp paired-end reads with 10% spike-in of PhiX Sequencing Control (Illumina) to ensure accurate cluster calling.

### 7.4 Sequencing data analysis

#### 7.4.1 CMC-seq analysis

Illumina sequencing adapters were trimmed and low-quality reads were removed using trim_galore(v0.3.7) powered by Cutadapt [82]. The remaining reads were aligned to the rDNA locus of the yeast genome (sacCer3; locus: chrXII:451000..459999) using bowtie(v1.1.1) allowing only uniquely mapping reads with no more than two mismatches per read (*-v 2 -m 1*) [67]. Per-base coverage was then calculated using bedtools(v.2.20.1) multicov, which was then normalized using DESeq(v1.20.0) in Bioconductor for R [99].

CMC-stat was derived by first calculating the median normalized coverage at each position for treated and mock-treated libraries. The $\log_2$-transformed ratio of mock-treated coverage to treated coverage was calculated, adding a pseudocount of 1 to both the numerator and the denominator to avoid division by 0. CMC-stat plots were generated using the R package ggplot2(v.2.1.0).

For mismatch analysis alignment files were parsed using the Python module pysam(v.0.8.1) to determine the frequency of nonreference nucleotide incorporation at each position and the total number of reads mapping to each position. A mismatch rate (MR) was calculated by dividing the number of mismatched reads by the total number of reads at each position. The median MR was then calculated for treated and mock-treated libraries and the $\log_2$-transformed ratio of treated to mock-treated median MRs was determined. Further analysis of nonreference nucleotide incorporation profiles was restricted to positions with a median treated MR of greater than 1.5% and a $\log_2$-

transformed MR ratio of greater than 2. Nonreference nucleotide incorporation profiles were determined by dividing the number of each nonreference nucleotide incorporated at a given position by the total number of reads covering that position. Ternary plots were then drawn with the ggplot2 extension ggtern(v.2.1.4) in R.

### 7.4.2 Ψ-seq analysis

The Unix command grep was utilized to extract 'left-hand' reads containing the last eight nucleotides of the common priming sequence, followed by six random nucleotides and the ligation linker (i.e. GCGTTCGT……ACAG for adapter A). One mismatch was allowed within the adapter sequence. The SeqIO module from BioPython(v.1.63) was used to extract barcode sequences to create an index file with read name and barcode sequence used later for deduplication. The adapter sequence was then trimmed from the remaining left-hand reads, and corresponding paired 'right-hand' reads were extracted.

Right-hand reads were first aligned to the rDNA locus (genome Tb927v5.1; genes: Tb927.2.1389, Tb927.2.1398, Tb927.2.1407, Tb927.2.1416, Tb927.2.1425, Tb927.2.1434, Tb927.2.1443, Tb927.2.1452) using bowtie2(v.2.1.0) in end-to-end alignment mode [66]. Reads unaligned to the rDNA locus were then aligned to the whole genome. A custom R script utilizing the 'data.table' package was then used to deduplicate reads by discarding copies of those that map to the same position with the same sequence and barcode.

Following deduplication, the Python module pysam was used to calculate the number of reads initiating at a given position, the total number of reads covering that position, and the number of nonreference nucleotides incorporated for every mapped position. For each replicate, the Ψ-ratio was calculated for every position by dividing the

number of read starts by the total number of reads covering a given position, adding a pseudocount of 1 to both the numerator and denominator to avoid division by 0. The Ψ-fc was then calculated for every position by $\log_2$-transforming the ratio of the median Ψ-ratio for treated libraries to the median Ψ-ratio for mock-treated libraries. Only positions covered by all replicates were considered. Putative Ψ sites were called for positions 5′ to a position with a treated Ψ-ratio greater than 0.1 and a Ψ-fc greater than 3.

Ψ-fc plots were generated using the R package ggplot2. Receiver operating characteristic (ROC) curves were generated for rDNA Ψ-ratio and Ψ-fc values using ggplot2 with extension plotROC(v.2.0.1).

### 7.4.3 Modified Ψ-seq analysis

Analysis was similar to that detailed in 7.4.2, except that the Unix command grep was utilized to extract 'right-hand' reads containing the last eight nucleotides of the common priming sequence, followed by six random nucleotides and the ligation linker. The SeqIO module from BioPython(v.1.63) was used to extract barcode sequences to create an index file and then to trim adapters from the FASTQ reads. 100bp sequencing was utilized, and several reads corresponded to cDNA fragments less than 100bp in size. Consequently, adapters were trimmed from both ends of the reads, and only the right-hand reads were used for alignment. Reads were aligned to the rDNA locus using bowtie1(v1.1.1) allowing only uniquely mapping reads with no more than two mismatches per read (*-v 2 -M 1*). The resulting alignment files were analyzed as in 7.4.2.

## 7.5 Primer sequences

**Table 7.1. Primer sequences.**

| Name | Sequence (5′ → 3′) |
|---|---|
| $tRNA^{Tyr}$ oligo | TGG TGG TGG GGG AAG GAT TCG AAC CTT CGA AGT CTG TGA CGG CAG ATT TAC AGT CTG CTC CCT TTG GCC GCT CGG GAA CCC CAC C |
| Ψ-seq Adapter A | /5Phos/ CUG UNN NNN NAC GAA CGC AAU CAG CUU GCC G/3ddC/ |
| Ψ-seq Adapter B | /5Phos/ CUG UNN NNN NGU CAG GAU CAG GAG GCC GU G/3ddC/ |
| Ψ-seq Adapter C | /5Phos/ CUG UNN NNN NCG ACG CCG GAU UAC GGG A G/3ddC/ |
| FSS-A-2N | NNG GCA AGC TGA TTG CGT TCG T |
| FSS-A-1N | NGG CAA GCT GAT TGC GTT CGT |
| FSS-A | GGC AAG CTG ATT GCG TTC GT |
| FSS-B-2N | NNA CGG CCT CCT GAT CCT GAC |
| FSS-B-1N | NAC GGC CTC CTG ATC CTG AC |
| FSS-B | ACG GCC TCC TGA TCC TGA C |
| FSS-C-2N | NNT CCC GTA ATC CGG CGT CG |
| FSS-C-1N | NTC CCG TAA TCC GGC GTC G |
| FSS-C | TCC CGT AAT CCG GCG TCG |
| mod-Ψ-seq FSS | /5Phos/ GUC UAU CGU CCG GAG /3ddC/ |
| mod-FSS-2N | NNC TNC GGA CGA TAG AC |
| mod-FSS-1N | NCT NCG GAC GAT AGA C |
| mod-FSS-0N | CTN CGG ACG ATA GAC |
| mod-Ψ-seq SSS | /5Phos/ CUG UNN NNN NAC GAA CGC AAT CNN GG /3ddC/ |
| mod-SSS-2N | NNC AGC GAT TGC GTT CGT |
| mod-SSS-1N | NCA GCG ATT GCG TTC GT |
| mod-SSS-0N | CAG CGA TTG CGT TCG T |

## REFERENCES


[1] B. Addepalli, P.A. Limbach, Mass spectrometry-based quantification of pseudouridine in RNA, J. Am. Soc. Mass Spectrom. 22 (2011) 1363–1372.

[2] G.C. Alter, D. Banks, S.D. Bowman, S. Buck, C. Chambers, Transparency and Openness Promotion (TOP) Guidelines, (2016) 1–24.

[3] S. Anders, W. Huber, Differential expression analysis for sequence count data, Genome Biol. 11 (2010) R106.

[4] C.J. Anderson, Š. Bahník, M. Barnett-Cowan, F.A. Bosco, J. Chandler, C.R. Chartier, et al., Response to Comment on "Estimating the reproducibility of psychological science", Science. 351 (2016) 1037–1037.

[5] M. Baker, 1,500 scientists lift the lid on reproducibility, Nature. 533 (2016) 452–454.

[6] A. Bakin, B.G. Lane, J. Ofengand, Clustering of pseudouridine residues around the peptidyltransferase center of yeast cytoplasmic and mitochondrial ribosomes, Biochemistry. 33 (1994) 13475–13483.

[7] A. Bakin, J. Ofengand, Four newly located pseudouridylate residues in Escherichia coli 23S ribosomal RNA are all at the peptidyltransferase center: analysis by the application of a new sequencing technique, Biochemistry. 32 (1993) 9754–9762.

[8] A. Bakin, J. Ofengand, Mapping of the 13 pseudouridine residues in Saccharomyces cerevisiae small subunit ribosomal RNA to nucleotide resolution, Nucleic Acids Research. 23 (1995) 3290–3294.

[9] A.V. Bakin, J. Ofengand, Mapping of pseudouridine residues in RNA to nucleotide resolution, Methods Mol. Biol. 77 (1998) 297–309.

[10] S. Barth, A. Hury, X.-H. Liang, S. Michaeli, Elucidating the role of H/ACA-like RNAs in trans-splicing and rRNA processing via RNA interference silencing of the Trypanosoma brucei CBF5 pseudouridine synthase, J. Biol. Chem. 280 (2005) 34558–34568.

[11] A. Basak, C.C. Query, A pseudouridine residue in the spliceosome core is part of the filamentous growth program in yeast, Cell Rep. 8 (2014) 966–973.

[12] A. Baudin-Baillieu, C. Fabret, X.-H. Liang, D. Piekna-Przybylska, M.J. Fournier, J.-P. Rousset, Nucleotide modifications in three functionally important regions of the Saccharomyces cerevisiae ribosome affect translation accuracy, Nucleic Acids Research. 37 (2009) 7665–7677.

[13] L. Bazak, A. Haviv, M. Barak, J. Jacob-Hirsch, P. Deng, R. Zhang, et al., A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes, Genome Res. 24 (2014) 365–376.

[14] H.F. Becker, Y. Motorin, R.J. Planta, H. Grosjean, The yeast gene YNL292w encodes a pseudouridine synthase (Pus4) catalyzing the formation of psi55 in both mitochondrial and cytoplasmic tRNAs, Nucleic Acids Research. 25 (1997) 4493–4499.

[15] C.G. Begley, L.M. Ellis, Drug development: Raise standards for preclinical cancer research, Nature. 483 (2012) 531–533.

[16] I. Behm-Ansmant, A. Urban, X. Ma, Y.-T. Yu, Y. Motorin, C. Branlant, The Saccharomyces cerevisiae U2 snRNA:pseudouridine-synthase Pus7p is a novel

multisite-multisubstrate RNA:Psi-synthase also acting on tRNAs, Rna. 9 (2003) 1371–1382.
[17] J.M. Berg, N. Bhalla, P.E. Bourne, M. Chalfie, D.G. Drubin, J.S. Fraser, et al., Preprints for the life sciences, Science. 352 (2016) 899–901.
[18] M. Berriman, E. Ghedin, C. Hertz-Fowler, G. Blandin, H. Renauld, D.C. Bartholomeu, et al., The genome of the African trypanosome Trypanosoma brucei, Science. 309 (2005) 416–422.
[19] T.M. Carlile, M.F. Rojas-Duran, B. Zinshteyn, H. Shin, K.M. Bartoli, W.V. Gilbert, Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells, Nature. 515 (2014) 143–146.
[20] J.A. Casbon, R.J. Osborne, S. Brenner, C.P. Lichtenstein, A method for counting PCR template molecules with application to next-generation sequencing, Nucleic Acids Research. 39 (2011) e81–e81.
[21] M. Charette, M.W. Gray, Pseudouridine in RNA: what, where, how, and why, IUBMB Life. 49 (2000) 341–351.
[22] H.M. Chen, S.H. Wu, Mining small RNA sequencing data: a new approach to identify small nucleolar RNAs in Arabidopsis, Nucleic Acids Research. 37 (2009) e69–e69.
[23] S.H. Chen, G. Habib, C.Y. Yang, Z.W. Gu, B.R. Lee, S.A. Weng, et al., Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon, Science. 238 (1987) 363–366.
[24] V. Chikne, T. Doniger, K.S. Rajan, O. Bartok, D. Eliaz, S. Cohen-Chalamish, et al., A pseudouridylation switch in rRNA is implicated in ribosome function during the life cycle of Trypanosoma brucei, (2016) 1–13.
[25] C.E. Clayton, Life without transcriptional control? From fly to man and back again, The EMBO Journal. 21 (2002) 1881–1888.
[26] W.E. COHN, Some results of the applications of ion-exchange chromatography to nucleic acid chemistry, J Cell Physiol Suppl. 38 (1951) 21–40.
[27] A. Das, Q. Zhang, J.B. Palenchar, B. Chatterjee, G.A.M. Cross, V. Bellofatto, Trypanosomal TBP functions with the multisubunit transcription factor tSNAP to direct spliced-leader RNA gene expression, Molecular and Cellular Biology. 25 (2005) 7314–7322.
[28] F.F. DAVIS, F.W. ALLEN, Ribonucleic acids from yeast which contain a fifth nucleotide, J. Biol. Chem. 227 (1957) 907–915.
[29] W.A. Decatur, M.J. Fournier, rRNA modifications and ribosome function, Trends in Biochemical Sciences. 27 (2002) 344–351.
[30] W.A. Decatur, M.N. Schnare, Different mechanisms for pseudouridine formation in yeast 5S and 5.8S rRNAs, Molecular and Cellular Biology. 28 (2008) 3089–3100.
[31] Y. Ding, Y. Tang, C.K. Kwok, Y. Zhang, P.C. Bevilacqua, S.M. Assmann, In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features, Nature. 505 (2014) 696–700.
[32] D. Dominissini, S. Moshitch-Moshkovitz, M. Salmon-Divon, N. Amariglio, G. Rechavi, Transcriptome-wide mapping of N(6)-methyladenosine by m(6)A-seq based on immunocapturing and massively parallel sequencing, Nat Protoc. 8 (2013) 176–189.

[33] A. Durairaj, P.A. Limbach, Improving CMC-derivatization of pseudouridine in RNA for mass spectrometric detection, Analytica Chimica Acta. 612 (2008) 173–181.
[34] T.M. Errington, E. Iorns, W. Gunn, F.E. Tan, J. Lomax, B.A. Nosek, An open investigation of the reproducibility of cancer biology research, Elife. 3 (2014) 5773.
[35] A. Fadda, M. Ryten, D. Droll, F. Rojas, V. Färber, J.R. Haanstra, et al., Transcriptome-wide analysis of trypanosome mRNA decay reveals complex degradation kinetics and suggests a role for co-transcriptional degradation in determining mRNA levels, Molecular Microbiology. 94 (2014) 307–326.
[36] D. Fanelli, "Positive" results increase down the Hierarchy of the Sciences, PLoS ONE. 5 (2010) e10068.
[37] I.S. Fernández, C.L. Ng, A.C. Kelley, G. Wu, Y.-T. Yu, V. Ramakrishnan, Unusual base pairing during the decoding of a stop codon by the ribosome, Nature. 500 (2013) 107–110.
[38] B.A. Flusberg, D.R. Webster, J.H. Lee, K.J. Travers, E.C. Olivares, T.A. Clark, et al., Direct detection of DNA methylation during single-molecule, real-time sequencing, Nat. Methods. 7 (2010) 461–465.
[39] G.K. Fu, J. Hu, P.-H. Wang, S.P.A. Fodor, Counting individual DNA molecules by the stochastic attachment of diverse labels, Proc. Natl. Acad. Sci. U.S.a. 108 (2011) 9026–9031.
[40] G.K. Fu, W. Xu, J. Wilhelmy, M.N. Mindrinos, R.W. Davis, W. Xiao, et al., Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations, Proc. Natl. Acad. Sci. U.S.a. 111 (2014) 1891–1896.
[41] P. Ganot, M.L. Bortolin, T. Kiss, Site-specific pseudouridine formation in preribosomal RNA is guided by small nucleolar RNAs, Cell. 89 (1997) 799–809.
[42] P.A. Genest, L. Baugh, A. Taipale, W. Zhao, S. Jan, H.G.A.M. van Luenen, et al., Defining the sequence requirements for the positioning of base J in DNA using SMRT sequencing, Nucleic Acids Research. (2015) gkv095.
[43] W. Gibson, M. Bailey, The development of Trypanosoma brucei within the tsetse fly midgut observed using green fluorescent trypanosomes, Kinetoplastid Biol Dis. 2 (2003) 1.
[44] W. Gibson, L. Peacock, V. Ferris, K. Williams, M. Bailey, The use of yellow fluorescent hybrids to indicate mating in Trypanosoma brucei, Parasit Vectors. 1 (2008) 4.
[45] D.T. Gilbert, G. King, S. Pettigrew, T.D. Wilson, Comment on "Estimating the reproducibility of psychological science," (2016) 1–3.
[46] G. Gilinger, V. Bellofatto, Trypanosome spliced leader RNA genes contain the first identified RNA polymerase II gene promoter in these organisms, Nucleic Acids Research. 29 (2001) 1556–1564.
[47] R.C. Gupta, B.A. Roe, K. Randerath, The nucleotide sequence of human tRNAGly (anticodon GCC), Nucleic Acids Research. 7 (1979) 959–970.
[48] K.D. Hansen, S.E. Brenner, S. Dudoit, Biases in Illumina transcriptome sequencing caused by random hexamer priming, Nucleic Acids Research. 38

(2010) e131–e131.
[49] R. Hauenschild, L. Tserovski, K. Schmid, K. Thüring, M.-L. Winz, S. Sharma, et al., The reverse transcription signature of N-1-methyladenosine in RNA-Seq is sequence dependent, Nucleic Acids Research. 43 (2015) 9950–9964.
[50] A. Hodgkinson, Y. Idaghdour, E. Gbeha, J.-C. Grenier, E. Hip-Ki, V. Bruat, et al., High-resolution genomic analysis of human mitochondrial RNA sequence variation, Science. 344 (2014) 413–415.
[51] J.P. Holdren, Increasing Access to the Results of Federally Funded Scientific Research, (2013) 1–6.
[52] R.W. HOLLEY, G.A. EVERETT, J.T. MADISON, A. ZAMIR, NUCLEOTIDE SEQUENCES IN THE YEAST ALANINE TRANSFER RIBONUCLEIC ACID, J. Biol. Chem. 240 (1965) 2122–2128.
[53] C. Huang, J. Karijolich, Y.-T. Yu, Post-transcriptional Modification of RNAs by Artificial Box H/ACA and Box C/D RNPs, in: Methods in Molecular Biology, Humana Press, Totowa, NJ, 2011: pp. 227–244.
[54] A. Hüttenhofer, M. Kiefmann, S. Meier-Ewert, J. O'Brien, H. Lehrach, J.P. Bachellerie, et al., RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse, The EMBO Journal. 20 (2001) 2943–2953.
[55] D. Incarnato, F. Anselmi, E. Morandi, F. Neri, M. Maldotti, S. Rapelli, et al., High-throughput single-base resolution mapping of RNA 2'-O-methylated residues, Nucleic Acids Research. (2016).
[56] K. Jack, C. Bellodi, D.M. Landry, R.O. Niederer, A. Meskauskas, S. Musalgaonkar, et al., rRNA pseudouridylation defects affect ribosomal ligand binding and translational fidelity from yeast to human cells, Molecular Cell. 44 (2011) 660–666.
[57] J.E. Jackman, J.D. Alfonzo, Transfer RNA modifications: nature's combinatorial chemistry playground, WIREs RNA. 4 (2013) 35–48.
[58] S.A. James, M.J.T. O'Kelly, D.M. Carter, R.P. Davey, A. van Oudenaarden, I.N. Roberts, Repetitive sequence variation and dynamics in the ribosomal DNA array of Saccharomyces cerevisiae as revealed by whole-genome resequencing, Genome Res. 19 (2009) 626–635.
[59] C.J. Janzen, F. van Deursen, H. Shi, G.A.M. Cross, K.R. Matthews, E. Ullu, Expression site silencing and life-cycle progression appear normal in Argonaute1-deficient Trypanosoma brucei, Mol. Biochem. Parasitol. 149 (2006) 102–107.
[60] S. Kabani, K. Fenn, A. Ross, A. Ivens, T.K. Smith, P. Ghazal, et al., Genome-wide expression profiling of in vivo-derived bloodstream parasite stages and dynamic analysis of mRNA alterations during synchronous differentiation in Trypanosoma brucei, BMC Genomics. 10 (2009) 427.
[61] J. Karijolich, Y.-T. Yu, Converting nonsense codons into sense codons by targeted pseudouridylation, Nature. 474 (2011) 395–398.
[62] K. Karikó, H. Muramatsu, F.A. Welsh, J. Ludwig, H. Kato, S. Akira, et al., Incorporation of Pseudouridine Into mRNA Yields Superior Nonimmunogenic Vector With Increased Translational Capacity and Biological Stability, Mol Ther. 16 (2008) 1833–1840.

[63] Z. Kiss-László, Y. Henry, J.P. Bachellerie, M. Caizergues-Ferrer, T. Kiss, Site-specific ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs, Cell. 85 (1996) 1077–1088.
[64] J. Korlach, C. He, T.A. Clark, L. Zhang, X. Lu, **Identification of 5-methyl-C in nucleic acid templates**, US20140004511, 2014.
[65] F. Krueger, S.R. Andrews, C.S. Osborne, Large scale loss of data in low-diversity illumina sequencing libraries can be recovered by deferred cluster calling, PLoS ONE. 6 (2011) e16607.
[66] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, Nat. Methods. 9 (2012) 357–359.
[67] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, Genome Biol. 10 (2009) R25.
[68] S. Li, C.E. Mason, The pivotal regulatory landscape of RNA modifications, Annu Rev Genomics Hum Genet. 15 (2014) 127–150.
[69] X. Li, P. Zhu, S. Ma, J. Song, J. Bai, F. Sun, et al., Chemical pulldown reveals dynamic pseudouridylation of the mammalian transcriptome, Nature Chemical Biology. 11 (2015) 592–597.
[70] X.-H. Liang, Q. Liu, M.J. Fournier, Loss of rRNA modifications in the decoding center of the ribosome impairs translation and strongly delays pre-rRNA processing, Rna. 15 (2009) 1716–1728.
[71] X.-H. Liang, S. Uliel, A. Hury, S. Barth, T. Doniger, R. Unger, et al., A genome-wide analysis of C/D and H/ACA-like small nucleolar RNAs in Trypanosoma brucei reveals a trypanosome-specific pattern of rRNA modification, Rna. 11 (2005) 619–645.
[72] X.-H. Liang, Y.-X. Xu, S. Michaeli, The spliced leader-associated RNA is a trypanosome-specific sn(o) RNA that has the potential to guide pseudouridine formation on the SL RNA, Rna. 8 (2002) 237–246.
[73] X.H. Liang, L. Liu, S. Michaeli, Identification of the first trypanosome H/ACA RNA that guides pseudouridine formation on rRNA, J. Biol. Chem. 276 (2001) 40313–40318.
[74] P.A. Limbach, M.J. Paulines, Going global: the new era of mapping modifications in RNA, WIREs RNA. (2016) 1–17.
[75] N. Liu, M. Parisien, Q. Dai, G. Zheng, C. He, T. Pan, Probing N6-methyladenosine RNA modification status at single nucleotide resolution in mRNA and long noncoding RNA, Rna. 19 (2013) 1848–1856.
[76] A.F. Lovejoy, D.P. Riordan, P.O. Brown, Transcriptome-wide mapping of pseudouridines: pseudouridine synthases modify specific mRNAs in S. cerevisiae, PLoS ONE. 9 (2014) e110799.
[77] C. Luo, D. Tsementzi, N. Kyrpides, T. Read, K.T. Konstantinidis, Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample, PLoS ONE. 7 (2012) e30087.
[78] X. Ma, X. Zhao, Y.-T. Yu, Pseudouridylation (Psi) of U2 snRNA in S. cerevisiae is catalyzed by an RNA-independent mechanism, The EMBO Journal. 22 (2003) 1889–1897.
[79] M.A. Machnicka, K. Milanowska, O. Osman Oglou, E. Purta, M. Kurkowska,

A. Olchowik, et al., MODOMICS: a database of RNA modification pathways--2013 update, Nucleic Acids Research. 41 (2013) D262–7.
[80] B.E. Maden, Mapping 2'-O-methyl groups in ribosomal RNA, Methods. 25 (2001) 374–382.
[81] B.E. Maden, M.E. Corbett, P.A. Heeney, K. Pugh, P.M. Ajuh, Classical and novel approaches to the detection and localization of the numerous modified nucleotides in eukaryotic ribosomal RNA, Biochimie. 77 (1995) 22–29.
[82] M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads, EMBnet Journal. 17 (2011) 10–12.
[83] K.R. Matthews, The developmental cell biology of Trypanosoma brucei, Journal of Cell Science. 118 (2005) 283–290.
[84] K.R. Matthews, C. Tschudi, E. Ullu, A common pyrimidine-rich motif governs trans-splicing and polyadenylation of tubulin polycistronic pre-mRNA in trypanosomes, 8 (1994) 491–501.
[85] M. Mayho, K. Fenn, P. Craddy, S. Crosthwaite, K. Matthews, Post-transcriptional control of nuclear-encoded cytochrome oxidase subunits in Trypanosoma brucei: evidence for genome-wide conservation of life-cycle stage-specific regulatory elements, Nucleic Acids Research. 34 (2006) 5312–5324.
[86] E.C. McKiernan, P.E. Bourne, C.T. Brown, S. Buck, A. Kenall, J. Lin, et al., How open science helps researchers succeed, Elife. 5 (2016) 372.
[87] M. McNutt, K. Lehnert, B. Hanson, B.A. Nosek, A.M. Ellison, J.L. King, Liberating field science samples and data, Science. 351 (2016) 1024–1026.
[88] U.T. Meier, Pseudouridylation goes regulatory, The EMBO Journal. 30 (2011) 3–4.
[89] S. Michaeli, T. Doniger, S.K. Gupta, O. Wurtzel, M. Romano, D. Visnovezky, et al., RNA-seq analysis of small RNPs in Trypanosoma brucei reveals a rich repertoire of non-coding RNAs, Nucleic Acids Research. 40 (2012) 1282–1298.
[90] M.I. Newby, N.L. Greenbaum, Investigation of Overhauser effects between pseudouridine and water protons in RNA helices, Proc. Natl. Acad. Sci. U.S.a. 99 (2002) 12697–12702.
[91] J. Ni, A.L. Tien, M.J. Fournier, Small nucleolar RNAs direct site-specific synthesis of pseudouridine in ribosomal RNA, Cell. 89 (1997) 565–573.
[92] D. Nilsson, K. Gunasekera, J. Mani, M. Osteras, L. Farinelli, L. Baerlocher, et al., Spliced leader trapping reveals widespread alternative splicing patterns in the highly dynamic transcriptome of Trypanosoma brucei, PLoS Pathog. 6 (2010) e1001037.
[93] D.P. Nolan, S. Rolin, J.R. Rodriguez, J. Van Den Abbeele, E. Pays, Slender and stumpy bloodstream forms of Trypanosoma brucei display a differential response to extracellular acidic and proteolytic stress, Eur. J. Biochem. 267 (2000) 18–27.
[94] B.A. Nosek, G. Alter, G.C. Banks, D. Borsboom, S.D. Bowman, S.J. Breckler, et al., Promoting an open research culture, Science. 348 (2015) 1422–1425.
[95] Open Science Collaboration, Estimating the reproducibility of psychological science, Science. 349 (2015) aac4716.

[96] M. Parisien, C. Yi, T. Pan, Rationalization and prediction of selective decoding of pseudouridine-modified nonsense and sense codons, Rna. 18 (2012) 355–367.
[97] K.G. Patteson, L.P. Rodicio, P.A. Limbach, Identification of the mass-silent post-transcriptionally modified nucleoside pseudouridine in RNA by matrix-assisted laser desorption/ionization mass spectrometry, Nucleic Acids Research. 29 (2001) 1–7.
[98] L.M. Powell, S.C. Wallis, R.J. Pease, Y.H. Edwards, T.J. Knott, J. Scott, A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine, Cell. 50 (1987) 831–840.
[99] A.R. Quinlan, I.M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features, Bioinformatics. 26 (2010) 841–842.
[100] B. Reuner, E. Vassella, B. Yutzy, M. Boshart, Cell density triggers slender to stumpy differentiation of Trypanosoma brucei bloodstream forms in culture, Mol. Biochem. Parasitol. 90 (1997) 269–280.
[101] D. Reynolds, L. Cliffe, K.U. Förstner, C.-C. Hon, T.N. Siegel, R. Sabatini, Regulation of transcription termination by glucosylated hydroxymethyluracil, base J, in Leishmania major and Trypanosoma brucei, Nucleic Acids Research. 42 (2014) 9717–9729.
[102] B.R. Rosenberg, C.E. Hamilton, M.M. Mwangi, S. Dewell, F.N. Papavasiliou, Transcriptome-wide sequencing reveals numerous APOBEC1 mRNA-editing targets in transcript 3' UTRs, Nat. Struct. Mol. Biol. 18 (2011) 230–236.
[103] G.M. Rubin, The nucleotide sequence of Saccharomyces cerevisiae 5.8 S ribosomal ribonucleic acid, J. Biol. Chem. 248 (1973) 3860–3875.
[104] P. Ryvkin, Y.Y. Leung, I.M. Silverman, M. Childress, O. Valladares, I. Dragomir, et al., HAMR: high-throughput annotation of modified ribonucleotides, Rna. 19 (2013) 1684–1692.
[105] P. Schattner, S. Barberan-Soler, T.M. Lowe, A computational screen for mammalian pseudouridylation guide H/ACA RNAs, Rna. 12 (2006) 15–25.
[106] D. Schulz, M. Zaringhalam, F.N. Papavasiliou, H.-S. Kim, Base J and H3.V Regulate Transcriptional Termination in Trypanosoma brucei, PLoS Genet. 12 (2016) e1005762.
[107] S. Schwartz, D.A. Bernstein, M.R. Mumbach, M. Jovanovic, R.H. Herbst, B.X. León-Ricardo, et al., Transcriptome-wide Mapping Reveals Widespread Dynamic-Regulated Pseudouridylation of ncRNA and mRNA, Cell. 159 (2014) 148–162.
[108] S. Schwartz, D.A. Bernstein, M.R. Mumbach, M. Jovanovic, R.H. Herbst, B.X. León-Ricardo, et al., Transcriptome-wide Mapping Reveals Widespread Dynamic-Regulated Pseudouridylation of ncRNA and mRNA, Cell. 159 (2014) 148–162.
[109] K. Shiroguchi, T.Z. Jia, P.A. Sims, X.S. Xie, Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes, Proc. Natl. Acad. Sci. U.S.a. 109 (2012) 1347–1352.
[110] B.S. Sibert, J.R. Patton, Pseudouridine synthase 1: a site-specific synthase without strict sequence recognition requirements, Nucleic Acids Research. 40 (2012) 2107–2118.

[111] T.N. Siegel, D.R. Hekstra, L.E. Kemp, L.M. Figueiredo, J.E. Lowell, D. Fenyo, et al., Four histone variants mark the boundaries of polycistronic transcription units in Trypanosoma brucei, 23 (2009) 1063–1076.
[112] F. Spenkuch, Y. Motorin, M. Helm, Pseudouridine: Still mysterious, but never a fake (uridine)! RNA Biology. 11 (2015) 1540–1554.
[113] Y. Tanaka, T.A. Dyer, G.G. Brownlee, An improved direct RNA sequence method; its application to Vicia faba 5.8S ribosomal RNA, Nucleic Acids Research. 8 (1980) 1259–1272.
[114] M. Taoka, Y. Nobe, M. Hori, A. Takeuchi, S. Masaki, Y. Yamauchi, et al., A mass spectrometry-based method for comprehensive quantitative determination of post-transcriptional RNA modifications: the complete chemical structure of Schizosaccharomyces pombe ribosomal RNAs, Nucleic Acids Research. 43 (2015) e115–e115.
[115] K. Tomita, T. Ueda, K. Watanabe, The presence of pseudouridine in the anticodon alters the genetic code: a possible mechanism for assignment of the AAA lysine codon as asparagine in echinoderm mitochondria, Nucleic Acids Research. 27 (1999) 1683–1689.
[116] S. Trindade, F. Rijo-Ferreira, T. Carvalho, D. Pinto-Neves, F. Guegan, F. Aresta-Branco, et al., Trypanosoma brucei Parasites Occupy and Functionally Adapt to the Adipose Tissue in Mice, Cell Host Microbe. 19 (2016) 837–848.
[117] E. Ullu, K.R. Matthews, C. Tschudi, Temporal order of RNA-processing reactions in trypanosomes: rapid trans splicing precedes polyadenylation of newly synthesized tubulin transcripts, Molecular and Cellular Biology. 13 (1993) 720–725.
[118] H.G.A.M. van Luenen, C. Farris, S. Jan, P.-A. Genest, P. Tripathi, A. Velds, et al., Glucosylated hydroxymethyluracil, DNA base J, prevents transcriptional readthrough in Leishmania, Cell. 150 (2012) 909–921.
[119] J.-J. Vasquez, C.-C. Hon, J.T. Vanselow, A. Schlosser, T.N. Siegel, Comparative ribosome profiling reveals extensive translational complexity in different Trypanosoma brucei life cycle stages, Nucleic Acids Research. 42 (2014) 3623–3637.
[120] E. Vassella, R. Braun, I. Roditi, Control of polyadenylation and alternative splicing of transcripts from adjacent genes in a procyclin expression site: a dual role for polypyrimidine tracts in trypanosomes? Nucleic Acids Research. 22 (1994) 1359–1364.
[121] V.E. Velculescu, L. Zhang, W. Zhou, J. Vogelstein, M.A. Basrai, D.E. Bassett, et al., Characterization of the yeast transcriptome, Cell. 88 (1997) 243–251.
[122] J. Venema, D. Tollervey, Ribosome Synthesis in Saccharomyces cerevisiae, Http://Dx.Doi.org/10.1146/Annurev.Genet.33.1.261. 33 (2003) 1–51.
[123] I.D. Vilfan, Y.-C. Tsai, T.A. Clark, J. Wegener, Q. Dai, C. Yi, et al., Analysis of RNA base modification and structural rearrangement by single-molecule real-time detection of reverse transcription, J Nanobiotechnology. 11 (2013) 8.
[124] G. Wu, M.K. Radwan, M. Xiao, H. Adachi, J. Fan, Y.-T. Yu, The TOR signaling pathway regulates starvation-induced pseudouridylation of yeast U2 snRNA, Rna. 22 (2016) 1146–1152.
[125] G. Wu, M. Xiao, C. Yang, Y.-T. Yu, U2 snRNA is inducibly pseudouridylated

at novel sites by Pus7p and snR81 RNP, The EMBO Journal. 30 (2010) 79–89.
[126] G. Wu, A.T. Yu, A. Kantartzis, Y.-T. Yu, Functions and mechanisms of spliceosomal small nuclear RNA pseudouridylation, WIREs RNA. 2 (2011) 571–581.
[127] C. Yang, D.S. McPheeters, Y.-T. Yu, Psi35 in the branch site recognition region of U2 small nuclear RNA is important for pre-mRNA splicing in Saccharomyces cerevisiae, J. Biol. Chem. 280 (2005) 6655–6662.
[128] J.-H. Yang, X.-C. Zhang, Z.-P. Huang, H. Zhou, M.-B. Huang, S. Zhang, et al., snoSeeker: an advanced computational package for screening of guide and orphan snoRNA genes in the human genome, Nucleic Acids Research. 34 (2006) 5112–5123.
[129] Y.T. Yu, M.D. Shu, J.A. Steitz, A new method for detecting sites of 2'-O-methylation in RNA molecules, Rna. 3 (1997) 324–331.
[130] Y.T. Yu, M.D. Shu, J.A. Steitz, Modifications of U2 snRNA are required for snRNP assembly and pre-mRNA splicing, The EMBO Journal. 17 (1998) 5783–5795.
[131] M. Zaringhalam, F.N. Papavasiliou, Pseudouridylation meets next-generation sequencing, Methods. (2016).