The Dissertation Committee for Jeffrey Alan Hussmann
certifies that this is the approved version of the following dissertation:

Expanding the Applications of High-throughput DNA Sequencing

Committee:

William H. Press, Supervisor

Sara Sawyer, Co-supervisor

Inderjit Dhillon

Ron Elber

Oscar Gonzalez

Edward Marcotte

Expanding the Applications of High-throughput DNA Sequencing

by

Jeffrey Alan Hussmann, B.S.E., M.S.C.A.M.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2015

Dedicated to Mom and Dad.

Acknowledgments

First and foremost, I thank my parents, Glen and Cathie Hussmann, for their love and support throughout my education. I would like to thank my advisor, Bill Press, for introducing me to the world of computational biology and for providing guidance and wisdom as I explored it. I would also like to thank my co-advisor, Sara Sawyer, for invaluable mentorship, perspective, and opportunities. I thank Edward Marcotte for letting me be a stray member of his lab and absorb the culture of experimental biology by osmosis. I would also like to thank Inderjit Dhillon, Ron Elber, and Oscar Gonzalez for their service on my committee and for their advice and insights. I have been fortunate to work with great experimental collaborators at UT, particularly Dianne Lou, Sandie Shan, Ross McBee, Stephanie Patchett, and Arlen Johnson. Special thanks to Sandie for her patience teaching a computer scientist how to pipette. I also thank Premal Shah and Joshua Plotkin for many productive conversations about ribosomes, and Ashely Acevedo and Raul Andino, our collaborators on circle sequencing. I would like to thank all of the members of the Press lab, but most especially John Hawkins, for many helpful conversations. Thanks to all the members of the Sawyer lab - Alex, Maryska, Nick, Paul, Scott, and Ann - and to all other friends for making my time in Austin so enjoyable and memorable.

Expanding the Applications of High-throughput DNA Sequencing

Publication No. _____

Jeffrey Alan Hussmann, Ph.D.
The University of Texas at Austin, 2015

Supervisors: William H. Press
Sara Sawyer

DNA sequencing is the process of determining the identities of the nucleotides that make up a molecule of DNA. The rapid pace of advancements in sequencing technologies in recent years have made it possible to simultaneously determine the sequences of hundreds of millions of short DNA fragments. The ability to perform sequencing with such high throughput has revolutionized the study of biological systems, but the types of questions that can be answered through sequencing-based experiments can be limited by the presence of different kinds of noise and biases in these experiments.

One class of applications of high-throughput sequencing involves identifying genetic variation, such as finding rare mutations in the genomes of cancerous cells. In these applications, the sensitivity with which rare genetic variants can be detected is limited by the relatively high rate with which

current DNA sequencing technologies incorrectly identify nucleotides. In the first half of this thesis, we present a method for dramatically reducing the rate at which these incorrect identifications occur. Our method, called circle sequencing, creates redundant copies of the sequence of each input molecule of DNA. This is accomplished by circularizing each DNA fragment and performing rolling circle amplification on these circles with a strand-displacing polymerase. The resulting products consist of several physically linked copies of the original sequence in each fragment. When these products are sequenced, this informational redundancy protects against random errors introduced during sequencing, allowing for highly accurate recovery of the original sequence of each input molecule. By eliminating the vast majority of incorrectly identified nucleotides from the resulting data, our method enables the sensitive detection of rare variants and opens up exciting new questions involving such variants to direct measurement by sequencing.

An entirely different application of high-throughput sequencing is to selectively capture and sequence stretches of DNA or RNA that are participating in a process of interest within a cell. The accuracy of quantitative inferences made by this type of experiment can be severely impacted, however, by biases introduced during the experimental manipulations used to isolate biologically relevant fragments of DNA from cells. Ribosome profiling is an experimental technique that consists of sequencing short stretches of messenger RNAs that are protected from nuclease digestion by the presence of a bound ribosome. The resulting data represents millions of snapshots of the locations of actively

translating ribosomes. In theory, these snapshots can be used to determine how long ribosomes take to translate each type of codon by quantifying how often ribosomes are observed positioned over that codon. In practice, different studies in yeast attempting to do this have reached contradictory and counterintuitive conclusions. In the second half of this thesis, we perform a large-scale comparative analysis of data from many different ribosome profiling experiments in order to resolve these contradictions. We identify a previously unappreciated source of systematic bias in a subset of these experiments. This bias prevents these experiments from accurately measuring ribosomes in proportion to how long they spend at each position *in vivo*. Understanding this bias provides insight into the true signatures of translation dynamics in yeast and offers important guidance for the future design and interpretation of sequencing-based approaches to measuring these dynamics.

Table of Contents

List of Figures

Chapter 1

Introduction

The genetic information of every living organism is stored in the precise sequence of nucleotide identities in the DNA of its genome. Over the last forty years, determining the sequence of bases that make up a molecule of DNA has become one of the fundamental tools of experimental biology. In the last decade, a rapid series of technological advances have made it possible to carry out this process massively in parallel in order to sequence incredibly large numbers of molecules of DNA simultaneously [7, 27, 75, 95]. Thanks to the advent of these high-throughput sequencing technologies, it is now a matter of routine to sequence billions of bases from hundreds of millions of short stretches of DNA.

To leverage this technological progress, a enormous array of clever biochemical methods have been developed for capturing information about biological processes in the form of libraries of short DNA fragments [56, 105]. Computational and mathematical tools can then be used to extract information from the massive amounts of data generated by sequencing these fragments. Many interesting questions remain outside the reach of this paradigm, however. The types of signals that can be accurately measured by high-throughput sequenc-

ing are limited in many cases by the presence of errors and biases introduced by both the sequencing technologies themselves and by the biochemical manipulations used to construct DNA fragment libraries. To attack these limitations, we can take advantage of the fact that the digital nature and enormous scale of sequencing data are fundamentally different from the diagnostic information that has historically been available in molecular biology. Deep analysis of sequencing data can provide unprecedented mechanistic insight into what is actually happening during experimental manipulations of DNA. These insights can then be used to understand and eliminate confounding signals, either by computationally correcting for sources of noise and bias or by informing the design of new experimental protocols. The theme of this thesis is to identify and push back at the limitations of existing high-throughput sequencing technologies and experimental designs in order to accurately measure new kinds of biological signals.

Sensitive detection of genetic variants

In the first half of the thesis, we explore methods for dramatically reducing the base-calling error rate of sequencing in order to allow the sensitive detection of rare genetic variants. Many important open questions in biomedical research involve determining exactly which nucleotides in one copy of the genome of an organism are different from those in other nearly-identical copies. Examples of such questions include measuring how quickly mutations accumulate in different cells over the course of cell divisions[74], identifying

the spectrum of mutations present in genetically heterogeneous tumors [15], cataloging emerging variation in rapidly evolving populations of viruses [1], identifying the different antibody sequences produced by an adaptive immune system [103], or quantifying the relationship between the received dosage of a mutagen such as ionizing radiation and the rate of induced mutations [36]. In theory, high-throughput sequencing could be a powerful tool to search for rare genetic variants on a genome-wide scale, but the ability to sensitively detect rare variants is hampered by the relatively high base-calling error rate of current sequencing technologies. Because bases that are misidentified during sequencing are indistinguishable from any true genetic variants that may be present, the base-calling error rate imposes a lower bound on the frequency of variants that can be detected without being overwhelmed by false positives. The types of base-calling errors made and the rates at which these errors occur vary considerably between different sequencing technologies [72], but Illumina sequencing technologies currently offer the lowest rate of substitution errors, with measured error rates ranging from 0.05% to 1% [55]. For many applications, variants of interest are expected to be present at frequencies orders of magnitude lower than this. This means that any naive attempt to use high-throughput sequencing to search for rare variants is doomed to fail: true biologically relevant sequence variants will be needles lost in a haystack of artifactual apparent variants caused by sequencing errors.

In order to enable the use of high-throughput sequencing to answer questions that involve identifying rare variants, therefore, the effective error

rate of base calling therefore needs to be dramatically improved. In chapter 2, we describe the development of a method called circle sequencing for accomplishing this. Our method takes as input a library of short DNA fragments and encodes the sequence information of each fragment into a simple error-correcting code by circularizing single-stranded DNA templates and performing rolling circle amplification on each template with a strand-displacing polymerase. The resulting products consist of concatamers of several physically linked copies of the original sequence in each input template. After sequencing these specially-constructed products, the different redundant copies of information contained in each sequencing read can be compared to each other. Random errors introduced by the sequencing process can be identified and corrected, allowing for highly accurate base calling. Exploiting this redundancy presents several novel computational challenges. In order to identify the structure of the repeated information in the sequence of each concatamer, we developed software tools to efficiently compute discrete auto- and cross-correlations and to perform rotation-insensitive mapping of inferred consensus sequences to a reference genome. The combination of our experimental design and computational processing achieves a dramatic reduction in sequencing error rates and has both theoretical and practical cost-efficiency advantages over alternative error-correction strategies.

Accurate measurement of translation dynamics

In the second half of the thesis, we explore the use of high-throughput sequencing to accurately measure the dynamics of translation. Translation is the process by which proteins are assembled based on the instructions provided by the sequence of codons in a messenger RNA. Ribosomes carry out the conversion of information from codons into amino acids through the sequential binding of tRNAs according to the genetic code, the mapping of codons into amino acids [107]. Because this mapping is not one-to-one, a particular amino acid may be encoded by one of several synonymous codons. Although the choice of synonymous codon used to encode an amino acid does not change the composition of the protein produced, synonymous codons are not used with equal frequencies in genomes. In many organisms, synonymous codons corresponding to more abundant tRNAs are known as preferred or optimal codons because of a tendency for such codons to be used more often in highly expressed genes [100]. This tendency implies that the use of optimal codons provides a selective advantage to organisms. A large body of theoretical work hypothesizes that differences in the speed with which each type of codon is translated by ribosomes provides the mechanism for this advantage[88], but the ability to test these hypotheses experimentally has lagged behind the theory.

Chapter 0 serves as a motivating example for why accurate genome-wide measurements of *in vivo* translation dynamics are necessary. Recent theoretical work [13] has observed that pairs of occurrences of the same amino acid that are nearby each other in a coding sequence tend to use the same codon

more often than expected given the genome-wide frequencies with which each codon is individually used. This genomic signature has been interpreted as evidence that the second amino acid in such a pair of occurrences is translated more quickly if it uses the same codon as the first because decoding of second occurrence can be performed by the same tRNA molecule as the first, a proposed mechanism called tRNA recycling. When we examine the evidence for this hypothesis more closely, however, we find that these statistical signatures cannot be taken as specific support for this novel proposed mechanism. Instead, by straightforward mathematical arguments involving Jensen's inequality, we show that the apparent excess use of nearby identical codon pairs is an inevitable consequence of any non-uniformity in codon usage across a genome. We identify a simple negative control to test if pressure to exploit tRNA recycling contributes to these signals in excess of the contributions from generic variation in codon preferences. When we compute this control, we find no evidence for any such contribution, and conclude that the observed patterns in codon usage do not by themselves support a substantial role for tRNA recycling in translation dynamics.

This debate is just one of many theoretical questions involving the connection between translation speed and selection on synonymous codon usage that could be answered if it were possible to measure how long ribosomes actually spend translating every codon across all of an organism's coding sequences *in vivo* [88]. A recently developed sequencing-based experimental technique called ribosome profiling has the potential to produce measurements

of this kind. Ribosome profiling consists of selectively sequencing short ∼28 nucleotide regions of mRNAs that are protected from nuclease digestion by the presence of a single bound ribosome. The resulting data consists of measurements the locations of millions of ribosomes at a snapshot in time with single-nucleotide resolution. This method was initially developed by Ingolia et al. [46] in 2009 and has been applied to study a wide variety of different aspects of translation in different organisms by many different groups since then [44].

Viewed from a high level, ribosome profiling is an example of a large class of sequencing-based approaches to studying the diverse set of cellular processes that DNA in a genome and RNA derived from this DNA participate in. The general form of these experiments is to use a clever series of biochemical manipulations to capture fragments of sequence information that are participating in a process of interest within cells. Massively parallel sequencing is used to identify all of the fragments that were captured, and statistical analysis of how often different sequences appear in the resulting data can then be used to make quantitative inferences about the process of interest. A vast array of of experimental schemes of this type have been developed, including methods for measuring levels of transcription [80], variation in transcript isoforms [84], chromatin structure [71], histone occupancy patterns [10], and a variety of interactions between DNA and proteins [8], amongst many others.

In order for the inferences produced by this paradigm to be accurate, the number of times each particular sequence fragment makes it through the

whole experimental pipeline must be representative of how often the sequence was actually involved in the process being studied. In practice, the experimental manipulations used can prefer some sequences over others for artifactual reasons that have nothing to do with the underlying biology, distorting the number of times each sequence appears in the final data. If these biases are severe enough and go unrecognized, quantitative inferences made using the biased data can produce misleading overall pictures of the process being studied [31, 49, 108]. Identifying these biases is a major challenge for the use of high-throughput sequencing to produce accurate insights into biological processes. Understanding the mechanistic causes of such biases can lead to ways to overcome them, either by modifying experimental designs to avoid them [102, 115] or by computationally correcting for their presence [83].

In chapter 4, we explore the ability of ribosome profiling to accurately measure how long ribosomes spend positioned over each codon during the process of translation in *Saccharomyces cerevisiae*. Through a comparative analysis of publicly available data from a large body of recent studies, as well as new data produced by our experimental collaborators, we identify a previously unappreciated source of bias in many ribosome profiling experiments that interferes with the accuracy of these measurements. Most of these experiments have used a chemical inhibitor of translation called cycloheximide to attempt to arrest ribosomes in place before measuring their locations. We present evidence that this process does not irreversibly arrest ribosomes in their steady-state distribution, but instead has the net effect of redistributing

ribosomes across coding sequences so that their locations do not reflect how much time they spend at each position *in vivo*. The bias that this previously-unappreciated behavior of cycloheximide introduces is particularly insidious - roughly speaking, it flips which codons appear to be fast and which appear to be slow, completely disrupting the general conclusions about translation that data produced by these experiments appear to support. By uncovering and characterizing this bias, we provide a principled resolution to the contradictory claims made by experiments performed with and without cycloheximide and clarify the general principles of elongation speeds in yeast. Understanding this bias establishes an important principle for the future design and interpretation of sequencing-based experiments that hope to accurately measure codon-resolution signatures of translation dynamics.

Chapter 2

High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing

2.1 Introduction

One of the simplest conceptual uses of high-throughput sequencing is to sequence genomic DNA from a population in order to characterize the genetic variation present. Because bases that are misidentified during sequencing are indistinguishable from true variants, however, the base-calling error rate of the sequencing technology used represents a lower bound on the frequency of variants that can be sensitively detected in this way. We describe here the development and performance characteristics of a new library preparation strategy called circle sequencing that allows for computational identification

This chapter is based in part on work reported in D. I. Lou*, J.A. Hussmann*, R. M. McBee, A. Acevedo, R. Andino, W. H. Press, and S. L. Sawyer, "High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing," *Proceedings of the National Academy of Sciences*, 110 (49), 19872–19877, 2013 (* co-first authors). JAH performed all computational analysis, DIL and RMM performed experiments, and all authors conceived and designed experiments.

and correction of sequencing errors. In this strategy, short DNA fragments in an input library are circularized and then copied multiple times in tandem by a so-called rolling circle amplification process. The resulting DNA molecules each consist of several physically linked copies of the original sequence in a particular input molecule. After sequencing these resulting molecules, we can computationally decompose each sequencing read produced into the redundant copies of information that it contains and form a consensus out of these copies. This informational redundancy protects against stochastic errors introduced by the sequencing process and allows for highly accurate recovery of the sequence of each starting molecule. The fact that copies of information are physically packaged into the same molecule before delivery to the sequencing machine gives this strategy important cost-efficiency advantages over alternative schemes for correcting sequencing errors.

We first give general background on the Illumina sequencing technologies used by current implementations of our method on and describe other existing library preparation strategies for correcting errors in high-throughput sequencing data. We then outline the experimental design of the circle sequencing library preparation process and describe the computational strategies used to analyze data produced by this process. To test the method, we apply it to sequence a nearly-genetically-identical population of yeast cells. We analyze the performance characteristics of the resulting data in terms of the error rates achieved and the cost-efficiency of the overall process, and we compare these characteristics to those of alternative error correction strate-

gies. After completing this main narrative, we loop back to highlight several unexpected features of circle sequencing data that revealed interesting and unexpected properties of the biochemical manipulations involved. We conclude with a basic proof-of-concept of a possible future direction of improvement to the experimental design.

2.2 Background

2.2.1 Illumina sequencing technologies

Illumina technologies are the dominant force in the current landscape of experimental applications of massively parallel sequencing, offering the highest throughput of sequencing base calls per machine run and the lowest cost per base sequenced of all the major sequencing platforms [72]. Because some features of the experimental designs and data analysis we present in this chapter rely on a detailed understanding of how the Illumina sequencing process works, we give a brief overview of this process here. For a more thorough description, see [7].

The mechanics of Illumina sequencing make it possible to sequence a stretch of bases from both ends of each input fragment of DNA, a process called paired-end sequencing. To prepare a library of DNA fragments for paired-end sequencing on an Illumina machine, special adapter sequences are ligated to each end. Two different identities of adapter sequences are used, each of which consists of a flow cell attachment sequence followed by a sequencing primer. The two adapter sequences are designed to share a short stretch of bases in

common at one of their ends. This means that when single-stranded copies of one of the sequences and the reverse complement of the other are synthesized and annealed to each other, the complementary portions will base-pair with each other, producing a partially-complementary y-shaped construct. When this construct is ligated on to both sides of double-stranded input fragments, this ensures that each strand of the fragment receives each of the adapter identities on exactly one of its ends.

The sequencing machine contain a flow cell surface that is covered in a lawn of many copies of two different short DNA oligonucleotides anchored to the surface. These oligonucleotides are complementary to the flow cell attachment portion of the adapter sequences that were ligated on to opposite ends of each molecule to be sequenced during the library preparation process. When adapter-ligated input DNA fragments are denatured and washed over this surface, the adapter sequences on their ends hybridize to the complementary anchors. A process called bridge amplification is then carried out that uses polymerases to grow each attached fragment into a cluster of around 1000 double stranded copies of the fragment, each of which is attached to the flow cell on both ends.

To determine the sequence of the first end of the fragment from which each cluster was grown, a restriction enzyme is washed over the cell that recognizes and cuts a specific sequence in one of the flow cell attachment sequence identities. The strand that is no longer anchored to the flow cell because of this cut is denatured and washed away. The clusters are now ready for the

first sequencing reaction, called the R1 read. A oligonucleotide complementary to the R1 sequencing primer region located just downstream of the flow cell attachment sequence that was cut is annealed. The sequencing process then consists of resynthesizing the strand that was washed away in controlled, single base cycles, starting from the sequencing primer and growing downwards towards the flow cell. This polymerization is performed in such a way that the identity of the base incorporated at each cycle can be measured. To do this, in each cycle, microfluidics are used to flow a mixture of specially modified forms all four nucleotides over the clusters. The modifications consists of the addition of reversible terminators with a nucleotide-identity-specific fluorescent label. Whichever base identity is complementary to the next base in the anchored fragments will be incorporated into every fragment by polymerases, and the terminating modifications to these nucleotide will temporarily block any further incorporations. The fluorescent labels in each just-incorporated nucleotide are then excited by lasers, and the fluorescent intensity emitted by each cluster at frequencies corresponding to each nucleotide-identity-specific label is captured by optics to determine which base was incorporated. The fluorescent labels on the set of nucleotides incorporated are theoretically identical across all copies of a sequence in a cluster, but stochastic over- or under-incorporation of nucleotides at each fragment inevitably leads the different copies in each clusters to drift out of phase with each other and fluorescent signals emitted by each cluster to become less clear-cut. Sophisticated base-calling software interprets these signals to produce a best guess as to which

base was incorporated in each cluster as well as a quality score attached to this base call that quantities how confident the machine is in its identification. The rate at which the wrong base is identified by this process depends on many factors, but typical estimates of this rate are around 0.1% [72]. After reading a single base from all clusters, a chemical is washed over the flow cell that cleaves the labelled terminators to permit further elongation, and the set of modified nucleotides are washed over the clusters again. This cycle of steps is carried out many times, with each repetition of the cycle reading another base in the sequence of the fragment in every cluster.

Once the targeted number of cycles in the R1 read have been carried out, the second strand in every fragment that was synthesized by the sequencing reaction is denatured and washed away. The ends of each fragment are then reannealed to the lawn of flow cell attachment oligos, and the second strand is resynthesized from this priming, restoring every fragment in the cluster to a state in which it is attached at both ends to the flow cell. A restriction enzyme that recognizes and cuts the opposite flow cell adapter sequence is then washed over the cell. The net result is that every cluster has been flipped relative to its orientation during the R1 read. The R2 sequencing primer is then washed over the clusters and cycles of the sequencing reaction are carried out to produce the R2 read of the other end of each cluster.

At the time of its initial development, Illumina's cluster generation and sequencing-by-synthesis paradigm was severely limited in the length of continuous reads that it could produce before base-calling accuracy degraded to

unacceptably low levels. Incremental improvements over time in the engineering of the polymerases used and the chemistry of the modified nucleotides and reaction conditions have steadily erased this limitation, however, with read lengths increasing from 35 bases in 2008 to 300 bases in late 2013.

2.2.2 Existing methods for error-correction in high-throughput sequencing

A natural strategy for protecting information when its needs to be transmitted through a noisy channel is to send multiple redundant copies of the information and then compare the different copies received to identify where errors have occurred. High-throughput sequencing technologies can be viewed as a noisy channel through which we are attempting to transmit information consisting of the sequence of each input molecule. A closely related class of methods for creating and sequencing redundant copies of each molecule in a library of short fragments of DNA have recently been developed by several groups [33, 48, 55]. We will collectively call these barcoding methods. The main idea of these methods is to create a large number of short, artificial stretches of DNA to be used as identifying barcodes. These barcodes are then randomly attached to the ends of each molecule in a library of DNA fragments that is to be sequenced (figure 2.1, step 1A). Polymerase chain reaction (PCR) is then used to exponentially amplify the randomly labeled library, producing a large number of physically distinct copies of each labeled starting molecule (figure 2.1, step 2A). This amplified library is then sequenced. In the set of sequencing reads produced by this process, reads that represent different

copies of the same original input molecule creating during the amplification process can be grouped together by matching the sequences of their random labels (grey boxes in 2.1, step 4A). We will call such a grouping of reads that are amplification products of the same starting molecule a 'read family'. After grouping together sequences in this way, the sequences in each member of a read family represent votes as to what the sequence of the original input molecule was. Stochastic errors introduced during PCR or sequencing (blue circles throughout figure 2.1A) are expected to be scattered randomly throughout sequencing reads (that is, to not align vertically across multiple reads in a grey box). True genetic variants that were present in an input molecule (red circles in 2.1A), on the other hand, should be present in each copy.

Schmitt et al.[96] recently developed a clever enhancement to this basic scheme, called 'duplex barcoding'. They noticed a useful property of the fact that Illumina sequencing primers are attached to each side of a double-stranded DNA molecule in the form of partially-complementary y-adapters. If the resulting construct is amplified via PCR, the set of double stranded molecules produced consist of two similar but distinguishable forms. Both forms have the same sequence in between the primers, but all downstream products of one of the strands will have the R1 primer on side and the R2 primer on the other, while all products of the other strand will have the placement of these two primers flipped. (See [96] for a diagram of this process.) In other words, attaching y-adapters and amplifying has the net effect of asymmetrically labelling PCR products derived from each of the two strands of the

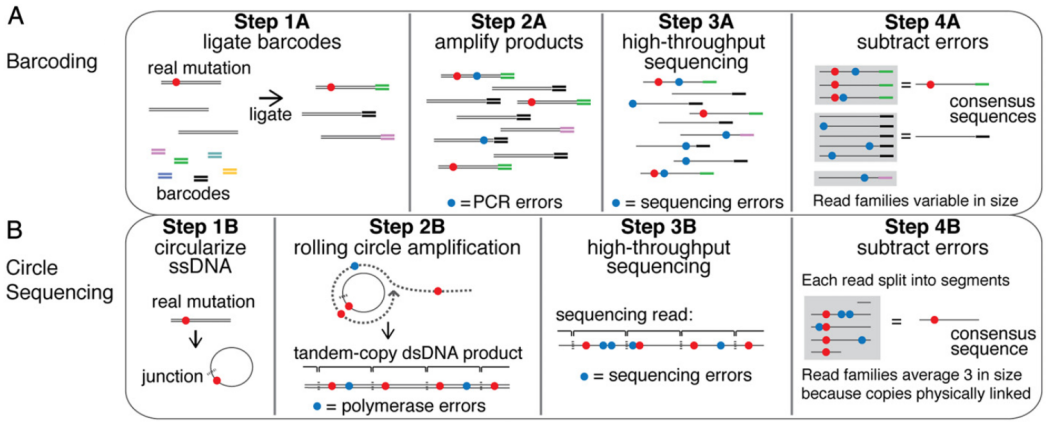

Figure 2.1: **Overview of barcoding methods and circle sequencing.**
(A) In barcoding methods, adapters containing randomized nucleotide regions (barcodes) are ligated to each molecule in the DNA sample (step 1A). The library is then amplified by PCR (step 2A). Products are sequenced (step 3A), individual reads containing the same barcode are grouped into read families (grey boxes), and consensus sequences are derived (step 4A). Errors generated during PCR amplification (step 2A, blue circles) and during the sequencing process (step 3A, blue circles) can be computationally identified.
(B) In circle sequencing, DNA is denatured and single-stranded DNA is circularized (step 1B). Random primers are annealed to circles, and Phi29 polymerase is used to perform rolling circle replication (step 2B). Products consisting of tandemly linked copies of the information in the circle are sequenced (step 3B). Each read (or paired-end read pair) is computationally split into the individual copies of the original circle (grey box) and used to generate a consensus sequence (step 4B).

starting molecule. To take advantage of this fact, Schmitt et al. modified standard Illumina y-adapters to include a stretch of randomized nucleotides that serves as a barcode. Performing paired end sequencing on a library prepared by attaching these adapters to both sides of input molecules and amplifying allows for the identification of a distinct read family that is independently derived from each strand of a particular input molecule. If read families from both strands can be recovered, they can be compared to each other, providing an additional level of informational redundancy. This extra redundancy protects against errors that could only affect families derived from one of the two strands at a time, such as single-stranded DNA damage to starting templates.

2.3 Results

2.3.1 Experimental design of circle sequencing

There are two major potential drawbacks to barcoding methods. The first drawback limits how much useful error-corrected sequence information can be produced for a given investment of sequencing resources. In order to correct errors, barcoding methods rely on forming read families by stochastically sampling multiple amplified copies of a given input molecule from a large pool of molecules. For both theoretical and practical reasons, this sampling process is inherently inefficient in the sense that many read families produced are either larger or smaller than they need to be. This limits the total amount of useful error-corrected data produced for a given investment of sequencing resources. (See section 2.3.5 below for a thorough discussion of this point.)

The second drawback limits how accurately the methods can recover the sequence information in input molecules. Barcoding methods have the undesirable property that any base-incorporation errors during the amplification process create error-containing fragments that are themselves used as templates during all future rounds of amplification. If such an error happens at an early cycle of PCR (a so-called jackpot error), a large fraction of the copies of a starting molecule can all contain the same incorrect base at the same position. The use of redundancy to protect sequence information assumes that error events are independent from each other, so that multiple rare error events are unlikely to strike multiple copies of the same piece of information. Because a single error event during amplification can change the base identity reported at the same position in multiple redundant copies of an input sequence, this represents an error process that redundancy can't protect against. (In the interest of completeness, we note that duplex barcoding offers protection against this second drawback, at the cost of increased vulnerability to the first.)

Both of the drawbacks stem from the fact that the redundant copies of each starting molecule are produced in physically distinct molecules. This necessitates inefficiently forming read families by randomly sampling from a large pool of molecules and allows for the possibility that members of a read family are copies of an error-containing amplification intermediate rather than direct copies of the input molecule. One of the dimensions on which sequencing technologies have made remarkable recent advances, however, is the length of

sequences that can be continuously read. Motivated by these rapid advances in read lengths, we reasoned that these drawbacks could be avoided if each set of copies of sequence information was packaged and delivered to the sequencing machine as a single molecule. Increased read lengths make this possible by allowing multiple copies of reasonably sized starting molecules to fit in a single sequencing read or paired-end read pair. In collaboration with Ashely Acevedo and Raul Andino, we developed a novel biochemical protocol that we call circle sequencing [73] for producing such physically linked copies. We provide a high-level description of the experimental design here; a more detailed technical description of the biochemical protocol can be found in [73].

Briefly, a library of short double-stranded DNA fragments is produced by randomly shearing genomic DNA or by producing amplicons with appropriately spaced primers. Fragments of approximately 100 base pairs are typically targeted. The two strands of each fragment are denatured into single-stranded DNA, and an enzyme called CircLigase is used to attach the two ends of each single-stranded DNA molecule to each other to form circular templates (figure 2.1, step 1B). A process called rolling circle amplification (RCA) is then carried out. Random hexamers of DNA are added that bind to the circular templates and form a short double-stranded stretch that can prime polymerization by the phi29 polymerase. The closed circular topology of each template means that such polymerization will eventually make its way around the circle and run into the double-stranded stretch of DNA that it recently synthesized. When it does, this polymerase has a special strand-displacing activity that allows it

to evict the recently-synthesized second strand to continue its way around the circle again (figure 2.1, step 2B, top). The net result of many repeated trips by the polymerase around the circular template is a single-stranded concatamer consisting of many tandem repeats of the information in the original fragment. Random hexamers can then bind to each single-stranded concatamer and prime polymerization of these linear templates by the polymerase, filling in the opposite strand to create double-stranded concatamers (figure 2.1, step 2B, bottom). The resulting products are typically many kilobases long. These products are then sheared to produce fragments of an appropriate length for compatibility with paired-end sequencing, typically around 1000 base pairs. Illumina sequencing adapters are ligated to the ends of these sheared fragments and paired-end sequencing is performed (figure 2.1, step 3B).

2.3.2 Detecting structure in circle sequencing reads

The data produced by paired-end sequencing of a circle-sequencing library consists of several million pairs of strings of the letters {A, G, T, C, N} representing the called identities of the bases for a continuous stretch on each end of a fragment (figure 2.2A). These string are of fixed length for a particular configuration of the sequencing machine; in our experiments, this length was either 150 or 250 characters. Each string of base calls is accompanied by a list of quality scores representing the confidence of the base-calling algorithm in its assignment of an identity at each position.

For each such read pair, our first goal is to identify the structure of

the repeats of information within and between the pair of sequences so that we can decompose the two sequences into the different copies of an original input molecule that they are made up of (figure 2.2B). Once this is done, we will be able to compare these different copies to each other to form a best consensus estimate of what the original sequence was (figure 2.2C and D). In the schematics of figure 2.2, grey tickmarks are drawn to mark each location in the read pair that corresponds to the junction of circularization in the original template, but in reality, no such direct markings of these locations exist. In the absence of such markings, we can infer the structure of repeats by finding periodicity in each of the reads and by aligning the two reads to each other.

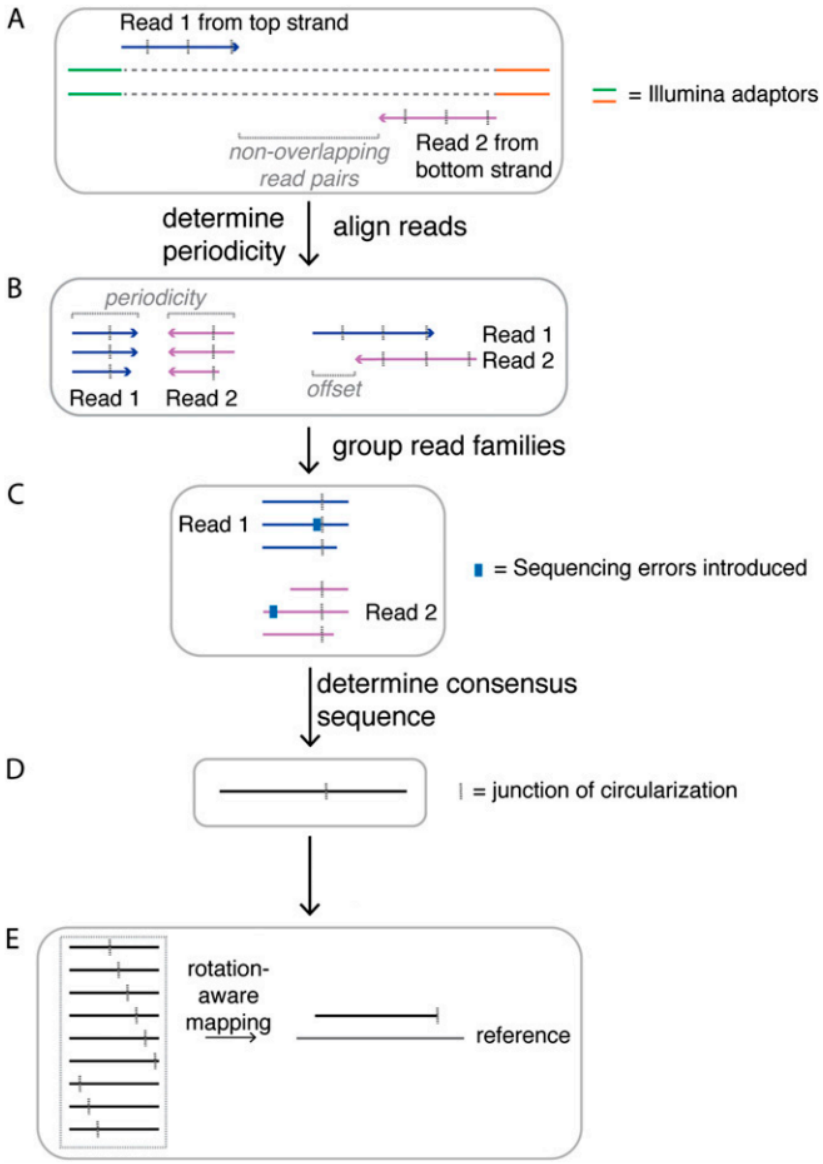

Figure 2.2: **Processing circle-sequencing read pairs.**

Figure 2.2 (Continued): **Processing circle-sequencing read pairs.**
A. Data consists of pairs of reads from opposite ends of a double-stranded DNA fragment of unknown length generated by rolling-circle amplification of a circular template of unknown length.
B. The length of the original circular template is inferred by computing the discrete autocorrelation of each read. The offset into a repeat relative to the first read that the second read begins at is inferred by computing the discrete cross-correlation of the two reads.
C. These inferences allow the read pair to be decomposed into independent copies of the sequence information in the circular template. When these copies are lined up, each column represents a group of interrogations of the identity of a particular base in the circular template.
D. The set of probabilistic base calls in each column are aggregated into a consensus base call and confidence estimate for each base in the circular template.
E. The location in the resulting consensus sequences at which the circular ligation junction occurred is determined by mapping all rotations of the consensus sequence to a reference genome.

If we assume for now that there are no insertions or deletions introduced by the phi29 polymerase as it performs rolling circle amplification, the structure of repeats in a read pair is completely determined by two parameters: the length of the original circular template, and the offset into this length that the second read in the read pair begins at relative to the first read. If the circular template had length P, each sequence in the pair is expected to be periodic with period length P. The base calls made at every pair of positions within a read separated by distance P should therefore be identical unless a sequencing error or phi29 misincorporation has changed one of the bases. Inferring the true value of P from the read pair therefore consists of computing the discrete autocorrelation of each sequence in the read pair - that is, computing the fraction of pairs of base calls in a sequence separated by distance p that are identical for all values of p from a minimum physically reasonable circle size up to a detection limit where the number of eligible pairs is too small to reliably distinguish true periodicity from chance. Conceptually, for each sequence s, this consists of forming the upper half of a symmetric matrix M whose (i, j)th entry is 1 if $s[i] = s[j]$ and 0 otherwise (figure 2.3). For each value of p, the pth upper diagonal of this matrix consists of all comparisons of pairs of bases in the sequence separated by a distance of exactly p. The sum of the pth upper diagonal divided by its length is therefore the fraction of such pairs that are identical.

For a sequence of length n, explicit computation of the autocorrelation in this way requires $O(n^2)$ operations. It could, of course, be done in $O(n \log n)$

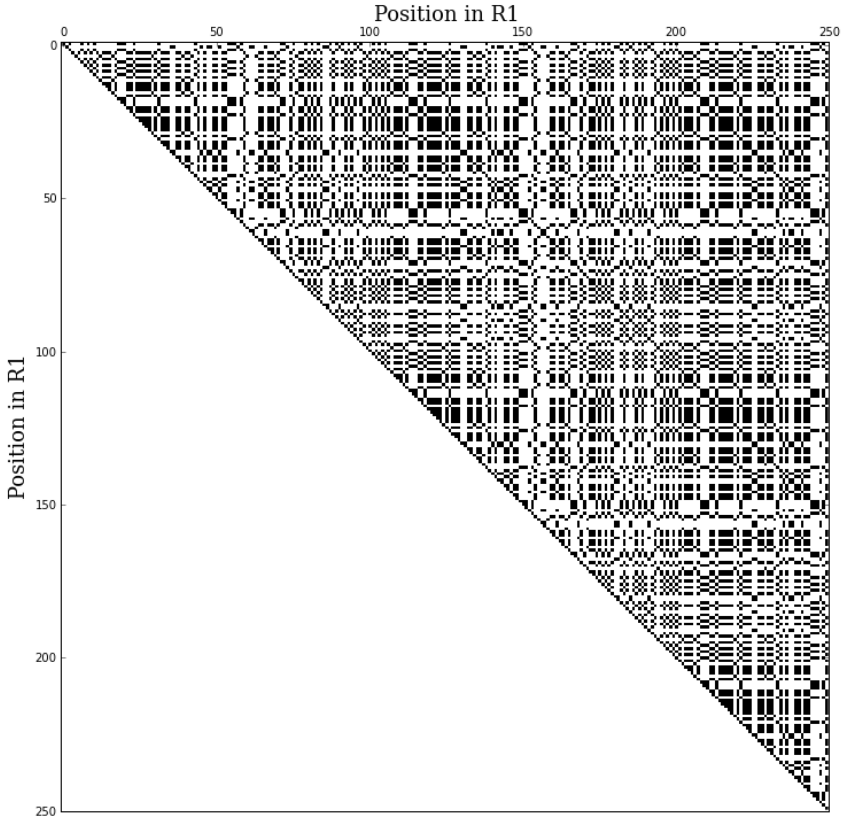

Figure 2.3: **Visualizing autocorrelation in circle sequencing reads** For each sequencing read in a read pair of a concatamer, the length of the original circular template is inferred by detecting periodicity in the read. Shown is the binary matrix of comparisons of base identities at positions i (row index) and j (column index) in such a sequence. A clear line of almost perfect identity on the 95th upper diagonal indicates that this sequence consists of repeating units 95 bases long.

operations with Fourier transforms via the Weiner-Khinchin theorem [91]. In practice, this does not offer any substantial performance improvements for the read lengths used so far, particularly since an optimized implementation of the explicit autocorrelation computation is not a bottleneck in the overall processing pipeline. The explicit $O(n^2)$ formulation is also easier to generalize to consider situations where the rolling circle amplification process has introduced an insertion or deletion, as discussed below.

In principle, the value of p that has the highest fraction of identical distance-p-separated pairs should be the inferred period length. In practice, any sequence that is periodic with period length p is, of course, also periodic with periodc length np for any integer n, and the chance positioning of sequencing errors may cause a multiple of the true circular template length to have a slightly higher fraction of identical pairs than the true length. To recover the true period length in these cases, all factors of the value of p with the highest identical fraction are reexamined, and the smallest such factor that also has a sufficiently high fraction of identical pairs is taken to be the inferred period length.

More precisely, define

$$f_p = \frac{\sum_{i=0}^{r_l - p} \mathbf{1}_{\{s[i]=s[i+p]\}}}{r_l - p}. \tag{2.1}$$

Let

$$P_{\max} = \underset{10 \le p \le r_l - 25}{\operatorname{argmax}} f_p. \tag{2.2}$$

Then the inferred period length is

$$P = \min\{p : 10 \leq p \leq r_l - 25, f(p) > 0.6, P_{\max} \equiv 0 \mod p\}. \tag{2.3}$$

Because the rolling circle amplification product is randomly sheared before sequencing, the information in the R2 read will begin at a random offset into the repeat structure relative to the information in the R1 read (figure 2.2B). Inferring this offset consists of computing the discrete cross-correlation between the first sequence in the pair and the reverse complement of the second sequence in the pair. This can also be viewed as aligning the two members of the pair to each other. To infer this offset, define

$$f_o = \frac{\sum_{i=0}^{r_l - p} \mathbf{1}_{\{s_2[i] = s_1[i+o]\}}}{r_l - o}. \tag{2.4}$$

Let

$$O_{\max} = \underset{0 \leq o \leq r_l - 25}{\operatorname{argmax}} f_o. \tag{2.5}$$

Then the inferred offset is

$$O = O_{\max} \mod P. \tag{2.6}$$

With the structure of repeats in a read pair determined, the base calls in the read pair can be organized into groups that each consist of multiple copies of a particular base in the starting template. The information in each such group can then be aggregated to produce a consensus base call. Each constituent base call in a group comes with a Phred quality score that represents the confidence that the sequencing platform assigns to it and therefore, in

some appropriate sense, the relative weight that should be assigned to it during consensus formation. Interpreting each base call/quality score pair (b_i, q_i) in a consensus group as independent, probabilistic data about the identity I of the consensus base in a Bayesian sense, the probability that the consensus has true identity b given that a member of the consensus group with identity b_i and quality score of q_i was observed is proportional to

$$\mathbb{P}[I = b|(b_i, q_i)] \propto \begin{cases} 1 - 10^{-\frac{q}{10}} & : b = b_i \\ \frac{10^{-\frac{q}{10}}}{3} & : b \neq b_i \end{cases} \tag{2.7}$$

$$= p(b, b_i, q_i). \tag{2.8}$$

Call this expression $p(b, b_i, q_i)$ for convenience. The posterior distribution of the consensus identity given the entire consensus group is then

$$\mathbb{P}[I = b|\{(b_i, q_i)\}] = \frac{\prod_i p(b, b_i, q_i)}{\sum_{b' \in T,C,A,G} \prod_i p(b, b_i, q_i)}. \tag{2.9}$$

The inferred consensus base call b_c is taken to be the value of b with maximum posterior probability, and the corresponding consensus quality score is

$$q_c = -10 \log_{10}(1 - \mathbb{P}[I = b_c|\{(b_i, q_i)\}]). \tag{2.10}$$

For compatibility with downstream processing steps, this value is capped at 93, the largest encodable value in the standard Phred ASCII encoding. Highest-confidnce consensus base calls therefore correspond roughly to a consensus group consisting of three quality-score-30 base calls with unanimous identity.

2.3.3 Mapping circle sequencing data to reference genomes

Because the creation of a rolling circle amplification product is primed at a random location in a circular template and because rolling circle ampli-

fication products are randomly sheared to produce a final sequencing library, the consensus sequence produced by the initial processing outlined above represents an arbitrary rotation of the original input fragment sequence - that is, the original sequence with some length removed from the beginning and appended to the end. Information about the exact location of the junction of circulation has been irreversibly lost by the process and can only be recovered by exploiting other knowledge about the expected structure of pre-circularized sequences. In particular, for pre-circularized sequences that consist of randomly sheared genomic DNA from an organism with known reference genome, we expect some rotation of a consensus sequence to be similar to a stretch of this reference genome. The processes of inferring the junction of circularization in order to 'unrotate' the consensus sequence and mapping the consensus sequence to the genome are therefore inextricably linked.

Identifying the genomic location that a short sequencing read is derived from is a well-studied problem. From a computational standpoint, this mapping process boils down to repeatedly searching a very large target string of the characters {A, T, C, G} for substrings matching each of a series of many short query strings. The target strings being searched, typically the complete genomes of organisms, can be up to billions of nucleotides long. A single sequencing experiment can produce up to hundreds of millions of short query strings to be mapped. The sheer size of these inputs demands the development of computational approaches that scale well. As a further complication, inexact matches allowing for substitutions, insertions, and deletions are typically

required in order to accommodate errors introduced by the sequencing process or the presence of true variants relative to reference genomes. A variety of algorithms and data structures have been explored for attacking this problem over the last decade [29]. A common theme of many successful approaches is to perform a one-time preprocessing of a reference genome to produce an auxiliary index data structure, potentially much larger than the size of the unprocessed genome, that permits faster query searches [70, 109]. Such approaches involve a practical trade-off between the query performance of an index design and the size of the computed index, and trading time for space in this way has practical limitations. In particular, the computed index must be small enough to fit in memory on commodity machines in order to produce acceptable performance.

The use of the Burroughs-Wheeler transform [12, 28] has emerged as a particularly effective navigation of this trade-off. It produces indices with small memory footprints that can be queried quickly and can, with some algorithmic tweaks, accommodate inexact matching. Two groups independently made the connection between this somewhat obscure (at the time) algorithmic idea and the challenges posed by short read mapping, leading to the nearly simultaneous release of Bowtie [63] and BWA [68] in 2009. Bowtie2 [62] was later developed to allow for insertions or deletions in mappings by splitting reads into short segments that are mapped via the BWT machinery to produce seedings of mappings that are then expanded by a more versatile dynamic programming-based local alignment process in the neighborhood of a seed.

The goal of the circle sequencing mapping process is to find all possible substrings of a reference genome that are close enough in Hamming or Levenshtein distance to some rotation of each consensus sequence. When framed in this way, a simple brute-force approach is apparent. We can simply form every possible rotation of each consensus sequence and independently map each rotation to the reference with Bowtie2, configured to allow up a desired number of substitutions or indels and to report all possible mappings. The set of all mappings produced for all rotations of a particular sequence are then sorted and separated into groups whose leftmost mapped position form connected stretches. Each such group represents a single distinct mapping, up to possible ambiguity in the assignment of bases to either side of the inferred circularization junction.

While this is a viable approach that is straightforward to implement and to reason about the sensitivity of, it is computationally wasteful. Precircularized fragments are typically targeted to be around 100 to 150 bases long, so this brute force strategy takes on the order of 100 times the computational effort of mapping an equivalent number of conventionally produced sequences. Of course, the set of comparisons between query and reference implied by this process contains considerable redundancy that can be exploited. To do this, we implemented a seed-and-extend strategy in which a small number of short segments of the consensus sequence are extracted and mapped to the reference using a BWT-based mapper. Any such segment that doesn't span the junction of circularization should in theory represent a continous stretch

of the reference genome and therefore be mappable. Unrolling the remaining consensus sequence around the seed provided by this mapping can then be done via a dynamic programming-based alignment of the remaining consensus sequence to the reference in the vicinity of the seed. For computational efficiency and convenience, we want to arrange for Bowtie2 do as much of this work as possible. To accomplish this, each consensus sequence is augmented by appending a copy of the first half of the sequence onto the end. Because this guarantees that the augmented string contains every possible rotation of the original string as a substring, the correct rotation corresponding to the original orientation of the pre-circularized fragment will be contained somewhere in this augmented string. The augmented string can then be mapped using Bowtie2's local mode to handle both the seeding and extension processes. Incidentally, using Bowtie2 in this way required fixing an obscure bug in Bowtie2's local mode that caused it to erroneously discard multiple local alignments to the same genomic location from a single read, preventing it from being used to reliably recover the full original template from the augmented consensus sequence.

As a final note, one potential pitfall during rotation-insensitive mapping to be on guard against is underestimation of true mismatch rates in the vicinity of the circularization junction. A true mismatch adjacent to one side of the ligation junction could be incorrectly identified as a non-variant on the other side of the junction if the genomic sequence there happens to agree with the variant base identity. To rule out this possibility, we exclude a small number

of bases from either end of each final mapping from all downstream variant calling.

2.3.4 Error correcting properties of circle sequencing data

To measure the ability of this method to correct sequencing errors, we sequenced genomic DNA from a *Saccharomyces cerervisiae* cell culture. While rare difference between the different cells in such a culture will always exist due to mutations that occur in cell divisions during the growth of the culture, such differences are expected to be many orders of magnitude rarer than the base-calling error rate of conventional Illumina sequencing. There is therefore plenty of room to demonstrate improvement over conventional sequencing by comparing the number of apparent such differences that appear in conventional sequencing data to the number that appear in error-corrected circle sequencing data under the assumption that any such differences represents an error.

We first needed to determine if there were any locations in the genome of the specific strain of yeast we were sequencing that were clonally different than the yeast reference genome. If these population-wide variants are not identified, consensus base calls that correctly report the input sequence will disagree with the reference at these positions and be incorrectly flagged as sequencing errors. To do this, we performed conventional Illumina sequencing of the strain of yeast to be used, producing approximately 50-fold coverage of 12 megabase yeast genome. We analyzed the resulting data with the GATK pipeline [21], following the Broad Institutes Best Practice Variant Detection

with the GATK v4 workflow. This process identified 514 potential variant sites in our strain. Any bases mapping to these sites were excluded from any subsequent analysis of error rates. To minimize other potential sources of artifactual mismatches introduced by the mapping process, reads mapping to the incompletely assembled rDNA locus (chromosome XII, positions 451,000 to 491,000), nonuniquely mapping reads, and any mappings containing insertions or deletions were also excluded from analysis of error rates.

With these filters in place, we applied circle sequencing to yeast genomic DNA. To determine the extent to which the redundant information created by the circle sequencing process actually corrects errors, we performed the following proof-of-concept analysis. For each mapped consensus sequence, we returned to the individual repeats of information that went into creating the consensus and artificially restricted ourself to the information present in the first repeat. We computed the fraction of high quality base calls that differed from the reference genome. We then incrementally incorporated information from each subsequent repeat, recomputing a consensus base call and consensus quality score at each position using all the information incorporated so far, and computed the fraction of high quality consensus base calls that differed from the reference genome at each step. As successive repeats are incorporated, high-confidence but incorrect base calls have a chance to either have the consensus quality assigned to them degraded (so that they are no longer confidently wrong) or to have the consensus base identity assigned to them corrected by subsequent correct base calls. Only positions for which

the incorrect base is seen repeatedly without any dissenting votes survive this process to remain errors in the final consensus sequence. Plotting the resulting mismatch rate versus number of repeats incorporated shows the extent to which error detection and correction are happening. High-quality bases in the first repeat of each sequencing read had an error rate of 5.8×10^4 (Fig. 2A). As expected, incorporating the subsequent tandem repeats reduced this error rate, but the effect was surprisingly small, with the error rate asymptotically converging to around 2.7×10^{-4} with all information used, a substantially higher overall mismatch rate than expected (figure 2.4A).

This suggested the presence of processes in which a single error event is able to affect multiple copies of information in a concatamer. One potential such process is single-stranded DNA base damage to starting circular templates. Such damage could be caused any of a number of the experimental manipulations used to create short circles out of genomic DNA, including acoustic energy from the shearing process or something as seemingly innocuous as heat. Single-stranded base damage represents corruption of information before rolling circle amplification has had a chance to protect it. Certain kinds of damage to a base are known to cause polymerases to preferentially incorporate a base other than the standard Watson-Crick pair of the damaged base when using it as a template. This would result in a change in the identify of every copy of the base in a concatamer produced from a damaged template, and therefore in a high confidence incorrect consensus base call.

To explore this possibility, we stratified mismatches by type by asking

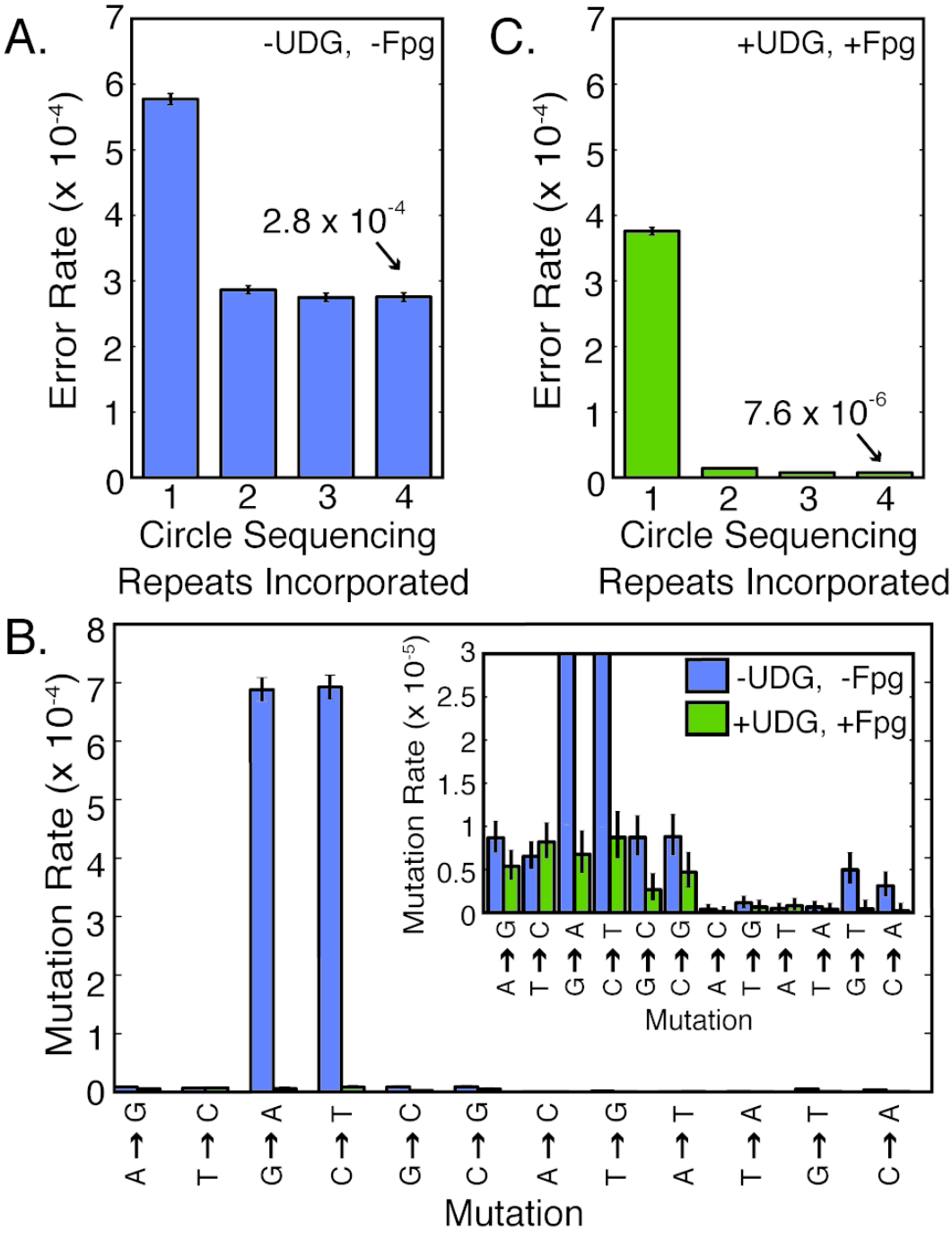

Figure 2.4: **Error correction in circle sequencing.**

Figure 2.4 (Continued): **Error correction in circle sequencing.**
A. Each circle sequencing read consists of several redundant copies of information. As a proof of concept that the full set of copies acts as a check for identifying and correcting errors that occur in any single copy, we calculated the fraction of inferred high-confidence consensus base identities that differed from the reference genome if we artificially restricted ourselves to only using information in the first n copies of information in each read for $n = 1, 2, 3,$ and 4. For the initial experimental design, decreasing error rate with the incorporation of additional information is a demonstration that error correction is occurring, but this rate asymptotes to an unexpectedly high value.
B. When stratified by type, mismatches in data from the initial experimental design (blue bars) are dominated by G→A and C→T mismatches, consistent with induced cytosine deamination in circular templates during the experimental process. Modifying the experimental protocol to treat this damage mechanism dramatically reduced the rate at which these types of mismatches occurred (green bars, shown in more detail in zoomed-in inset).
C. Proof-of-concept demonstration as in **A** for the modified experimental design. Using all redundant copies of information results in error rates in high-confidence consensus base calls below 8×10^{-6}.

what fraction of the time each reference base identity was called as each other base identity (figure 2.4B). The spectrum of mismatches in high-quality consensus base of our initial experimental protocol was strikingly dominated by C→T and G→A mismatches. Such changes are consistent with the spontaneous deamination of cytosines in circular templates into uracils [5]. During rolling-circle amplification, uracils will behave like thymines, base-pairing with adenine instead of guanine (figure 2.5B, middle). Reads derived from the two strands of the resulting double-stranded RCA product will therefore incorrectly report a C in place of a T or a G in place of an A with high confidence (figure 2.5B, right).

To test whether cytosine deamination was responsible for the observed mismatches, we modified our experimental protocol to add uracil DNA glycosolase (UDG) during the rolling circle amplification process. This enzyme is a part of repair pathways in many organisms that prevent the relatively high rate of spontaneous cytosine deamination in genomic DNA (particularly of 5-methylcytosines) from resulting in unsustainable somatic mutation rates [101]. UDG recognizes uracils attached to a DNA backbone, which should not be present in genomic DNA, and excises the nucleotide, leaving an abasic site (figure 2.5C, left). We hypothesized that these abasic sites would not be read through by the phi29 polymerase (figure 2.5C, middle), effectively removing templates which had randomly undergone a cytosine deamination event from the pool of templates that can produce RCA products (figure 2.5C, right). As a test for the presence of additional types of base damage, we also added

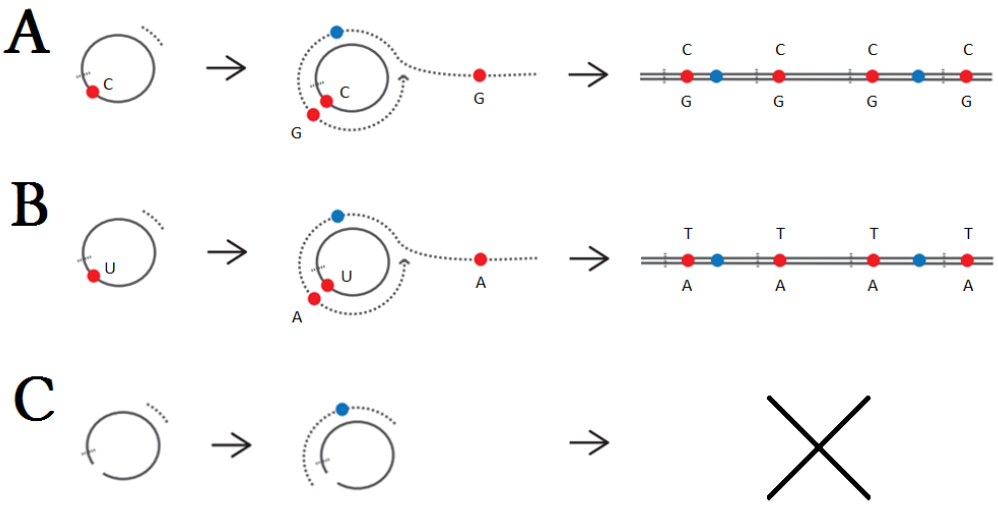

Figure 2.5: **Eliminating artifactual variants created by cytosine deamination.**
(A) Undamaged cytosines in circular template base pair with guanine during rolling circle amplification, producing double-stranded concatamers with C opposite G at the corresponding position in every repeat.
(B) Cytosines that become uracil through spontaneous deamination base pair with adenine during rolling circle amplification, producing double-stranded concatamers with T opposite A at the corresponding position in every repeat.
(C) Excising uracils from backbones using UDG leaves an abasic site that prevents rolling circle amplification from producing a concatamer.

formamidopyrimidine-DNA glycosolase, an enzyme that excises guanines that have undergone oxidative damage that causes them to preferentially base-pair with A instead of C [18, 96].

Examining the mismatch spectrum of data produced by this modified experimental protocol showed a striking reduction in C→T and G→A mismatches (figure 2.4B), and a less dramatic but still clear reduction in G→T and C→A mismatches. This confirms that cytosine deamination was the dominant cause of mismatches in high-confidence consensus bases in data produced without the repair enzymes, and suggests that the remaining mismatches may also represent other types of damage to vulnerable single-stranded circular templates. Performing the same proof-of-concept analysis as above on this data showed asymptotic convergence to the more impressive overall error rate of 7.6×10^{-6} (figure 2.4C). The ability of circle sequencing to filter out the majority of sequencing errors is therefore clear, but this ability is contingent on preventing templates that have undergone cytosine deamination from producing sequencing reads. The rate at which other types of base damage to single-stranded circular templates occur during library preparation most likely represents the limiting factor preventing this error rate from being even lower.

2.3.5 Comparisons of efficiency of error-correction schemes

By creating and sequencing redundant copies of information from each starting molecule, both circle sequencing and barcoding methods inevitably trade throughput for accuracy. Every time a redundant copy of information is

sequenced represents a lost opportunity to instead sequence something new. In order to characterize rarely occurring variants, an experimental method not only needs to eliminate false positives caused by sequencing errors. It also needs to produce large enough quantities of error-corrected data to observe the rarely occurring variants a sufficient number of times.

The amount of useful error-corrected sequence that is produced by sequencing redundant depends on the efficiency with which families of redundant copies are formed. A large variance in the size of families means that raw sequencing reads are wasted on families which do not end up with enough copies to produce high-confidence consensus sequences and on families that are large enough to exceed the point of diminishing returns of informational redundancy. By producing read families that are packaged into single reads rather than recovered from an amplified mixture by sampling, circle-sequencing allows tighter control over the size of redundant families produced, offering both theoretical and practical advantages in efficiency over barcoding methods.

Recall that the central paradigm of barcoding is to amplify uniquely labeled fragments and then sample a large number of reads from the amplified mixture. When multiple reads originating from the same fragment are seen, they can be pooled to form a family from which a consensus sequence can be derived. The efficiency with which large enough families of reads derived from the same starting fragment are seen depends on the relative sizes of the pool of uniquely labeled input fragments (sometimes referred to as the complexity of the input library) and the pool of reads used to sample from the amplified

products of these inputs. This is intuitively straightforward to see in extreme cases. For input libraries containing very few distinct molecules, virtually every input molecule will be seen many times. The cost of achieving this is that the average number of times each molecule is seen is much higher than the minimum number of times needed to form an accurate consensus sequence. The excess times provide no new information and represent wasted reads, and efficiency is low. On the other extreme, for input libraries containing many more distinct molecules than the number of eventual sequencing reads, it is rare to happen to see an input molecule multiple times. Most reads are not seen enough times to form an accurate consensus sequence. The reads spent on these incomplete families are wasted, and efficiency is low. The theoretical efficiency expected when moving across the spectrum of input library complexities spectrum from one extreme to the other is dictated by the Poisson statistics that govern how often the same item is expected to be seen when sampling repeatedly from a large set of items with replacement.

To model the barcoding process, barcoded fragments are assumed to be amplified uniformly to a level such that there are many more copies of an input molecule in the final amplified mixture than are expected to be sampled. Drawing reads from the population of amplified fragments can therefore be modeled as sampling uniformly from the input population with replacement. Let n be the number of distinct double-stranded fragments that receive strand-asymmetric barcodes. There are therefore $2n$ distinguishable inputs, but we will discard information about strand-specificity to reduce this to n inputs,

labeled 1 through n arbitrarily. Let r be the total number of reads. Let X_i be the number of times that input i is seen. X_i is binomially distributed with r trials and success probability $1/n$. For large r and n, this is effectively Poisson distributed with mean r/n. The expected number of inputs seen at least t times is therefore

$$\mathbb{E}\left[\sum_{i=1}^{n} \mathbf{1}_{\{X_i \geq t\}}\right] = \sum_{i=1}^{n} \mathbb{P}[X_i \geq t] \tag{2.11}$$

$$\approx n e^{-\frac{r}{n}} \sum_{j=t}^{\infty} \frac{r^t}{(n)^t t!}. \tag{2.12}$$

Efficiency is defined to be the expected number of consensus bases produced per total base calls (equivalently, the number of eligible consensus families per total reads), which is this expected value divided by r. Every occurrence of n or r in this expression is then in the linked form n/r - the ratio of the number of distinct successfully barcoded input molecules to the total number of reads. The expected efficiency with which consensus families containing at least $t = 3$ reads are produced as a function of n/r for $n = 10^6$ is shown in purple in figure 2.6. In particular, note that while ideal efficiency of $1/3$ would be achieved if every read family contained exactly 3 members, unavoidable variance in the size of read families due to the sampling process caps efficiency at $\sim 19\%$ even for an optimally targeted number of input molecules.

Now consider Schmitt et al.'s duplex barcoding scheme. In this setting, information about strand-specificity is retained so that there are $2n$ distinguishable inputs. We are interested in the probability of sampling both members of a particular input pair at least t times. The random vector $\{X_i\}$

is multinomially distributed with r trials and success probabilities $\{p_i = 1/2n\}$. The components of this vector are not strictly independent, but for large r and n and $t \ll n$, the events $\{X_i \geq t\}$ and $\{X_j \geq t\}$ are close to independent for $i \neq j$, and the probability of such events is negligible if t is not much smaller than n. The expected number of strand pairs for which both strands are seen at least t times is therefore well-approximated by

$$\mathbb{E}\left[\sum_{i=1}^{n} \mathbf{1}_{\{X_i \geq t\} \cap \{X_{n+i} \geq t\}}\right] = \sum_{i=1}^{n} \mathbb{P}[X_i \geq t, X_{n+i} \geq t] \tag{2.13}$$

$$\approx \sum_{i=1}^{n} \mathbb{P}[X_i \geq t]\mathbb{P}[X_{n+i} \geq t] \tag{2.14}$$

$$\approx n\left(e^{-\frac{r}{2n}} \sum_{j=t}^{\infty} \frac{r^t}{(2n)^t t!}\right)^2 \tag{2.15}$$

$$\approx n e^{-\frac{r}{n}} \left(\sum_{j=t}^{\infty} \frac{r^t}{(2n)^t t!}\right)^2. \tag{2.16}$$

The expected efficiency with which consensus families containing at least 3 reads of both members of a pair are produced as a function of n/r for $n = 10^6$ is shown in green in figure 2.6. This efficiency is capped at $\sim 8\%$. Just a notable as the peak levels obtained by the green and purple curves in this plot is the relatively narrow range of input values for which efficiency stays close to this peak. Any imprecision in the titration of the number of successfully barcoded input molecules that enter the amplification process of a barcoding experiment leads to a sharp drop in the amount of useful error-corrected data that the experiment will produce.

In contrast, because circle sequencing delivers families packaged into

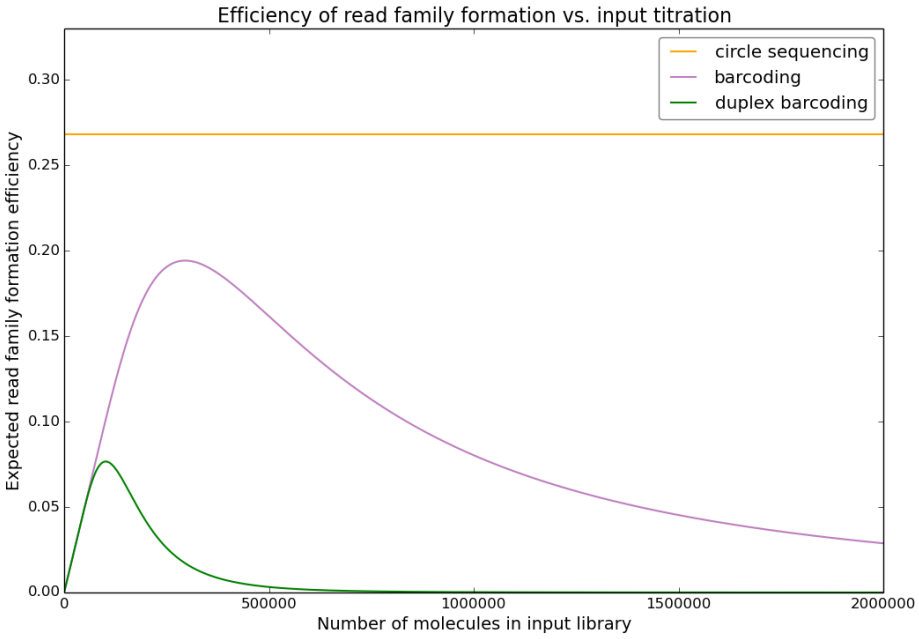

Figure 2.6: **Dependence of efficiency of error correction schemes on input library size.**
The efficiency of read family formation for an error correction scheme is defined to be the expected number number of starting molecules for which at least 3 copies (circle sequencing and barcoding) or 3 copies derived from each strand (duplex barcoding) are recovered divided by the total number of sequencing reads used. Theoretical efficiency of circle sequencing, standard barcoding, and duplex barcoding as a function of the number of distinct molecules in the input library is plotted for a hypothetical experiment producing exactly one million reads. Because barcoding methods assemble families of redundant copies of information by sampling randomly from a pool of copies of each starting molecule, unavoidable variance in the number of times each starting molecule is sampled limits the maximum efficiency with which groups of sufficient size can be produced. The sampling process also strongly couples the efficiency of barcoding methods (y-axis) to the ratio of the number of distinct starting molecules to the total number of sequencing reads (x-axis, since the hypothetical number of reads is fixed). Because circle sequencing physically links copies instead of sampling from a pool, its efficiency does not depend on the number of input molecules used.

single reads, the expected efficiency of circle sequencing is independent of the complexity of the input library. Instead, the efficiency of circle sequencing depends on the size of circular templates relative to the length of sequencing reads used. In an idealized world in which we could produce very long continuous reads of fixed length and in which we were only interested in circular templates that were all exactly one-third the length of our reads, each base in a circular template would be seen exactly three times in the read of its rolling circle amplification product and the number of consensus bases produced would be exactly one-third of the total number of bases read. In reality, given variability in the lengths of circular templates and the fact that the bases read are split into two reads which are separated by a uniformly random offset ranging from 0 to the length of the circular template, the exact way in which edges line up for a given combination of read length, circular template length, and offset will cause some bases to be seen more than or less than three times (figure 2.7A). For a fixed read length and a given circular template length, averaging this efficiency over all possible offsets gives the expected efficiency with which that template length will produce consensus bases. This expected efficiency as a function of pre-circularized fragment length for 2x250 bp read lengths is shown in figure 2.7B. For applications where the length of the circular templates is essentially uniform, such as when an input library is created by amplifying a target region of a genome with primers on either side of it (so-called amplicon libraries), the overall efficiency can be read directly off the value of this curve at the length of the amplicon. For applications where the

distribution of lengths of circular templates is variable, such as when genomic DNA has been sheared and size selected to produce an input library, the overall efficiency is this curve integrated with respect to the circular template length distribution.

To demonstrate these points in practice, we compared data from a circle sequencing experiment performed on yeast to data from a duplex barcoding experiment by Schmitt et al. using genomic DNA from the M13mp2 phage [96]. For each experiment, we computed the ratio of the number of error-corrected consensus bases produced to the number of raw sequencing base calls that went in to producing them, excluding temporarily from consideration any reads that didn't participate in this process, such as phiX contaminants, non-periodic reads, or reads without well-formed barcodes. In order to form a consensus base, we required that at least 3 copies of the base (circle sequencing and barcoding) or at least 3 copies of the base from each strand (duplex barcoding) be observed. As discussed above, the maximum possible theoretical efficiency achievable by the three methods in this case are 26%, 19%, and 8%, respectively. Figure 2.8A gives the actual efficiences achieved in these experiments. As expected, for circle sequencing, the distribution of lengths of sheared input templates around the optimal target length leads to a slight decrease in efficiency to 20.2%, but a reasonably high fraction of the peak theoretical efficiency is obtained. In constrast, for barcoding and duplex barcoding, difficulty in precisely controlling the number of succesfully barcoded input molecules causes actual efficiencies to be substantially lower than their

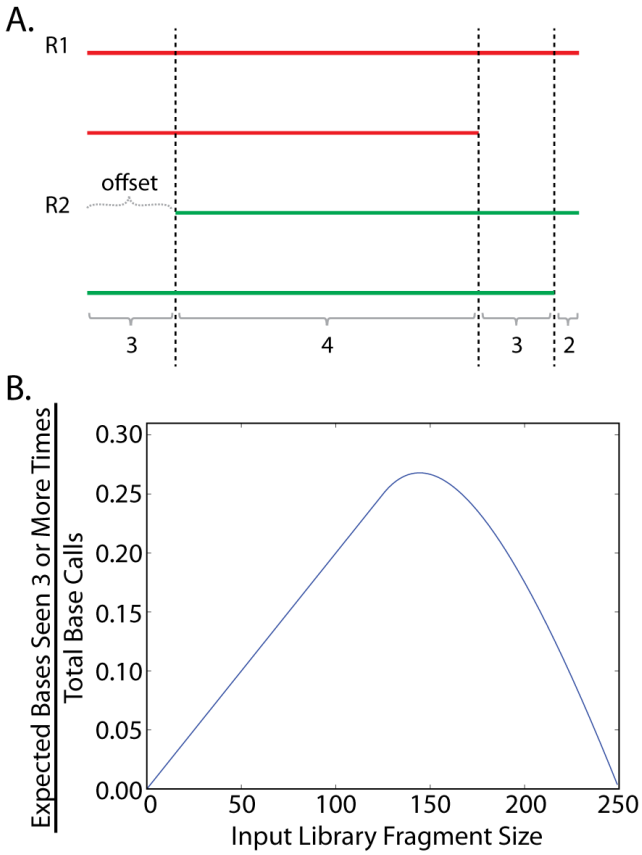

Figure 2.7: **Dependence of circle sequencing efficiency on input fragment length.**
A. The number of copies of each base in the original template present in paired end reads of a concatamer (that is, the number of horizontal lines crossed by a vertical line at any offset in the diagram) is determined by the length of the circular template, the length of each read in the read pair, and the random positioning of the offset of the R2 read in the repeat structure relative to the R1 read.
B. The efficiency of read family formation, defined as the expected number of bases seen at least three times for read pairs consisting of 250 bases each, assuming uniformly random R2 offset values, is plotted as a function of circular template length. Note that the peak value achieved is somewhat less than 1/3.

theoretical maximums (3.0% and 0.8%, respectively).

To directly demonstrate the impact of the number of input molecules used on the efficiency of barcoding methods, we then carried out duplex barcoding on a series of samples of yeast genomic DNA produced by serial 10-fold dilutions. For each sample, the same number of sequencing reads was targeted. The resulting efficiencies of formation of standard barcoding read families or of duplex barcoding read families from the resulting data are shown in figure 2.8B. Moving down the rows of this table from the highest number of input molecules to the lowest represents moving from right to left along the theoretical model of efficiency as a function of input library size in figure 2.6. As expected from this model, we see that efficiency is lowest for very high or very low numbers of input molecules. For the two intermediate input library sizes that produce the highest efficiencies (40 amol and 4 amol), the efficiencies obtained are still substantially lower than the theoretical maximum. Examining the distributions of sizes of read families produced for these samples (figure 2.8C) demonstrates why. For 40 amol (blue), there are too many different read families trying to be sampled, resulting in many read families ending up with fewer than 3 members. For 4 amol (green), there are not enough different read families being sampled, resulting in most families having dramatically more than 3 members.

The theoretical efficiency with which read families can be formed is an important factor in determining how much error-corrected data will be produced by an error-correcting sequencing scheme, but it is not the only such

A

	Bases In	Bases Out	Efficiency
Circle Seq	413M	83M	20.2%
Barcoding^S	2,210M	66M	3.0%
Duplex Bar^S	2,210M	18M	0.8%

B

Input DNA Molecules	Eligible Reads	Barcoding # Read Families	Barcoding Efficiency	Duplex Barcoding # Read Families	Duplex Barcoding Efficiency
4000 amol	304k	11	0.004%	0	0.0%
400 amol	351k	978	0.3%	0	0.0%
40 amol	254k	18,020	7.1%	236	0.1%
4 amol	372k	29,036	7.8%	4,651	1.3%
0.4 amol	431k	2,906	0.7%	495	0.1%

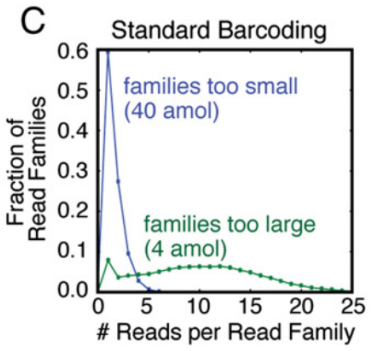

Figure 2.8: **Efficiency of read family formation in actual realizations of different error-correction strategies.**
(A) The table shows the efficiency of read family formation in real data for three error-correction strategies: circle sequencing, standard barcoding, and duplex barcoding. Bases in refers to the total number of bases used to build read families. For barcoding-based approaches, these are bases in well-formed, uniquely mapping reads. For circle sequencing, these are bases in reads showing clear periodicity. Bases out refers to consensus bases. Consensus bases are produced from read families with at least three members (at least three members derived from each strand for duplex barcoding). Efficiency is calculated as the number of consensus bases produced divided by the total number of bases used to produce them. Standard and duplex barcoding values (S superscript) are reanalysis of a dataset from [96].
(B) Standard barcoding and duplex barcoding were used to sequence yeast genomic DNA. Tenfold serial dilutions of the input material were made before amplification. The number of eligible reads refers to the number of reads used to build read families. Also shown are the number of read families consisting of at least three members (standard barcoding) or at least three members from each strand (duplex barcoding), and the efficiency of consensus sequence formation (ratio of read families produced to total eligible reads).
(C) The distribution of sizes of read families (number of reads per read family) produced by standard barcoding with 40-attomol input (blue) and 4-attomol input (green).

factor. The amount of data produced by an actual experiment depends on the practical difficulty of tuning the relevant parameter (either library complexity or library fragment size distribution) to the value that will achieve the optimal theoretical efficiency. Raw sequencing reads will also be wasted on sequencing products that do not have the desired structure but are an unavoidable by-product of the library preparation and sequencing process, such as adapter dimers or phiX spike-ins. Within reads with the appropriate structure, bases may be wasted forming consensus bases with low consensus quality scores due to mechanisms such as PCR-mediated recombination or polymerase errors during barcoding amplification. The fraction of reads that map uniquely to the reference genome may also differ between schemes due to differences in consensus read lengths.

A fair comparison of the cost-effectiveness of different error correction schemes must consider all of these factors together. To make this comparison, we define the yield of an experiment to be the number of high-quality consensus bases in uniquely mapped consensus sequences divided by the total raw number of sequencing base calls. We define the error rate an experiment to be the fraction of such high-quality uniquely mapped consensus bases that disagree with the reference. (Of course, this quantity actually represents the sum of the error rate of the method and the amount of allelic heterogeneity present in the sample being sequenced.) Figure 2.9 shows how standard barcoding, duplex barcoding, and circle sequencing navigate the trade-off between error rate and yield in practice. Circle sequencing achieves error rates equal to or lower than

all realizations of standard barcoding methods with a consistent yield that is several times higher than that of any standard barcoding experiment. While duplex barcoding is able to achieve the lowest error rate of any method because its ability to filter out all types of single-stranded base damage, this comes at the cost of a drop in yield of two orders of magnitude compared to circle sequencing.

2.3.6 Unexpected phenomena in circle sequencing data

Having presented an overall view of circle sequencing and its place in the current landscape of error-correction strategies, we now circle back to describe further technical details of the computational analysis of data produced by the method. Specifically, we will discuss three different unexpected features of the data that emerged over the course of our analysis. Tracking down the sources of these features reveals several interesting mechanisms that go on during the enzymatic manipulations of library preparation and during the sequencing process. Understanding these features informs the interpretation of circle sequencing data and, in one case, provides a potential direction for future improvement to the experimental design.

2.3.6.1 phiX contamination

Each sequence in a read pair of a concatamer should theoretically consist of exact repeats of the sequence of the circular template from which the pair was generated. As discussed above, there should therefore exist a value

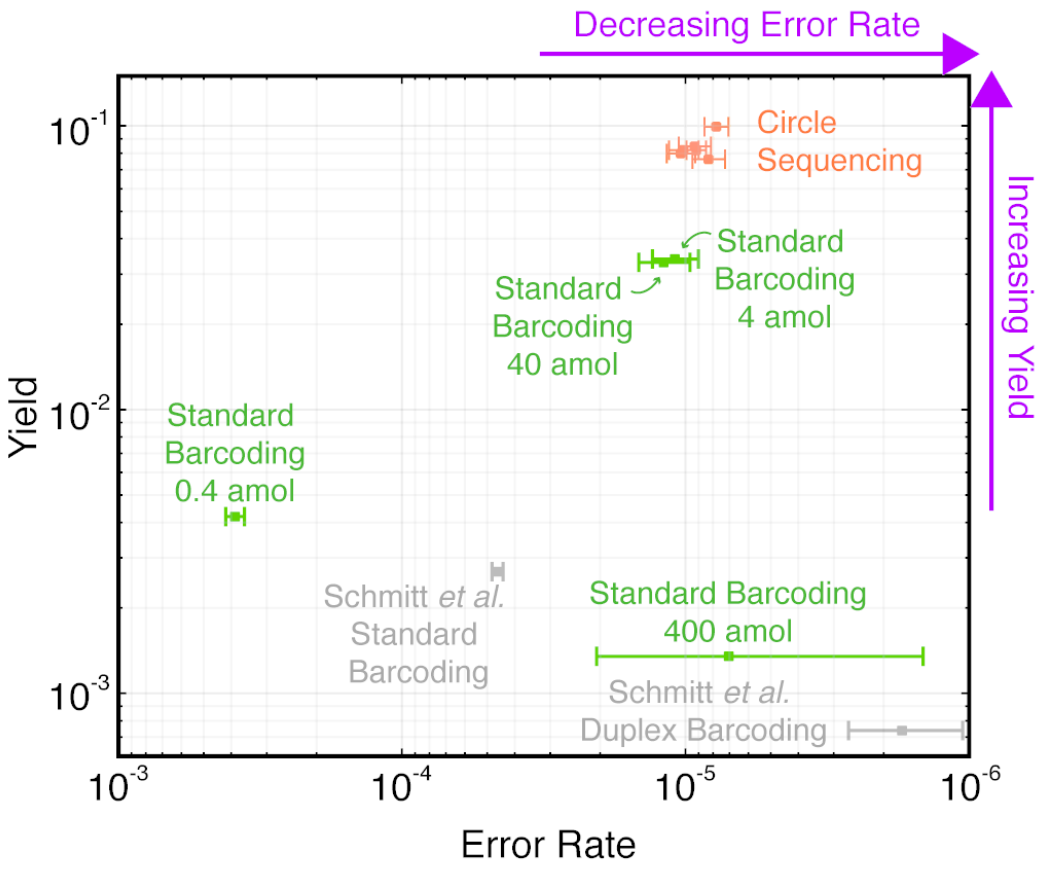

Figure 2.9: **Yield vs. error rate for different error correction strategies.**
Yield is defined to be the number of error-corrected bases produced by an experiment divided by the total number of raw sequencing base calls used to produce them. Error rate is defined to be the fraction of high-confidence consensus bases that disagree with the relevant reference genome. An ideal experiment produces low error rate with high yield and therefore occupies the upper right corner of this plot. Green dots represent our implementation of standard barcoding at different input titrations. Grey dots represent our analysis of Schmitt el al.'s duplex barcoding data, either ignoring strand information as if it were standard barcoding or using strand information. Yellow dots represent different experimental replicates of circle sequencing.

P (the period length) such that almost every pair of base calls separated by distance p in each sequence are identical. To determine if this is the case, we can examine the distribution of f_P - the fraction of pairs of base calls in a read pair separated by distance P that are identical for the value of P that maximizes this fraction - across all of the read pairs produced by a circle sequencing experiment. We expect this distribution to be be peaked at or near 1 and drop off sharply below this. When we first examined the distribution for real data, however, a substantial fraction of reads exhibited essentially no periodicity in excess of random expectation - that is, with f_P only slightly higher than 0.25.

BLAST search of the NCBI nucleotide collection revealed that the sequence of these reads was from the genome of the bacteriophage phiX174. A spike-in of phiX DNA is typically added to sequencing libraries on Illumina machines to allow for internal calibration of the base-calling software. Samples from different libraries are multiplexed on a single run of an Illumina machine by incorporating a unique six nucleotide index sequence in between the R2 sequencing primer and the flow cell attachment sequence that follows this primer. This index is read by a separate sequencing reaction after the R1 read has finished but before clusters have been flipped around to perform the R2 read. This index read uses a sequencing primer that targets the reverse complement of part of R2 sequencing primer. The sequencing reads produced from each cluster can be demultiplexed into the different samples that they came from based on the identity of the index sequence read. In principle, no phiX spike-in reads should exist in the demultiplexed data corresponding to

any indexed sample since Illumina's PhiX Control V3 library 'is not indexed' [42]. In practice, we found that up to 5% of read pairs in our samples mapped concordantly to the phiX genome. As a matter of general interest, we were able to determine the mechanism by which this misassignment occurs.

When the insert between the sequencing primers in a sequencing read is shorter than the read length, the primer sequence on the far end of the insert is itself sequenced. Although the distribution of lengths of inserts in the phiX control library is peaked around 350 base pairs, the process by which these libraries is fragmented and size selected has enough variance that a small fraction of the inserts are less than 250 base pairs long. To determine the precise (proprietary) meaning of 'is not indexed', we examined the adapter sequences that were read through when phiX inserts happened to be shorter than the 250 base pair read length of our data. We found that the adapters on the R2 end of the phiX V3 library are different than the standard 'TruSeq' R2 sequencing primer used to prepare standard Illumina libraries. They instead use the older so-called 'PE' primer. (As an aside, this implies that this older primer must be mixed in to standard Illumina sequencing primer reagents to allow R2 reads of the phiX library.) Because the index read sequencing primer targets the TruSeq R2 sequencing primer, this means that clusters containing the phiX library have no region complementary to the indexing read primer and are not expected to fluoresce during the indexing read.

How, then, are index sequences being read at the phiX clusters in order to assign them to an indexed sample during demultiplexing? One possible

explanation is that bleed-over fluorescence from a nearby indexed cluster is detected at an otherwise-dark phiX clusters during the index sequencing reaction. Two lines of evidence support this hypothesis. The first is that quality scores for the supposed index reads at phiX clusters are dramatically lower on average than those for non-phiX clusters (data not shown). This is expected if these base calls are based on low levels of fluorescence intensity from a nearby cluster rather than on direct fluorescence from the phiX cluster itself. The second is that phiX reads that have a particular index sequence assigned to them are systemically closer to another cluster with that index sequence than phiX reads that correctly have no index sequence assigned to them. To determine this, we computed the distribution of distances from each indexed phiX cluster to the nearest indexed non-phiX cluster by extracting the locations of each cluster from the corresponding read names and using a k-d tree for efficient nearest neighbor searches (figure 2.10A). We compared this to the distribution of distances from each non-indexed phiX cluster to the nearest indexed non-phiX cluster. We found that phiX clusters that received index assignments were on the whole strikingly closer to indexed non-phiX clusters than phiX clusters that did not receive index assignments (figure 2.10B), supporting the model of bleed-over index assignment.

The implications of phiX contamination on circle sequencing are minor. Once their existence is known, they are trivial to filter out, and only have the effect of slightly reducing the apparent efficiency of experiments by adding a small amount of useless data to the denominator of the yield calculations

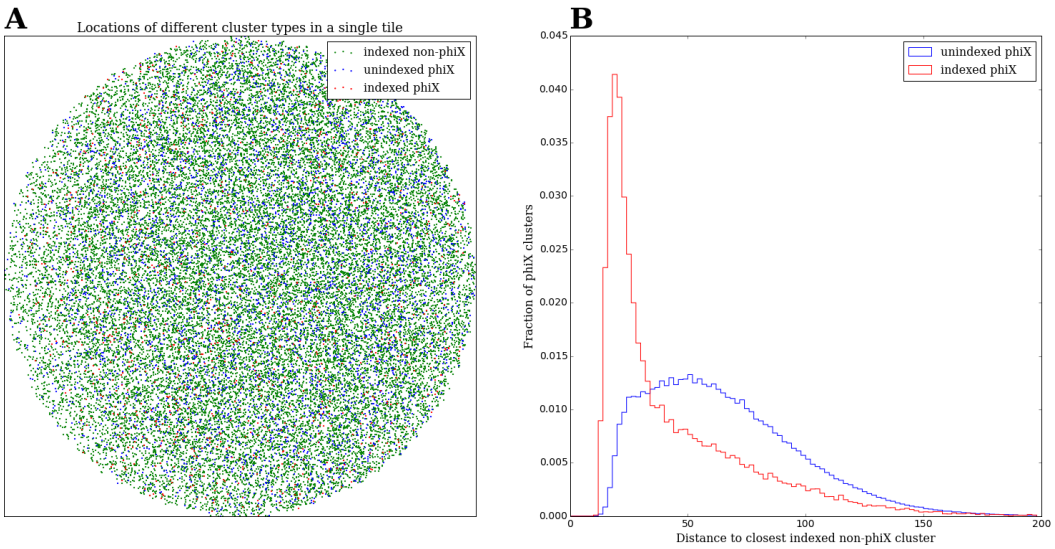

Figure 2.10: **phiX clusters are incorrectly assigned index sequences when too close to indexed clusters.**
(A) Illumina read names contain coordinates of the corresponding cluster on the surface of the flow cell. To determine if phiX reads that have mysteriously been assigned a particular index sequence tend to be closer to another cluster with that index sequence than phiX reads that have not, we can compute the distribution of distances to the nearest indexed non-phiX cluster (green points) for every indexed phiX cluster (red points) and for every unindexed phiX cluster (blue points).
(B) Indexed phiX clusters are strikingly closer to indexed non-phiX clusters (red distribution) than unindexed phiX clusters are (blue distribution), supporting a model in which fluorescence from a nearby indexed cluster during the index read is misinterpreted as the presence of an index sequence at nearby phiX clusters.

described above. For applications involving *de novo* assembly of a genome or transcriptome, however, it is useful to be aware that these reads can exist to avoid incorrectly including a copy of the phiX genome in the resulting assembly [79].

2.3.6.2 PCR-mediated recombination

Once non-periodic contaminants are filtered out, we expected the periodicity in remaining reads to be nearly perfect. However, the distribution of values of f_P across the remaining read pairs has a surprisingly heavy tail of values substantially less than 1 (figure 2.11, blue). Such reads are clearly still concatamers of some original template sequence - for values of P in the neighborhood of 100, as in this data, values of f_P higher than e.g. 0.5 are vanishingly unlikely to occur by chance. Deviations from perfect periodicity in these reads therefore represent a global measure of the fidelity with which each repeat reflects the original template sequence. The fidelity implied by the blue distribution in figure 2.11 is worrisome. If deviations are caused by independent random errors introduced during the propagation of information from a starting circular template through to the sequencing of a base in a concatamer, the rate at which they appear to be occurring could limit the accuracy of consensus base calls. Roughly speaking, the consensus error rate from n copies of information that are each independently wrong with probability p can't be better than p^n. The width of the tail in the blue distribution implies values of p on the order of 5%, potentially placing a lower bound of

$\sim 10^{-4}$ on consensus error rates. This 5% value is substantially higher than the expected error rate of Illumina sequencing or of the phi29 polymerase. We therefore needed to determine what additional mechanisms were introducing deviations from perfect peridoicity. If they represented random independent errors, the overall performance of circle sequencing could be limited by them. If, on the other hand, the deviations exhibit some predictable structure, we can account for this structure when assigning confidence to consensus base calls to avoid being confidently wrong.

One potential source of large deviations from perfect periodicity in a read pair is an insertion or deletion during RCA or during sequencing. Such insertions or deletions would represent extended excursions from the upper diagonal path that a perfectly periodic sequence moves along in figure 2.3 to the next diagonal up or down. To determine if such events contribute substantially to the tail, we generalized the autocorrelation computations above to a full dynamic-programming search of alignments of each sequence to itself using a Cython implementation of the Smith-Watterman algorithm [104]. This revealed a handful of cases of clear indels, typically occurring in the middle of long homopolymer stretches, consistent with general lore about sequence contexts that polymerases tend to slip on [59]. The net contribution of these instance to the tail, however, was negligible.

A general clue was provided by comparing the distribution of quality scores assigned to base calls over the length of sequencing reads from concatamers to those of conventional (i.e. non-periodic) samples on the same se-

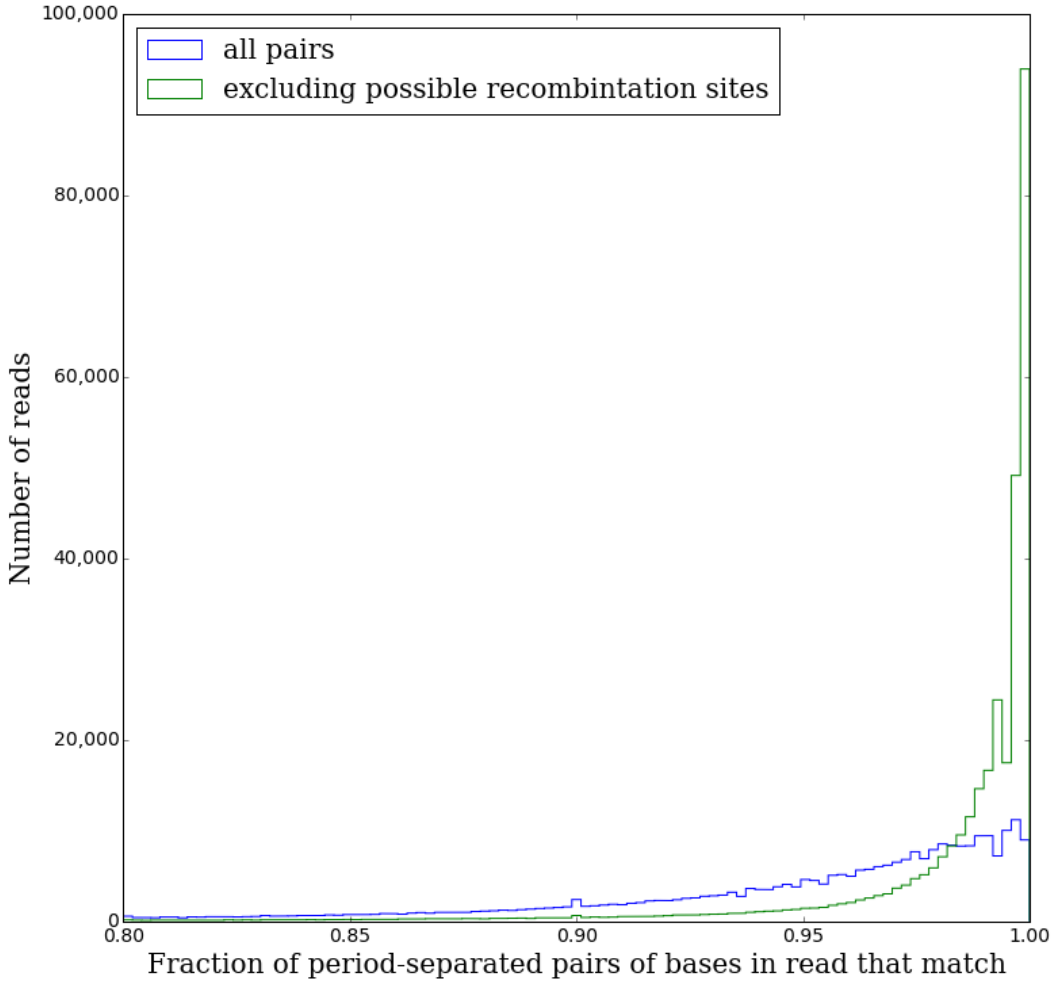

Figure 2.11: **Recombination explains unexpectedly high rates of disagreement between different copies in sequencing reads of concatamers.**
For a sequencing read of a concatamer, there should be a period P such that essentially all pairs of base calls separated by distance P agree. Let f_P denote the fraction of such pairs that are identical for the value of P that maximizes this fraction for each read. Across all reads in a real dataset, many reads produce values of f_P lower than expected given expected sequencing error rates (blue). If possible recombination positions are excluded from the numerator and denominator of this fraction (see text), the distribution shifts dramatically towards one, indicating that recombination is the source of most of these excess sequencing errors.

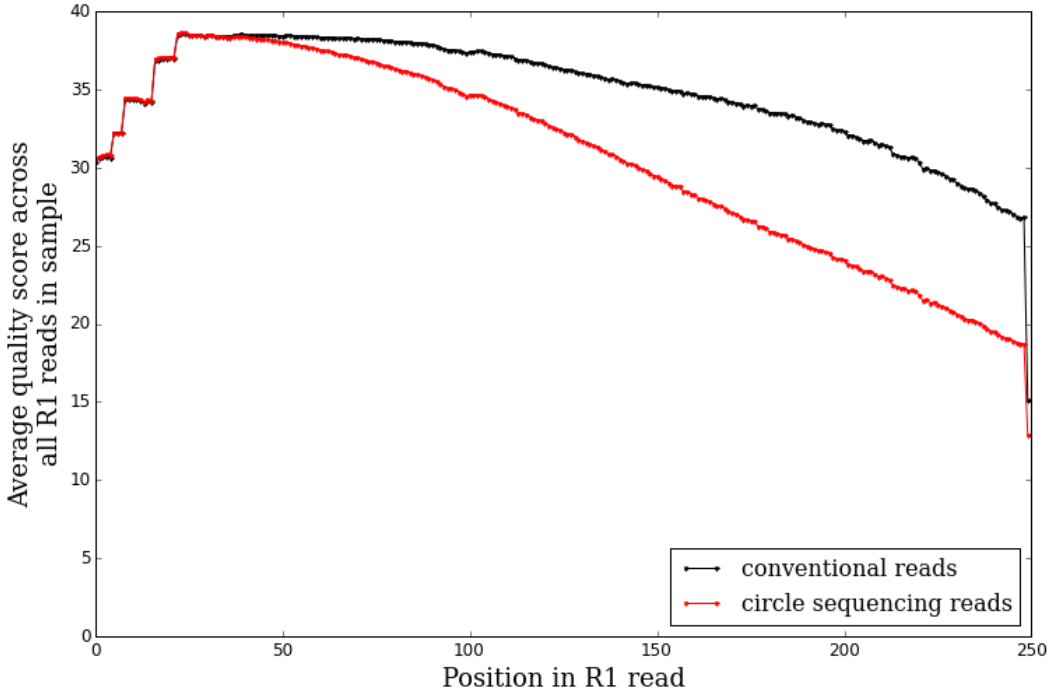

Figure 2.12: **Excess drop in quality scores of base calls over the length of sequencing reads of concatamers.**
Distributions of quality scores at each position across all reads in a conventional sample (black) and in a circle sequencing sample (red) from the same sequencing run. Quality scores degrades substantially more towards the end of reads in concatamers than in conventional reads, suggesting that some property of concatamers presents difficulties for the sequencing process.

quencing run (figure 2.12). When sequencing any kind of sample, these quality scores are expected to decay over the length of reads as the different sequences in a cluster go slightly out of phase with each other. The extent of this drop, however, was substantially more pronounced in circle sequencing reads then in conventional reads. This suggests that some property of concatamers interacts with the sequencing process to make the fluorescent signals presented to the sequencing machine more ambiguous than normal towards the end of reads.

A serendipitous literature encounter [118] brought such a process to our attention. The formation of chimeric sequences via PCR-mediated recombination (also known as template swapping) is a well-studied phenomenon [78]. The fundamental principle of PCR is that any molecule that is flanked by both designed primer sequences will be exponentially amplified. If a polymerase incompletely extends a template during a round of PCR, the partial product produced will lack a primer on one end. No additional copies of the partial product can be made, and the single copy that exists will represent a negligible fraction of the final amplified pool of molecules. If, however, the incompletely extended template ends in a sequence stretch that exists in a different molecule in the pool, the incompletely extended template can hybridize to this alternative location and act as a primer. When extended, this priming results in the creation of a chimeric sequence which was not present in the original input but is flanked by both primers. The presence of both primers allows this chimeric sequence to be amplified in all subsequent PCR cycles and to potentially constitute a substantial fraction of the final pool.

The spatially-localized amplification of a rolling-circle amplification product consisting of several tandem repeats of long, near-identical sequences is in theory particularly vulnerable to this effect. If a polymerase incompletely extends any such template during amplification (figure 2.13A), the premature end is guaranteed by construction to have several alternative locations to which it can hybridize. In particular, it can hybridize to any position that differs from its true position by a multiple of the period length (figure 2.13B). Any such hybridization primes the creation of a chimeric template which consists of the original template with some whole number of complete periods added to or removed from it (figure 2.13C). Instead of a clonal population of copies of the original template, the final population produced will consist of a heterogeneous mixture of the original sequence and a modification of the original sequence that cuts off prematurely and switches to adapter sequence. The proportions of each sequence in the final mixture will depend on how early in the cycles the chimera initially forms and the relative efficiency with which the chimera is amplified compared to the original sequence. If this phenomenon occurs at appreciable levels during cluster generation on Illumina flow cells, it could potentially explain the decrease in quality score over the length of concatamer reads and the elevated levels of deviations from perfect periodicity. Towards the ends of reads, when recombination has caused a cluster to consist of a mixture of continued repeats of the original template sequence and premature adapter sequence, the fluorescence being produced will consists of a mixture of the 'true' base identity and the adapter base. At best, this will cause the

base-caller to identify the true base but with low quality score because of ambiguity introduced by the mixed fluorescence. At worst, fluorescence from the adapter base could outweigh fluorescence from the true base, leading to an incorrect base call.

The fact that the sequences in a read pair contain implicit information about the position of the two reads relative to each other in the sequenced template can be exploited to determine exactly where possible artifactual signatures of PCR mediated recombination could occur in the sequences of the reads. Specifically, once the period length and offset in a concatamer read pair have been determined, a sequence in the read pair can potential consist of a superposition of the real periodic sequence and adapter sequence starting at positions $a = o + l_r \mod p$ and at $\{a + nP : n \in \mathbb{N}, a + nP < l_r\}$ (figure 2.13D). To diagnose the extent to which recombination is leading to the appearance of adapter sequence stochastically superimposed on the expected sequence repeats, we can take successfully mapped consensus sequences and ask, 'of all base calls at position $a + nP + i$ in the raw sequence that went into the creation of this consensus, what fraction had each base identity when this was not identity of the corresponding mapped reference base?' Figure 2.14 shows this diagnostic in R1 reads for $n = 0$ (i.e. the first opportunity in each read for adapter sequence to appear) and $n = 1$ (the second such opportunity) on the top and bottom, respectively. Colored lines show the fraction of all reads that differ from the mapped reference base at each offset into the potential adapter region, and the background is shaded according to the

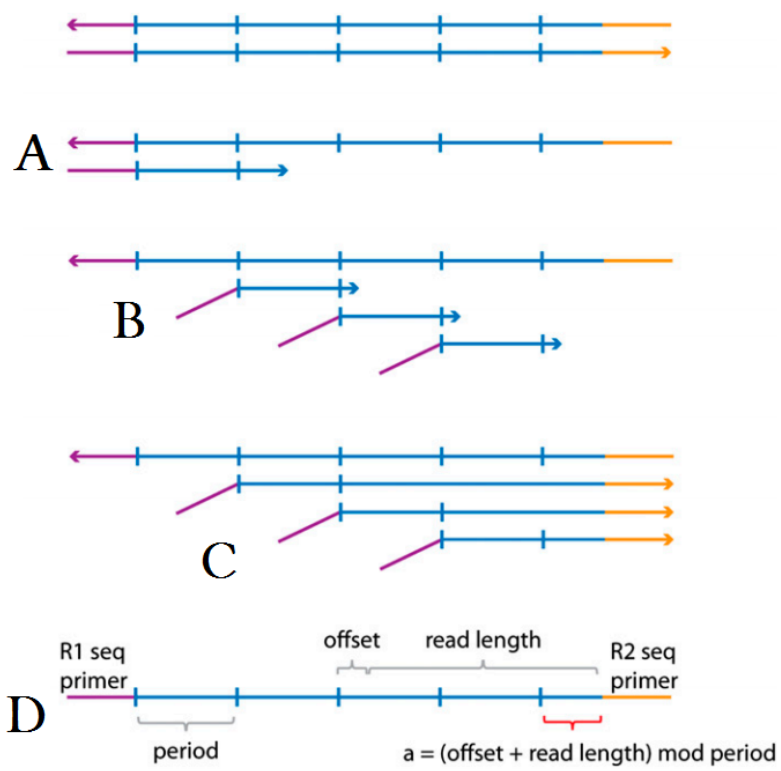

Figure 2.13: **Schematic of PCR-mediated recombination during amplification of concatamers.**
During amplification of a double stranded sequence (top) consisting of several repeats of the same sequence (blue) flanked by primer sequences (purple and orange), incomplete extension during an amplification cycle (A) leaves a truncated sequence that can hybridize to the other strand at several other locations separated by exact multiples of the repeat length (B). Extension of these hybridizations results in chimeric sequences that have primers on both ends but have had some number of repeats subtracted (C).

expected adapter sequence (specifically, the reverse complement of the R2 sequencing primer followed by the flow cell attachment sequence). On a separate scale, the black line shows the average quality score at each aligned position across all reads. Excess incorrect base calls track perfectly with the expected adapter sequence (i.e., each colored line spikes up exactly where the background color predicts it should). Average quality scores drop sharply at the beginning of the potential adapter region and recover moderately after leaving it, indicating that clusters are heterogeneous in this region. Finally, each of these effects is more pronounced at the second opportunity than it was at the first, indicating that less fragments in a cluster have managed to recombine enough to remove the additional repeating unit necessary to place the adapter sequence that much closer. Together, these constitute overwhelming evidence that PCR-mediated recombination happens during cluster generation.

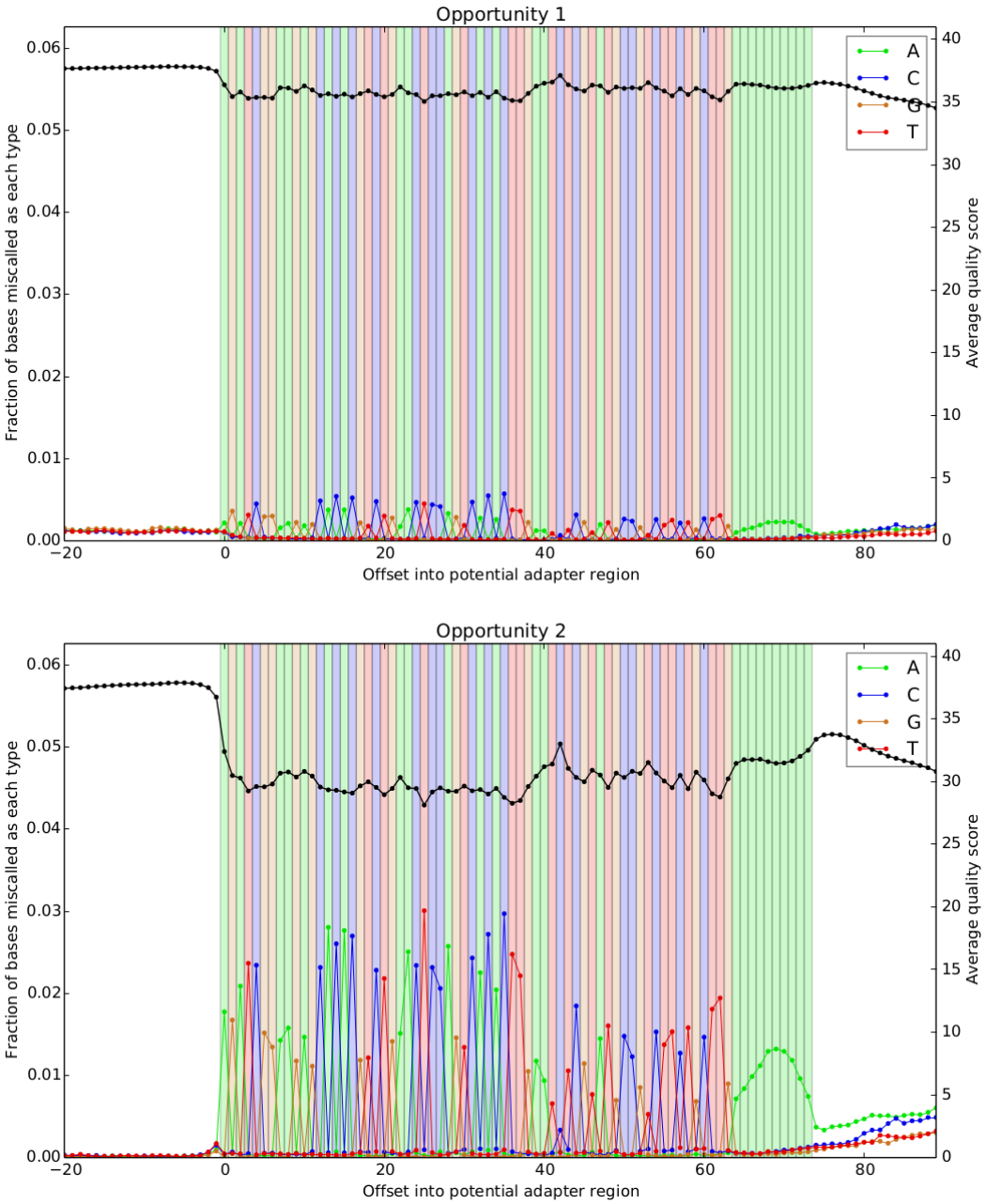

Figure 2.14: **Mismatch profiles and quality scores at possible recombination sites are consistent with recombination during cluster generation.**

Figure 2.14 (Continued): **Mismatch profiles and quality scores at possible recombination sites are consistent with recombination during cluster generation.**
Recombination can potentially cause the sequence at a specific set of positions in each read to consist of a mixture of base calls from the original circular template and from adapter sequence. Starting at the first (top) or second (bottom) such potential adapter sequence location in each R1 sequencing read, we examined 20 bases upstream and 90 bases downstream. At each offset (x axis), we plotted the fraction of bases miscalled as each type (A, C, G, and T) (y axis). Independently, we shaded the background at each distance corresponding to the sequence of the adapter that could be introduced by recombination. The fact that the color of the line that is highest at each position perfectly matches the shaded background color confirms that the signature of mismatches seen agrees with the signature expected to be produced by PCR-mediated recombination. Black lines plot average quality scores across the same positions. Quality scores dip in the adapter regions, consistent with heterogeneity in clusters due to recombination. Both phenomena are more pronounced at the second opportunity than at the first.

To confirm that sequencing errors introduced by recombination are the dominant source of the heavy tail in the distribution of f_P in figure 2.11, we can compute a modified form of this statistic. Recall that f_P consists of the fraction of all pairs of base calls separated by the inferred period length P that are identical. In each such pair, if either base call is at a position that could reflect recombination and is reporting the base identify from the adapter sequence that recombination could place there, we can exclude the comparison of that pair from both the numerator and denominator of the calculation of f_P. If the heavy tail were caused by errors distributed randomly throughout reads, excluding this particularly structured set of comparison from the fraction would have no systematic effect on the distribution of f_P values. Instead, we see a striking shift to the right in the distribution of the modified statistic (figure 2.11, green), confirming that for many reads, predictably located sequencing errors caused by recombination during cluster generation are the dominant source of deviations from perfect periodicity.

A plausible argument can be made that recombination is unlikely to lead to incorrect high-confidence consensus base calls. A high-confidence consensus base call requires at least three high-quality constituent base calls of the same base identity. In order for recombination to affect more than one constituent base call in a group, the total population of templates in a cluster must be distributed across more than two recombined forms. The mixture of signals from these multiple forms should make high-quality base calls at any given position unlikely. To confirm this plausibility argument, we artificially

set the quality of any position consistent with expected adapter sequence to zero. Explicitly disallowing the formation of high-quality consensus bases at positions possible affected by recombination had no significant impact on the overall high-confidence mismatch rate.

While recombination has no measurable impact on the accuracy of high-confidence consensus base calls, it can negatively impact the error rate of low-confidence base calls. It can also affect overall efficiency by reducing the number of positions that are able to achieve high consensus quality score. It could conceivably also affect the ability of circle-sequencing to efficiently use longer read lengths or to transfer to non-Illumina platforms. The actual amount of PCR-mediated recombination that occurs - that is, the distribution across all clusters of the fraction of each cluster that consists of chimeric products - depends on the number of incompletely extended templates that form, which in turn depends on the effective processivity of the polymerase used and on the accessibility of templates to each other in order for undesirable priming to occur. These factors could vary considerably across the amplification schemes used by the different high-throughput sequencing platforms on the market. Once formed, the extent to which heterogeneous clusters are a problem depends on the mechanics of the sequencing process used and on the ability of base calling algorithms to deal with the ambiguous signals that the heterogeneity will present. The amount of recombination could be potentially be minimized by optimizing the amplification process for processivity [61]. Alternatively, recombination-aware base calling software could be developed

that re-evaluates the raw fluorescence intensity data and explicitly models the heterogeneity, potentially recovering more signal.

2.3.6.3 Duplex circles

Once the repeats in a concatamer have been combined into a consensus sequence, we expect every such consensus sequence to consist of a rotation of a region from a single strand of the reference genome that the library was constructed from. In section 2.3.3 above, we discussed strategies for mapping rotated consensus sequences to reference genomes to identify to these regions. In real data, we observed that a small fraction of consensus sequences failed to map the yeast genome in the expected way - that is, there exist well-formed concatamers made up of clear repeats of a circular starting template such that there exists no continuous region in the yeast genome that is a near match to any rotation of the sequence of this circular template.

Identifying the source of mysterious sequences like these is a common procedure in the analysis of data from high-throughput sequencing experiments. These experiments typically involve manipulating a starting pool of fragments of DNA or RNA with various ligations, hybridizations, and reverse transcriptions in order to produce carefully designed sequencable libraries. If every stage in these manipulations works as intended, each molecule in the library will consist of a combination of various payloads of genomic or transcriptomic sequence and various synthetic oligonucleotide sequences laid out in a particular order. These manipulations are not always perfect, however.

Unanticipated enzymatic side effects of the manipulations can produce sequence constructs that do not have the structures they are expected to. The ability to diagnose and catalogue these side effects is frequently necessary for troubleshooting experimental designs in order to increase the rate of production of the intended sequence structures. Even when experimental designs work sufficiently well enough to produce enough useful data, tracking down anomalies can provide insights into unappreciated enzymatic activities that can potentially be harnessed.

To facilitate this kind of analysis, we developed a tool to produce text-based visualizations of the different possible ways each sequencing read can be decomposed into component pieces consisting of stretches of genomic or transcriptomic sequences or of specific synthetic sequences that were introduced. By doing so in a way that makes no assumptions about each read's layout, this enables detection of novel structures. This can be viewed as the converse of the process of simultaneously visualizing the mappings of many reads to a single stretch of a reference genome - instead, the simultaneous alignment of multiple stretches of a reference genome and of various oligonucleotide sequencs to each individual sequencing read are visualized. Alternatively, it can be viewed as a high-throughput and flexible version of the visualizations produced by the NBCI BLAST web server. For each sequencing read or read pair, the tool uses Bowtie2 to produce a comprehensive set of local alignments of the read to genome-scale targets in SAM format. The tool also uses a Cython implementation of the Smith-Waterman algorithm to produce local alignments of

the read to smaller sequence targets, such as adapter sequences. These two sources of alignments are then merged, and text-based representations of each alignment are laid out around the read for visualization.

We applied this tool to understand the source of the circular templates that could not be explained by the standard mechanisms of our experimental protocol. To orient the reader before moving on to these results, we first demonstrate what we expect circle sequencing concatamer reads to look like when visualized with the tool (figure 2.15A). Because the data involved consists of paired-end reads of 250 bases each, printed page dimensions require the text of alignments across a read pair to be split across two horizontal bands in this figure. In the actual output, these two bands are joined at the ellipses. To simultaneously visualize both members of a read pair, R1 and R2 reads are first offset relative to each other to maximize sequence identity between them and then printed on consecutive lines. All local alignments to R1 are then stacked above the aligned read pair, and all local alignments to R2 are stacked below it. The bounds of each local alignment of a continuous region from one of the strands of a double-stranded reference genome sequence to a stretch of a sequencing read are marked by | characters. Text above (for alignments to R1) or below (for alignments to R2) the line marking these bounds annotates the name and coordinates of the reference sequence involved, and the space between the | characters is filled with characters indicating whether the alignment is to the forward strand (>) or reverse strand (<) of the reference. Although none are present in this first example, any mismatches,

insertions, or deletions in these alignments are annotated by replacing these strand-indicating characters with `x`, `-`, or `\/`. If different alignments overlap each other, they are vertically offset from each other; otherwise, they are packed next to each other. In this first example, we see that the read pair consists of repeating units of the region from coordinates 31,307 to 31,434 on the forward strand of chromosome III of the yeast genome (figure 2.15A). (The stretch at the beginning of R1 with no alignment is too short to be detected by Bowtie2, but by inspection is seen to represent the tail end of a preceding copy of the same repeating unit.) This means that, as expected, this read pair represents the result of a single stranded fragment of DNA (figure 2.15B) that was ligated end-to-end to produce a circular template for rolling-circle amplification (figure 2.15C).

When we examined the entire set of read pairs, however, a non-trivial number of read pairs could not be explained as repeats of a single genomic stretch. Figure 2.16A shows a representative such read pair. The sequence of the aligned read pairs begins with a stretch from the forward strand of chromosome XII, ending at position 456,817. The sequence following this, however, is not a second repeat of the same sequence as is expected. Instead, it aligns to the opposite strand of the same genomic extent, beginning at the same coordinate as the end of the forward-strand alignment and extending back to position 456,731. The end of this reverse-strand alignment is followed by a second forward-strand alignment that begins just upstream of (and eventually overlaps the same extent as) the initial forward strand alignment. The

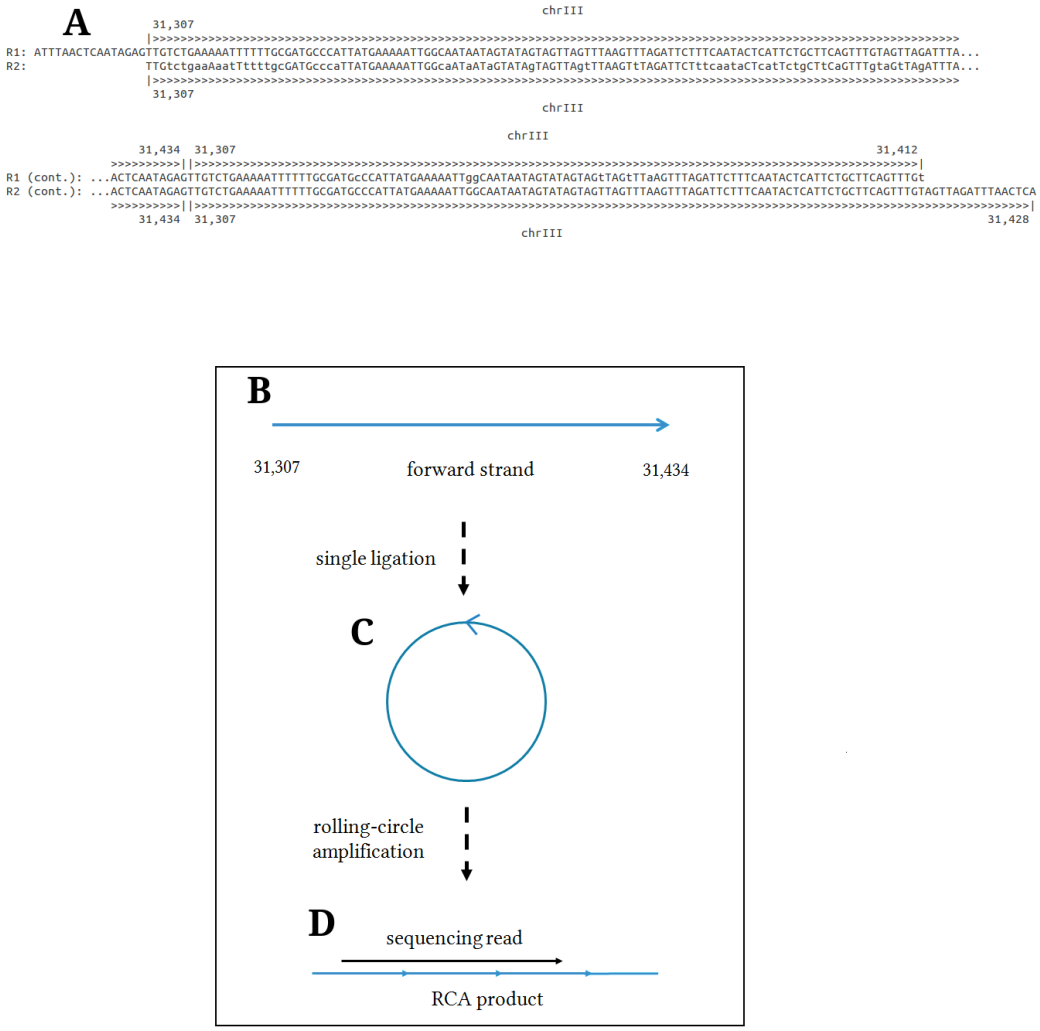

Figure 2.15: **Expected structures in concatamers**
(A) A comprehensive set of local mappings of a read pair of a concatamer to the yeast genome indicates that each member of the read pair consists of repeated copies of a region from a single strand of chromosome III.
This means that, as expected according to the canonical activity of CircLigase, the concatamer was produced by a single ligation of the ends of a single-stranded template (**B** and **C**) followed by rolling circle amplification and sequencing (**D**).

repeating unit in this concatamer therefore consists of end-to-end stretches of a region from the forward strand and a subset of this region from the reverse strand. The rare existence of such concatamers reveals a previously uncharacterized (to our knowledge) enzymatic activity of CircLigase. In order to produce this concatamer, CircLigase must have acted on a template consisting of a mostly double-stranded stretch of genomic DNA that had a short-single stranded overhang on the 5' end (figure 2.16B). A circular structure must have been formed involving two ligations - one between the 5' end of the forward strand extent and the 3' end of the reverse strand extent, and one between the 5' end of the reverse strand extent and the 3' end of the forward strand extent (figure 2.16C).

Because concatamers produced from templates that have been formed in this way contain multiple copies of sequence information independently derived from each strand of a double stranded starting molecule, they have the same potential to detect generic single-stranded damage events offered by Schmitt et al.'s duplex barcoding method. We set out to systematically examine the 'duplex circle sequencing' data we had accidentally produced for a proof-of-principle that this kind of duplex error correction was possible. While the visualization tool described above allowed us to identify that these structures exist, it is not suitable for systematically identifying them and decomposing them their constituent pieces. To do this, the defining feature of concatamers produced by this two-ligation mechanism is that there exist long stretches of sequence in them that are exact or near-exact reverse complements

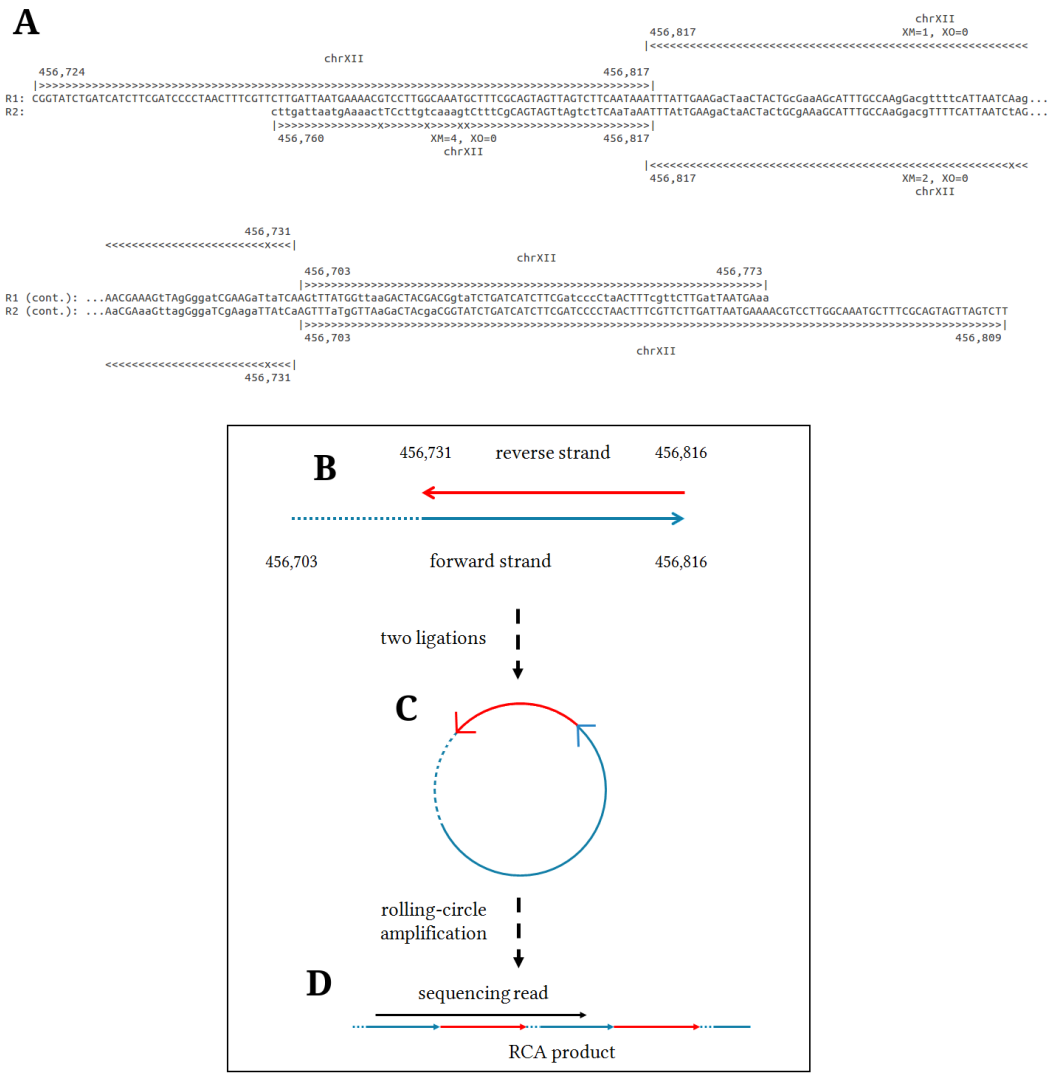

Figure 2.16: **Unexpected structures in concatamers**
(A) Rarely, read pairs of concatamers whose repeating unit consists of a genomic region from one strand (in this case, the forward strand of chromosome XII from positions 456,703 to 456,817) followed by a subset of this region from the other strand (in this case, the reverse strand of the same chromosome from positions 456,817 to 456,731) are observed.
This indicates that, unexpectedly, CircLigase can act on a double-stranded template with a short single-stranded overhang (**B**), performing two ligations to produce a circle consisting of the two strands joined end-to-end (**C**). Rolling circle amplification of this circle produces the read pair observed (**D**).

of each other. The detection of such stretches is conceptually similar to searching for secondary structure in a sequence. Let s be the sequence of a potential duplex concatamer. If it was produced from a template with the two-ligation structure described above, it will be made up of four different component sequence stretches: c, the sequence of one strand of the center region for which both strands exist (solid blue line in figure 2.16); c', the sequence of the other strand of this center region (red line in figure 2.16B); l, the sequence of the potential single stranded overhang on the left of the double stranded center (dotted blue line in figure 2.16B); and r, the sequence of the potential single stranded overhang on the right (not present in figure 2.16). We will collectively call l and r the turnaround sequences. Note that either of these can in theory be of length zero if the corresponding end of the starting template was blunt (that is, double stranded all the way to its end, as in the right side of figure 2.16). s will consist of repeating units of the form $c - r - c' - l$.

The ultimate goal is to identify and group together all copies of c and of c'. The existence of reverse-complementary copies of c and c' separated by a turnaround sequence means that there exists a series of positions in the sequence that we will call reflection points. These points will have the property that if we compare the sequence extending backwards from the reflection point to the complement of the sequence extending forwards from it, there will be a long stretch of identical bases somewhere in this comparison. More precisely, in the middle of each turnaround, there be an index M_j such that if t is the turnaround sequence involved (i.e. one of either l or r), and T is the length of

t, then if T is even,

$$s[M_j - \frac{T}{2} - i] = s_{\text{complement}}[M_j + \frac{T}{2} + 1 + i],$$

and if T is odd,

$$s[M_j - \left\lceil \frac{T}{2} \right\rceil - i] = s_{\text{complement}}[M_j + \lceil \frac{T}{2} \rceil + i]$$

for all $i \in \{0, \ldots, C-1\}$, except where single-stranded damage or sequencing errors have changed the identity of a base call.

To conceptualize this, consider a matrix in which the (i, j)th entry is 1 if the ith base in s is equal to the jth base in the complement of s (figure 2.18). This matrix is symmetric since

$$s[i] = s_{\text{complement}}[j] \text{ if and only if } s_{\text{complement}}[i] = s_{[}j], \tag{2.17}$$

so only the upper triangular region needs to be shown. Around each reflection point, we expect a stretch of identical bases as we move simultaneously backwards in the sequence and forwards in the complement of the sequence. This correspond to moving upwards along an anti-diagonal in the matrix. The existence of long anti-diagonal stretches in this matrix containing almost exclusively ones therefore indicates that the sequence being considered represents a duplex concatamer. The values of the reflection points are given by the indices of the anti-diagonals that contain these stretches. The offsets from the main diagonal at which these stretches begin are the lengths of the turnaround sequences. If we require stretches of perfect identity, we would be

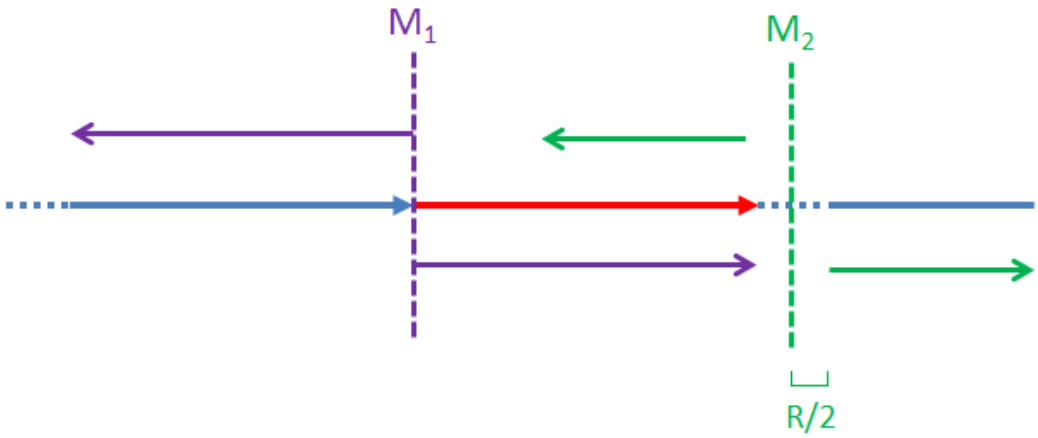

Figure 2.17: **Concatamers from duplex templates contain reflection points.**
In a concatamer produced from the starting template and circle in figure 2.16B and C, the ligation between the two strands on the right of the starting template produces a point M_1 with the property that a stretch of the sequence moving backwards from M_1 (purple arrow to the left) will be nearly identical to the complement of the sequence moving forward from M_1 (purple arrow to the right). The ligation involving the single-stranded overhang on the left of the starting template produces a point M_2 with a similar property, with the modification that a total distance equal to the length of the overhang (blue dotted line) must be skipped before sequence identity moving backward from M_2 (green arrow to the left) will be nearly identical to the complement of sequence moving forward from M_2 (green arrow to the right).

unable to detect any instances of duplex error correction, since base damage to one strand but not the other will results in a pair of matched bases that are not Watson-Crick complements of each other in the middle of an otherwise causal anti-diagonal stretch. We therefore introduce heuristics in which long enough stretches on either side of such a gap are joined together.

Once the reflection points and turnaround lengths have been identified, a read pair can be decomposed into the different copies of c, c', l, and r that it is made up of, and each set of copies can be aggregated into a consensus sequence. These consensus sequences can then be mapped to the yeast reference genome. In contrast to the mapping of normal circle sequencing data, there is no ambiguity about the location of ligation junctions. By construction, c should represent an (unrotated) stretch of the reference genome and can therefore be mapped directly with bowtie2. As a consistency check, we can then confirm that l and r each map to a strand on either side of the mapped location of c.

To search for potential instances of duplex concatamers in which one strand had undergone a base damage event, we applied this processing strategy to a group of 8 samples of yeast genomic DNA produced using the initial design of our experimental protocol. Importantly, because this protocol did not include either of the base-damage excision enzymes (UDG and Fpg) used in the final protocol during rolling circle amplification, we expect relatively high levels of single-stranded base damage. If information from both strands can detect this damage, we can 'excise' all such instances of damage informatically

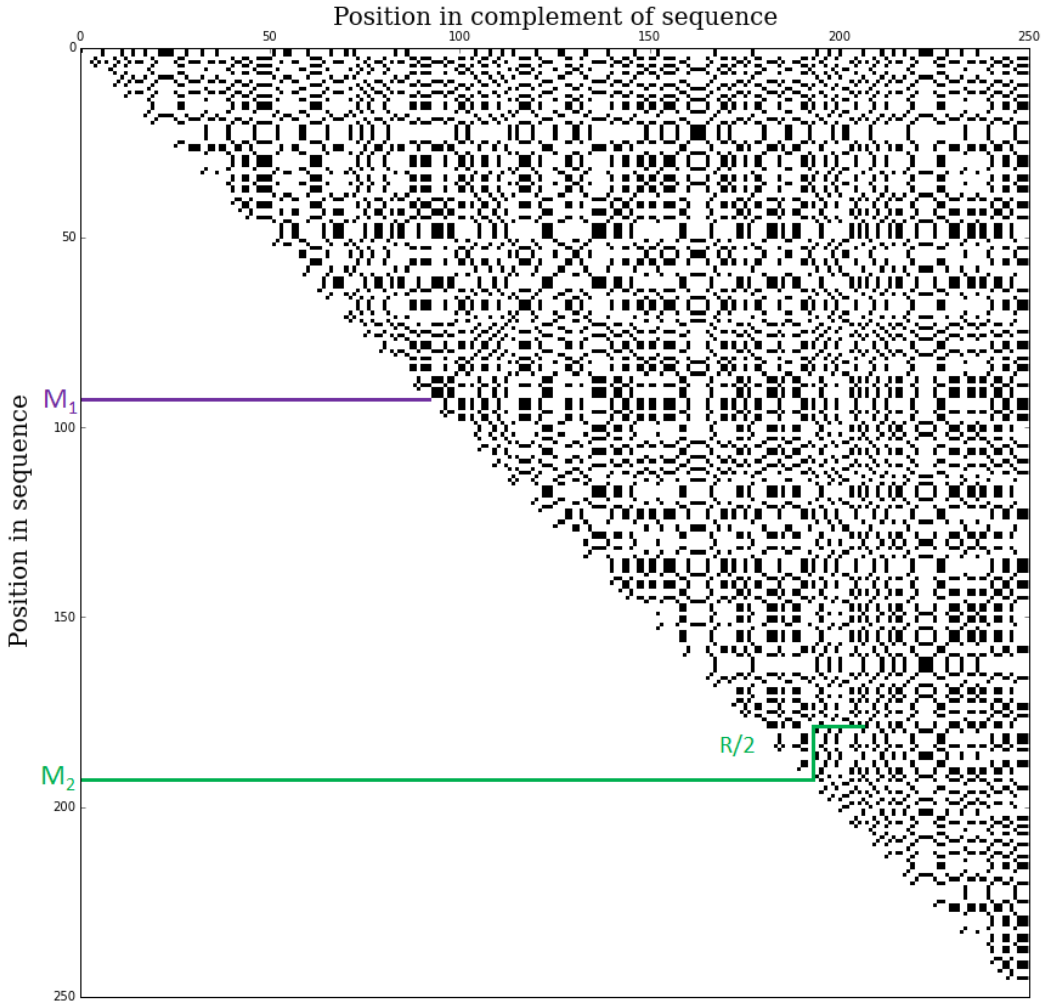

Figure 2.18: **Detecting duplex structures in concatamers.**
In the binary matrix of comparisons of base identities between positions in a concatamer sequence (rows) and positions in the complement of this sequence (columns), long anti-diagonal stretches indicate that the concatamer contains long regions that are reverse complements of each other. The presence of such stretches therefore identifies concatamers from circles produced by two end-to-end ligations of a double-stranded starting template. The indices of these anti-diagonal stretches identify reflection points (M_1 and M_2; see figure 2.17), and gaps between the main diagonal and these stretches identify the lengths of any overhang sequences ($R/2$).

instead of biochemically, just as is done in Schmitt et al.'s duplex barcoding. We identified several thousand unambiguous duplex circular structures, producing several hundred thousand bases of error-corrected consensus sequences. We then analyzed the mismatch rates in the consensus sequences from these structures, first treating each c and c' sequence as if it were a standard circle sequencing consensus and counting how often each type of high-confidence but incorrect identification of a reference sequence base occurred (figure 2.19, blue). When treated as standard data, the mismatch profile has large peaks in G→A and C→T mismatches consistent with cytosine deamination, and, to a lesser extent, G→T and C→A mismatches consistent with oxidative damage. We then took advantage of the consistency check offered by both strands of information by forming duplex consensus sequences out of each c/c' pair, with high-confidence quality scores assigned to a duplex consensus base only if both c and c' agree on the identity of the base. Strikingly, every single high-confidence mismatch disappears after this duplex processing (figure 2.19, green). That is, each mismatch is correctly flagged as artifactual by the absence of a corresponding mismatch on the opposite strand of its duplex partner. Although the total number of duplex consensus bases observed is far too small to place tight bounds on how low the mismatch rate is, as is seen by the large error bar around zero in each duplex column, this represents both additional evidence that single-stranded base damage events are the source of remaining errors made by circle sequencing and an encouraging proof-of-principle that duplex concatamers could be used to correct these errors.

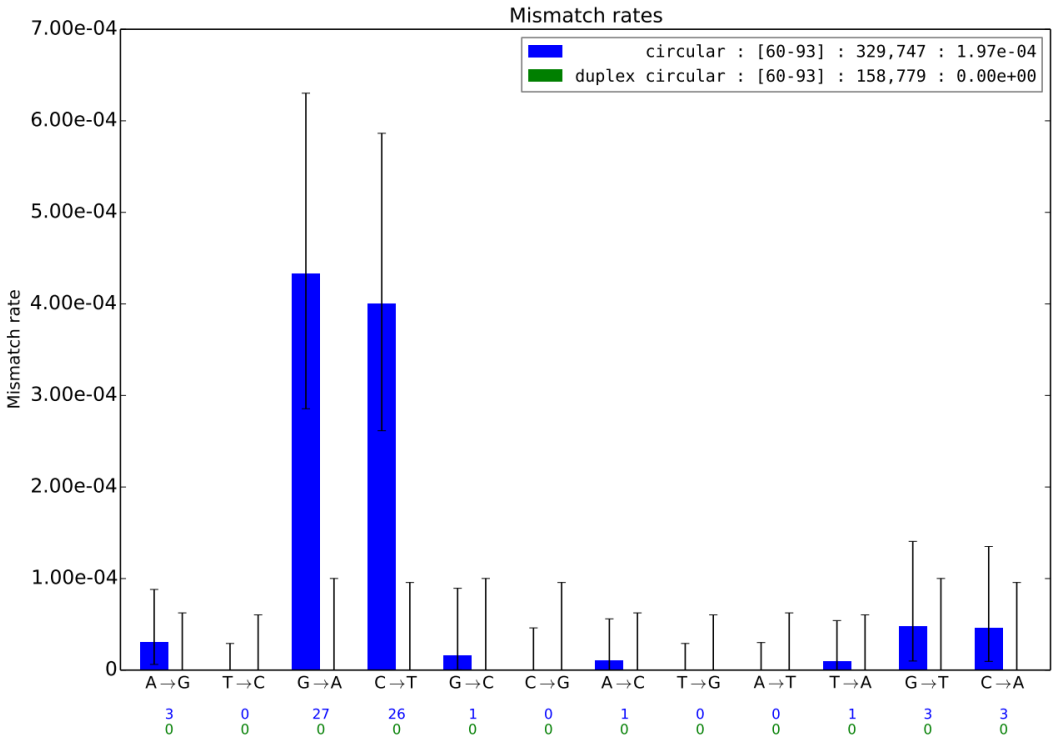

Figure 2.19: **Duplex circular sequences correct errors from base damage.**
Pairs of consensus sequences representing multiple independently derived copies of each strand in double-stranded input molecules were mapped to the yeast genome, and the rates at which each possible type of mismatch between a consensus base and the corresponding reference base occurred were calculated. Each column shows mismatches of a particular type, with bars displaying the mismatch rate (i.e. the fraction of all reference bases of one type misread as another type) and numbers below showing the absolute number of mismatches. Each consensus sequence was first treated as normal circle sequencing data, ignoring strand information (blue). High rates of G→A/C→T mismatches are observed, consistent with cytosine deamination in single strands previously observed. Consensus sequences from each pair of strands were then combined to form duplex consensus sequences (green). Every mismatch in an single strand's consensus sequence is flagged as artifactual by its strand partner, leaving no high-quality duplex mismatches (green bars that would be next to each blue bar if not zero, and green zeros in text below).

2.3.7 Conclusion

In this chapter, we presented a new library preparation strategy called circle sequencing for reducing error rates in high-throughput sequencing by producing and sequencing tandemly linked repeats of every molecule in an input library. We described the computational strategies used to process and analyze data produced by this strategy, and benchmarked the performance of circle sequencing by applying it to sequence genomic DNA from a clonal population of yeast. Circle sequencing achieves a dramatic reduction in sequencing error rates while offering substantial efficiency advantages over alternative barcoding-based methods for performing error correction.

By pushing through the floor of sequencing error rates, library preparation strategies such as circle sequencing and barcoding methods have revealed that single-stranded damage to input DNA templates represent the next hurdle that is preventing even lower error rates from being achieved. Schmitt et al.'s duplex barcoding scheme overcomes this obstacle, but practical inefficiencies in the scheme make it difficult to reliably produce large amounts of error-corrected data with it. We presented a basic proof-of-concept that it may be possible to incorporate Schmitt et al.'s key insight into circle sequencing by constructing concatamers containing multiple copies of each strand in double-stranded input molecules. This could improve the error rate of circle sequencing further by protecting against errors caused by damaged bases in starting templates while retaining the efficiency advantages that delivering physically-linked copies of sequence information provides.

Chapter 3

Local correlations in codon usage do not support a model of tRNA recycling

3.1 Introduction

It has been proposed that patterns in the usage of synonymous codons provide evidence that individual tRNA molecules are recycled through the ribosome, translating several occurrences of the same amino acid before diffusing away. The claimed informatic evidence is based on counting the frequency with which pairs of synonymous codons are used at nearby occurrences of an amino acid, as compared to the frequency expected if each codon were chosen independently from a single genome-wide distribution. Here, we show that such statistics simply measure variation in codon preferences across a genome and do not provide specific evidence for tRNA recycling. An apparently striking pattern observed in such statistics is a universal excess in pairs of occurrences

This chapter is based in part on work reported in J. A. Hussmann and W. H . Press, "Local correlations in codon preferences do not support a model of tRNA recycling," *Cell Reports*, 8 (6), 1624–1629, 2014. JAH performed all computational analysis, and all authors conceived and designed the analysis.

of the same amino acid encoded by the same codon. We show that this pattern is, by a straightforward mathematical argument, a necessary consequence of the existence of any variation in codon preferences across a genome. As a simple negative control on the contribution of pressure to exploit tRNA recycling on these signals, we examine local correlations in the usage of pairs of codons that encode different, rather than identical, amino acids. Such correlations cannot be caused by selection for tRNA recycling and therefore measure the extent to which statistics of this kind are shaped by all other mechanisms affecting codon usage. We find that these negative control signals are statistically as strong as the claimed evidence. We conclude that there is no specific informatic evidence that tRNA recycling is a force shaping codon usage.

3.1.1 tRNA recycling hypothesis

Due to degeneracies in the genetic code, sets of synonymous codons are translated into the same amino acid. Despite the fact that substitutions between synonymous codons in a coding sequence do not change the amino acid sequence of the translated protein, synonymous codons are not used with equal frequencies in the genomes of many organisms, a phenomenon known as codon usage bias[2, 99]. The extent and directions of codon usage biases vary between organisms, between genes within an organism's genome, and within genes [88]. Many theories have been advanced that invoke the mechanics of the complex chain of processes that lead from packaged DNA to translated protein to explain the observed trends, including, but not limited

to, mutational bias[11], bias in repair mechanisms[26], selection for enhanced translational elongation speed or translational accuracy via the coupling of codon usage frequencies to tRNA abundance differentials [24, 65], selection to enhance mRNA stability [53] or to minimize mRNA secondary structure in the neighborhood of binding sites for the translation initiation complex [58], and selection to maintain control over splicing [14]. The relative importance of these mechanisms in shaping the structure of codon usage biases remains poorly understood.

Just as existing biological knowledge can be used to make sense of patterns in codon usage, the detection of patterns in codon usage across and between genomes can in principle be used to make novel inferences about biological process. In a recent paper[13], Cannarozzi et al. make such an inference about the dynamics of translation. They examine all coding sequences of the genomes of several organisms and measure several related statistics which are based on counting the frequency with which a given pair of codons is used to encode pairs of occurrences of the same amino acid that are located close to each other in a coding sequence. They observe that the same codon is used for two nearby occurrences more often than would be expected if every codon choice was drawn independently from a single genome-wide distribution. Furthermore, they observe that nearby pairs consisting of two distinct codons which occur more often than expected tend to be codons which are translated by the same isoaccepting tRNA species. They interpret these results as evidence for the intriguing hypothesis that consecutive codon choices are not

made independently but instead have a tendency to use codons from the same isoaccepting class in order allow a single tRNA molecule to translate multiple codons before diffusing away from the ribosome. When making a novel inference such as this, care must be taken to disentangle other potential sources of the observed supporting evidence. In particular, it is important to determine whether the statistical evidence presented by Cannarozzi et al. offers specific support for their proposed tRNA recycling hypothesis over other previously established mechanisms influencing codon usage.

3.2 Results

3.2.1 Positive diagonal entries are a generic indicator of nonuniform codon preferences

The main line of Cannarozi et al.'s informatic evidence for the tRNA recycling hypothesis consists of a set of statistics that we will call the local covariances in codon preference relative to genome-wide preferences. To compute these statistics, an ordered pair of codons translating the same amino acid is selected. The locations of all occurrences of the amino acid in all coding sequences of a genome are extracted, and the number of times that a sequential pair of occurrences are encoded by the pair of codons of interest is counted (Figure 3.1A and 3.1B). The count recorded is then compared to the number expected under a null model in which the codon used at each occurrence of the amino acid is an independent draw from a genome-wide codon preference distribution for the amino acid, estimated by the genome-wide frequencies with

which each codon is used (Figure 3.1C). An amino acid encoded by d synonymous codons has d^2 possible ordered pairs of codons and therefore produces d^2 of these statistics, which can be naturally arranged in a $d \times d$ matrix. Terms on the diagonal of the matrix correspond to pairs consisting of repeated uses of the same codon, while terms off of the diagonal correspond to pairs consisting of two distinct codons. Cannarozi et al. compute this set of statistics for several amino acids in *Saccharomyces cerevisiae* and find that diagonal terms are universally positive, corresponding to more occurrences of pairs of the same codon than expected under the null model. They interpret this observation as evidence that successive codon choices are not made independently but instead preferentially reuse the same codon.

The set of statistics considered do not provide specific support for this interpretation. The statistics are unable to distinguish between a model of codon usage in which the choices of codon used at consecutive occurrences of an amino acid are not independent and a model in which consecutive choices are conditionally independent given the location of the pair in the genome but drawn from distributions whose parameters vary across the genome with any spatial structure at scales longer than the distance between amino acid occurrences but shorter than the entire genome.

To see this, consider an arbitrary amino acid translated by d synonymous codons and pick one of these codons. Let p_{local} be the location-specific probability with which the codon is used. Suppose that p_{local} varies as a function of location in the genome on scales longer than the typical distance

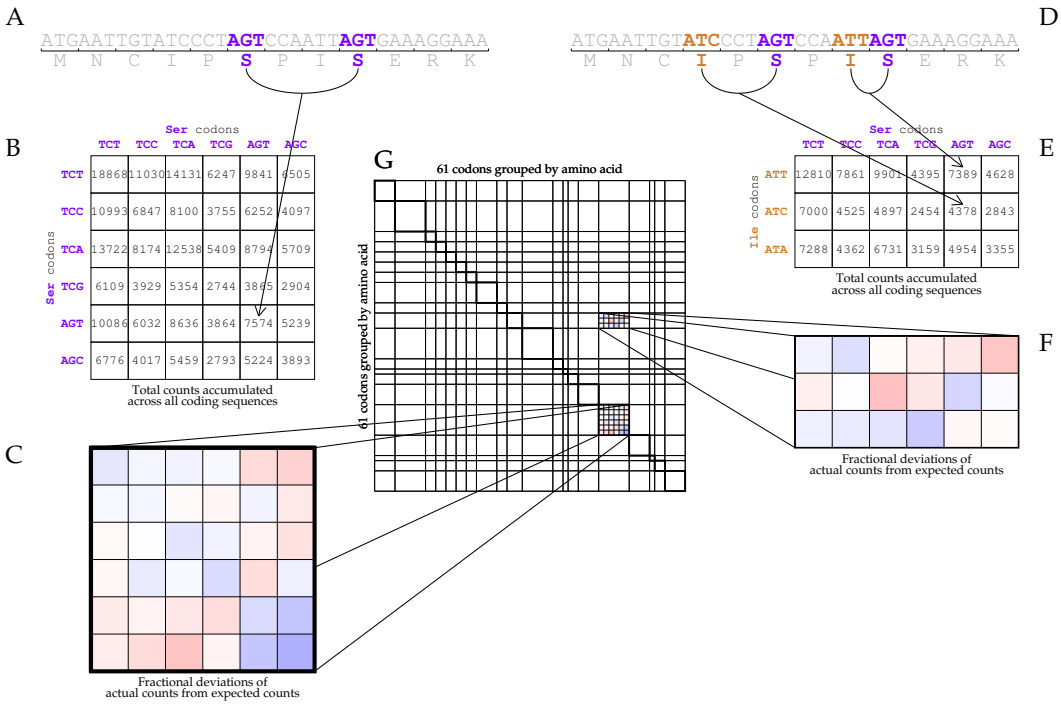

Figure 3.1: **Arbitrary codon pairs exhibit comparable local covariance in usage to same-amino-acid pairs.**
Case 1: For codons encoding the same amino acids (the cases considered by Cannarozzi et al.), all sequential pairs of occurrences of an amino acid (in this case, serine) are identified (A) and the pairs of codons used to encode each pair of occurrences are counted (B). The total counts recorded over all coding sequences are then compared to the counts expected under a null model to produce fractional deviations of actual counts from expected counts (C).
Case 2: For codons encoding different amino acids, all ordered sequential pairs of occurrences of an ordered pair of amino acids (in this case, isoleucine and serine) are identified (D) and the pairs of codons used to encode each pair of occurrences are counted (E). Fractional deviations of actual counts from expected counts are produced as in the previous case (F).
Collectively, the first case makes up the block diagonal of Figure 3.3, and the second case makes up the block off-diagonal portion of Figure 3.3 (G). Deviations of comparable size are seen in (C) and (F).

between occurrences of the amino acid, so that the values of p_{local} at the two locations which make up a sequential pair of amino acid occurrences can be viewed as two nearby samples of a locally approximately constant function. Let n_{genome} be the number of sequential pairs of occurrences of the amino acid in the genome, and let $\mathbb{E}_{genome}$ denote taking the expected value across all such pairs. Then (neglecting edge effects stemming from the fact that the first and last occurrence of an amino acid in each gene only participate in one pair, and neglecting terms that become asymptotically negligible as the number of total occurrences of the amino acid becomes large) a null model of independent draws from a genome-wide codon preference distribution predicts that

$$n_{genome}\mathbb{E}_{genome}[p_{local}]^2$$

pairs of the codon will be observed, while the actual number expected is given by

$$n_{genome}\mathbb{E}_{genome}[p_{local}^2].$$

The statistic of interest, the deviation of observed counts from the genome-wide null model prediction, therefore has expected value

$$n_{genome}\left(\mathbb{E}_{genome}[p_{local}^2] - \mathbb{E}_{genome}[p_{local}]^2\right)$$

and has a clear interpretation as a measure of the variance across the genome in the local independent probability with which the codon is used. In particular, the fact that this expression consists of the difference between the expected value of the square of a function and the square of the expected value of the

function means that it is guaranteed (by Jensen's inequality) to be positive if p_{local} is not simply constant across the genome. Intuitively, the application of Jensen's inequality tells us that while variation in p_{local} leads to the accumulation of excess consecutive pairs of the codon in regions where p_{local} is higher than its genome-wide average and the depletion of pairs in regions where p_{local} is lower than its average, the nonlinearity (more specifically, the strict convexity) of the function of p_{local} in question (namely, squaring) guarantees that the gains will always more than offset the losses. The fact that universally positive values of the statistics are observed on the diagonal of matrices is now seen to be unremarkable. It is expected under any model of codon usage in which codon preferences are not uniform across a genome.

Of course, codon preferences are not uniform across genomes. In particular, the existence of gene-specific codon preferences is a well studied and well accepted (if not completely well understood) phenomenon [100, 120]. Cannarozi et al. correctly identify the need to control for gene-specific codon preferences and correctly identify that shuffling the assignments of codon choices to amino acid occurrences within each gene provides a way to do this. They do not, however, carry out the computation of the expected numbers of pairs of sequential occurrences of each amino acid encoded by each pair of codons under such a gene-by-gene shuffle. The striking feature of their controls, which compare their statistics computed on real data to statistics computed on a single shuffle of the data, is not that some signal survives the shuffle but that most of the signal does not.

Replacing the values expected under a single genome-wide codon preference distribution null model with these gene-specific expected values in the construction of the statistics of interest is a necessary first step in disentangling a potential signature of tRNA recycling from other sources of codon preference variation; as an immediate side effect, this replacement also allows an assessment of the extent to which the magnitudes of the positive diagonal values observed by Cannarozzi et al. in *Saccharomyces cerevisiae* are simply due to gene-specific variation in codon preferences (Figure 3.2). To compute these expectations, pick an arbitrary amino acid that is translated by d codons. Let N be the number of genes in the genome, n_g be the number of occurrences of the amino acid in gene g, and $c_{g,i}$ be the number of occurrences of codon i in gene g. What is the expected number of consecutive occurrences of the pair of choices (i, j) in a synthetic assignment of codon choices to occurrences produced by randomly shuffling the actual set of codon choices within each gene? First note that for any gene g and pair of amino acid occurrences k and $k+1$,

$$\mathbb{P}_{\text{shuffle}}[\text{occurrence } k \text{ is codon } i \text{ and occurrence } k+1 \text{ is codon } j] = \begin{cases} \frac{c_{g,i}}{n_g}\frac{c_{g,i}-1}{n_g-1} & \text{if } i = j \\ \frac{c_{g,i}}{n_g}\frac{c_{g,j}}{n_g-1} & \text{if } i \neq j \end{cases}. \tag{3.1}$$

Let $\mathbf{1}_{g,k}^{(i,j)}$ be 1 if the kth pair of consecutive occurrences of the amino acid in gene g (that is, occurrences k and $k+1$) consists of codon choices i and j, zero

otherwise. Then

$$\mathbb{E}_{\text{shuffle}}[\text{number of pairs } i,j] = \mathbb{E}_{\text{shuffle}}\left[\sum_{g=1}^{N}\sum_{k=1}^{n_g-1}\mathbf{1}_{g,k}^{(i,j)}\right] \tag{3.2}$$

$$= \sum_{g=1}^{N}\sum_{k=1}^{n_g-1}\mathbb{E}_{\text{shuffle}}[\mathbf{1}_{g,k}^{(i,j)}] \tag{3.3}$$

$$= \begin{cases} \sum_{g=1}^{N}\sum_{k=1}^{n_g-1}\frac{c_{g,i}}{n_g}\frac{c_{g,i}-1}{n_g-1} & \text{if } i=j \\ \sum_{g=1}^{N}\sum_{k=1}^{n_g-1}\frac{c_{g,i}}{n_g}\frac{c_{g,j}}{n_g-1} & \text{if } i\neq j \end{cases} \tag{3.4}$$

$$= \begin{cases} \sum_{g=1}^{N}\frac{c_{g,i}(c_{g,i}-1)}{n_g} & \text{if } i=j \\ \sum_{g=1}^{N}\frac{c_{g,i}c_{g,j}}{n_g} & \text{if } i\neq j \end{cases}. \tag{3.5}$$

At first glance, it might seem like the fact that pairs overlap (that is, that the identity of the second member of pair k is required to be the same as the identity of the first member of pair $k+1$) would make computing this expectation more complicated, but the linearity of expectation prevents this non-independence from being a problem.

With formulas to compute the expected number of pairs given gene-specific codon preferences in hand, we see that fractional deviations of the data over a gene-specific null model are substantially less extreme (Figure 3.2A) and less uniformly statistically significant (Figure 3.2B) than deviations over a genome-wide model. In other words, almost all of the excess in usage of the consecutive pairs of the same codon is simply due to the fact that codons are used with different rates in different genes. Having presented this control, it should be noted that Cannarozzi et al.'s argument that "if the correlation effect was simply due to the accumulation of frequent codons in genes with biased codon composition, this effect should also be highest for frequent

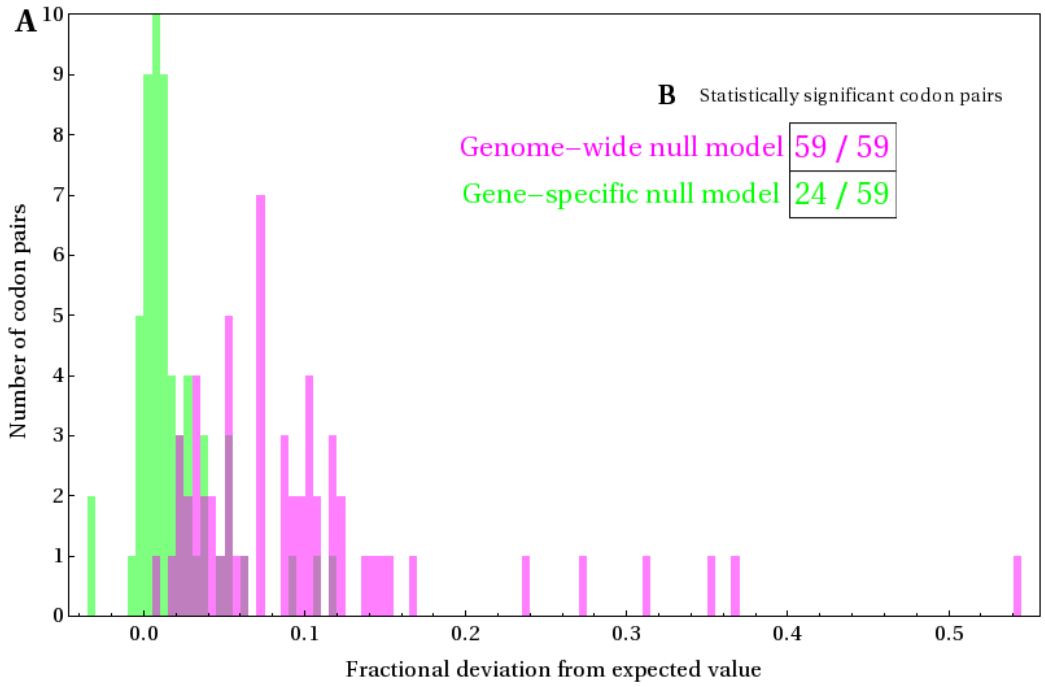

Figure 3.2: **Most of Cannarozzi et al.'s signal is due to gene-specific codon preferences.**
(A) Fractional deviations of actual counts of concordant codon pairs (diagonal entries in matrices) from expected counts under null models of (i) a single genome-wide codon preference distribution (magenta) or (ii) gene-by-gene distributions (green). Most of the strength of the signal present relative to a genome-wide model disappears relative to a gene-by-gene model.
(B) Fractions of concordant codon pairs with statistically significant deviations from expected counts under null models at Benjamini−Hochberg false discovery rate of $\alpha = 0.05$. P-values for each codon pair to input to the Benjamini−Hochberg prescription were computed as (i) the fraction of 10,000 shuffles of codon assignments to amino acids within the entire genome (magenta) or (ii) fraction of 10,000 shuffles of codon assignments to amino acids within each gene (green) for which the shuffled term was more extreme relative to the appropriate expected value than that of the actual data. Much of the statistical significance present relative to a genome-wide model disappears relative to a gene-by-gene model.

codons and not observed for rare codons" misstates the effect that local bias in codon composition has on correlation effects. The effect will be highest for codons whose location-specific frequency exhibits the most variation around its average frequency in the genome, not those whose average frequency is highest.

3.2.2 Signal that survives gene-by-gene shuffling is also nonspecific

The existence of statistically significant (but substantially reduced) residual positive diagonal values after replacing Cannarozzi et al.'s genome-wide null model with a gene-specific null model is no more specific evidence for tRNA recycling than the original signal was. By repeating the same argument as above with the phrase "gene-specific" substituted for "genome-wide", the expected value of the modified statistic is approximately (up to edge effects)

$$\sum_{genes} n_{gene}\big(\mathbb{E}_{gene}[p_{local}^2] - \mathbb{E}_{gene}[p_{local}]^2\big),$$

where n_{gene} is the number of pairs of occurrences in a given gene, and positive values of the modified statistic are generic evidence for the existence of structure in codon preferences at scales larger than the distance between occurrences but smaller than genes.

The existence of intragenic codon preference structure in many organisms is well established [93], and several models of sources for such structure have been proposed [37, 111]. (See section 3.2.3 below for a demonstration of a particularly simple source of intragenic structure in human coding sequences.)

A simple observation allows us to assess the amount of sub-gene-scale structure in codon preference that is due to sources that are not tRNA recycling. The set of statistics considered by Cannarozzi et al. can be extended in a natural way to consider pairs of codons encoding distinct amino acids. To construct this generalized set of statistics, label the 61 non-stop codons and select an arbitrary ordered pair (i, j). Let a_i be the amino acid translated by the first codon and a_j be the amino acid translated by the second codon. In each coding sequence in the genome, identify every sequential pair of occurrences of a_i and a_j (that is, a pair such that the occurrence of a_i is before that of a_j and there are no other occurrences of either amino acid in between the two) (Figure 3.1D). Record the number of such pairs which are encoded by codons i and j (Figure 3.1E). The count produced can then be compared to the number expected under gene-specific shuffling of codon assignments to amino acid occurrences (Figure 3.1F).

To compute these expected values, consider codons i and j encoding amino acids a_i and a_j with $a_i \neq a_j$. Let $n_g^{(a_i)}$ be the number of occurrences of a_i, $n_g^{(a_j)}$ be the number of occurrences of a_j, let $n_g^{(a_i,a_j)}$ be the number of pairs of occurrences of a_i followed at some distance by a_j such that there is no other occurrence of a_i or a_j between the two in gene g, and let $\mathbf{1}_{g,k}^{(i,j)}$ be one if the

kth such pair in gene g consists of codons i and j and zero otherwise. Then

$$\mathbb{E}_{\text{shuffle}}[\text{number of pairs } i,j] = \mathbb{E}_{\text{shuffle}}\left[\sum_{g=1}^{N}\sum_{k=1}^{n_g^{(a_i,a_j)}} \mathbf{1}_{g,k}^{(i,j)}\right] \tag{3.6}$$

$$= \sum_{g=1}^{N}\sum_{k=1}^{n_g^{(a_i,a_j)}} \mathbb{E}_{\text{shuffle}}[\mathbf{1}_{g,k}^{(i,j)}] \tag{3.7}$$

$$= \sum_{g=1}^{N} n_g^{(a_i,a_j)} \frac{c_{g,i}}{n_g^{(a_i)}} \frac{c_{g,j}}{n_g^{(a_j)}}. \tag{3.8}$$

The statistics produced by comparing the observed counts for all possible pairs of codons to these expected values can be naturally arranged in a 61×61 matrix (Figure 3.1G). If codons are grouped according to the amino acid they translate, the original subset of codon pairs considered by Cannarozi et al. (the special cases for which $a_i = a_j$) occupy blocks on the diagonal. Significantly non-zero values for pairs of codons that do not encode the same amino acid cannot be caused by selective pressure for tRNA recycling because such pairs are neither translated by the same isoaccepting tRNA species nor forced to offset a potentially disproportionate share of expected counts taken up a pair that is. Such values are, however, easily explained by models of location-specific variation in codon preferences. Following a similar framework to arguments made above, if $p_{local}^{(i)}$ and $p_{local}^{(j)}$ are the local probabilities with which codons i and j, respectively, are used as a function of location in the genome, the (i,j)th entry in the matrix has an expected value approximately

equal to

$$\sum_{genes} n_{gene}^{(a_i,a_j)} \left(\mathbb{E}_{gene}[p_{local}^{(i)} p_{local}^{(j)}] - \mathbb{E}_{gene}[p_{local}^{(i)}] \mathbb{E}_{gene}[p_{local}^{(j)}] \right),$$

where $n_{gene}^{(a_i,a_j)}$ is the number of sequential pairs of occurrences of a_i and a_j in a given gene. The motivation for calling this set of statistics the local covariance in codon preference is now clear.

Positive values of such an off-diagonal term indicate that regions in which codon i is used more often than its gene-wide frequency tend to overlap with regions in which codon j is used more often than its gene-wide frequency. Of course, this argument is unchanged if codons i and j are distinct but encode the same amino acid. As Cannarrozi et al. observe, in this case, positive values tend to be i, j pairs which are translated by the same tRNA species, an observation that survives the switch to a gene-specific null model. While this signature could be caused by tRNA recycling, it could also simply indicate that local codon preferences are coupled, by selection, to the identities of tRNA species. For example, translation may be locally slowed down in portions of genes to prevent ribosomal "traffic jams" [111] or to allow time for co-translational folding of the nascent polypeptide [57] via the usage of codons translated by scarce tRNA species. Such mechanisms would create positive covariances in location-specific preferences for codons translated by a given tRNA.

We now establish the plausibility of the second interpretation. Examining the strength and significance of local covariances between pairs of codons

translating distinct amino acids, which can be caused by local independent codon preference variation but not by tRNA recycling, and comparing these to pairs translating the same amino acid allows us to determine if tRNA recycling is plausibly a major influence on codon usage. The 61×61 matrix of fractional deviations for all codon pairs in *S. cerevisiae* shows widespread structure (Figure 3.3). In particular, the block-diagonal segment corresponding to pairs encoding the same amino acid is not a visually or statistically distinct subset of the entire matrix. The distributions of fractional deviations and statistical significances corresponding to terms inside of the block-diagonal subset but off of the main diagonal and those corresponding to terms outside of the block-diagonal subset are strikingly qualitatively similar (Figure 3.4A and 3.4B). Comparable fractions of terms from each class are indisputably statistically significant. The largest positive and negative values for pairs encoding distinct amino acids are as extreme as those for pairs of distinct codons encoding the same amino acid. Taken together, these observations suggest that values in the diagonal blocks can be explained entirely by local preference structure induced by non-tRNA recycling mechanisms and therefore cannot be taken as specific evidence that tRNA recycling is a major force shaping codon choices.

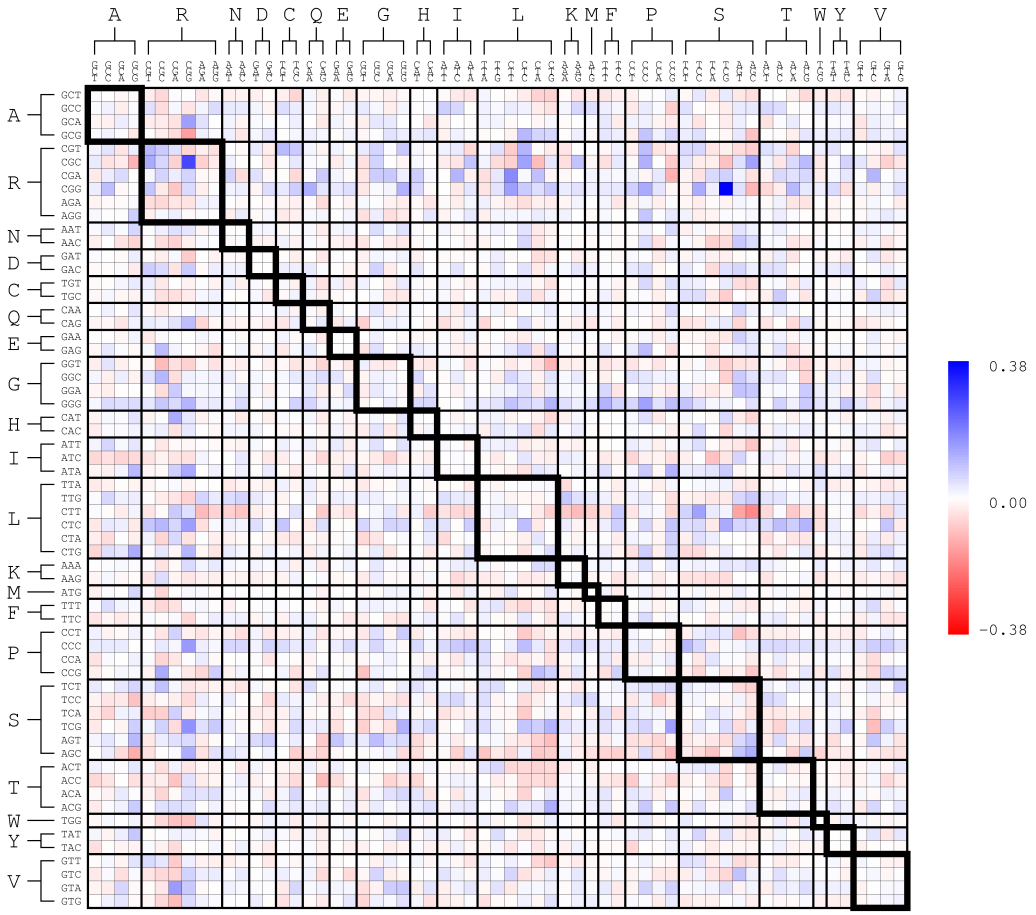

Figure 3.3: **Complete data for the framework shown in Figure 3.1G, generated according to the process outlined in Figure 3.1.**
Fractional deviations of counts of actual usage of codon pairs in all coding sequences of *S. cerevisiae* with respect to counts expected under a shuffling of assignments of codons to amino acids within each gene. The thickly bordered diagonal blocks contain those pairs of codons that encode the same amino acid. These diagonal blocks are not a visually distinct subset of the full matrix. (See Figure 3.4.)

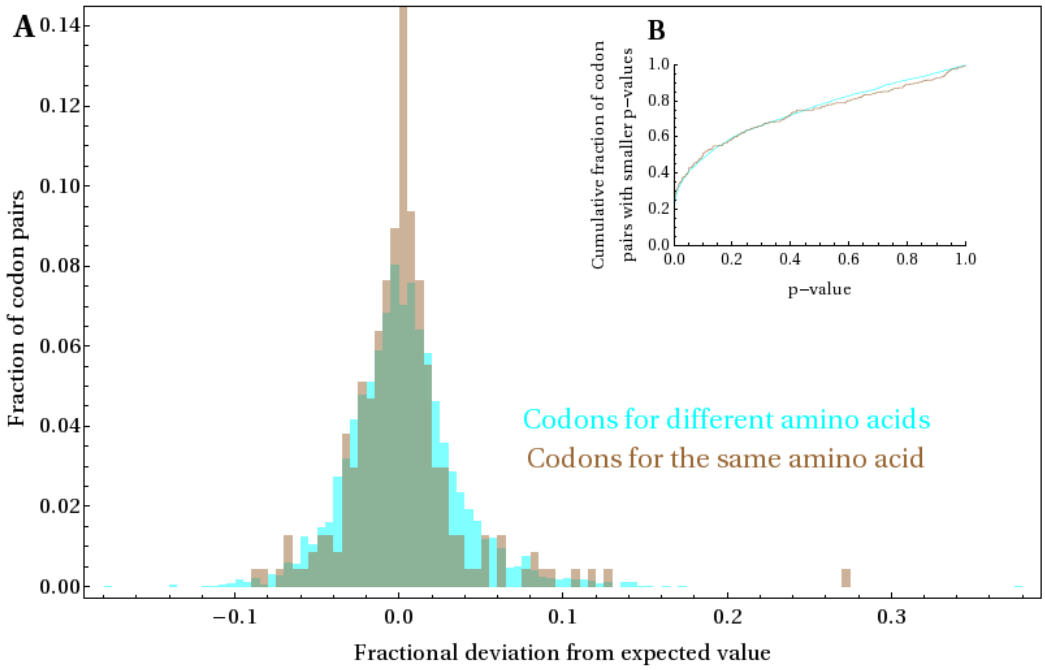

Figure 3.4: **Comparison of signals observed for codon pairs encoding the same amino acid and codon pairs encoding different amino acids.** **(A)** Distributions of fractional deviations shown in Figure 3.3 for (i) terms inside of the block diagonal, representing pairs of codons encoding the same amino acid (brown) and (ii) terms outside of the block diagonal, representing pairs of codons encoding different amino acids (cyan). (Overlapping bars show as darker green.) The distributions of signal strengths for these two classes of codon pairs are strikingly similar.
(B) Empirical CDFs of p-values for the fractional deviations in Figure 3 for (i) terms inside of the block diagonal but off of the main diagonal (brown) and (ii) terms outside of the block diagonal (cyan). P-values for each term were computed as the fraction of 10,000 shuffles of codon assignments to amino acids within each gene for which the shuffled term was more extreme relative to the expected value than that of the actual data.

3.2.3 Pattern in *Homo sapiens* coding sequences confirms that codon preference correlations have a diverse set of causal mechanisms.

As a postscript to these results, we repeated the same analysis as above on representative transcript models for every coding sequence in *Homo sapiens*. We were surprised to see that the local codon preference covariance matrix for *H. sapiens* exhibits a distinctive checkerboard-like pattern of alternating red and blue (Figure 3.5). Reexamining Figure 3.3, this pattern is not present in *S. cerevisiae*. The source of this puzzling pattern became clear when we noticed that because the different codons for each amino acid were ordered lexicographically with a nucleotide order of T-C-A-G along the rows and columns, the change in codon that happens when moving between almost any pair of adjacent entries in the matrix involves switching at least one nucleotide from an A or T to a C or G (or vice versa). In light of this, the checkerboard pattern indirectly suggests that for any given pair of codons, the total AT or CG content of each codon in the pair influences whether the pair covaries positively or negatively with each other.

To test this theory, we permuted the rows and columns of the matrix in order to group codons together by the total number of Cs and Gs that they contain, breaking the original grouping of codons together by the amino acid that they encoded. This new clustering produces a striking block pattern (Figure 3.6). Blocks in the upper left and lower right corners, which consist of covariances between codon pairs containing no GCs or consisting entirely of GCs, respectively, are almost exclusively positive. Blocks in the lower left

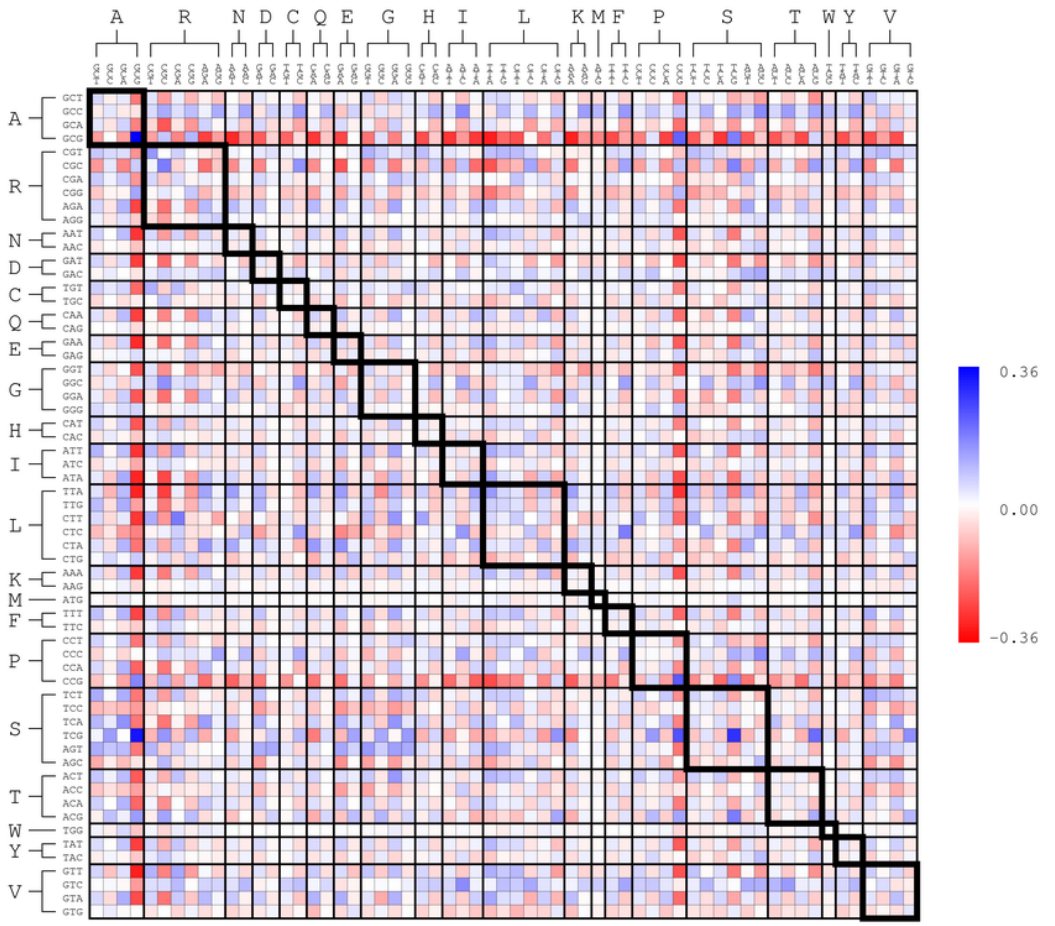

Figure 3.5: **Local covariances in codon preferences in *H. sapiens* with codons grouped by amino acid**
Fractional deviations of counts of actual usage of codon pairs in representative transcript models of each coding sequence of *H. sapiens* with respect to counts expected under a shuffling of assignments of codons to amino acids within each gene. A distinctive checkerboard pattern of alternating red and blue is apparent.

and upper right corner, which contain covariance between codon pairs with opposite CG contents, are almost exclusively negative. These patterns suggest that in human coding sequences, covariances in codon usage within genes are mainly an indirect reflection of patterns in nucleotide composition. Performing the same regrouping in *S. cerevisiae* does not exhibit the same block pattern (data not shown). This provides further evidence that values of covariances are shaped predominantly by diverse, species-specific mechanisms rather than by a universal preference for tRNA recycling.

3.3 Conclusion

In this chapter, we presented mathematical arguments and statistical controls suggesting that observed patterns in the use of synonymous codons at nearby occurrences of the same amino acid cannot be taken as specific support for the hypothesis that ribosomes move more quickly along coding sequences if individual tRNA molecules are recycled through ribosomes multiples times. This chapter has represented a minor detour from the overall theme of the thesis, in the sense that there was no high-throughput sequencing data involved in it at all! Instead, it served to introduce the topics of synonymous codon usage and translation dynamics, and to demonstrate the subtleties that can arise when attempting to use evidence left behind by evolution in the form of codon usage patterns to make indirect inferences about the speed of translation *in vivo*. In this sense, it can be viewed as an extended motivating example for why the ability to directly and accurately measure how long

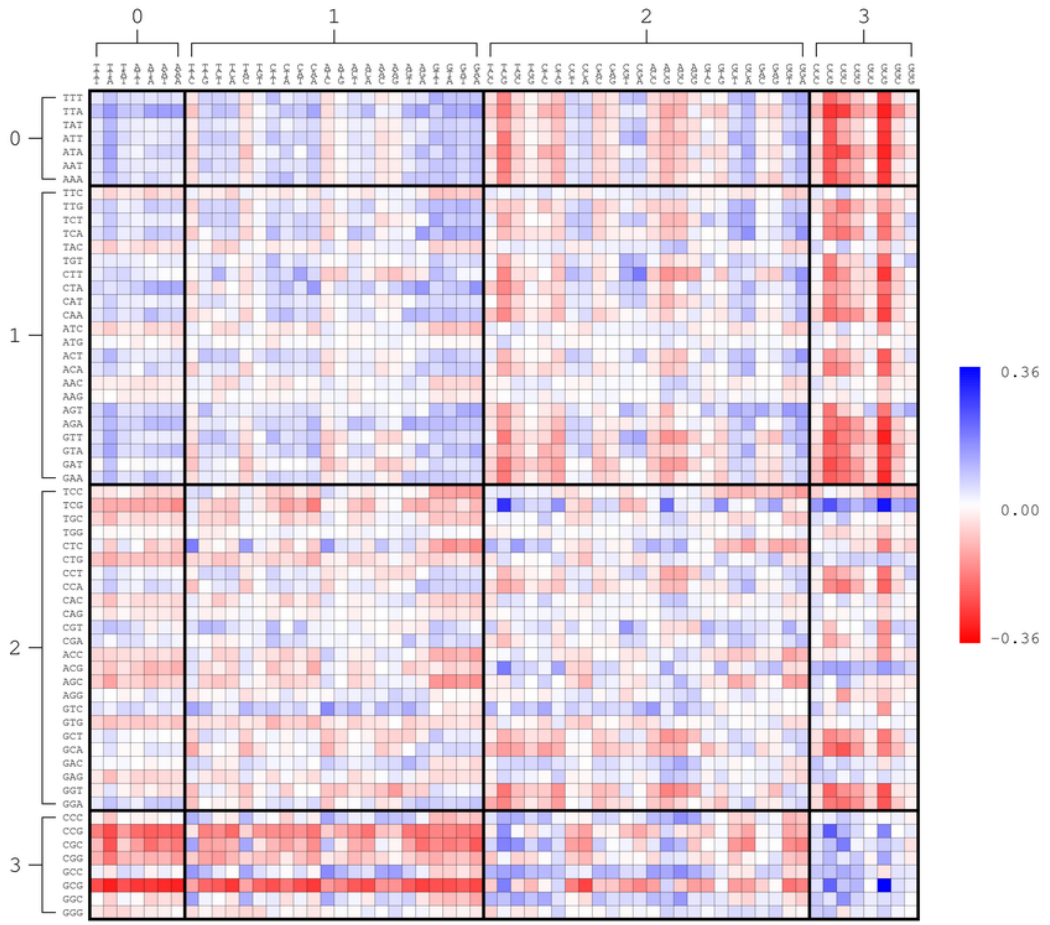

Figure 3.6: **Local covariances in codon preferences in *H. sapiens* with codons grouped by GC content**
The same covariance values as in Figure 3.5 are shown, but codons are grouped along each axis by the total number of Gs and Cs they contain, rather than by amino acid. A general trend of positive covariance between codons with similar GC content and negative covariance between codons with dissimilar GC content is observed.

ribosomes spend translating each codon on transcriptome-wide scales under natural conditions would be a useful thing to have. The use of sequencing to directly experimentally explore these topics will be the subject of the next chapter.

Chapter 4

Understanding biases in ribosome profiling experiments reveals signatures of translation dynamics

4.1 Introduction

Ribosome profiling is an experimental technique in which high-throughput sequencing is used to produces snapshots of the locations of actively translating ribosomes on messenger RNAs. These snapshots can be used to make inferences about translation dynamics. Recent ribosome profiling studies in yeast, however, have reached contradictory conclusions regarding the average translation rate of each codon. Some experiments have used cycloheximide (CHX) to stabilize ribosomes before measuring their positions, and these studies all counterintuitively report a weak negative correlation between the translation rate of a codon and the abundance of its cognate tRNA. In contrast, some experiments performed without CHX report strong positive correlations. To explain this contradiction, we identify unexpected patterns in ribosome density downstream of each type of codon in experiments that use CHX. These patterns are evidence that elongation continues to occur in the presence of CHX but with dramatically altered codon-specific elongation rates. The measured positions of ribosomes in these experiments therefore do not reflect the

amounts of time ribosomes spend at each position *in vivo*. These results suggest that conclusions from experiments using CHX may need reexamination. In particular, we show that in all such experiments, codons decoded by less abundant tRNAs were in fact being translated more slowly before the addition of CHX disrupted these dynamics.

4.1.1 Ribosome profiling

Translation is the process by which the assembly of a protein is directed by the sequence of codons in a messenger RNA. Ribosomes mediate this conversion of information from codons into amino acids through the sequential binding of tRNAs [107]. During the incorporation of each successive amino acid, there are several stages at which the identity of the codon being translated may potentially influence the speed with which a ribosome advances along a coding sequence. When a codon is presented in the A-site of a ribosome, an appropriate tRNA must diffuse into the A-site and successfully form a codon-anticodon base pairing interaction [34, 52]. tRNAs decoding different codons are expressed at different abundances [22, 111], suggesting that ribosomes could spend longer waiting for less abundant tRNAs to arrive [113]. Because translation is accomplished with fewer tRNA identities than there are codon identities, some codon-anticodon interactions involve non-Watson-Crick base-pairings [35, 86]. These so-called wobble pairings are thought to modulate the speed of decoding [51, 67, 106, 119]. Once a tRNA has arrived and base-paired, the speed of peptide bond formation between the C-terminal

amino acid in the nascent chain and an incoming amino acid may be influenced by chemical properties of these amino acids [3]. The relative contributions of these effects to overall rates of translation remain poorly understood.

Because the genetic code that governs the process of translation maps 61 codon identities to only 20 standard amino acids, multiple synonymous codons can be used to encode most amino acids. There is a rich body of theoretical work on the role of translation speed as a selective force shaping synonymous codon usage [88], but the ability to directly measure the speed with which each codon is translated *in vivo* in order to test these theories has historically been lacking. The recent development of ribosome profiling, the massively parallel sequencing of footprints that actively translating ribosomes protect from nuclease digestion on messenger RNAs [43–46], presents exciting opportunities to close this gap. Ideally, the millions of sequencing reads produced by a ribosome profiling experiment are snapshots of translation representing samples drawn from the steady state distribution of ribosomes across all coding sequences. The statistical properties of these snapshots can in theory be used to measure the relative speed with which each codon position is translated: the more often ribosomes are observed at a position, the longer ribosomes are inferred to spend at that position.

In practice, ribosome profiling studies in *Saccharomyces cerevisiae* using different experimental protocols have reached contradictory conclusions about the average decoding times of codon identities. Because yeast rapidly regulate translation when stressed and ribosomes cannot be instantaneously

harvested from cells, the original ribosome profiling protocol of Ingolia et al. [46] pretreats cells with cycloheximide (CHX) for several minutes to stabilize ribosomes in place before the harvesting process begins. CHX is a small-molecule translation inhibitor that has been a staple of experimental approaches to the study of translation for decades. However, the exact mechanism of this inhibition is not completely understood, with recent studies suggesting that CHX binds to a ribosome's E-site along with a deacylated tRNA to block further translocation [20, 97]. The majority of the rapidly growing body of ribosome profiling experiments in yeast have followed this original CHX-pretreatment protocol [3, 4, 9, 25, 32, 64, 77, 81, 121]. Several groups have applied a variety of conceptually similar computational methods to the data produced by these experiments to infer the average speed with which each codon identity is translated. Counterintuitively, these groups have found that, on the whole, codons decoded by rare tRNAs appear to be translated faster than those decoded by more abundant tRNAs [16, 92]. Different theories have been advanced to contextualize these unexpected results, hypothesizing that the measured elongation rates reflect a co-evolved balance between codon usage and tRNA abundances [92], or that translation dynamics are dominated by interactions involving the nascent chain rather than the actual decoding process [16].

More recently, however, several groups have produced data using an optimized harvesting and flash-freezing protocol that allows CHX pretreatment to be omitted [30, 31, 39, 64, 81, 89]. This omission was motivated by obser-

vations that treatment with CHX affects several high-level characteristics of footprinting data, including the distribution of lengths of nuclease-protected fragments in mammalian cells [47] and the amount of enrichment in ribosome density at the 5' end of coding sequences in yeast [31]. In contrast to data produced using CHX pretreatment, several studies using this alternative protocol have reported that non-optimal codons are in fact translated more slowly [30, 115]. The source of this discrepancy between the statistical properties of measured ribosome positions with and without CHX pretreatment has been unclear, leading to uncertainty as to which measurements correspond to actual properties of *in vivo* translation dynamics.

Here, we present analysis of data from a large body of ribosome profiling studies to resolve these contradictory results. We find consistent differences between experiments performed with and without CHX pretreatment in how often ribosomes are measured with specific codon identities positioned in their tRNA binding sites. We also find unexpected patterns in how often ribosomes are found downstream of specific codon identities in experiments using CHX pretreatment. Together, these observations suggest that translation elongation continues for many cycles after the introduction of CHX, but that the amount of time ribosomes spend translating each codon under these perturbed conditions is quite different from the unperturbed dynamics.

4.2 Results

4.2.1 Treatment with cycloheximide consistently changes enrichments of codon identities at ribosomal tRNA binding sites

To characterize differences in the measured positions of ribosomes when cells are pretreated with CHX and when they are not, we compared the relative amount of time ribosomes spend at each codon position in data from many different ribosome profiling experiments. For each experiment, we mapped footprint sequencing reads to yeast coding sequences and assigned each read to the codon position in the A-site of the associated ribosome (the sixth codon from the 5' end in a canonical 28nt footprint [46]). The raw count of reads assigned to each codon position is affected by levels of transcription and of translation initiation as well as the average speed with which the codon position is translated. As a simple control for these expression level effects, for each coding sequence, we divided the read count at each codon position by the average across the coding sequence, producing relative enrichment values for each codon position. To measure the average relative amount of time a codon identity spends in the A-site of a ribosome each time it is translated, we computed the mean of these relative enrichment values at all occurrences of each codon across all yeast coding sequences (figure 4.2A). If the mean relative enrichment of a codon is higher than 1, translation of the codon is inferred to be slower than the average speed of its surroundings. Conversely, a mean relative enrichment value lower than 1 indicates that translation of a codon is faster than its surroundings.

This straightforward computation ignores the potentially confounding influence of patterns in amino acid composition and synonymous codon usage across different genes [30, 89]. To evaluate the impact of ignoring these effects, we considered a more sophisticated process for inferring relative elongation times that takes the full sequence of codons of each gene into account; see section 4.3.4 for details. Mean relative enrichment values for each codon identity produced by the naive computation agreed almost perfectly with maximum likelihood estimates of mean relative elongation times from this more principled inference process (figure 4.1), suggesting that these particular biases are negligible for our purposes.

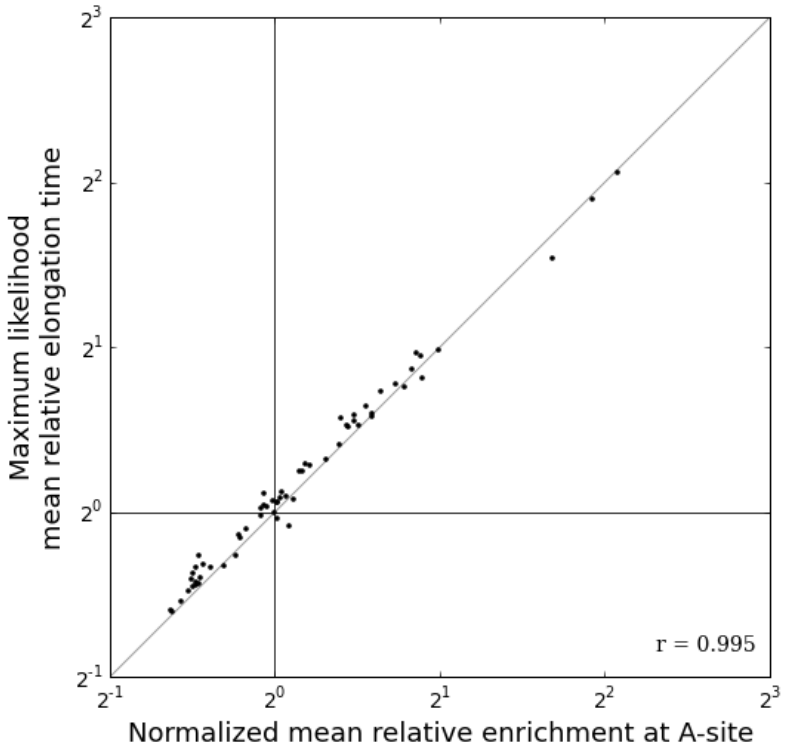

Figure 4.1: **Biases introduced by failing to account for the exact codon composition of each gene are a minor effect in A-site occupancy estimates.**
In data from an experiment by Weinberg [115], maximum likelihood mean relative elongation times for each codon identity computed using a model that accounts for the full codon sequence of each gene are virtually identical to mean relative enrichments at the A-site for each codon identity. Because the maximum likelihood values are only determined up to an arbitrary scaling factor, they are scaled so that AAA = 1. For this comparison, mean relative enrichment values are normalized in the same way.

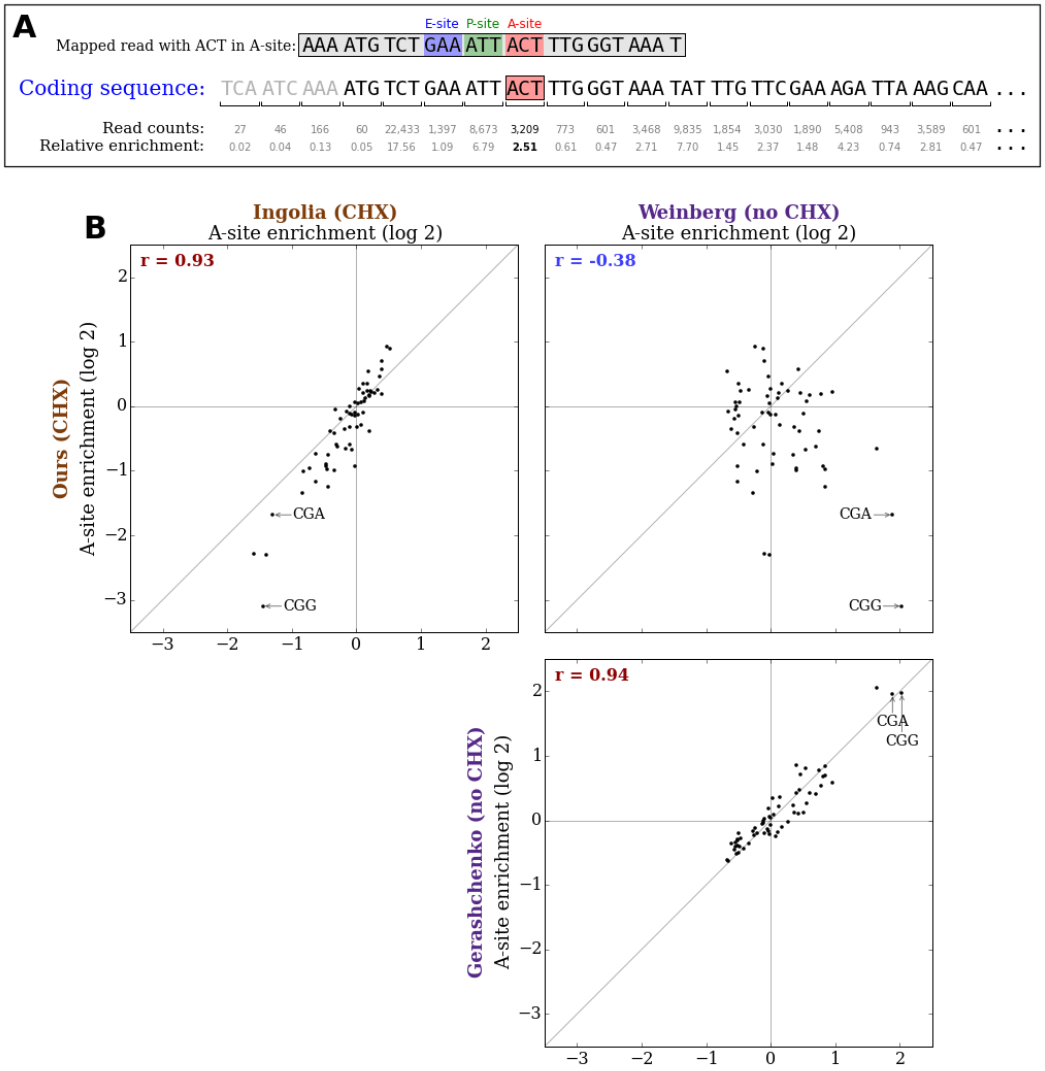

Figure 4.2: **Experiments with and without CHX report different A-site occupancies.**

Figure 4.2 (Continued): **Experiments with and without CHX report different A-site occupancies.**
(A) To measure how long a ribosome spends on average with each codon identity in its A-site, footprint sequencing reads are mapped to the yeast transcriptome and assigned to the codon position that the A-site of the ribosome was positioned over. For each coding sequence, read counts at each position are divided by the average count across the coding sequence to produce relative enrichments. The A-site occupancy of each codon identity (in this example, ACT) is computed by averaging the relative enrichment at all occurrences of the codon identity across all coding sequences.
(B) Comparisons of measured A-site occupancies of all 61 non-stop codons between different experiments. A pair of experiments using CHX (upper left) and a pair of experiments done without CHX (lower right) report A-site occupancies with strong positive Pearson correlations, but these two internally consistent sets of values are strikingly different from each other (upper right).

We compared the mean relative A-site enrichments for all 61 non-stop codons between the original CHX-pretreatment data of Ingolia [46], an experiment we performed using the same CHX-pretreatment protocol, and data from experiments without CHX pretreatment by Gerashchenko [31] and by Weinberg [115] (figure 4.2B). A-site occupancies are strongly positively correlated between experiments that use CHX pretreatment (upper left panel) and between experiments that do not (lower right panel). The two sets of values reproducibly reported by each experimental protocol are inconsistent with each other, however, with a moderate negative correlation between them (upper right panel). To test the generality of these comparisons, we computed Pearson correlations between the A-site occupancies in representative experiments from many different studies in yeast and performed unsupervised hierarchical clustering on the resulting matrix of correlation values (figure 4.2B). Experiments with and without CHX pretreatment separate into two distinct clusters, confirming that the two experimental conditions produce two reproducible but different pictures of translation dynamics. We note that somewhat greater variability is observed between subclusters of the experiments without CHX. A subset of these experiments correlate weakly positively, rather than weakly negatively, with the CHX experiments; see discussion in section 4.2.8 below.

The computation of mean relative enrichment at the A-site described above can be naturally generalized to measure the impact on elongation times of the codon identity situated in the P- or E-sites of ribosomes. We computed

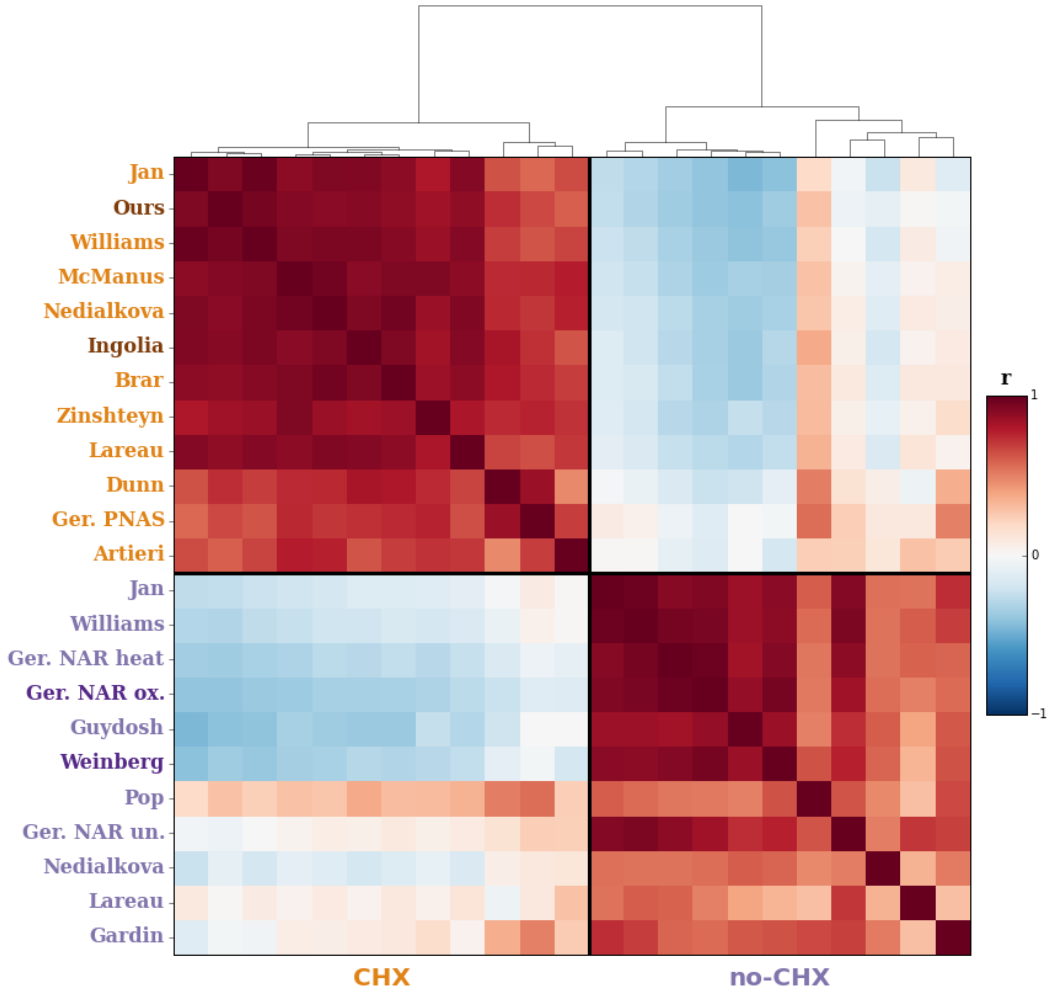

Figure 4.3: **Hierarchical clustering of A-site occupancies separates experiments by protocol.**
Pearson correlations of measured A-site occupancies between representative experiments from many different studies in yeast, grouped by hierarchical clustering. Clustering separates experiments using the standard CHX pretreatment protocol (labeled in orange) from experiments done without CHX pretreatment (labeled in purple), confirming the generality of the conclusion in Figure 4.2B. Darker labels of each color correspond to those samples compared in (B). Clusterings were computed via UPGMA using Euclidean distances.

the average P- or E- site occupancies of a codon identity by taking the mean of the relative enrichment values at all positions one codon (P-site) or two codons (E-site) downstream of an occurrence of the codon identity (figure 4.4A and 4.5A). Clustering the same set of experiments by P-site occupancy values recapitulates the groupings produced by A-site occupancies almost identically (figure 4.4). Clustering by E-site occupancies also separates experiments with and without CHX pretreatment, but there is less dynamic range in the E-site occupancy levels of different codon identities (figure 4.5B) and less consistency within experimental condition (figure 4.5C) compared to the A- and P-sites.

4.2.2 CHX-induced changes in ribosomal A- and P-site enrichments are concentration dependent

The fact that tRNA binding site enrichment values from experiments with and without CHX pretreatment separate into two clusters represents two incompatible claims about how long ribosomes spend translating each codon identity. To test how well each of these two apparent phenotypes agreed with intuitive expectations about elongation times, for each experiment, we computed the Spearman rank correlation between each codon identity's mean relative A-site enrichment and the inverse of its tRNA adaptation index (tAI) [23, 111]. The tAI of each codon identity is the weighted sum of the genomic copy numbers of the different tRNA genes that can decode the codon identity, with empirically determined weights penalizing wobble base pairings. This calculation quantifies the expectation that tRNAs expressed at lower abundances or that involve non-standard base pairing in their codon-anticodon

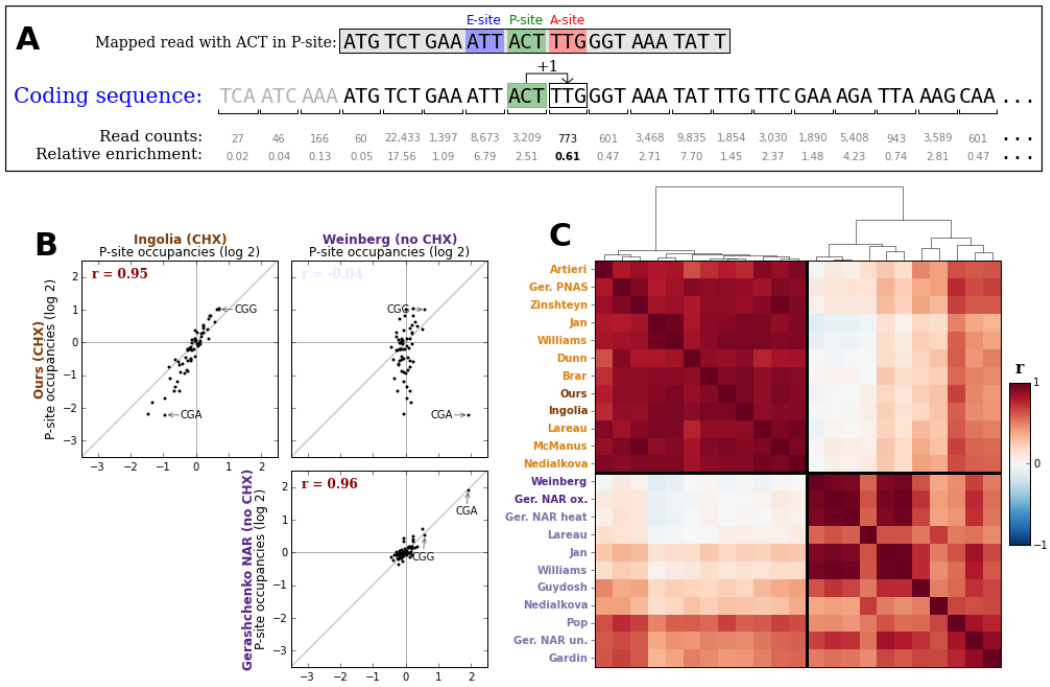

Figure 4.4: **Comparisons of P-site occupancies between experiments.** **(A)** To measure how frequently ribosomes are observed with a particular codon identity (in this example, ACT) in the P-site, marginalizing over all other sequence features, the mean of the relative enrichments at all codon positions one codon downstream of an occurrence of the codon identity is computed. Panels **(B)** and **(C)** are constructed as in figures 4.2 and 4.3, respectively, but report P-site enrichments. Clustering by P-site occupancy separates CHX (orange) from no-CHX (purple) experiments, but there is substantially less dynamic range in the P-site occupancies of different codon identities in no-CHX experiments than in CHX experiments. Relative P-site occupancies in no-CHX experiments are tightly grouped around one, with the sole exception of CGA, which is consistently a high occupancy outlier.

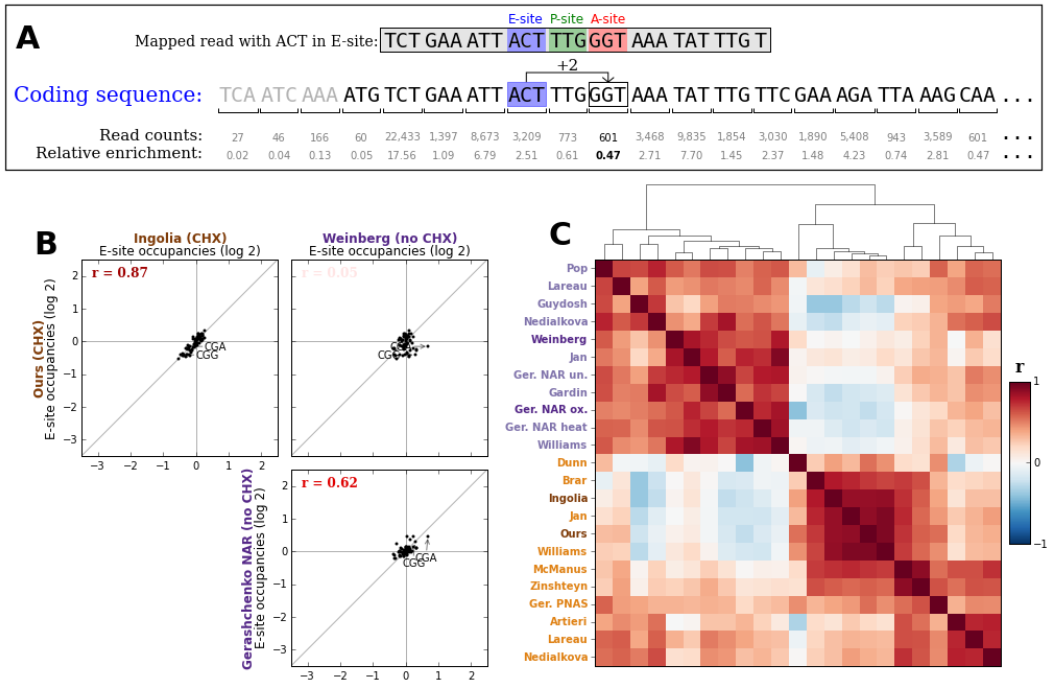

Figure 4.5: **Comparisons of E-site occupancies between experiments.** **(A)** To measure how frequently ribosomes are observed with a particular codon identity in the E-site (in this example, ACT), the mean of the relative enrichments at all codon positions two codons downstream of an occurrence of the codon identity is computed.
Panels **(B)** and **(C)** are constructed as in figures 4.2 and 4.3, respectively, but report E-site enrichments. E-site occupancies cluster by experimental condition but have have little dynamic range in either experimental condition and less coherence within experimental condition compared to the A- and P-sites.

interaction should require longer to translate. Consistent with previous reports [3, 16, 92, 121], all CHX experiments report weak to moderate negative correlations (figure 4.3C, orange labels), representing apparent translation dynamics in which less abundant tRNAs are actually translated faster. Experiments without CHX, on the other hand, report positive correlations of varying magnitude (figure 4.3C, 0x Gerashchenko NAR points and purple labels). Experiments by Pop [89], Lareau [64], Nedialkova [81], Guydosh [39] and Gardin [30] produce weak to moderate correlations, but experiments by Gerashchenko [31], Jan [50], Williams [117], and Weinberg [115] produce fairly strong and highly statistically significant correlations.

Serendipitously, a series of experiments by Gerashchenko[31] performed to measure the effect of CHX concentration on the observed ramp in ribosome density at the 5' end of coding sequences provide a way to confirm that CHX is directly responsible for these contradictory results. Gerashchenko produced datasets using pretreatment with a gradient of seven different CHX concentrations (0x, 1/64x, 1/16x, 1/4x, 1x, 8x, and 100x, expressed in multiples of the original protocol's concentration of 100 μg/ml) for two different cellular conditions (unstressed and oxidatively stressed cells), and using two different concentrations (0x and 1x) for heat shocked cells. Intriguingly, the rank correlation of A-site enrichment with 1 / tAI in these experiments moves smoothly from moderately negative with the highest CHX concentration to strongly positive with no CHX across each sets of samples (figure 4.3C), with only one sample (1/16x unstressed) deviating from perfect monotonicity. This

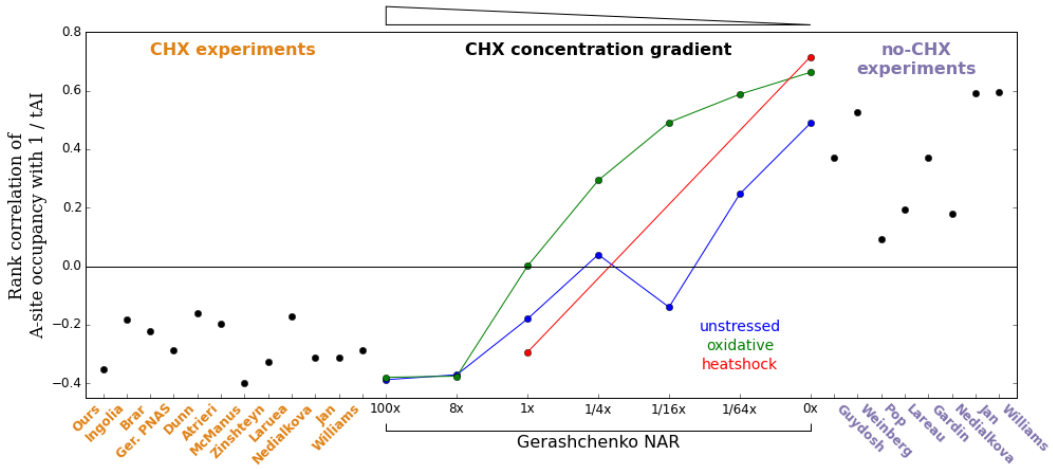

Figure 4.6: **Rank correlation of A-site occupanices with 1 / tAI is disrupted by CHX.**
Spearman rank correlations of codon identities' A-site occupancies with the inverse of tRNA adaption index (tAI) in different experiments. A positive correlation represents translation dynamics in which codons decoded by less abundant tRNAs take longer to translate, while a negative correlation implies that codons decoded by less abundant tRNAs are counter-intuitively faster to translate. Experiments using the standard CHX pretreatment protocol report a negative correlation, experiments done without CHX pretreatment report a positive correlation, and three sets of experiments across a gradient of CHX concentrations produced by Gerashchenko[31] each interpolate between these two phenotypes.

concentration-dependence is a strong confirmation that CHX systematically shifts statistical properties of where ribosomes are measured.

To further explore the effect of CHX on A-site occupancies, we plotted the movement of the mean relative A-site enrichment of each codon identity across the concentration gradient. Figure 4.7 shows data from the oxidatively stressed set of samples and figure 4.8 shows data from all three sets. Strikingly, a set of the codons with the highest enrichments (that is, the codons that are slowest to translate) when there is no CHX undergo consistent, gradual depletion with increasing concentration until they become among the fastest. Two prominent examples are CGA and CGG, codons encoding arginine. Mean relative enrichment at CGA codons is approximately four with no CHX, but this steadily decreases to a final value of less than 1/2 at the highest CHX concentration. CGA is translated by a tRNA identity with a moderate genomic copy number. However, its first anti-codon nucleotide is postranscriptionally modified to an inosine, making it the only codon in yeast that is decoded exclusively by an I-A wobble pairing [86]. Several studies have demonstrated that this leads to substantial translational pausing at occurrences of CGAs, particularly at CGA-CGA dicodons [67, 85, 119]. CGG, which is decoded by a tRNA with only a single genomic copy and therefore also expected to be slowly translated, undergoes an even larger shift from apparently slow with no CHX to apparently fast with high CHX concentration. For both codons, a CHX-concentration-dependent, and therefore almost certainly CHX-mediated, mechanism drives measured translation speeds away from their intuitively ex-

pected values.

We also examined changes in occupancy at the P- and E-sites across the concentration gradient (figure 4.8). A smaller number of codons undergo substantial changes in mean relative P-site enrichment, with the dominant effect being a dramatic reduction in CGA enrichment with increasing CHX concentration. Compared to the A- and P-sites, there is less concentration-linked change in occupancy at the E-site.

4.2.3 Experiments using CHX exhibit consistent patterns in ribosome density downstream of different codon identities

The tendency for ribosomes to be found at any particular offset upstream or downstream of a particular codon identity can be measured by computing the mean of the relative enrichments at all codon positions which are located exactly that offset away from an occurrence of that codon identity. The A-, P-, and E-site occupancies discussed above are the special cases of offsets of 0, +1 and +2 downstream, respectively, but these computations can be generalized to offsets arbitrarily removed from the codon identity of interest. We computed mean enrichment values for a wide range of offsets around each codon identity in data from our (CHX pretreatment) experiment. Although there is no *a priori* biological reason to expect mean enrichments to deviate substantially from one at offsets that are far removed from tRNA binding sites, we unexpectedly observed prominent peaks and dips in enrichments downstream of many codon identities. Figure 4.9B shows enrichment profiles

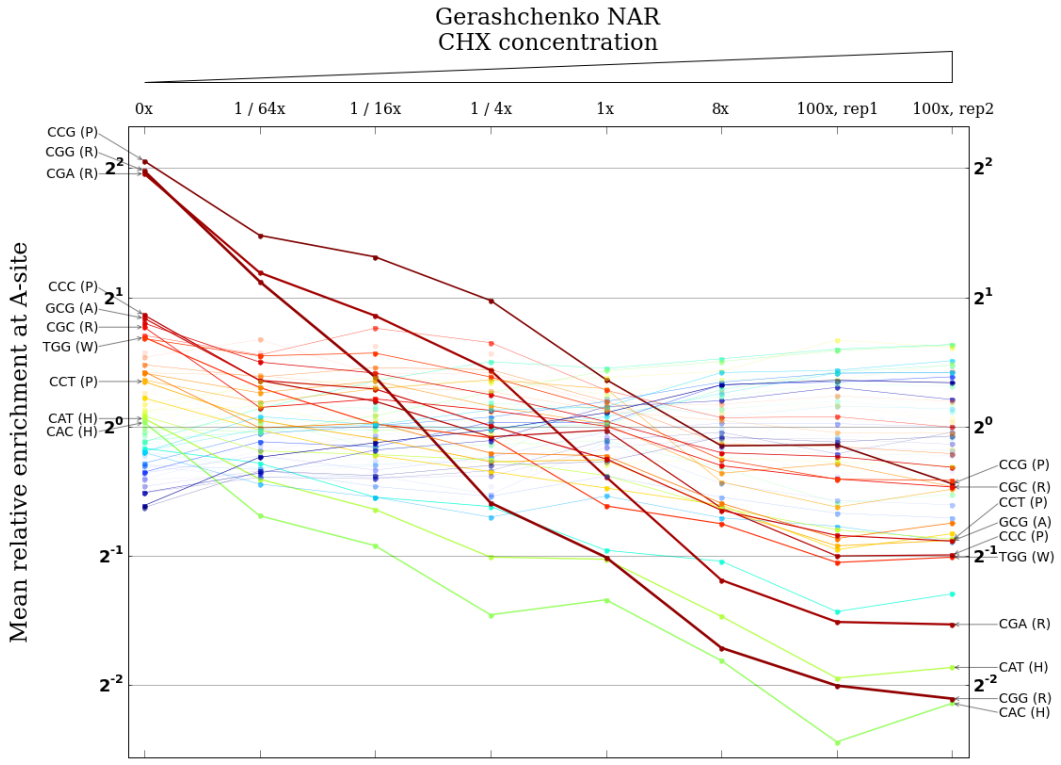

Figure 4.7: **CHX treatment affects A-site occupancies in a coherent concentration-dependent manner.**
Columns correspond to a series of experiments by Gerashchenko [31] using a gradient of concentrations of CHX, starting from no CHX on the far left and increasing to 100 times the standard concentration on the far right. Each column plots the measured A-site occupancies of all 61 non-stop codons for that concentration on a log scale. The width of the line connecting each codon identity across concentrations is scaled by the codon's net change from no CHX to 100x CHX. The ten codon identities with the largest net changes are labeled. Most notably, the codon identities with the highest enrichments in the experiment with no CHX undergo dramatic, concentration-dependent depletions over the course of the gradient.

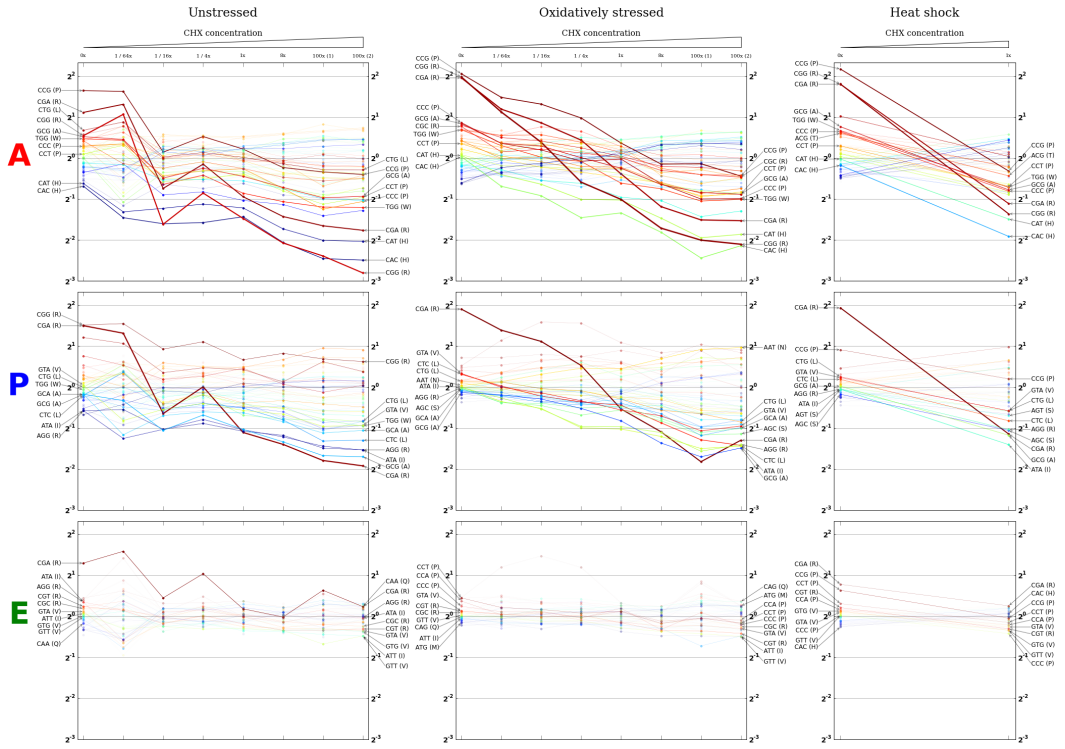

Figure 4.8: **A-, P-, and E-site occupancy changes across CHX concentration gradients.**
Each panel is constructed as in figure 4.7. Each row reports occupancies of a different tRNA binding site (top, A-site; middle, P-site; bottom, E-site). Each column reports occupancies for samples from Gerashchenko [31] under different conditions (left, unstressed; middle, oxidatively stressed; right, heat shock).

around codon identities encoding arginine and figure 4.10 shows enrichment profiles around codon identities for several other individual amino acids.

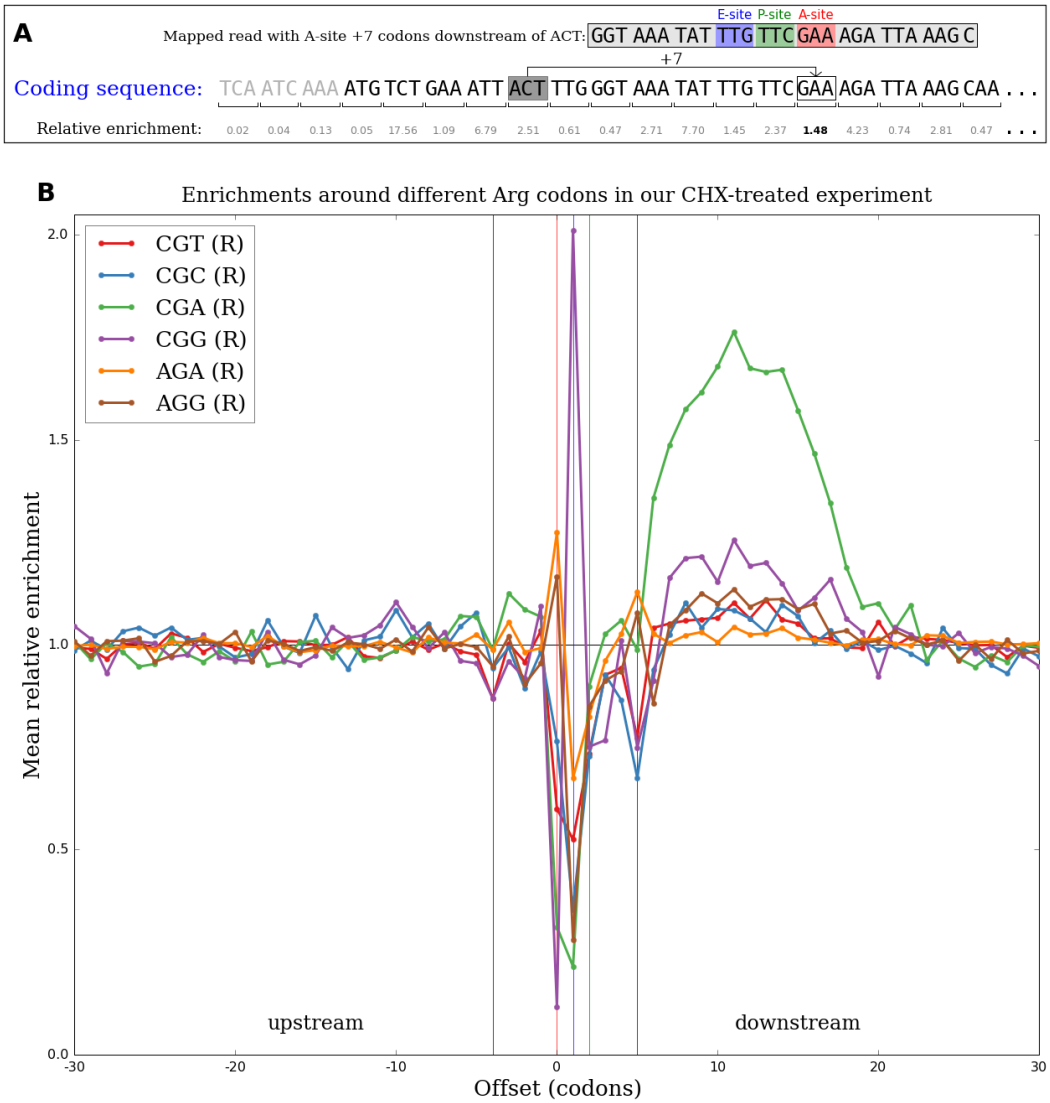

Figure 4.9: **Experiments using CHX exhibit patterns in ribosome density downstream of different codon identities.**

Figure 4.9 (Continued): **Experiments using CHX exhibit patterns in ribosome density downstream of different codon identities.**
(A) To measure how frequently ribosomes are observed with their A-site positioned at a particular offset upstream or downstream of a given codon identity (for example, 7 codons downstream of ACT, like the boxed footprint sequence on the top line), relative enrichments at each position are first calculated as in figure 4.3A. The enrichment values at all positions located exactly 7 codons downstream of an ACT (such as the bolded value) across all coding sequences are then averaged.
(B) Profiles of mean relative enrichments at a range of offsets around all six arginine codons in our CHX-pretreatment experiment. Grey vertical lines mark the boundaries of a canonical 28 nt footprint, and red, blue, and green vertical lines (corresponding to offsets of 0, +1, and +2) mark the A-, P-, and E-sites. Unexpected peaks of enrichment at downstream offsets outside of the grey lines are observed. The magnitudes of peaks vary substantially between different codon identities encoding the same amino acid, but the horizontal extents of peaks are roughly the same across all codon identities.

After observing these peaks in our data, we examined data from many other experiments in yeast for evidence of similar peaks. Peaks are ubiquitous in data from experiments using CHX pretreatment (blue, green, and purple lines in Figures 4.11 and 4.14), but are almost entirely absent in data from experiments that do not use CHX pretreatment (Figure 4.13, and red lines in Figures 4.11 and 4.14, but see discussion of Pop et al. data below). For data from a particular experiment, the peaks corresponding to different codon identities occupy roughly the same range of offsets downstream (Figures 4.9B and 4.10), but across experiments carried out by different groups, the locations and shapes of the set of peaks change considerably (figure 4.11). The centers of peaks vary from as close as ∼10 codons downstream in our data to as far away as ∼50 codons downstream in data from McManus [77], with other CHX experiments densely populating the range of offsets between these observed extremes. Peaks become broader in width and smaller in maximum magnitude the further downstream they are located.

To test if CHX treatment had a concentration-dependent effect on the locations and shapes of these peaks, we again turned to data from the CHX concentration gradient experiments of Gerashchenko. Figure 4.14 shows enrichment profiles downstream of CGA in the oxidatively stressed series of samples. Peaks were absent in the samples with no CHX and minimal in the samples with concentrations below 1/4x the standard concentration (with the notable exception of unstressed 1/16x, which is also a clear outlier in figures 4.3C and 4.8). For samples with concentrations greater than or equal to 1x,

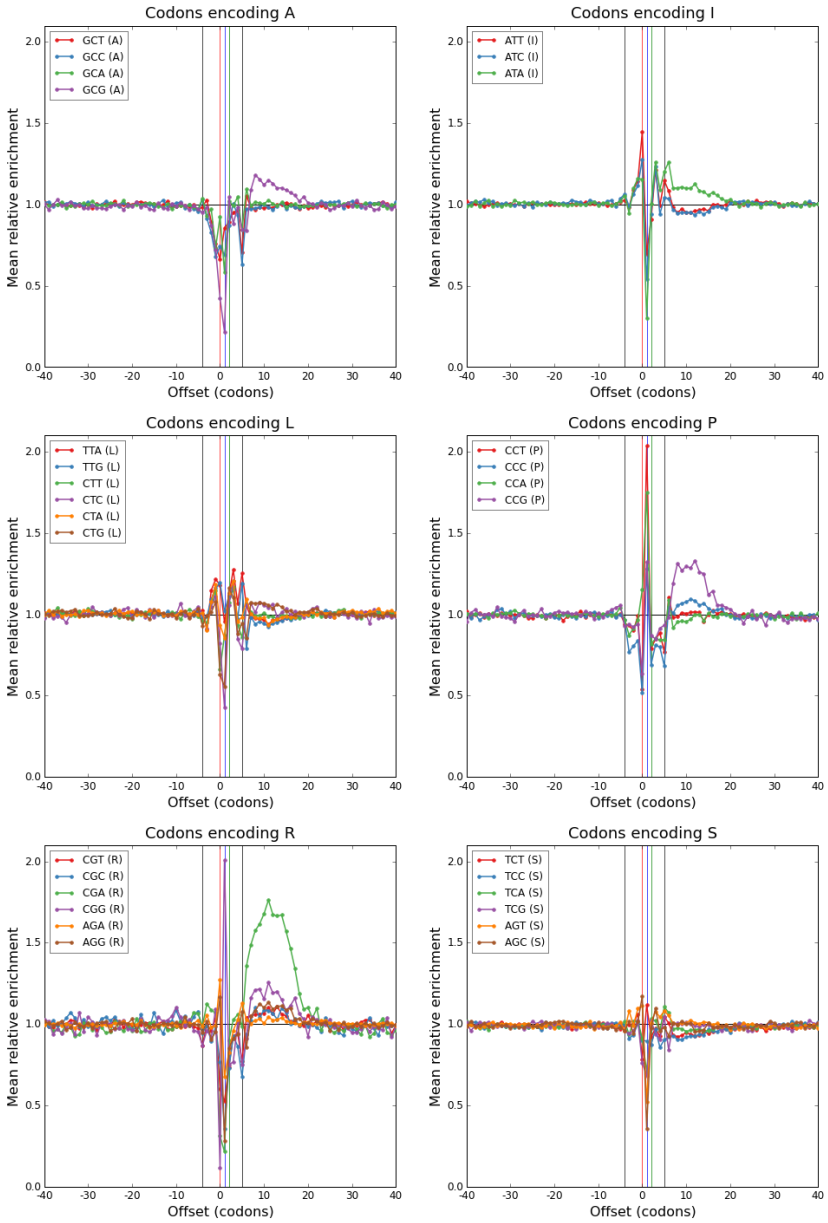

Figure 4.10: **Enrichment profiles around codons for different amino acids in our CHX-pretreatment experiment.**
Each panel is constructed as in figure 4.9B but shows codons encoding a different amino acid. Several amino acids show substantial variation in the magnitude and direction of downstream peaks between different codons.

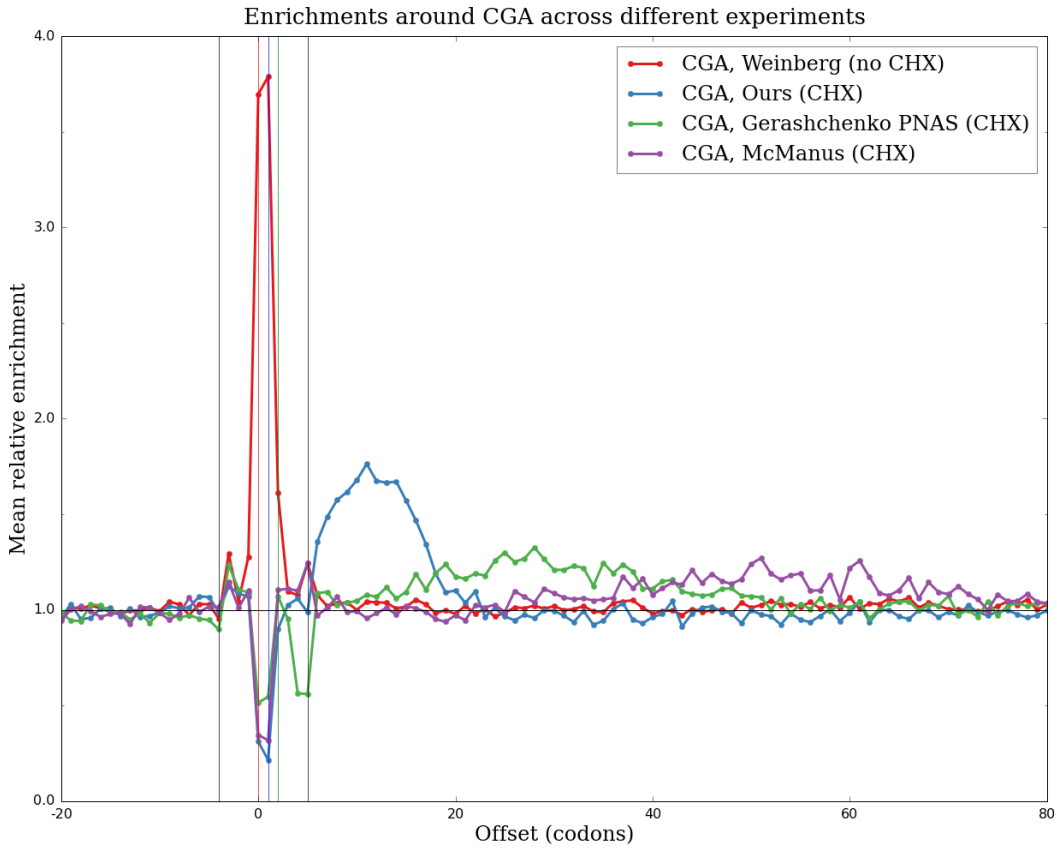

Figure 4.11: **Locations of downstream waves vary between experiments from different studies.**
Profiles of mean relative enrichments around a single codon identity (CGA) in experiments from different studies. There is no downstream peak in the no-CHX experiment of Weinberg et al. (red), but downstream peaks are ubiquitous in experiments using CHX (all other colors). Peaks are centered at a wide range of different offsets in CHX experiments by different groups and become broader and lower when located farther downstream.

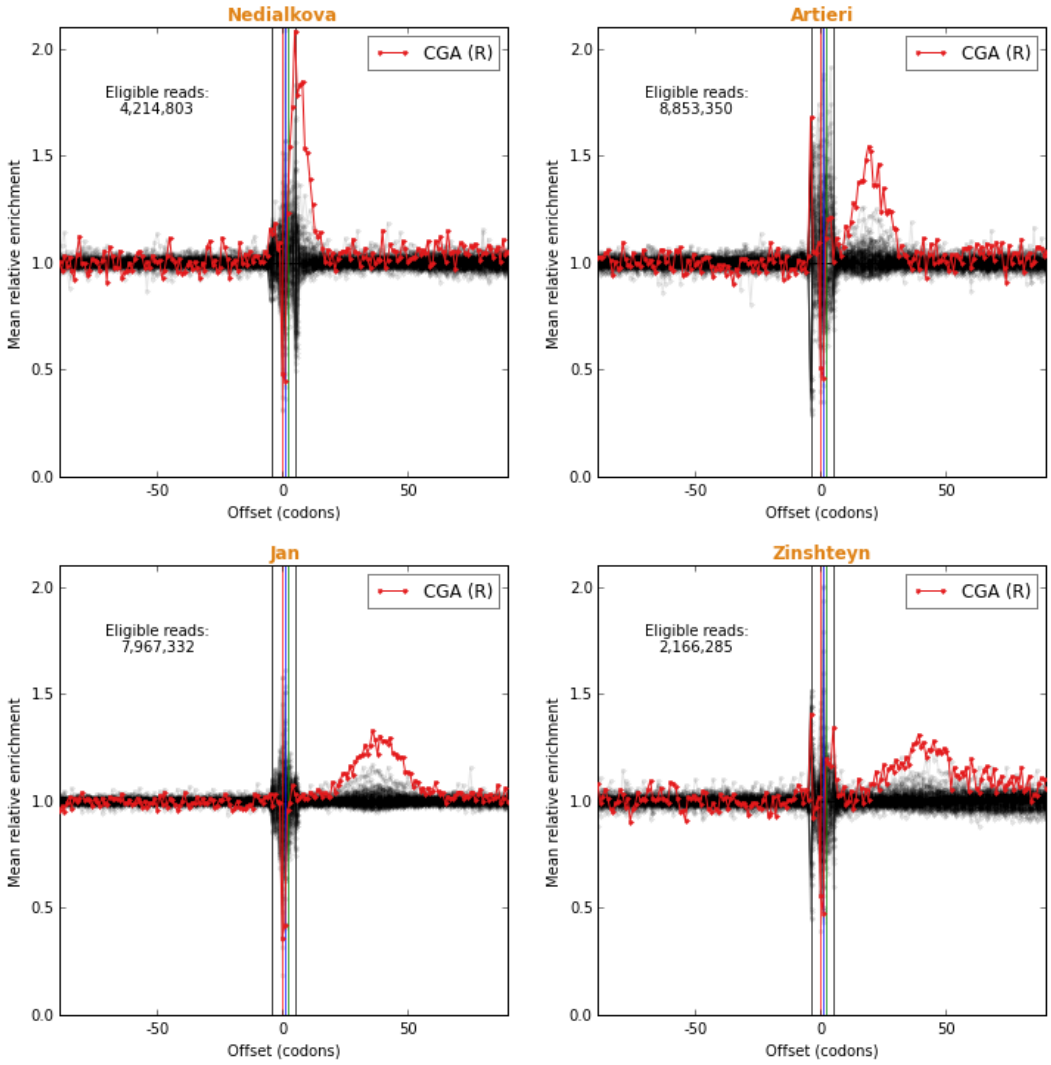

Figure 4.12: **Downstream peaks in representative experiments using CHX pretreatment from additional studies.**
Each panel shows enrichment profiles around all 61 non-stop codons for an experiment using CHX pretreatment from a different study, with CGA highlighted in red. Data sources are given in 4.24. Additional experiments confirm that clear downstream peaks are a ubiquitous feature of CHX-pretreatment, but that the exact range of offsets occupied by the peaks varies considerably between experiments.

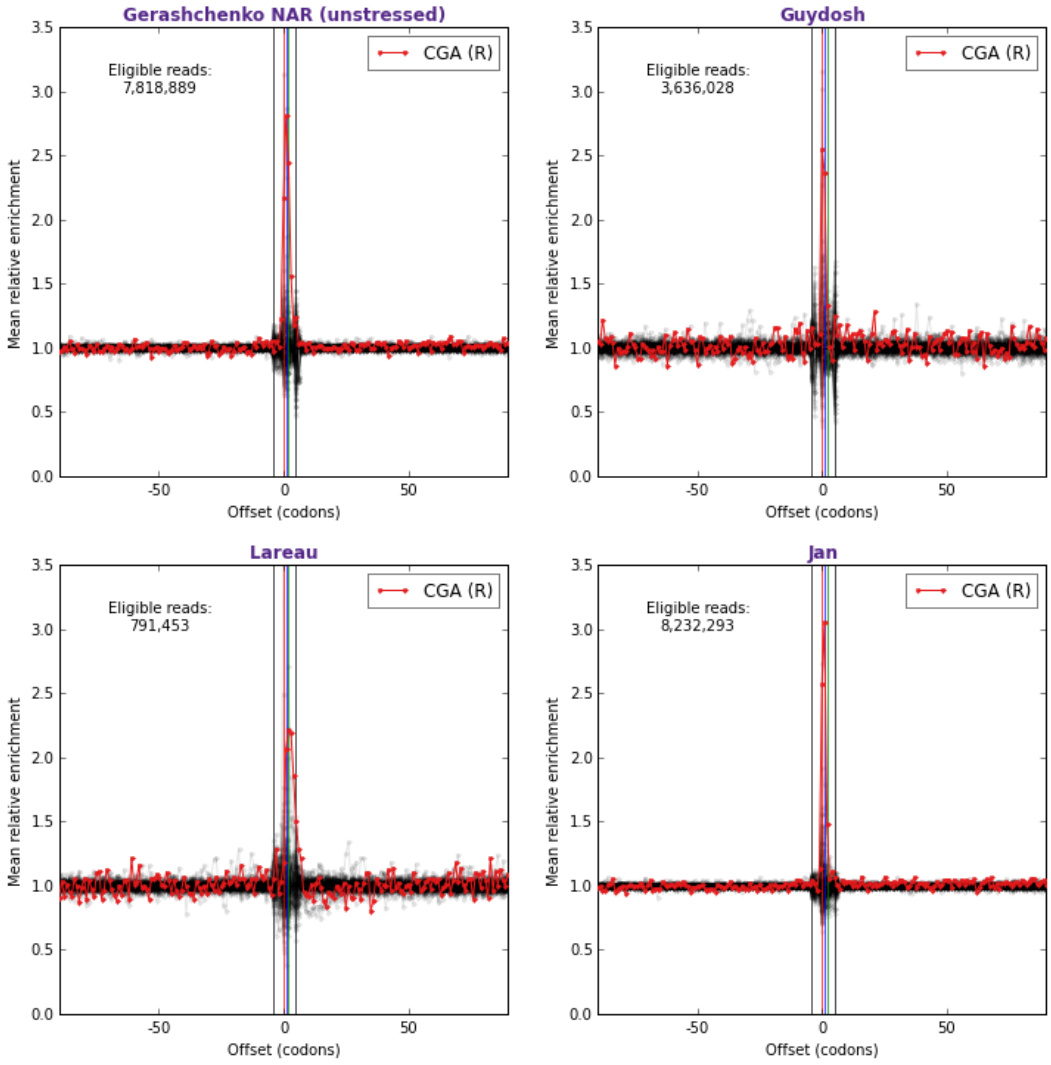

Figure 4.13: **Downstream peaks for representative samples from additional no-CHX experiments.**
Figure is constructed as in figure 4.12 but panels shows experiments performed without CHX pretreatment from different studies. There are no appreciable downstream peaks in any such experiment. In the two experiments shown here that correspond to rows in Figure 4.3 that correlate less stongly with the main cluster of no-CHX experiments (Gerashchenko NAR unstressed and Lareau), however, a slight downstream shift in the CGA profiles can be seen. See section 4.2.8 below for a discussion of this point.

for which clear peaks were observed, peaks are located less far downstream and become narrower and taller as CHX concentration increases (figure 4.14).

Other studies[16, 17] have hypothesized that interactions between recently incorporated amino acids and the ribosome exit tunnel lead to slower ribosome movement downstream of occurrences of certain amino acids. There are two lines of evidence that the downstream peaks observed here are not simply the result of these effects. First, within a single sample, the magnitudes of peaks vary substantially between different codons encoding the same amino acid. As examples, the peak downstream of CGA in our experiment is substantially higher than the peaks downstream of other codons encoding arginine (figure 4.9B), and GCG is the only codon encoding alanine for which there is an appreciable downstream peak (figure 4.10). If these peaks were caused by interactions involving an amino acid in the nascent polypeptide chain, they should be agnostic to the codon identity used to encode the amino acid. Second, the facts that the locations of peaks change in response to changes in CHX concentration and that peaks disappear in the absence of CHX strongly suggest that the peaks are a consequence of CHX treatment rather than a genuine feature of translation.

4.2.4 Disrupting steady-state elongation rates causes downstream peaks in analytical and simulation models

Having observed large shifts in tRNA binding site occupancies between experiments with and without CHX and the appearance of downstream peaks

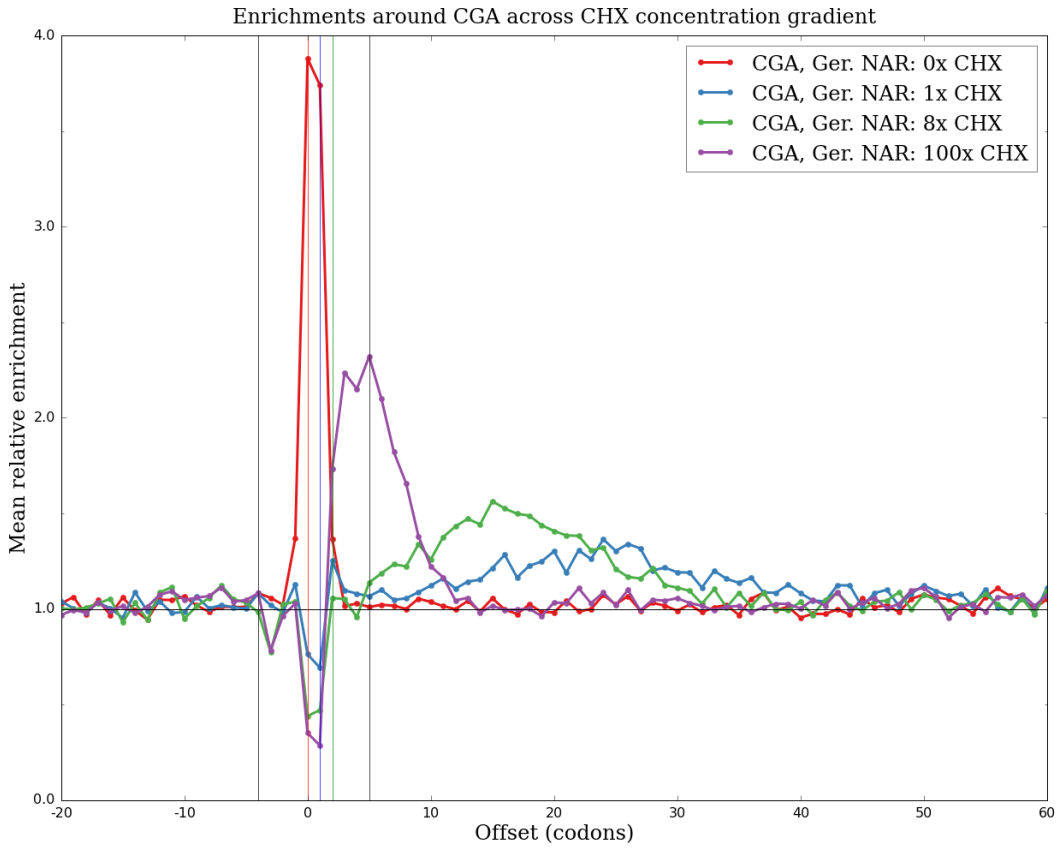

Figure 4.14: **Locations of downstream waves move coherently in response to CHX concentration.**
Profiles of mean relative enrichments around a single codon identity (CGA) in experiments by Gerashchenko[31] using different concentrations of CHX. With no CHX (red), there is a strong enrichment at the A- and P-sites and no downstream peak. With CHX (all other colors), there are depletions at the A- and P-sites and downstream peaks that become closer, narrower, and higher with increasing concentration.

in CHX experiments, we sought a model for how CHX treatment disrupts the measured positions of ribosomes that could parsimoniously explain both phenomena. To test potential models, we developed a software simulation of the movement of ribosomes along the repertoire of yeast coding sequences; see section 4.3.3 for simulation details. By incorporating different possible mechanistic effects of the introduction of CHX into these simulations, we could evaluate the ability of different models to explain the observed features of the experimental data.

A natural first hypothesis is that each ribosome waits an exponentially distributed amount of time until a CHX molecule diffuses into the ribosome's E-site and irreversibly arrests it, with the timescale of this exponential varying inversely with CHX concentration. If ribosomes continue to spend the same relative amounts of time on each codon identity while waiting for CHX to arrive, however, the position of each ribosome at the random instant of CHX arrival samples from the same steady state distribution that ribosomes occupied before CHX was introduced. We confirmed by simulation that this potential mechanism produces neither downstream peaks nor substantial changes in A-site occupancies.

Next, we noticed that the codon identities that undergo the largest changes in measured A- and P-site occupancies between no-CHX and CHX experiments consistently tend to be the same codon identities that exhibit the largest downstream peaks. The most obvious example is CGA, which has the largest downstream peak in virtually all CHX experiments and also has both

one of the largest decreases in A-site occupancy and by far the largest decrease in P-site occupancy over the increasing CHX concentration gradients. This potential link between binding site changes and downstream peak size suggested a model in which elongation continues in the presence of CHX but with codon-specific relative elongation rates that are substantially changed from their pre-CHX treatment values. To model this possible behavior, we evolved a mechanistic simulation of translation to steady state with the mean relative elongation time of each codon identity set to its mean relative A-site enrichment in the no-CHX experiment of Weinberg [115]. We then instantaneously switched the relative elongation time of each codon identity to its mean relative A-site enrichment in our CHX-pretreatment experiment and allowed translation to proceed under these new elongation rates for a short period of time. At the end of this short period of continued elongation, we recorded the positions of all ribosomes and processed the resulting simulated ribosome footprints identically to the real experimental datasets. Interestingly, the resulting simulated enrichment profiles qualitatively reproduce both major phenomenon observed in data from CHX experiments (figure 4.15A).

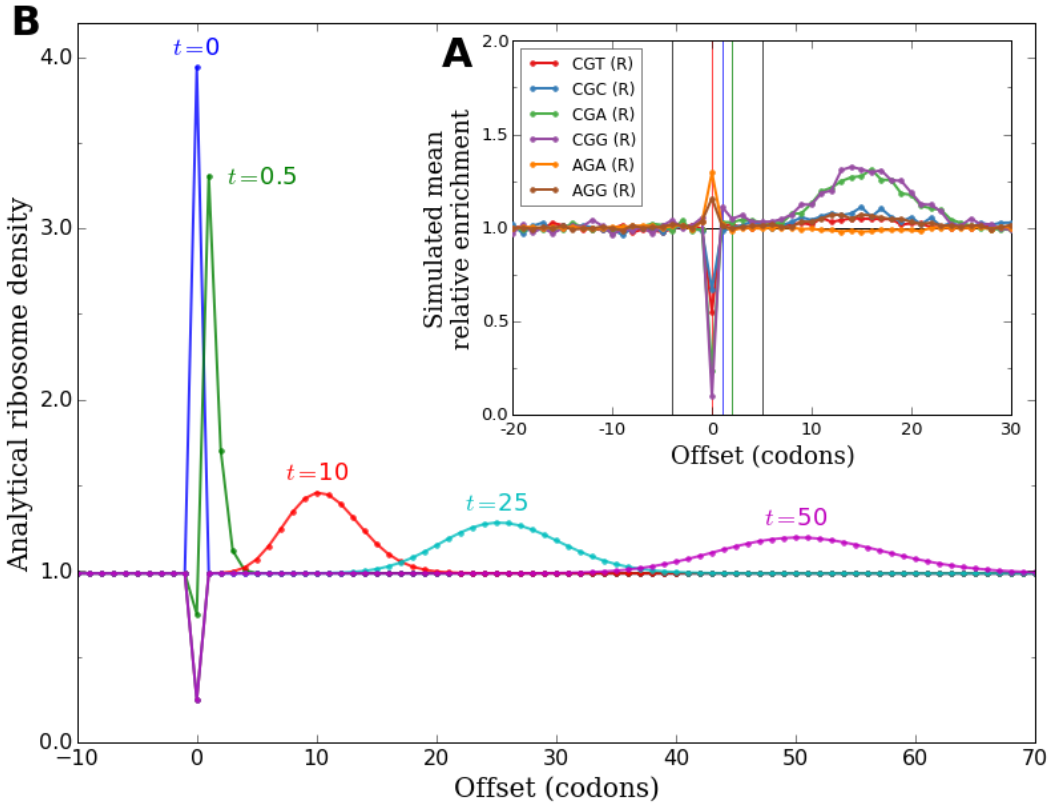

Figure 4.15: **A sudden change in the relative elongation rates of codon identities produces downstream waves in simulation and analytical models.**

Figure 4.15 (Continued): **A sudden change in the relative elongation rates of codon identities produces downstream waves in simulation and analytical models.**
(A) In a simulation of the translation of yeast coding sequences, the average relative elongation time of each codon identity was suddenly changed from the codon's A-site enrichment in a no-CHX experiment to its A-site enrichment in a CHX experiment. Elongation was allowed to proceed for a short time after this change, then enrichments in the positions of ribosomes were analyzed as in figure 4.9A. The resulting profiles of simulated mean enrichments qualitatively reproduce the downstream peaks in data from experiments using CHX.
(B) A continuous-time Markov chain model of the translation of a hypothetical coding sequence consisting of a single slow codon surrounded by long stretches of identically faster codons on either side was analyzed. At $t = 0$, the relative speed of the slow codon was suddenly changed to be faster than its surroundings. Ribosome density at offsets around the formerly-slow codon is plotted at several time points after this change. Immediately after the change, there is a temporary excess of ribosomes positioned at the formerly-slow codon relative to the eventual steady state of the new dynamics. As these excess ribosomes advance along the coding sequence, a transient wave of increased ribosome density moves downstream and spreads out over time.

To better understand why changing relative elongation rates shortly before measuring ribosome positions produces these patterns, we constructed a simple analytical continuous-time Markov chain model of the process of translation. See section 4.3.4 below for a more detailed description of this model. In this idealized model, ribosomes wait an exponentially distributed amount of time at each codon position before moving on to the next, with the rate parameter of this exponential distribution depending only on the codon identity in the A-site of the ribosome. Across many copies of a particular coding sequence being translated by many ribosomes, the instantaneous rate of flow of ribosome density from codon position i to codon position $i + 1$ is therefore equal to the current ribosome density at position i times the elongation rate of the codon identity at position i. This implies that the steady-state distribution of ribosome density at each position is proportional to the mean elongation time of the position; see section 4.3.4 for details.

In this analytical model, we considered a hypothetical coding sequence consisting of one codon that is translated slowly surrounded on either side by many identical copies of a codon identity that is translated faster. We numerically evolved the density of ribosomes on many copies of this coding sequence under these elongation dynamics to their steady state distribution. Then, at $t = 0$, we instantaneously changed the relative elongation rates of the two codon identities so that the previously slow codon was now faster than its surroundings. We plotted the evolution in ribosome density across the hypothetical coding sequence over time following this change (figure 4.15B). A

sufficiently long time after the change, the system will have reached the steady state distribution of the new elongation dynamics, in which ribosome density is lower at the now-faster codon than its uniform level at all of the surrounding codons. Immediately after the rates are changed, however, ribosomes are still distributed at the steady state densities implied by the relative speeds before the change and are therefore out of equilibrium under the new dynamics. There is a temporary excess of ribosomes at the formerly-slow codon. The process of relaxing from the old steady state to the new steady state manifests as these excess ribosomes advancing along the coding sequence over time. Stochastic variation in the exponential wait times of each individual ribosome at each subsequent codon position causes the excess to gradually spread out as it advances. Hypothetical measurements of the positions of all ribosomes at a series of increasing times after the change to the new dynamics would therefore produces patterns that look like an advancing wave of enrichment, as is seen around e.g. several arginine codons in real (figure 4.9B) and simulated (figure 4.15) data. Importantly, as this wave advances downstream, density at the formerly slow codon itself quickly equilibrates to its new steady state level.

We also considered a hypothetical coding sequence in which a single special codon undergoes an instantaneous increase, rather than decrease, in mean relative elongation time compared to stretches of identical codons on either side (figure 4.16). In this case, the time period immediately following the change in dynamics is spent filling the formerly-faster codon position up to its newly increased steady state density. During this time, there are temporarily

less ribosomes being promoted onward to downstream positions than there were before the change. This results in a transient wave of depletion, rather than enrichment, that advances away from the formerly-faster codon position and spreads out over time (figure 4.16A). This is qualitatively consistent with the profile of depletions downstream of e.g. two isoleucine codons in real (figure 4.10) and simulated (figure 4.16B) data.

4.2.5 Magnitudes of downstream peaks are quantitatively consistent with predictions made by wave hypothesis

The hypothesis that changes in measured tRNA binding site occupancies and the appearance of downstream peaks are both caused by continued elongation with disrupted dynamics in the presence of CHX makes a testable prediction about the quantitative link between these two phenomenon. If downstream peaks are transient waves moving downstream after a change in the relative amounts of time ribosomes spend positioned over each codon identity, the total CHX-induced excess or deficit in enrichment downstream of each codon identity should exactly offset the total CHX-induced change in enrichments at the tRNA binding sites. To test whether experimental data agreed with this prediction, we analyzed several matched pair of experiments performed with and without CHX by Jan [50] (figure 4.17), Williams [117], and Gerashchenko [31]. For each codon identity, we compared the sum of the differences in enrichment between the experiments at the A-, P-, and E-sites (green area in insets) to the sum of the difference in enrichment across the range of downstream offsets occupied by the putative waves (red area in

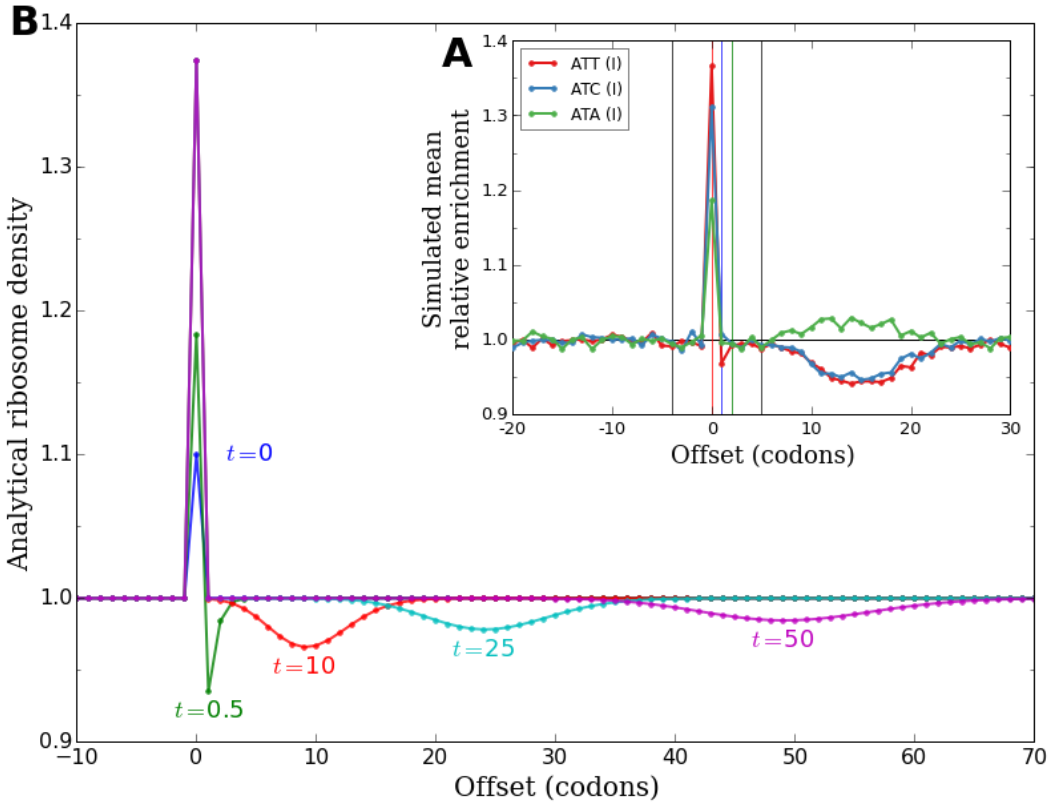

Figure 4.16: **Decreasing the relative elongation rate of a codon creates downstream waves of depletion.**
(A) In a simulation of translation, the average relative elongation time of each codon identity was changed from its A-site enrichment in the no-CHX experiment of Weinberg to its A-site enrichment in our CHX experiment. Allowing translation to proceed for a brief period of time under these new dynamics results in a negative peak of depletion downstream of those codons that become relatively slower in the new dynamics, such as ATT and ATC.
(B) In an analytical model of the translation of a single special codon surrounded by long stretches of codons that are identically slightly faster than it, suddenly changing the dynamics so that the special codon is even slower causes a transient wave of depletion in ribosome density to move downstream from the special codon over time.

insets). In all matched pairs of experiments, the area of each codon identity's downstream peak is strongly predicted by its tRNA binding site changes ($r^2 = 0.85$ to 0.93, slope of best fit line $\beta = -1.00$ to -1.20). Insets in figure 4.17 and 4.18 demonstrate the full dynamic range of the agreement between the two phenomenon. CGA, which undergoes comparably large decreases in enrichment at both the A- and P-sites, produces a downstream wave with approximately twice the area of CCG, which undergoes a large decrease in enrichment at the A-site but not the P-site. Codons with similar enrichments at all three tRNA binding sites between the two experiments, such as ACT, produce no appreciable downstream waves, while several codons that undergo modest increases in enrichment at the binding sites, such as TTG, produces proportionally modest net deficits of enrichment downstream. This close correspondence strongly suggests that the downstream peaks are in fact transient waves, and therefore that tRNA binding site enrichments in CHX experiments do not reflect natural translation dynamics.

4.2.6 Disrupted elongation in the presence of CHX explains counterintuitive results in CHX experiments

The fact that net changes in tRNA binding site enrichments between each pair of experiments with and without CHX match the net areas of downstream waves in the CHX experiment suggests that the areas of downstream waves can be used to recover indirect information about translation dynamics in each CHX experiment before these dynamics were disrupted by CHX. Specifically, given data from a CHX experiment, we can produce a reasonable

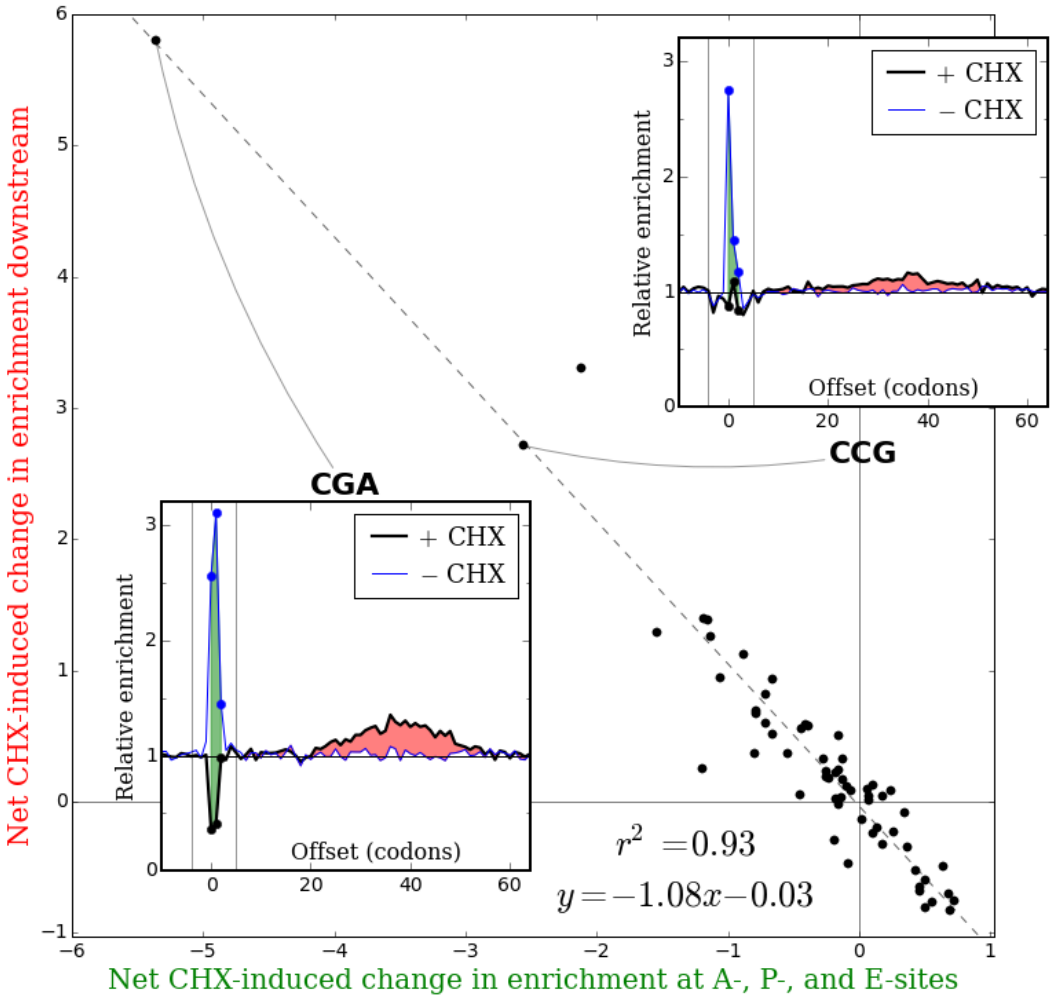

Figure 4.17: **Changes in tRNA binding site enrichments between a pair of experiments with and without CHX are matched by areas of downstream waves in the CHX experiment.**
For a pair of experiments with and without CHX by Jan [50], the sum of each codon identity's changes in mean relative enrichment at the A-, P-, and E-sites between the two experiments (green area in insets) is plotted against the total excess or deficit of enrichment in the CHX experiment from 6 to 65 codons downstream (red area in insets). The area of each codon identity's downstream peak is strongly predicted by changes in enrichment at the tRNA binding sites, consistent with the hypothesis that downstream peaks are transient waves caused by continued elongation with disrupted dynamics in the presence of CHX.

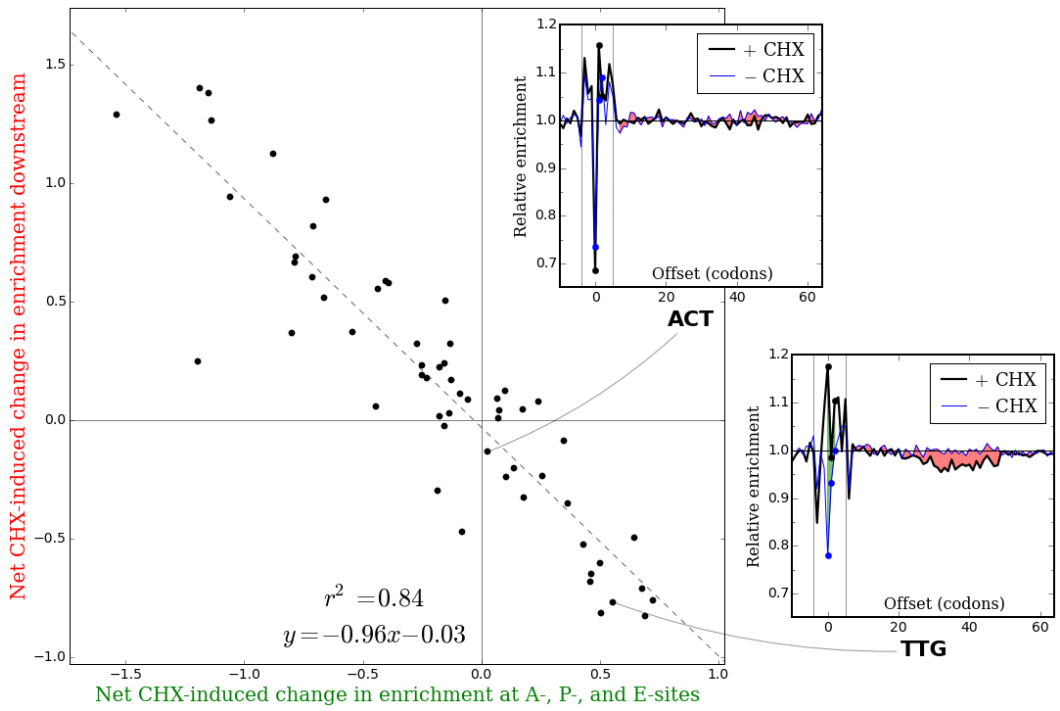

Figure 4.18: **Zoomed-in view of Figure 4.17, excluding CGA, CGG, and CCG.**
Figure is identical to 4.17 but excludes CGA, CGG, and CCG from the regression. Insets highlight examples of codons with no substantial change (ACT) or a moderate increase (TTG) in net tRNA binding site enrichments in the presence of CHX. The correlation between net tRNA binding site changes and downstream area remains strong ($r^2 = 0.84$) even after excluding the three codons that participate most in these phenomena.

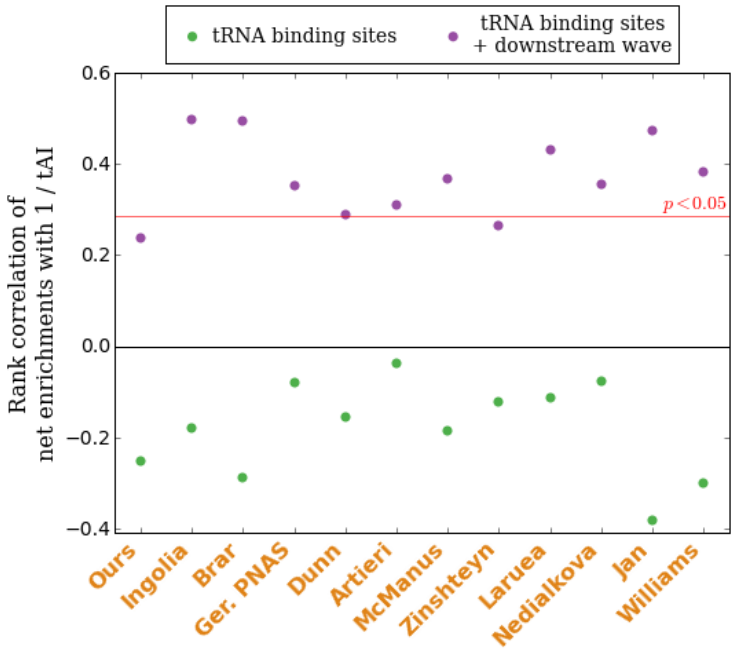

Figure 4.19: **Downstream waves recover positive correlations of estimated elongation times with 1 / tAI.**
In experiments using CHX, the combined enrichment of each codon identity at the A-, P-, and E-sites before the introduction of CHX can be indirectly estimated by adding the net area of the codon's downstream wave back to the total remaining enrichments at the three tRNA binding sites. This sum correlates positively with 1 / tAI in all CHX experiments (purple dots), recovering the positive correlations counterintutively absent at the tRNA binding sites alone in these experiments (green dots). Positive correlations are statistically significant ($p < 0.05$, one-tailed) in all but two experiments. This suggests that non-optimal codons were being translated less quickly than optimal codons in CHX experiments before the introduction of CHX disrupted these dynamics.

estimate of what the sum of the enrichments at the A-, P-, and E-sites was for each codon identity before the introduction of CHX by adding the net area of the wave that moved downstream during elongation in the presence of CHX back to the sum of the enrichments that remain at the binding sites. We will call this quantity the corrected aggregate enrichment of each codon. It can be interpreted as the average relative amount of time that a ribosome took to decode each occurrence of a codon before CHX was introduced, from when the codon was presented in the A-site to when it left the E-site. While we would of course prefer to recover how long each codon spent in each individual tRNA binding site in these experiments, this single-codon-resolution information has been irreversibly lost. As the CGA and CCG insets in figure 4.17 demonstrate, changes in enrichment at the A-site or at the P-site result in downstream waves that occupy the same large range of downstream offsets, so the area in each wave cannot be unambiguously assigned back to a particular tRNA binding site.

To test if codons decoded by less abundant or wobble-paired tRNAs tended to be translated more slowly than more abundant tRNAs in CHX experiments before the introduction of CHX, we computed the Spearman rank correlation between the corrected aggregate enrichment of each codon identity and 1 / tAI. Corrected aggregate enrichment correlates positively with 1 / tAI for every CHX experiment analyzed (figure 4.19, purple dots), recovering an intuitively expected signature of translation dynamics that is absent in CHX experiments if the total elongation time is estimated by the sum of the tRNA

binding sites enrichments alone (figure 4.19, green dots).

Continued elongation with disrupted dynamics after the introduction of CHX also offers a potential explanation for counterintuitive results in a set of experiments by Zinshteyn et al. [121]. Zinshteyn performed ribosome profiling on yeast strains that lacked different genes required to post-transcriptionally add mcm^5s^2 groups to a uridine in the anticodons of tRNAs that decode codons ending in AA and AG. These anticodon modifications are thought to enhance codon-anticodon recognition and speed up translation of these codons [94]. Surprisingly, Zinsheyn found that measured changes in tRNA binding site occupancies between deletion strains and the wild type were much smaller than expected given the phenotypic consequences of lacking these modifications. These experiments followed the standard CHX pretreatment protocol, however, and we observe clear downstream waves in enrichment in all of them (data not shown). According to our model, therefore, tRNA binding site occupancy levels in these experiments reflect properties of elongation in the presence of CHX rather than of *in vivo* dynamics. To test if CHX-disrupted elongation was masking the true impact of the absence of anticodon modification in these experiments, we compared the profiles of mean enrichment around all codon identities decoded by the modification-deficit tRNA species between the deletion strains and the wild type. Intriguingly, the profiles of mean enrichment around AAA showed consistently increased downstream wave areas in all of the deletion strains compared to wild type (figure 4.20A). To quantify this increase, we computed the corrected aggregate enrichment of each codon iden-

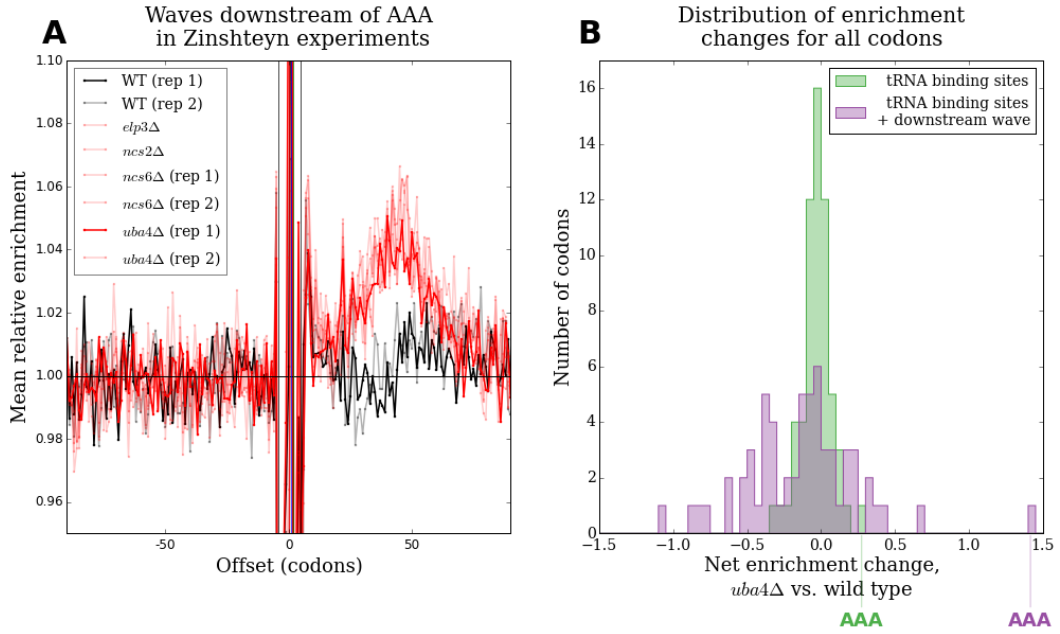

Figure 4.20: **Downstream waves recover the expected effects of lacking tRNA modifications.**
(A) Profiles of mean relative enrichments around AAA in two wild type experiments (black lines) and six experiments with different components of the mcm^{5}s^{2}U-pathway deleted (red lines) from Zinshteyn [121]. All mcm^{5}s^{2}U deletion strains produce clearly increased waves downstream of AAA compared to wild type. Darker lines correspond to the experiments compared in (B).
(B) Histograms of the net change in enrichment for each codon identity between *uba*4Δ and wild type at the A-, P-, and E-sites (green) or at the A- P, and E-sites plus 7 to 90 codons downstream (purple). AAA shows a modest increase in net enrichment at the tRNA binding sites, but a dramatically larger increase in net enrichment if the area of downstream waves is also taken into account. This suggests that AAA does take substantially longer to decode *in vivo* in *uba*4Δ than in wild type, but that most of this difference disappears during continued elongation in the presence of CHX.

tity as above by adding the downstream wave area to the sum of the binding site enrichments. We then computed the change in corrected aggregate enrichment for each codon identity between each of the deletion strains and the wild type (figures 4.20B). In each of the deletion strains, but not in a replicate of the wild type, AAA undergoes a dramatically larger increase in aggregate tRNA binding site enrichment when corrected to include downstream wave area (purple) than if wave area is not included (green). This argues that AAA does in fact take substantially longer to decode *in vivo* in cells lacking the ability to modify its tRNA, but that most of this difference disappears during continued elongation in the presence of CHX.

4.2.7 Mechanism of continued elongation in the presence of CHX

Although the exact mechanistic details of how disrupted elongation in the presence of CHX occurs remains unclear, there are several key features of observed patterns in the data and of known properties of CHX that any potential mechanism must accommodate. The first is that the disruption in dynamics is concentration-dependent. The second is that relative elongation rates in the new dynamics are still coupled to codon identities. Mean relative A- and P-site enrichments do not simply collapse towards being uniform in CHX experiments, but instead reproducibly take on a wide dynamic range of codon-specific values. The third is that absolute elongation rates must be dramatically slower in the presence of standard concentrations of CHX. Any model implying the contrary is not plausible; CHX has been successfully

used as a translation inhibitor for decades. We suggest that this inhibition is accomplished by a large reduction in the rate of elongation rather than a complete halt.

A possible mechanism with all three of these properties is that CHX repeatedly binds and unbinds to ribosomes, preventing advancement when bound but allowing elongation to proceed when unbound. In this model, the global rate of CHX binding to all ribosomes increases with increasing CHX concentration, leading to a decrease in the amount of time each ribosome spends unbound and therefore globally decreasing the rate of continued elongation. This accounts for the fact that downstream peaks move less far downstream in the same amount of time with increasing CHX concentration. Because the distance that peaks have moved downstream is the product of the total duration of disrupted elongation and the CHX-concentration-dependent average rate of this elongation, the magnitude of the reduction in elongation rate can be roughly estimated. Although there is broad agreement between downstream peak offset and annotated pretreatment time in experiments using the standard CHX concentration, with the longest pretreatment time (5 minutes in McManus[77]) corresponding to the farthest peaks and the shortest pretreatment time (1 minute in Lareau [64] and Nedialkova [81]) corresponding to the closest peaks, we note that a range of different peak locations are observed across experiments using the standard 2 minute pretreatment. These differences could conceivably reflect variation in the duration of harvesting or in effective CHX concentrations. Conservatively assuming that no elongation

occurs during harvesting, the range of peak centers with standard CHX concentration and 2 minutes of pretreatment implies absolute elongation rates of 0.1 to 0.3 aa/s. Because natural elongation rates in yeast are 7 to 9 aa/s [60], this represents an approximately 20- to 90-fold reduction in the speed of elongation.

To explain the reproducible range of codon identity-specific elongation rates in the presence of CHX, changes in the conformation of the ribosome as a result of differences in the geometry or base-pairing interactions of the tRNAs occupying the A- and P-sites could modulate the rates of CHX binding and unbinding. Conformational changes in the ribosome as a result of codon-anticodon interactions are known to be an integral part of the elongation cycle [107]. Given the unique presence of I-A wobble pairing in the decoding of CGA codons, the outsize role that CGA plays in these phenomenon suggests that base-pairing interactions could play a major role in determining CHX affinity. This offers an elegant potential explanation for the negative correlation between A-site enrichments with and without CHX: codon identities that produce unusual ribosome conformations tend to slow down elongation when tRNA binding is rate-limiting, but tend to speed up elongation when CHX disassociation is rate-limiting. In this model, the concentration-dependent interpolation between these two regimes observed in figure 4.7 reflects the fact that as CHX concentration decreases, each ribosome spends an increasing fraction of time unbound by CHX and therefore elongating according to the unperturbed dynamics.

4.2.8 Heterogeneity in experiments without CHX pretreatment

Heterogeneity in measured A- and P-site enrichment values between experiments that avoid CHX pretreatment presents complications for this model. Except for a single experiment by Guydosh [39], however, all such experiments still include CHX in the lysis buffer into which cells are harvested. The exact point in the harvesting process at which the elongation factors and charged tRNAs required for elongation are no longer accessible to ribosomes is unclear. One explanation for the observed heterogeneity could be that, under some conditions, the same continued elongation with disrupted dynamics that occurs during CHX pretreatment could also occur during the harvesting process once ribosomes have been exposed to CHX in the lysis buffer. The A- and P-site occupancies in the non-pretreated experiments by Pop[89], Lareau[64], Gardin[30], and Nedialkova[81] can be interpreted as an intermediate phenotype halfway in between the two tighter clusters consisting of CHX-pretreatment experiments and of the no-pretreatment experiments by Gerashchenko, Weinberg, Jan, and Williams in figure 4.3B, potentially reflecting a small amount of CHX-mediated elongation in these intermediate experiments. Consistent with this interpretation, enrichment profiles around CGA appear to be shifted slightly downstream in these intermediate-phenotype no-CHX experiments (figures 4.13 and 4.22).

The complete set of five non-pretreated experiments produced by Pop et al.[89] are particularly heterogeneous. Three of these experiments (WT-URA_footprint, AGG-OE_footprint, and AGG-QC_footprint) report A-site oc-

cupancies strikingly similar to values reported by CHX pretreatment experiments and less similar to the other non-pretreated experiments (figure 4.21). These same three experiments also have distinct peaks located approximately 20 codon positions downstream in the enrichment profiles around each codon identity that are absent in the other two experiments from the study (figure 4.22). These are the only non-pretreated experiments we examined for which clear downstream peaks are observed. Such extreme heterogeneity between non-pretreated experiments is difficult to account for in our model, but suggests that a wide range of different amounts of elongation after exposure to the lysis buffer are possible across different implementations of harvesting protocols.

One of the experiments performed by Pop et al. consisted of overexpressing the tRNA decoding AGG in order to test if increased availability of these tRNAs reduces the average elongation time of occurrences of AGG. They found a surprisingly small change in the relative frequency with which AGG was located in the A-site of footprints in the AGG-overexpression (OE) strain compared to the wild type. Given the strong evidence that substantial CHX-disrupted elongation occurred in these experiments, we wondered if this continued elongation was obscuring the true impact of tRNA overexpression. Waves downstream of AGG in the wild-type and AGG-OE experiments are too noisy to provide clear visual evidence of a difference in wave magnitudes between the two strains (figure 4.23A). When we summed the downstream wave areas of each codon and added these sums to the active-site enrichments

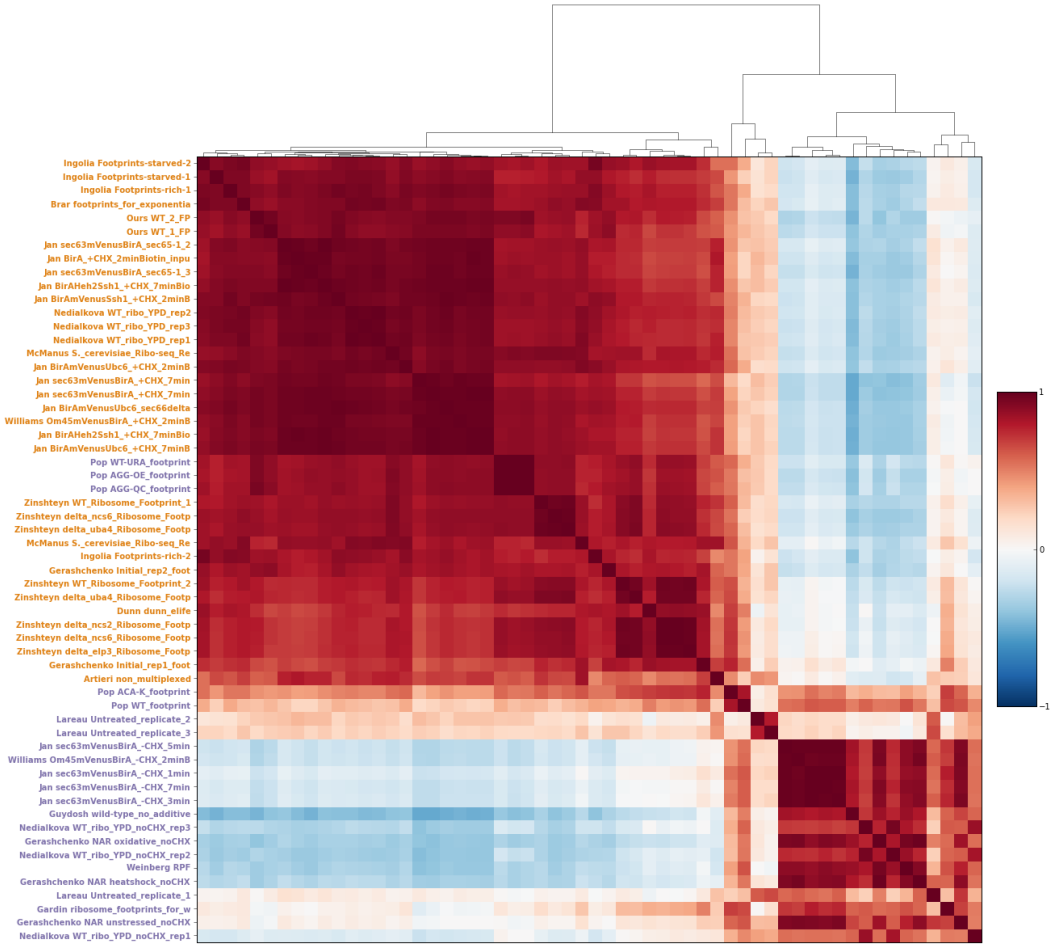

Figure 4.21: **A-site enrichments in some experiments from Pop [89] cluster with CHX-pretreatment experiments.**
Figure is constructed as in figure 4.3 but includes additional experiments from each study. Three no-CHX-pretreatment experiments from Pop [89] are more similar to CHX-pretreatment experiments (orange text) than they are to all other no-CHX-pretreatment experiments (purple text).

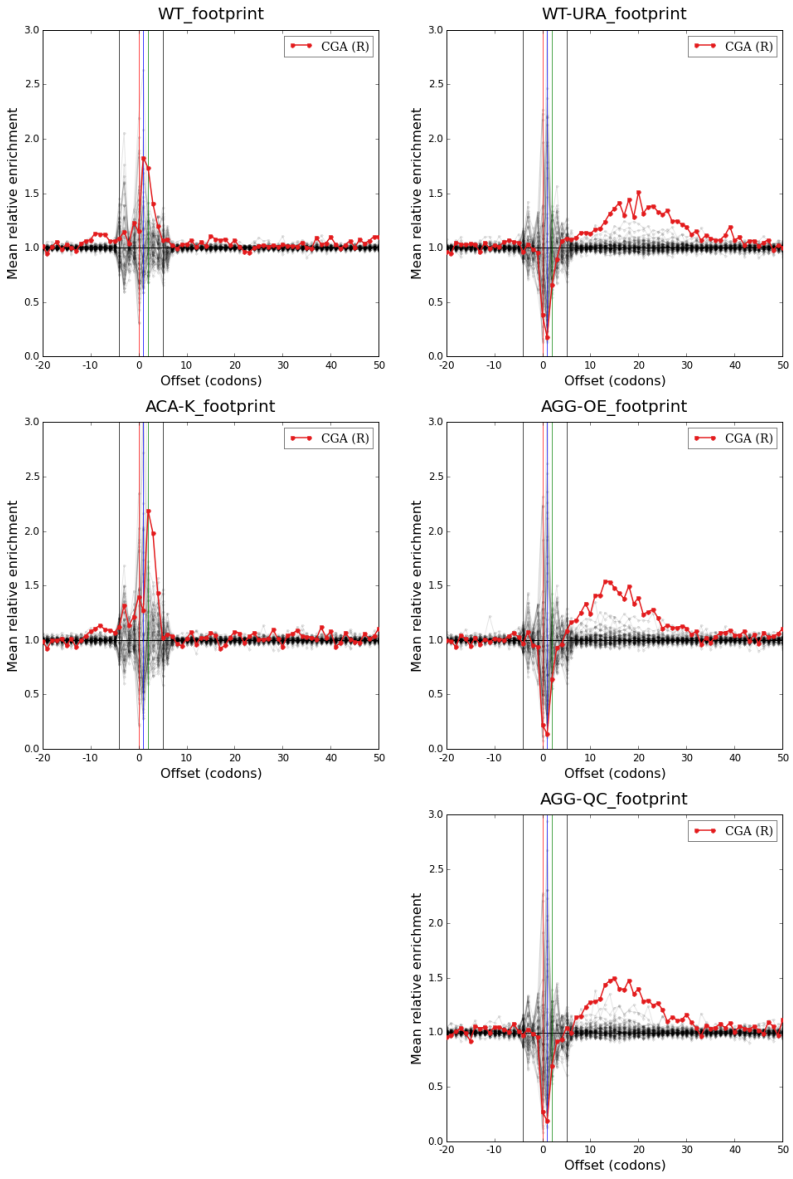

Figure 4.22: **Downstream peaks in experiments from Pop [89].** Figure is constructed as in figure 4.12 but shows five experiments from Pop [89]. Although these experiments were not performed with CHX pretreatment, WT-URA_footprint, AGG_OE_footprint, and AGG-QC_footprint (right column) all show clear peaks ∼20 codons downstream, and WT_footprint and ACA-K_footprint (left column) have profiles of enrichment around CGA shifted slightly downstream compared to non-pretreated experiments from other studies.

to produce corrected aggregate enrichments for each codon, however, AGG is a clear outlier to the left in the distribution of these corrected values. In other words, there is substantially less net area in the wave downstream of AGG in the AGG-OE experiment than the wild-type, arguing that overexpression did in fact cause a substantial speed-up in the translation of AGG in the pre-disruption dynamics.

4.3 Methods

In this section, we provide further technical details of the analyses described above.

4.3.1 Details of initial processing and mapping of footprinting data

All ribosome profiling experiments analyzed involve attaching a known sequence to the $3'$ end of RNA footprints to which a reverse transcription primer can be annealed. Some experiments use polyA tailing for this purpose, while others attach an oligonucleotide linker sequence. For experiments using polyA tailing, reads were trimmed from the end back to the first base that wasn't an `A` or an `N`. For experiments using linker sequences, linkers were located in reads by local alignment with the expected sequence and trimmed. Trimmed reads were first mapped to yeast rRNA sequences with bowtie2[62], and any reads that mapped were filtered out. Remaining reads were mapped with tophat2[54] to the yeast genome (version EF4) and spliced transcriptome (using transcript models from the Saccharomyces Genome Database's

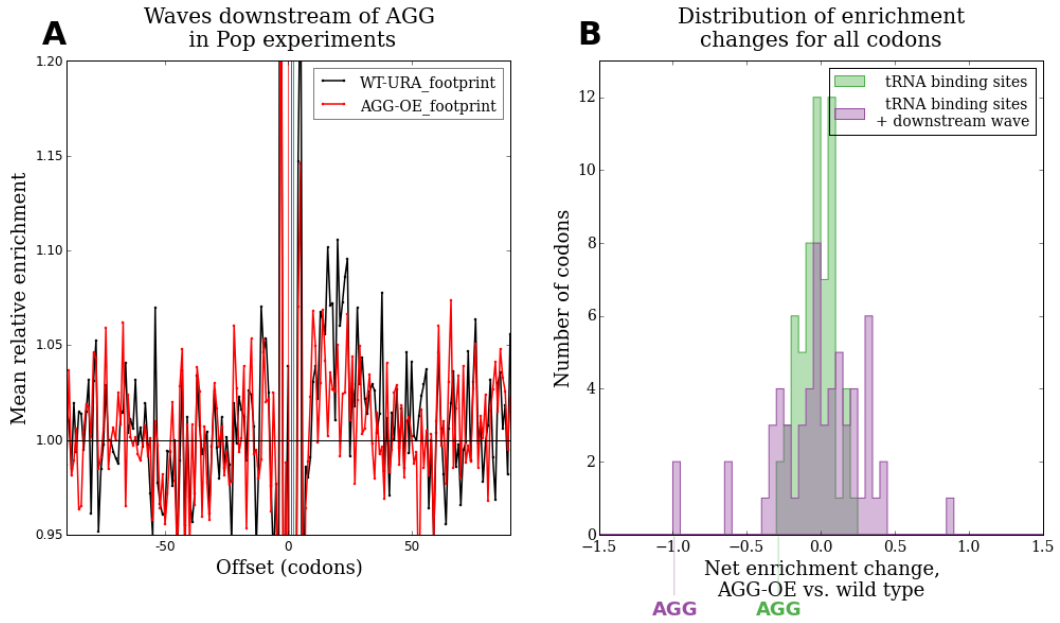

Figure 4.23: **Downstream waves recover the expected effects of overexpressing a tRNA.**
(A) Profiles of mean relative enrichments around AGG in a wild type experiment (black lines) and an experiment in which the tRNA decoding AGG was overexpressed on a plasmid from Pop [89].
(B) Histograms of the net change in enrichment for each codon identity between AGG-OE and wild type at the A-, P-, and E-sites (green) or at the A-P, and E-sites plus 7 to 90 codons downstream (purple). AGG shows a modest decrease in net enrichment at the tRNA binding sites, but a dramatically larger decrease in net enrichment if the area of downstream waves is also taken into account.

First author of study	Accession number	Representative sample(s) used
Ingolia [46]	GSE13750	Footprints-rich-1
Brar [9]	GSE34082	footprints_for_exponential_vegetative_cells_of_the_strain_gb15_used_for_the_traditional_timecourse
Gerashchenko (PNAS) [32]	obtained from authors	Initial_rep2_foot
Artieri [3]	GSE50049	non_multiplexed
Mcmanus [77]	GSE52119	S._cerevisiae_Ribo-seq_Rep_1
Zinshteyn [121]	GSE45366	WT_Ribosome_Footprint_1
Dunn [25]	obtained from authors	only one sample
Ours	n/a	WT_2_FP
Lareau [64]	GSE58321	Cycloheximide_replicate_1 (CHX) Untreated_replicate_1 (no CHX)
Gerashchenko (NAR) [31]	GSE59573	many samples used
Pop [89]	GSE63789	WT_footprint
Gardin [30]	GSE51164	ribosome_footprints_for_wildtype
Weinberg	GSE53268	Cerevisiae_RPF
Guydosh [39]	GSE52968	wild-type_CHX
Nedialkova [81]	GSE67387	WT_ribo_YPD_rep1 (CHX) WT_ribo_YPD_noCHX_rep2 (no CHX)
Jan [50]	GSE61012	sec63mVenusBirA_+CHX_7minBiotin_input (CHX) sec63mVenusBirA_-CHX_7minBiotin_input (no CHX)
Williams [117]	GSE61011	Om45mVenusBirA_+CHX_2minBiotin_input (CHX) Om45mVenusBirA_-CHX_2minBiotin_input (no CHX)

Figure 4.24: **Data sources.**

`.gff` dated `Fri Apr 11 19:50:03 2014`). Unmapped reads had any terminal stretches of `A` trimmed and were put through tophat2 again to recover potential mappings overlapping transcript polyA tails, although this has minimal impact on the analysis presented here. The reverse transcription process used to convert footprints to DNA can add untemplated bases to the end of intermediate anti-sense DNA products, which ultimately end up located at the beginning of sequencing reads [45]. We observed that the rate at which this happens varies considerably between different experiments. To prevent these untemplated bases from potentially shifting the codon positions that reads end up assigned to, bases that mismatch the reference sequence are trimmed from the beginning of all mappings up to the first matching base. For every annotated coding sequence, uniquely mapped reads of length 28 or 29 were assigned to the in-frame codon closest to the nucleotide at (0-based) offset +15 from the $5'$ end of the read; reads of length 30 were assigned to the in-from codon closest to offset +16.

This work involves data from a large number of studies deposited in the GEO and SRA databases. Manual acquisition of data for many experiments via the GEO web interface can be a tedious and error-prone process. To automate this process, we developed a software tool that takes as input a GSE accession number. The tool scrapes and parses XML from the NCBI website to determine URLs for data from the different samples associated with that accession number, then downloads these samples via ascp and dumps fastqs from the resulting .sra files using the sra-toolkit. This tool is available at

github.com/jeffhussmann.

Other software tools used to process or visualize data include IPython [87], pysam/samtools [69], numpy [112], scipy [82], matplotlib [41], cython [6], pandas [76], and seaborn [114].

4.3.2 Computing stratified mean enrichments

After assigning mapped read counts to the codon position in the A-site as described above, the goal is to estimate the impact that the presence of a particular codon identity at a particular offset from the A-site has on the relative frequency with which ribosomes are measured, marginalizing over all other nearby sequence features. To do this, uniquely mapped read counts were normalized by dividing all counts for each coding sequence by the mean across that coding sequence to control for the total number of ribosomes observed on the coding sequence, which is informative about mRNA levels and translation initiation levels but not about relative elongation times. The set of all codon positions across all coding sequences is then stratified to select those positions that are located at the offset of interest from an occurrence of the codon identity of interest. The mean of the relative enrichment values at all positions in this stratified set is then computed. To exclude the influence of poorly-understood structure in measured ribosome density at the $5'$ end of coding sequence, we excluded 90 codons at the beginning and end of each coding sequence from all stratified mean enrichment computations. For all calculations of downstream wave areas, we increased the number of codons excluded

from the edges of genes from 90 to 200. For some experiments, excluding this wider range substantially improves the agreement between CHX-induced tRNA binding site changes and downstream wave areas. This suggests that patterns in codon composition, which are particularly pronounced at the beginning of genes [111], may introduce small confounding biases that aggregate when adding relative enrichments over a wide range of offsets to calculate downstream wave areas.

More formally, let g be an index over genes. Let l_g be the length in codons of gene g's coding sequene. Let $c_{g,i}$ be the codon identity at position i in gene g, Let $r_{g,i}$ be the count of uniquely mapped reads assigned to this codon position. Let d be the number of codons to be exclude from the edge of each gene. For each gene consisting of at least $2d+1$ codons, so that there is something left after excluding d from the beginning and the end, compute the mean read count over all eligible positions in a gene

$$M_g = \frac{\sum_{i=d}^{l_g-d} r_{g,i}}{l_g - 2d} \tag{4.1}$$

and define the relative enrichment at each positions as

$$e_{g,i} = \frac{r_{g,i}}{M_g}. \tag{4.2}$$

For a given codon identity I and offset F, the stratified set of all eligible positions located exactly that offset downstream of an occurrence of that codon identity is

$$s_{I,F} = \{(g,i) : d < i < l_g - d, c_{g,i-F} = I\}. \tag{4.3}$$

The mean relative enrichment at the stratified set of such positions is therefore

$$\frac{\sum_{(g,i)\in s_{I,F}} e_{g,i}}{|s_{I,F}|}. \tag{4.4}$$

When a gene has a small number of reads mapped to it, the denominator in the expression for $e_{g,i}$ is small and the values produced by this expression are noisy. Maximizing signal-to-noise in mean relative enrichments is therefore a balancing act between including as many genes as possible in order to maximize the number of codon positions being averaged over while minimizing the effect of noisy relative enrichment values from lowly-expressed genes. This issue is particularly pronounced in the mean relative enrichment profiles around non-optimal codons (e.g. CGA), for which a disproportionate share of occurrences of the codon identity are in lowly-expressed genes. To navigate this balance, for each experiment, we excluded genes for which $M_g < 0.1$ - that is, genes with an average read density of less than 1 read per 10 codons across the eligible region of the gene. Because the number of useful sequencing reads produced by each experiment varies considerably due to differences in the number of raw reads produced and in the efficiency with which uninteresting rRNA contaminants are removed, the exact set of genes passing this filter varies from experiment to experiment. Profiles of mean relative enrichments in all experiments are qualitatively unchanged but noisier if we instead include every gene with a nonzero number of mapped reads in each experiment.

4.3.3 Simulation details

In order to evaluate the ability of different models of CHX activity to produce patterns observed in experimental data, we developed a simple event-driven simulator of the movement of ribosomes along coding sequences. We made several simplifying assumptions about translation in this simulation. First, we assume the elongation time at each position depends only on the codon identity in the A-site of a ribosome. Second, we assume that the rate of initiation for each mRNA is a constant (but potentially gene-specific) value - that is, we do not model competition for a pool of ribosomes between different mRNAs. Third, we measured time in arbitrary units not grounded in any absolute measurements.

The central object in the simulation is a representation of a single copy of a particular mRNA copy of a coding sequence. For each such mRNA object, multiple ribosomes are tracked as they simultaneous advance along the coding sequence. A priority queue of future events indexed by the time at which each event is scheduled to occur is maintained to determine the ordering of events. Evolution of the system is carried out by popping events off off the priority queue, processing the events, and then inserting any consequent events into the queue.

Simulation for each mRNA object begins with an immediate initiation event at $t = 0$. After each initiation event, the time interval until the next attempted initiation is drawn from an exponential distribution with the rate parameter set to a user specified, potentially gene-specific value. Although we

carried out simulations in which the initiation rate of each gene is proportional to the ratio of footprint RPKM to mRNA-seq RPKM from matched experiments (the so-called translational efficiency of the gene [46]), the simulation results shown in the main text have the initiation rate of every gene set uniformly to 0.01. Ribosomes are always assigned to the single codon identity in their A-site, but each ribosome occludes 5 codon positions upstream and 4 codon positions downstream of this. After initiation, the amount of time a ribosome waits at each codon position before attempting to advance is exponentially distributed with a rate parameter determined by the codon identity in the A-site. Ribosomes are prevented from advancing if doing so would cause its A-site to be within 4 codons of the next downstream ribosome's left edge. If this occurs, a new waiting time is drawn, after which the ribosome will attempt to advance again. To efficiently evolve a single instance of a coding sequence to steady state, events are processed until the first ribosome hits the stop codon. If t_{runoff} is the point in time at which this happens, a stopping time is chosen uniformly at random from the interval $[t_{\text{runoff}}, 2t_{\text{runoff}}]$. This stopping time is added to the priority queue as an event, and events are processed until this event is reached.

After steady state is reached, different potential CHX mechanisms can be introduced. Two such mechanisms are considered here. In the first, at the instant CHX is introduced to the system, each ribosome is assigned an amount of time to wait until a CHX molecule first arrives at it and irreversibly halts it. The mean of this waiting time distribution is the mechanistic knob that is as-

sumed to change with CHX concentration. Every ribosome that initiates after CHX is introduced is also assigned a waiting time in the same way. The system is evolved until every ribosome has been arrested and the initiation site is occluded by an arrested ribosome so that no further initiation is possible. The resulting positions of ribosomes are then recorded as simulated read counts. The only way in which this model of CHX action produces sampled positions that differ from the pre-CHX steady state is when stochastic differences in the arrival times of CHX at sequential ribosomes cause the upstream ribosome to be halted by running into the arrested ribosome in front of it instead of by the arrival of CHX. The average spacing between ribosomes is determined by the ratio between the rate of initiation and elongation rates. The extent to which stalling occurs can be tuned by controlling the ratio between this average spacing and the mean time until CHX arrival. If this ratio is small, pairs of sequential ribosomes frequently experience a large enough difference in CHX arrival times for the trailing ribosome to close the gap between them. This results in spikes in mean enrichment at offsets that are multiples of 10 upstream (i.e. at negative offsets in the profiles of mean enrichment plotted throughout the text) and broad, low-level enrichment downstream of any slow codon identity but no coherent downstream peaks. Mean enrichments at the A-site experience a contraction towards one, reflecting the fact that the codon identity in the A-site of a ribosome that was stopped by running into the ribosome ahead of it is essentially drawn uniformly from the codon identities in a coding sequence, rather than being drawn in proportion to the elonga-

tion times of codon identities. We are unable to find any region of parameter space for which this mechanism produces behavior that qualitatively hints at the changes in active site occupancies and appearance of downstream peaks present in real data.

In the second potential mechanism, at the instant of CHX arrival, the means of the exponential distributions from which the elongation waiting time of ribosomes at each codon identity are changed. Every ribosome with a pending elongation event in the priority queue has this event discarded and redrawn from the new distributions. After this shift in codon-identity-specific elongation rates, a user-specified interval of time is allowed to proceed before the locations of all ribosomes are measured. As discussed in the main text, a potential mechanistic basis for this behavior is the CHX molecules repeatedly bind and unbind from each ribosome, so that the mean time a ribosome spends at a position reflects the influence of the codons located in the tRNA binding sites of the ribosome on the rates of CHX association and dissociation.

For either model of CHX action, a template (real) experiment is used to guide the number of simulated reads produced for each gene in order to accurately reflect the dynamic range of expression in the yeast transcriptome. To do this, for each gene, copies of the coding sequence are evolved to steady state and put through simulated CHX treatment before recording simulated read positions until the total number of reads produced for the gene just exceeds the count of reads mapped to that gene in the template experiment.

4.3.4 Inferring codon-specific elongation rates accounting for gene-specific codon usage biases

In this section, we describe the continuous time Markov chain model of translation that was used above to determine the influence of gene-specific codon usage patterns on elongation rate estimates and to analyze transient behavior of ribosome density patterns following a change in codon-specific relative elongation rates.

Using mean relative enrichment as a measure of the average time spent decoding a particular codon identity has the benefit of simplicity but could in theory be biased by the presence of covariation in the usage of codons between different genes. To intuitively motivate this concern, if codon A and codon B tend to be used more frequently than their genome-wide averages in the same genes and codon B is extremely slow, codon A will appear to be slightly faster than it really is. To evaluate the size of this effect, we computed the mean elongation time of each codon identity in a way that takes the codon composition of each gene into account. For this calculation, our model of translation is that the amount of time a ribosome spends at a particular position depends only on the identity of the codon positioned at the A-site of the ribosome and that these time intervals are independent and exponentially distributed with a codon-identity specific rate parameter. We assume that rates of initiation are small enough relative to elongation rates that collisions between ribosomes can be ignored. In order to determine exactly what a set of observations of ribosomes at particular positions tells

us in the Bayesian sense about these codon-identity specific rate parameters, we need to compute the posterior distribution of these parameters given the data. In order to do this, we need to be able to compute the likelihood of a set of observed positions given a particular set of 61 values for the codon-specific rate parameters.

Suppose that the translated portion of an organism's transcriptome consists of n_c coding sequences. Each coding sequence is an ordered sequence of the 61 non-stop codons. Suppose that a particular coding sequence g consists of n codons with identities $\{c_i\}$ for $i = 1, \ldots, n$. Then the life cycle of a ribosome with respect to this coding sequence can be modelled as a simple continuous-time Markov chain with a dummy state 0 that represents the ribosome doing anything but translating this particular coding sequence. Transition from this state into the act of translating the first codon occurs at some coding-sequence specific initiation rate λ_{init} that is a nuisance parameter for the purposes of this calculation. After this, the ribosome transitions from each codon to the next at a rate determined by the identity of the codon it is currently translating. Assuming that the cell is in a steady state condition, the probability that a random point in time sampled from the lifetime of a ribosome will find it in the act of translating a particular codon, given that it was observed somewhere on this coding sequence, is given by the stationary distribution of this Markov chain, conditional on not being in state 0.

The infinitesimal generator matrix of this Markov chain is

$$\Lambda = \begin{bmatrix} -\lambda_{\text{init}} & \lambda_{\text{init}} & 0 & 0 & \dots & 0 \\ 0 & -\lambda_{c_1} & \lambda_{c_1} & 0 & \dots & 0 \\ 0 & 0 & -\lambda_{c_2} & \lambda_{c_2} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda_{c_n} & 0 & 0 & 0 & \dots & -\lambda_{c_n} \end{bmatrix}. \tag{4.5}$$

Because this Markov chain is irreducible, it has a unique stationary distribution

$$\mathbf{p} = \begin{bmatrix} p_0 & p_1 & \dots & p_n \end{bmatrix}. \tag{4.6}$$

The stationary distribution will satisfy the probability mass-balance equation

$$\mathbf{p}\Lambda = \mathbf{0}^T \tag{4.7}$$

and the normalization condition

$$\sum_{j=0}^{n} p_j = 1. \tag{4.8}$$

It is straightforward to verify that

$$p_0 = \frac{\frac{1}{\lambda_0}}{\sum_{k\in\{0,c_1,\dots,c_n\}} \frac{1}{\lambda_k}} \tag{4.9}$$

and

$$p_j = \frac{\frac{1}{\lambda_{c_j}}}{\sum_{k\in\{0,c_1,\dots,c_n\}} \frac{1}{\lambda_k}} \tag{4.10}$$

for $j = 1, \dots, n$ satisfy these equations. To produce the conditional stationary distribution given that not being in the dummy state, simply divide the other components by their sum. The net effect of this is to remove the term corresponding to the dummy state from the denominator, giving

$$p_j = \frac{\frac{1}{\lambda_{c_j}}}{\sum_{k=1}^{n} \frac{1}{\lambda_{c_k}}}. \tag{4.11}$$

For convenience, let $\beta_i = 1/\lambda_i$. Let $c_{g,i}$ be the number of occurrences of codon identity i in coding sequence g. Then given that a ribosome was observed on coding sequence g, the probability that is was observed at a specific occurrence of codon identity i is

$$\frac{\beta_i}{\sum_{j=1}^{61} c_{g,j}\beta_j}. \tag{4.12}$$

For an entire data set, let $r_{g,i}$ be the number of observations of a ribosome at any occurrence of codon identity i in coding sequence g. Then, making an independence assumption, the overall likelihood of the data given values of $\{\beta_i\}$ is

$$L = \prod_g \prod_{i=1}^{61} \frac{\beta_i^{r_{g,i}}}{\left(\sum_{j=1}^{61} c_{g,j}\beta_j\right)^{r_{g,i}}}. \tag{4.13}$$

We can then explore the posterior distribution of the rates given the data under this model via MCMC (Figure 4.25).

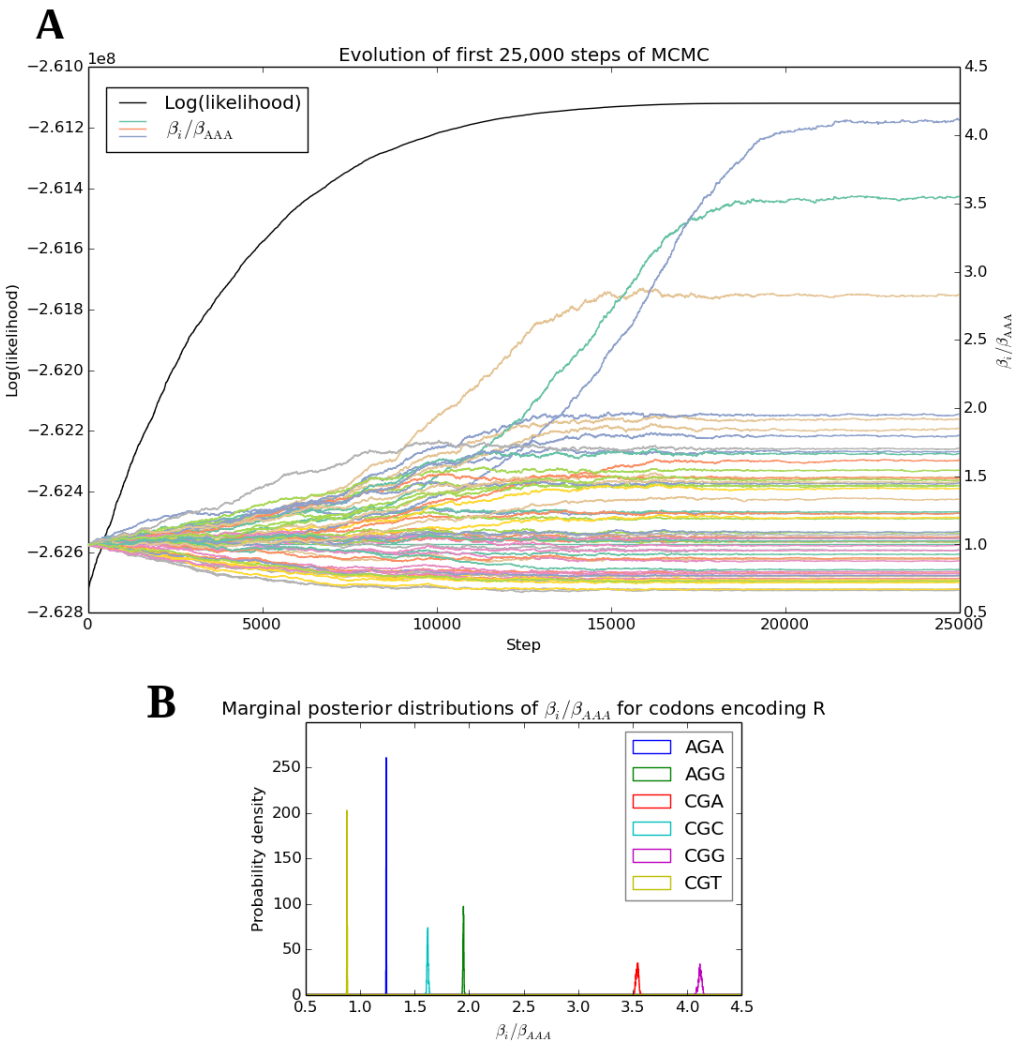

Figure 4.25: **Inferring codon-specific elongation rates via MCMC.**

Figure 4.25 (Continued): **Inferring codon-specific elongation rates via MCMC.**
(A) Exploration of the 61-dimensional posterior distribution of codon-identity specific elongation rates given the data from Weinberg [115] according to the forward statistical model in equation 4.13 via MCMC. Colored curves and right axis show movement of each component through the parameter space over the first 25,000 steps of a particular MCMC instantiation. The black curve and left axis shows the log-likelihood of the proposed parameter vector at each step, which climbs steadily during a burn-in period before leveling off.
(B) Marginal posterior distributions for the reciprocals of the elongation rates of the six codons that encode arginine as determined by the values visited in the subsequent 50,000 steps. Because rates are only determined up to a global multiplicative constant, the elongation rate of codon AAA is set to 1 as a arbitrary normalization.

Evaluation of the likelihood function is somewhat computationally expensive but can be parallelized with a simple message passing scheme in which the list of genes is split up between processes running on different cores of a single machine. A proposed sets of rates at which the likelihood is to be evaluated is passed to each process, and the contributions of each set of genes' data to the overall likelihood are collected from the processes and combined to give the overall likelihood. In practice, we observe that numerically maximizing the likelihood function using Powell's method converges to sets of rates that agree with the MAP values produced via MCMC within a few thousand evaluations of the likelihood function.

4.3.5 Transient behavior after changes in relative elongation rates

In the continuous time Markov model of translation presented above, if the probability distribution over states at $t = 0$ is given by $\mathbf{p}(0)$, then the evolution of the system of ordinary differential equations governing the flow of probability density between states over time is given by the matrix exponential of the generator matrix:

$$\mathbf{p}(t) = \mathbf{p}(0)e^{t\Lambda}. \tag{4.14}$$

Consider a coding sequence consisting of 100 copies of codon A, followed by a single copy of codon B, followed by 100 copies of codon A. Suppose that the two codon identities are translated with mean relative elongation times β_A and $\beta_{B,\text{before}}$. Let Λ_{before} be the infinitesimal generator matrix of the

Markov chain with these rates, and let $\mathbf{p}_{\text{before}}$ be the steady state distribution under Λ_{before}. Suppose that the system is at steady state and then at time 0 the dynamics of translation are instantaneously changed so that the relative elongation rates of the two codon identities become β_A and $\beta_{B,\text{after}}$. Let Λ_{after} and $\mathbf{p}_{\text{after}}$ be the generator matrix and steady state distribution, respectively, under these new relative elongation rates. Then

$$\mathbf{p}(t) = \mathbf{p}_{\text{before}} e^{t\Lambda_{\text{after}}}. \tag{4.15}$$

To understand the transient behavior as the system relaxes to the new steady state, decompose $\mathbf{p}_{\text{before}}$ into

$$\mathbf{p}_{\text{before}} = \mathbf{p}_{\text{after}} + (\mathbf{p}_{\text{before}} - \mathbf{p}_{\text{after}}), \tag{4.16}$$

giving

$$\mathbf{p}(t) = \mathbf{p}_{\text{after}} e^{t\Lambda_{\text{after}}} + (\mathbf{p}_{\text{before}} - \mathbf{p}_{\text{after}}) e^{t\Lambda_{\text{after}}}. \tag{4.17}$$

By construction, $\mathbf{p}_{\text{after}}$ is in the left null space of Λ and is therefore a left eigenvector of $e^{t\Lambda_{\text{after}}}$ with eigenvalue 1 for any t, so this becomes

$$\mathbf{p}(t) - \mathbf{p}_{\text{after}} = (\mathbf{p}_{\text{before}} - \mathbf{p}_{\text{after}}) e^{t\Lambda_{\text{after}}}. \tag{4.18}$$

The left side of this equation represents how much the distribution at time t still differs from the eventual steady state. Except for slight differences in normalization, $\mathbf{p}_{\text{before}} - \mathbf{p}_{\text{after}}$ is essentially an impulse at the location of the single occurrence of codon B, scaled by $\lambda_{B,\text{before}} - \lambda_{B,\text{after}}$. For a particular

offset downstream of the occurrence of codon B and a particular value of t, therefore, the linearity of the expression on the right hand side implies that the transient change in magnitude of the downstream wave is proportional to $\lambda_{B,\text{before}} - \lambda_{B,\text{after}}$.

Figure 4.15 and 4.16 plot evaluations of this solution at a range of positions around codon B for series of increasing time points for the cases where codon B changes from being slower than codon A to being faster than codon A at $t = 0$ (that is, $\beta_{B,\text{after}} < \beta_A < \beta_{B,\text{before}}$) and where codon B is slightly slower than codon A before $t = 0$ but then becomes even slower (that is, $\beta_A < \beta_{B,\text{before}} < \beta_{B,\text{after}}$), respectively. In a window around codon B of length l on either side of codon B, the instantaneous rate of change in net ribosome density in the entire window is equal to the rate of flow into the leftmost codon position in the window minus the rate of flow out of the rightmost codon position. Until the downstream wave reaches this rightmost position (or any wave of global change in density caused by a relative change in elongation rates compared to the rate of initiation reaches the leftmost position), these two terms remain equal to each other. This implies that the net density across the window remains unchanged, so the net excess or deficit in the downstream wave must be equal in magnitude but opposite in sign to the change at codon B.

This 'conservation of ribosome density' argument motivates the expectation in figures 4.17 and 4.18 that the downstream wave areas for each codon identity should exactly offset tRNA binding site changes, and the closely re-

lated argument that aggregate tRNA binding site enrichments before a CHX-induced change in dynamics can be recovered by adding downstream wave areas back to the binding site enrichments in the presence of CHX. Applying this correction recovers the positive correlations with 1 / tAI expected if codons decoded by less abundant or wobble base-paired tRNAs are on the whole translated slower than average, although a somewhat wide range of positive correlation values are observed across different experiments in figure 4.19. While this could represent genuine differences in translation dynamics between the experiments, it seems likely that technical biases could account for much of the variation. When downstream waves have only moved a few codons downstream (as in our experiment), enrichments are affected by biases in how efficiently footprints with different nucleotides at the 5' edge are converted into sequenceable DNA [3]. When waves have moved far enough downstream that large ranges of offsets need to be summed to capture all of their area, patterns in codon usage could lead to small biases in enrichments around different codon identities that aggregate when large ranges of offsets are summed.

4.4 Conclusion

We have seen that ribosome profiling experiments conducted with and without CHX pretreatment make incompatible claims about the average amounts of time ribosomes spend with each codon identity in their tRNA binding sites during elongation. We have reported the existence of unexpected structure

in measured ribosomes density downstream of each codon identity in experiments using CHX. The most parsimonious model to explain both of these phenomenon is that elongation continues for many cycles after the introduction of CHX, but that during this continued elongation, the amount of time each ribosome takes to advance from codon to codon has a different quantitative dependence on the codons positioned in the tRNA binding sites of the ribosome than it did before the introduction of CHX.

The ability of a particular ribosome profiling experiment to produce accurate quantitative inferences about translation dynamics is contingent on ribosomes being measured at each codon position in proportion to how long they spend occupying that position *in vivo*. The interpretation of experiments using CHX to stabilize ribosomes has assumed that this stabilization occurs by a mechanism that leaves ribosomes positioned according to their steady state *in vivo* distribution, e.g. by irreversibly arresting the further elongation of each ribosome upon binding to it. The patterns we observe argue that this assumption does not hold. Instead, many cycles of continued elongation with disrupted dynamics leave ribosomes distributed in a way that does not directly reflect natural translation dynamics, although telltale signs of the pre-disruption dynamics can be indirectly discerned. Given the impact of CHX treatment on ribosome positioning, future ribosome profiling experiments aiming to directly measure the amount of time ribosomes spend at each position *in vivo* should therefore entirely avoid the use of CHX.

In light of our model, several counterintuitive results from previous ri-

bosome profiling studies in yeast take on new interpretations. Most notably, we offer an explanation for contradictory claims about whether so-called optimal codons corresponding to more abundant tRNAs are translated more rapidly. In many organisms, optimal codons are used with greater frequency in highly expressed genes, and a large body of theoretical work assumes that increased elongation speed drives this tendency [88]. If optimal codons are decoded faster, this tendency could lead directly to increased expression by increasing the rate of production of protein from each message or by avoiding mRNA-decay pathways linked to ribosome pausing [90]. Alternatively, if translation initiation is typically rate-limiting, using faster codons reduces the amount of time ribosomes spend sequestered on highly transcribed mRNAs, freeing up ribosomes to translate other messages and leading to more efficient system-wide translation [98]. If A-site enrichments measured in CHX pretreatment experiments reflected natural translation dynamics, however, optimal codons would not be elongated more quickly, and these theories fall apart. By offering a model for why the measured positions of ribosomes in CHX experiments appear to report that non-optimal codons are the fastest to be translated, and by showing evidence that optimal codons were in fact being translated more quickly before the introduction of CHX, we enable a principled resolution to this controversy. Earlier studies in this area have hypothesized that the A-site enrichments reported by CHX experiments could reflect an optimal balance between codon usage and tRNA abundance [92], or that potential heterogeneity in elongation times at different occurrences of the same codon identity

could conspire to produce these A-site enrichments [19]. By showing that the A-site occupancies fed into these models almost certainly do not represent the actual *in vivo* dynamics, our results argue against both of these conclusions.

The existence of continued but slower elongation after the introduction of CHX also sheds light on the observed ramps of increased ribosome density at the 5' end of coding sequences. These ramps were first noticed by Ingolia et al. [46], and a theoretical model was later advanced to explain why the initial period of slower translation implied by theses ramps could alleviate potential traffic jams of ribosomes further along coding sequences [111]. More recently, however, Gerashchenko showed that ramps extend over a smaller extent of the 5' end of coding sequences when higher concentrations of CHX are used [31], suggesting that at least some fraction of the apparent elevated ribosome density in the ramp structures is an artifact of CHX treatment. Gerashchenko hypothesized that this concentration ramp could reflect differences in the time necessary for different concentrations of CHX to permeate cells. Our model suggest an alternative explanation: if translation initiation continues at similar rates after the introduction of CHX while elongation is dramatically slowed but not halted, this would produce a transient, gradually spreading 5' ramp. In our model, therefore, the fact that the range of positions occupied by ramps decreases with increasing CHX concentration reflects the inverse relationship between CHX concentration and the global rate of continued elongation. Although a complete accounting for the source of ramps is complicated by incomplete understanding of the rate of continued initiation over the course of

experimental protocols, our model offers further support for Gerashchenko's hypothesis that a substantial fraction of the 5' ramp in density does not reflect an actual tendency toward slower translation at the beginning of coding sequences.

Finally, our model offers an explanation for the small apparent impacts of several experimental attempts to modify tRNA repertoires on measured relative elongation rates. CHX-pretreatment experiments by Zinshteyn et al. [121] on mcm^5s^2U-pathway deletion strains show surprisingly small changes in tRNA binding site occupancies at codons decoded by the modification-deficient tRNAs. Our observation that all of the deletion strains have substantially increased waves of enrichment downstream of AAA, one such codon identity, compared to wild type suggests that binding site occupancy changes in the presence of CHX dramatically underestimate the actual *in vivo* increase in the decoding time of AAA in the deletion strains. Pop et al. [89] evaluated the impact of overexpressing, deleting, or modifying the body sequence of tRNAs and also found surprisingly small changes in the rates at which the corresponding codon identities were translated. As discussed above, the experiments of Pop et al. did not pretreat with CHX, but a subset of these experiments show both clear downstream peaks and A-site occupancies shifted towards values reported by CHX pretreatment experiments. This suggests that enough CHX-disrupted elongation occurred in these experiments during the harvesting process that the resulting A-site occupancies may not be able to measure any potential effects of the tRNA repertoire modifications. Repeating these experiments

without any CHX in order to accurately sample from *in vivo* dynamics could clarify the consequences of these tRNA manipulations.

Chapter 5

Conclusions and future directions

5.1 Improving high-throughput sequencing error rates

In the first half of this thesis, we presented circle sequencing, a method for detecting and correcting sequencing errors by creating physically linked copies of the sequence of input DNA fragments. We described computational strategies for analyzing data produced by this method. We demonstrated that circle sequencing is capable of detecting and filtering out the vast majority of sequencing errors by applying it to sequence yeast genomic DNA. We also demonstrated that it accomplishes this reduction in error rates with substantially higher cost-efficiency than alternatively barcoding-based strategies for correcting errors. A specialized form of barcoding recently developed by Schmitt et al. [96] called duplex barcoding is able to achieve even lower error rates by protecting against errors caused by rare damage events to input DNA during the library preparation process, but this protection comes at the cost of a large decrease in cost-efficiency. There are therefore two clear directions for future development of error correcting sequencing methods: improving the accuracy of circle sequencing by incorporating the key insight of duplex barcoding into it, and improving the efficiency of duplex barcoding in order to allow it to reliably produce large quantities of error-corrected data.

In any error-correcting library preparation strategy, redundant copies of sequence information are powerless to protect against error processes that corrupt this sequence information before the redundant copies are made. Single-stranded DNA base damage introduced by the experimental manipulations used to extract and prepare genomic DNA represent one such process. The unexpectedly high rate at which these damage events occur had previously been masked by the even higher rate of generic sequencing errors. By reducing these rates, circle sequencing and barcoding methods have revealed that protecting information from these damage events represents the next hurdle to be overcome in order to apply high-throughput sequencing to detect ultra-rare variants. Schmitt et al. realized that double-stranded DNA represents a naturally occurring form of protection against such events because every piece of sequence information is already present in the two redundant copies represented by the two strands. The accidental discovery in our data of the ability of CircLigase to form circles out of double-stranded input templates suggests a way to incorporate this insight into circle sequencing. If circular templates that consist of end-to-end ligations of information from both strands of a double-stranded starting molecule can be reliably produced and sequenced, the resulting concatamers can be decomposed into downstream copies of each of these strands. Single-stranded base damage suffered by this template will only affect all copies of the damaged position in one of these two groups of copies, so that damage-induced artifactual variants can be filtered out by comparing the two strand's consensus sequences to each other. The instances of

such templates that we found in our data were rare, accidental byproducts of a library preparation strategy that was not designed to produce them. Deliberately creating these templates e.g. by attaching hairpin loops to both ends of double-stranded input templates as in Pacific Biosciences SMRTbell scheme [110] could potentially allow for the ultra-high accuracy of duplex barcoding while retaining the efficiency advantages of circle sequencing.

Although we demonstrated theoretical limits on how efficient any barcoding process can be, in practice, most implementations of barcoding do not come close to achieving this limit. A major obstacle is the practical difficulty in reliably controlling the precise number of successfully barcoded input molecules that enter the amplification reaction in order to control the average number of copies of each input molecule that will be produced by sampling from the amplification products. Conceptually, this consists of titrating the number of barcoded input molecules to hit the narrow peak in efficiency in the purple and green curves in figure 2.6. Recent developments in an experimental technique called digital PCR [40] could provide a way to do this. In theory, this technique provides a way to produce precise absolute counts of the number of input molecules that have had adapter sequences containing barcodes successfully ligated to them. To do this, input molecules are spread out into physically separate compartments, such as wells on a plate or droplets of oil in an emulsion, under dilute conditions so that each compartment is expected to receive either zero or one input molecules on average. The number of compartments that received a well-formed input molecule can then be di-

rectly counted by performing a PCR amplification within each compartment using primers corresponding to the adapter sequences and using fluorescent probes to determine whether amplification products were produced. Although this technology is still in its infancy and the extent to which it is capable of producing precise absolute quantifications of diverse input libraries remains unclear [116], it may be possible to apply digital PCR to control exactly how many input molecules are being put into the barcoding process and therefore reliably extract the maximum possible efficiency from the method.

Finally, our characterizations of PCR-mediated recombination during cluster generation of concatamers and of the unexpected formation of circles from double-stranded templates represent examples of the qualitatively new kinds of experimental diagnostics that deep analysis of high-throughput sequencing data can provide. The sequences produced by high-throughput sequencing experiments represent a per-molecule digital record of exactly what is happening during experimental biochemical manipulations. This digital record is a much more powerful readout to inform methods development than the types of diagnostic information that have historically been available in experimental biology. The software we developed for large-scale visualization of the different component pieces that high-throughput sequencing reads are made up of should be a broadly useful tool for forming this feedback loop between data analysis and experimental design, and is available at github.com/jeffhussmann.

5.2 Accurate measurements of translation dynamics with high-throughput sequencing

In the second half of the thesis, we explored how to use high-throughput sequencing to produce accurate measurements of the amount of time ribosomes spend translating each codon. We first presented purely theoretical work that examined whether patterns in the use of synonymous codons at nearby occurrences of the same amino acid could be interpreted as evidence for the novel hypothesis that individual tRNA molecules are recycled through ribosomes multiples times in order to speed up translation. The ambiguities that arise when trying to make this kind of indirect inference about translation dynamics served as a motivation for why more direct, transcriptome-scale measurements of ribosome speeds are necessary.

We then analyzed data from many ribosome profiling experiments to evaluate whether these experiments were capable of accurately producing these kinds of measurements. We found strong evidence that an unexpected side-effect of treatment with the translation inhibitor cycloheximide in some of these experiments disrupts the ability of these experiments to accurately measure *in vivo* translation dynamics. By comparing patterns in ribosome occupancy at and downstream of different codon identities across a large body of experiments performed both with and without the use of cycloheximide, we showed that cycloheximide does not irreversibly halt translation upon binding to ribosomes, as previous interpretations of this data had assumed. Instead, many cycles of continued elongation occur in the presence of cycloheximide,

but the relative amount of time ribosomes spend positioned over each type of codon during this continued elongation is dramatically different than during unperturbed translation. The reason that experiments performed with and without cycloheximide reported such dramatically different pictures of translation dynamics was previously unclear. By characterizing the mechanism that causes this difference, we resolve this mystery and establish that measurements produced without CHX more accurately reflect how translation proceeds under natural conditions.

In light of our results, it is clearly necessary to omit pretreatment of cells with CHX for long periods of time in order to accurately measure how long ribosomes spend at each position *in vivo*, but this may not be sufficient. We observed substantial heterogeneity in tRNA binding site enrichment levels between different experiments that all omitted CHX pretreatment. The most parsimonious explanation for this heterogeneity is that even in the absence of CHX pretreatment, some implementations of protocols for harvesting ribosomes from cells do so in a way that doesn't allow any elongation with disrupted dynamics to occur and that some do not. To consistently produce accurate measurements of steady-state translation dynamics, it will be important to identify and eliminate the remaining differences between these implementations. Checking for the existence of any downstream waves in data produced by future experiments will be a useful diagnostic for determining whether each experiment has succeeded in capturing steady-state dynamics.

Our work so far has focused on the effects of a single translation in-

hibitor on measurements produced in a single organism (namely, cycloheximide and yeast) because this pair represents the large majority of ribosome profiling experiments that have been performed to date. Searching for downstream wave patterns in ribosome profiling data using other translation inhibitors and from other organisms will be necessary to determine the generality of the phenomena we described. In particular, Lareau et al. [64] have recently demonstrated that treatment with an alternative inhibitor, anisomycin, preferentially captures ribosomes in an alternative conformation that protects ~ 21 (as opposed to ~ 28) nucleotides of mRNA. Data produced after treatment with anisomycin can therefore potentially be used to study the timing of movements between different ribosome conformations during each elongation cycle. Understanding whether anisomycin association and dissociation rates are modulated by codon identities in the same way as CHX will be necessary in order to correctly interpret this kind of data. Although they have not yet been studied as extensively as yeast, ribosome profiling data from several higher eukaryotes exists, include *C. elegans* [106] and human and mouse cell lines [38, 47, 66]. The interactions of translation inhibitors with ribosomes may be qualitatively different in these organisms than in yeast. The relative simplicity of the yeast transcriptome - in particular, the fact that there is virtually no alternative splicing of yeast coding sequences - made it an ideal model organism in which to identify patterns in ribosome density across long distances in coding sequences. In higher eukaryotes for which substantial amounts of alternative splicing occur, interpreting aggregate patterns in ribosome density around different codon identities may

present additional computational challenges.

Finally, the identification of downstream waves in ribosome density after treatment with CHX was largely made possible by of the unusual nature of the tRNA decoding CGA in yeast. The I-A wobble pairing in this codon-anticodon interaction appears to produce a substantial slowdown in translation when positioned in either the A- or the P-site and to produce a substantial relative speedup in translation in the presence of CHX, perhaps by increasing rates of CHX disassociation from ribosomes. Together, these facts combine to produce the large wave downstream of CGA that is consistently the clearest evidence for continued elongation in the presence of CHX. Letzring et al. [67] have demonstrated through alternative experimental methods that introducing a synthetically modified tRNA that has an exactly matched anticodon to CGA dramatically reduces translational pausing at CGAs. Given the apparent ease with which yeast could evolve to eliminate the slow translation of CGAs, the fact that evolution has avoided doing so suggests that this pausing may serve a functional role, perhaps by interacting with co-translational mRNA decay pathways [85, 90] or by providing control over co-translational folding of nascent polypeptides [81]. It would be interesting to perform ribosome profiling with and without CHX pretreatment on yeast cells in which Letzring's modified CGA-decoding tRNA has been introduced. This would serve as a powerful test of our continued-elongation hypothesis, since our model predicts that this modified tRNA should eliminate both the wave downstream of CGA and the depletion of ribosomes with CGA in the A- and P-sites that are other-

wise ubiquitous in CHX-pretreated data. Transcriptome-wide measurements of mRNA levels and ribosome locations after alleviating pauses at CGAs could also potentially provide insights into the functional roles that these pauses play.

Bibliography

[1] Ashley Acevedo, Leonid Brodsky, and Raul Andino. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature*, November 2013.

[2] S G Andersson and C G Kurland. Codon preferences in free-living microorganisms. *Microbiological reviews*, 54(2):198–210, June 1990.

[3] Carlo G Artieri and Hunter B Fraser. Accounting for biases in riboprofiling data indicates a major role for proline in stalling translation. *Genome Research*, pages 2011–2021, 2014.

[4] Carlo G Artieri and Hunter B Fraser. Evolution at two levels of gene expression in yeast. *Genome research*, 24(3):411–21, March 2014.

[5] Deborah E Barnes and Tomas Lindahl. Repair and genetic consequences of endogenous DNA base damage in mammalian cells. *Annual review of genetics*, 38:445–476, 2004.

[6] Stefan Behnel, Robert Bradshaw, Craig Citro, Lisandro Dalcin, Dag Sverre Seljebotn, and Kurt Smith. Cython: The Best of Both Worlds. *Computing in Science & Engineering*, 13(2), 2011.

[7] David R Bentley, Shankar Balasubramanian, Harold P Swerdlow, Geoffrey P Smith, John Milton, Clive G Brown, Kevin P Hall, Dirk J Evers, Colin L Barnes, Helen R Bignell, Jonathan M Boutell, Jason Bryant, Richard J Carter, R Keira Cheetham, Anthony J Cox, Darren J Ellis, Michael R Flatbush, Niall a Gormley, Sean J Humphray, Leslie J Irving, Mirian S Karbelashvili, Scott M Kirk, Heng Li, Xiaohai Liu, Klaus S Maisinger, Lisa J Murray, Bojan Obradovic, Tobias Ost, Michael L Parkinson, Mark R Pratt, Isabelle M J Rasolonjatovo, Mark T Reed, Roberto Rigatti, Chiara Rodighiero, Mark T Ross, Andrea Sabot, Subramanian V Sankar, Aylwyn Scally, Gary P Schroth, Mark E

Smith, Vincent P Smith, Anastassia Spiridou, Peta E Torrance, Svilen S Tzonev, Eric H Vermaas, Klaudia Walter, Xiaolin Wu, Lu Zhang, Mohammed D Alam, Carole Anastasi, Ify C Aniebo, David M D Bailey, Iain R Bancarz, Saibal Banerjee, Selena G Barbour, Primo a Baybayan, Vincent a Benoit, Kevin F Benson, Claire Bevis, Phillip J Black, Asha Boodhun, Joe S Brennan, John a Bridgham, Rob C Brown, Andrew a Brown, Dale H Buermann, Abass a Bundu, James C Burrows, Nigel P Carter, Nestor Castillo, Maria Chiara E Catenazzi, Simon Chang, R Neil Cooley, Natasha R Crake, Olubunmi O Dada, Konstantinos D Diakoumakos, Belen Dominguez-Fernandez, David J Earnshaw, Ugonna C Egbujor, David W Elmore, Sergey S Etchin, Mark R Ewan, Milan Fedurco, Louise J Fraser, Karin V Fuentes Fajardo, W Scott Furey, David George, Kimberley J Gietzen, Colin P Goddard, George S Golda, Philip a Granieri, David E Green, David L Gustafson, Nancy F Hansen, Kevin Harnish, Christian D Haudenschild, Narinder I Heyer, Matthew M Hims, Johnny T Ho, Adrian M Horgan, Katya Hoschler, Steve Hurwitz, Denis V Ivanov, Maria Q Johnson, Terena James, T a Huw Jones, Gyoung-Dong Kang, Tzvetana H Kerelska, Alan D Kersey, Irina Khrebtukova, Alex P Kindwall, Zoya Kingsbury, Paula I Kokko-Gonzales, Anil Kumar, Marc a Laurent, Cynthia T Lawley, Sarah E Lee, Xavier Lee, Arnold K Liao, Jennifer a Loch, Mitch Lok, Shujun Luo, Radhika M Mammen, John W Martin, Patrick G McCauley, Paul McNitt, Parul Mehta, Keith W Moon, Joe W Mullens, Taksina Newington, Zemin Ning, Bee Ling Ng, Sonia M Novo, Michael J O'Neill, Mark a Osborne, Andrew Osnowski, Omead Ostadan, Lambros L Paraschos, Lea Pickering, Andrew C Pike, Alger C Pike, D Chris Pinkard, Daniel P Pliskin, Joe Podhasky, Victor J Quijano, Come Raczy, Vicki H Rae, Stephen R Rawlings, Ana Chiva Rodriguez, Phyllida M Roe, John Rogers, Maria C Rogert Bacigalupo, Nikolai Romanov, Anthony Romieu, Rithy K Roth, Natalie J Rourke, Silke T Ruediger, Eli Rusman, Raquel M Sanches-Kuiper, Martin R Schenker, Josefina M Seoane, Richard J Shaw, Mitch K Shiver, Steven W Short, Ning L Sizto, Johannes P Sluis, Melanie a Smith, Jean Ernest Sohna Sohna, Eric J Spence, Kim Stevens, Neil Sutton, Lukasz Szajkowski, Carolyn L Tregidgo, Gerardo Turcatti, Stephanie Vandevondele, Yuli Verhovsky, Selene M Virk, Suzanne Wakelin, Gre-

gory C Walcott, Jingwen Wang, Graham J Worsley, Juying Yan, Ling Yau, Mike Zuerlein, Jane Rogers, James C Mullikin, Matthew E Hurles, Nick J McCooke, John S West, Frank L Oaks, Peter L Lundberg, David Klenerman, Richard Durbin, and Anthony J Smith. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–9, November 2008.

[8] Bradley E Bernstein, Ewan Birney, Ian Dunham, Eric D Green, Chris Gunter, and Michael Snyder. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012.

[9] GA Brar, Moran Yassour, Nir Friedman, and Aviv Regev. High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science*, 335(February), 2012.

[10] Kristin Brogaard, Liqun Xi, Ji-Ping Wang, and Jonathan Widom. A map of nucleosome positions in yeast at base-pair resolution. *Nature*, 486(7404):496–501, June 2012.

[11] Michael Bulmer. The selection-mutation-drift theory of synonymous codon usage. *Genetics*, (1987), 1991.

[12] M Burrows and DJ Wheeler. A block-sorting lossless data compression algorithm. 1994.

[13] Gina Cannarozzi, Gina Cannarrozzi, Nicol N Schraudolph, Mahamadou Faty, Peter von Rohr, Markus T Friberg, Alexander C Roth, Pedro Gonnet, Gaston Gonnet, and Yves Barral. A role for codon order in translation dynamics. *Cell*, 141(2):355–67, April 2010.

[14] J V Chamary, Joanna L Parmley, and Laurence D Hurst. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature reviews. Genetics*, 7(2):98–108, February 2006.

[15] Kyle Chang, Chad J Creighton, Caleb Davis, Lawrence Donehower, Jennifer Drummond, David Wheeler, Adrian Ally, Miruna Balasundaram, Inanc Birol, Yaron S N Butterfield, Andy Chu, Eric Chuah, Hye-Jung E Chun, Noreen Dhalla, Ranabir Guin, Martin Hirst, Carrie Hirst, Robert a

Holt, Steven J M Jones, Darlene Lee, Haiyan I Li, Marco a Marra, Michael Mayo, Richard a Moore, Andrew J Mungall, a Gordon Robertson, Jacqueline E Schein, Payal Sipahimalani, Angela Tam, Nina Thiessen, Richard J Varhol, Rameen Beroukhim, Ami S Bhatt, Angela N Brooks, Andrew D Cherniack, Samuel S Freeman, Stacey B Gabriel, Elena Helman, Joonil Jung, Matthew Meyerson, Akinyemi I Ojesina, Chandra Sekhar Pedamallu, Gordon Saksena, Steven E Schumacher, Barbara Tabak, Travis Zack, Eric S Lander, Christopher a Bristow, Angela Hadjipanayis, Psalm Haseley, Raju Kucherlapati, Semin Lee, Eunjung Lee, Lovelace J Luquette, Harshad S Mahadeshwar, Angeliki Pantazi, Michael Parfenov, Peter J Park, Alexei Protopopov, Xiaojia Ren, Netty Santoso, Jonathan Seidman, Sahil Seth, Xingzhi Song, Jiabin Tang, Ruibin Xi, Andrew W Xu, Lixing Yang, Dong Zeng, J Todd Auman, Saianand Balu, Elizabeth Buda, Cheng Fan, Katherine a Hoadley, Corbin D Jones, Shaowu Meng, Piotr a Mieczkowski, Joel S Parker, Charles M Perou, Jeffrey Roach, Yan Shi, Grace O Silva, Donghui Tan, Umadevi Veluvolu, Scot Waring, Matthew D Wilkerson, Junyuan Wu, Wei Zhao, Tom Bodenheimer, D Neil Hayes, Alan P Hoyle, Stuart R Jeffreys, Lisle E Mose, Janae V Simons, Mathew G Soloway, Stephen B Baylin, Benjamin P Berman, Moiz S Bootwalla, Ludmila Danilova, James G Herman, Toshinori Hinoue, Peter W Laird, Suhn K Rhie, Hui Shen, Timothy Triche, Daniel J Weisenberger, Scott L Carter, Kristian Cibulskis, Lynda Chin, Jianhua Zhang, Gad Getz, Carrie Sougnez, Min Wang, Huyen Dinh, Harsha Vardhan Doddapaneni, Richard Gibbs, Preethi Gunaratne, Yi Han, Divya Kalra, Christie Kovar, Lora Lewis, Margaret Morgan, Donna Morton, Donna Muzny, Jeffrey Reid, Liu Xi, Juok Cho, Daniel Dicara, Scott Frazer, Nils Gehlenborg, David I Heiman, Jaegil Kim, Michael S Lawrence, Pei Lin, Yingchun Liu, Michael S Noble, Petar Stojanov, Doug Voet, Hailei Zhang, Lihua Zou, Chip Stewart, Brady Bernard, Ryan Bressler, Andrea Eakin, Lisa Iype, Theo Knijnenburg, Roger Kramer, Richard Kreisberg, Kalle Leinonen, Jake Lin, Yuexin Liu, Michael Miller, Sheila M Reynolds, Hector Rovira, Ilya Shmulevich, Vesteinn Thorsson, Da Yang, Wei Zhang, Samirkumar Amin, Chang-Jiun Wu, Chia-Chin Wu, Rehan Akbani, Kenneth Aldape, Keith a Baggerly, Bradley Broom, Tod D Casasent, James Cleland, Chad Creighton, Deepti Dodda, Mary

Edgerton, Leng Han, Shelley M Herbrich, Zhenlin Ju, Hoon Kim, Seth Lerner, Jun Li, Han Liang, Wenbin Liu, Philip L Lorenzi, Yiling Lu, James Melott, Gordon B Mills, Lam Nguyen, Xiaoping Su, Roeland Verhaak, Wenyi Wang, John N Weinstein, Andrew Wong, Yang Yang, Jun Yao, Rong Yao, Kosuke Yoshihara, Yuan Yuan, Alfred K Yung, Nianxiang Zhang, Siyuan Zheng, Michael Ryan, David W Kane, B Arman Aksoy, Giovanni Ciriello, Gideon Dresdner, Jianjiong Gao, Benjamin Gross, Anders Jacobsen, Andre Kahles, Marc Ladanyi, William Lee, Kjong-Van Lehmann, Martin L Miller, Ricardo Ramirez, Gunnar Rätsch, Boris Reva, Chris Sander, Nikolaus Schultz, Yasin Senbabaoglu, Ronglai Shen, Rileen Sinha, S Onur Sumer, Yichao Sun, Barry S Taylor, Nils Weinhold, Suzanne Fei, Paul Spellman, Christopher Benz, Daniel Carlin, Melisssa Cline, Brian Craft, Kyle Ellrott, Mary Goldman, David Haussler, Singer Ma, Sam Ng, Evan Paull, Amie Radenbaugh, Sofie Salama, Artem Sokolov, Joshua M Stuart, Teresa Swatloski, Vladislav Uzunangelov, Peter Waltman, Christina Yau, Jing Zhu, Stanley R Hamilton, Scott Abbott, Rachel Abbott, Nathan D Dees, Kim Delehaunty, Li Ding, David J Dooling, Jim M Eldred, Catrina C Fronick, Robert Fulton, Lucinda L Fulton, Joelle Kalicki-Veizer, Krishna-Latha Kanchi, Cyriac Kandoth, Daniel C Koboldt, David E Larson, Timothy J Ley, Ling Lin, Charles Lu, Vincent J Magrini, Elaine R Mardis, Michael D McLellan, Joshua F McMichael, Christopher a Miller, Michelle O'Laughlin, Craig Pohl, Heather Schmidt, Scott M Smith, Jason Walker, John W Wallis, Michael C Wendl, Richard K Wilson, Todd Wylie, Qunyuan Zhang, Robert Burton, Mark a Jensen, Ari Kahn, Todd Pihl, David Pot, Yunhu Wan, Douglas a Levine, Aaron D Black, Jay Bowen, Jessica Frick, Julie M Gastier-Foster, Hollie a Harper, Carmen Helsel, Kristen M Leraas, Tara M Lichtenberg, Cynthia McAllister, Nilsa C Ramirez, Samantha Sharpe, Lisa Wise, Erik Zmuda, Stephen J Chanock, Tanja Davidsen, John a Demchok, Greg Eley, Ina Felau, Brad a Ozenberger, Margi Sheth, Heidi Sofia, Louis Staudt, Roy Tarnuzzer, Zhining Wang, Liming Yang, Jiashan Zhang, Larsson Omberg, Adam Margolin, Benjamin J Raphael, Fabio Vandin, Hsin-Ta Wu, Mark D M Leiserson, Stephen C Benz, Charles J Vaske, Houtan Noushmehr, Denise Wolf, Laura Van't Veer, Eric a Collisson, Dimitris Anastassiou, Tai-Hsien Ou Yang, Nuria

Lopez-Bigas, Abel Gonzalez-Perez, David Tamborero, Zheng Xia, Wei Li, Dong-Yeon Cho, Teresa Przytycka, Mark Hamilton, Sean McGuire, Sven Nelander, Patrik Johansson, Rebecka Jörnsten, Teresia Kling, Jose Sanchez, and Kenna R Mills Shaw. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

[16] Catherine A Charneski and Laurence D Hurst. Positively Charged Residues Are the Major Determinants of Ribosomal Velocity. *PLoS Biology*, 11(3), 2013.

[17] Catherine a. Charneski and Laurence D. Hurst. Positive charge loading at protein termini is due to membrane protein topology, not a translational ramp. *Molecular Biology and Evolution*, 31(1):70–84, 2014.

[18] Christopher J Chetsanga and Tomas Lindahl. Realse of 7-methylguanine residues whose imidazole rings have been opened from damaged DNA by a DNA glycosylase from Escherichia coli. *Nucleic acids research*, 6(11):3673–3684, 1979.

[19] Alexandra Dana and Tamir Tuller. The effect of tRNA levels on decoding times of mRNA codons. *Nucleic acids research*, July 2014.

[20] Nicolas Garreau de Loubresse, Irina Prokhorova, Wolf Holtkamp, Marina V. Rodnina, Gulnara Yusupova, and Marat Yusupov. Structural basis for the inhibition of the eukaryotic ribosome. *Nature*, 513(7519):517–522, September 2014.

[21] Mark a DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony a Philippakis, Guillermo del Angel, Manuel a Rivas, Matt Hanna, Aaron McKenna, Tim J Fennell, Andrew M Kernytsky, Andrey Y Sivachenko, Kristian Cibulskis, Stacey B Gabriel, David Altshuler, and Mark J Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5):491–498, 2011.

[22] Kimberly a. Dittmar, Jeffrey M. Goodenbour, and Tao Pan. Tissue-specific differences in human transfer RNA expression. *PLoS Genetics*, 2(12):2107–2115, 2006.

[23] Mario dos Reis, Renos Savva, and Lorenz Wernisch. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic acids research*, 32(17):5036–44, January 2004.

[24] D Allan Drummond and Claus O Wilke. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, 134(2):341–52, July 2008.

[25] Joshua G Dunn, Catherine K Foo, Nicolette G Belletier, Elizabeth R Gavis, and Jonathan S Weissman. Ribosome profiling reveals pervasive and regulated stop codon readthrough in Drosophila melanogaster. *eLife*, 2:e01179, January 2013.

[26] Laurent Duret. Evolution of synonymous codon usage in metazoans. *Current opinion in genetics & development*, 12(6):640–9, December 2002.

[27] J Eid, A Fehr, J Gray, K Luong, J Lyle, and G Otto. Real-time DNA sequencing from single polymerase molecules. *Science*, (January):133–138, 2009.

[28] P. Ferragina and G. Manzini. Opportunistic data structures with applications. *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 390–398, 2000.

[29] Nuno a Fonseca, Johan Rung, Alvis Brazma, and John C Marioni. Tools for mapping high-throughput sequencing data. *Bioinformatics (Oxford, England)*, 28(24):3169–77, December 2012.

[30] Justin Gardin, Rukhsana Yeasmin, Alisa Yurovsky, Ying Cai, Steve Skiena, and Bruce Futcher. Measurement of average decoding rates of the 61 sense codons in vivo. *eLife*, 3:1–20, January 2014.

[31] Maxim V Gerashchenko and Vadim N Gladyshev. Translation inhibitors cause abnormalities in ribosome profiling experiments. *Nucleic acids research*, 42(17):1–7, 2014.

[32] Maxim V Gerashchenko, Alexei V Lobanov, and Vadim N Gladyshev. Genome-wide ribosome profiling reveals complex translational regulation

in response to oxidative stress. *Proceedings of the National Academy of Sciences of the United States of America*, 109(43):17394–9, October 2012.

[33] Jean-François Gout, W Kelley Thomas, Zachary Smith, Kazufusa Okamoto, and Michael Lynch. Large-scale detection of in vivo transcription errors. *Proceedings of the National Academy of Sciences of the United States of America*, 110(46), October 2013.

[34] R Green and H F Noller. Ribosomes and translation. *Annual review of biochemistry*, 66:679–716, 1997.

[35] Henri Grosjean, Valérie de Crécy-Lagard, and Christian Marck. Deciphering synonymous codons in the three domains of life: Co-evolution with specific tRNA modification enzymes. *FEBS Letters*, 584(2):252–264, 2010.

[36] Saskia Grudzenski, Antonia Raths, Sandro Conrad, Claudia E Rübe, and Markus Löbrich. Inducible response required for repair of low-dose radiation damage in human fibroblasts. *Proceedings of the National Academy of Sciences of the United States of America*, 107(32):14205–10, August 2010.

[37] Wanjun Gu, Tong Zhou, and Claus O Wilke. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS computational biology*, 6(2):e1000664, March 2010.

[38] Huili Guo, Nicholas T Ingolia, Jonathan S Weissman, and David P Bartel. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, 466(7308):835–40, August 2010.

[39] Nicholas R. Guydosh and Rachel Green. Dom34 Rescues Ribosomes in 3' Untranslated Regions. *Cell*, 156(5):950–962, February 2014.

[40] Benjamin J. Hindson, Kevin D. Ness, Donald a. Masquelier, Phillip Belgrader, Nicholas J. Heredia, Anthony J. Makarewicz, Isaac J. Bright,

Michael Y. Lucero, Amy L. Hiddessen, Tina C. Legler, Tyler K. Kitano, Michael R. Hodel, Jonathan F. Petersen, Paul W. Wyatt, Erin R. Steenblock, Pallavi H. Shah, Luc J. Bousse, Camille B. Troup, Jeffrey C. Mellen, Dean K. Wittmann, Nicholas G. Erndt, Thomas H. Cauley, Ryan T. Koehler, Austin P. So, Simant Dube, Klint a. Rose, Luz Montesclaros, Shenglong Wang, David P. Stumbo, Shawn P. Hodges, Steven Romine, Fred P. Milanovich, Helen E. White, John F. Regan, George a. Karlin-Neumann, Christopher M. Hindson, Serge Saxonov, and Bill W. Colston. High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. *Analytical Chemistry*, 83(22):8604–8610, 2011.

[41] John D Hunter. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3), 2007.

[42] Illumina. Using a PhiX Control for HiSeq Sequencing Runs. Technical report, Illumina, Inc., 2012.

[43] Nicholas T Ingolia. *Genome-wide translational profiling by ribosome footprinting.*, volume 470. Elsevier Inc., 2 edition, January 2010.

[44] Nicholas T Ingolia. Ribosome profiling: new views of translation, from single codons to genome scale. *Nature reviews. Genetics*, 15(3):205–13, March 2014.

[45] Nicholas T Ingolia, Gloria a Brar, Silvia Rouskin, Anna M McGeachy, and Jonathan S Weissman. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nature protocols*, 7(8):1534–50, August 2012.

[46] Nicholas T Ingolia, Sina Ghaemmaghami, John R S Newman, and Jonathan S Weissman. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science*, 324(April), 2009.

[47] Nicholas T Ingolia, Liana F Lareau, and Jonathan S Weissman. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, 147(4):789–802, November 2011.

[48] Cassandra B Jabara, Corbin D Jones, Jeffrey Roach, Jeffrey a Anderson, and Ronald Swanstrom. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proceedings of the National Academy of Sciences of the United States of America*, 108(50):20166–71, December 2011.

[49] D. Jain, S. Baldi, a. Zabel, T. Straub, and P. B. Becker. Active promoters give rise to false positive 'Phantom Peaks' in ChIP-seq experiments. *Nucleic Acids Research*, pages 1–10, 2015.

[50] C. H. Jan, C. C. Williams, and J. S. Weissman. Principles of ER cotranslational translocation revealed by proximity-specific ribosome profiling. *Science*, 2014.

[51] Marcus J O Johansson, Anders Esberg, Bo Huang, Glenn R Björk, and Anders S Byström. Eukaryotic wobble uridine modifications promote a functionally redundant decoding system. *Molecular and cellular biology*, 28(10):3301–3312, 2008.

[52] Lee D Kapp and Jon R Lorsch. The molecular mechanics of eukaryotic translation. *Annual review of biochemistry*, 73:657–704, 2004.

[53] Luba Katz and Christopher B Burge. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome research*, 13(9):2042–51, September 2003.

[54] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, 14(4):R36, April 2013.

[55] Isaac Kinde, Jian Wu, Nick Papadopoulos, Kenneth W Kinzler, and Bert Vogelstein. Detection and quantification of rare mutations with massively parallel sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 108(23):9530–5, June 2011.

[56] Daniel C Koboldt, Karyn Meltz Steinberg, David E Larson, Richard K Wilson, and Elaine R Mardis. The next-generation sequencing revolution and its impact on genomics. *Cell*, 155(1):27–38, September 2013.

[57] Anton a Komar. A pause for thought along the co-translational folding pathway. *Trends in biochemical sciences*, 34(1):16–24, January 2009.

[58] G Kudla, AW Murray, D Tollervey, and JB Plotkin. Coding-sequence determinants of gene expression in Escherichia coli. *science*, (April):255–258, 2009.

[59] Thomas a Kunkel and Katarzyna Bebenek. DNA Replication Fidelity. *Annual review of biochemistry*, pages 497–529, 2000.

[60] F Lacroute. RNA and protein elongation rates in Saccharomyces cerevisiae. *Molecular & general genetics : MGG*, 125(4):319–327, 1973.

[61] Daniel J G Lahr and Laura a. Katz. Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. *BioTechniques*, 47(4):857–866, 2009.

[62] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–360, 2012.

[63] Ben Langmead, Cole Trapnell, Mihai Pop, and SL Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3), 2009.

[64] Liana F Lareau, Dustin H Hite, Gregory J Hogan, and Patrick O Brown. Distinct stages of the translation elongation cycle revealed by sequencing ribosome-protected mRNA fragments. *eLife*, 3:e01257, January 2014.

[65] Yizhar Lavner and Daniel Kotlar. Codon bias as a factor in regulating expression via translation rate in the human genome. *Gene*, 345(1):127–38, January 2005.

[66] Sooncheol Lee, Botao Liu, Soohyun Lee, Sheng-Xiong Huang, Ben Shen, and Shu-Bing Qian. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proceedings of the National Academy of Sciences of the United States of America*, 109(37):E2424–32, September 2012.

[67] Daniel P Letzring, Kimberly M Dean, and Elizabeth J Grayhack. Control of translation efficiency in yeast by codon anticodon interactions. *RNA*, pages 2516–2528, 2010.

[68] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14):1754–60, July 2009.

[69] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–9, August 2009.

[70] Heng Li and Nils Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in bioinformatics*, 11(5):473–83, September 2010.

[71] Erez Lieberman-aiden, Nynke L Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, and Leonid A Mirny. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, 33292(October):289–294, 2009.

[72] Nicholas J Loman, Raju V Misra, Timothy J Dallman, Chrystala Constantinidou, Saheer E Gharbia, John Wain, and Mark J Pallen. Performance comparison of benchtop high-throughput sequencing platforms. *Nature biotechnology*, 30(5):434–9, May 2012.

[73] D. I. Lou, J. A. Hussmann, R. M. McBee, A. Acevedo, R. Andino, W. H. Press, and S. L. Sawyer. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proceedings of the National Academy of Sciences*, November 2013.

[74] Michael Lynch. Rate, molecular spectrum, and consequences of human mutation. *Proceedings of the National Academy of Sciences of the United States of America*, 107(3):961–8, January 2010.

[75] Marcel Margulies, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa a Bemben, Jan Berka, Michael S Braverman, Yi-Ju Chen, Zhoutao Chen, Scott B Dewell, Lei Du, Joseph M Fierro, Xavier V Gomes, Brian C Godwin, Wen He, Scott Helgesen, Chun Heen Ho, Chun He Ho, Gerard P Irzyk, Szilveszter C Jando, Maria L I Alenquer, Thomas P Jarvie, Kshama B Jirage, Jong-Bum Kim, James R Knight, Janna R Lanza, John H Leamon, Steven M Lefkowitz, Ming Lei, Jing Li, Kenton L Lohman, Hong Lu, Vinod B Makhijani, Keith E McDade, Michael P McKenna, Eugene W Myers, Elizabeth Nickerson, John R Nobile, Ramona Plant, Bernard P Puc, Michael T Ronan, George T Roth, Gary J Sarkis, Jan Fredrik Simons, John W Simpson, Maithreyan Srinivasan, Karrie R Tartaro, Alexander Tomasz, Kari a Vogt, Greg a Volkmer, Shally H Wang, Yong Wang, Michael P Weiner, Pengguang Yu, Richard F Begley, and Jonathan M Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–80, September 2005.

[76] Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 51–56, 2010.

[77] C Joel McManus, Gemma E May, Pieter Spealman, and Alan Shteyman. Ribosome profiling reveals post-transcriptional buffering of divergent gene expression in yeast. *Genome research*, 24(3):422–30, March 2014.

[78] Andreas Meyerhans, Jean-Pierre Vartanian, and Simon Wain-Hobson. DNA recombination during PCR. *Nucleic Acids Research*, 18(7):1687–1691, 1990.

[79] Supratim Mukherjee, Marcel Huntemann, Natalia Ivanova, Nikos C Kyrpides, and Amrita Pati. Large-scale contamination of microbial isolate genomes by Illumina PhiX control. *Standards in Genomic Sciences*, 10(1):1–4, 2015.

[80] Ugrappa Nagalakshmi. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(June), 2008.

[81] Danny D Nedialkova and Sebastian a Leidel. Optimization of Codon Translation Rates via tRNA Modifications Maintains Proteome Integrity Article Optimization of Codon Translation Rates via tRNA Modifications Maintains Proteome Integrity. *Cell*, pages 1–13, 2015.

[82] Travis E Oliphant. Python for Scientific Computing. *Computing in Science & Engineering*, 9(3), 2007.

[83] Daechan Park, Yaelim Lee, Gurvani Bhupindersingh, and Vishwanath R. Iyer. Widespread misinterpretable ChIP-seq bias in yeast. *PLoS ONE*, 8(12):1–16, 2013.

[84] Vicent Pelechano, Wu Wei, and Lars M Steinmetz. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature*, 497(7447):127–31, May 2013.

[85] Vicent Pelechano, Wu Wei, and LarsM. Steinmetz. Widespread Co-translational RNA Decay Reveals Ribosome Dynamics. *Cell*, 161(6):1400–1412, 2015.

[86] R Percudani, a Pavesi, and S Ottonello. Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. *Journal of molecular biology*, 268(2):322–330, 1997.

[87] Fernando Pérez and Brian E Granger. IPython: A System for Interactive Scientific Computing. *Computing in Science & Engineering*, 9(3), 2007.

[88] Joshua B Plotkin and Grzegorz Kudla. Synonymous but not the same: the causes and consequences of codon bias. *Nature reviews. Genetics*, 12(1):32–42, January 2011.

[89] Cristina Pop, Silvi Rouskin, Nicholas T Ingolia, Lu Han, Eric M Phizicky, Jonathan S Weissman, and Daphne Koller. Causal signals between codon bias , mRNA structure , and the efficiency of translation and elongation. *Molecular Systems Biology*, pages 1–15, 2014.

[90] Vladimir Presnyak, Najwa Alhusaini, Ying-Hsin Chen, Sophie Martin, Nathan Morris, Nicholas Kline, Sara Olson, David Weinberg, KristianE. Baker, BrentonR. Graveley, and Jeff Coller. Codon Optimality Is a Major Determinant of mRNA Stability. *Cell*, 160(6):1111–1124, 2015.

[91] William H Press, Saul A Teukolsky, William T Vetterling, and Brian P Flannery. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA, 3 edition, 2007.

[92] Wenfeng Qian, Jian-Rong Yang, Nathaniel M Pearson, Calum Maclean, and Jianzhi Zhang. Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS genetics*, 8(3):e1002603, January 2012.

[93] Hong Qin, Wei Biao Wu, Josep M Comeron, Martin Kreitman, and Wen-Hsiung Li. Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. *Genetics*, 168(4):2245–60, December 2004.

[94] Vanessa Anissa Nathalie Rezgui, Kshitiz Tyagi, Namit Ranjan, Andrey L Konevega, Joerg Mittelstaet, Marina V Rodnina, Matthias Peter, and Patrick G a Pedrioli. tRNA tKUUU, tQUUG, and tEUUC wobble position modifications fine-tune protein translation by promoting ribosome A-site binding. *Proceedings of the National Academy of Sciences of the United States of America*, 110(30):12289–94, 2013.

[95] Jonathan M Rothberg, Wolfgang Hinz, Todd M Rearick, Jonathan Schultz, William Mileski, Mel Davey, John H Leamon, Kim Johnson, Mark J Milgrew, Matthew Edwards, Jeremy Hoon, Jan F Simons, David Marran, Jason W Myers, John F Davidson, Annika Branting, John R Nobile, Bernard P Puc, David Light, Travis a Clark, Martin Huber, Jeffrey T Branciforte, Isaac B Stoner, Simon E Cawley, Michael Lyons, Yutao Fu, Nils Homer, Marina Sedova, Xin Miao, Brian Reed, Jeffrey Sabina, Erika Feierstein, Michelle Schorn, Mohammad Alanjary, Eileen Dimalanta, Devin Dressman, Rachel Kasinskas, Tanya Sokolsky, Jacqueline a Fidanza, Eugeni Namsaraev, Kevin J McKernan, Alan Williams, G Thomas Roth, and James Bustillo. An integrated semiconductor

device enabling non-optical genome sequencing. *Nature*, 475(7356):348–52, July 2011.

[96] Michael W Schmitt, Scott R Kennedy, Jesse J Salk, Edward J Fox, Joseph B Hiatt, and Lawrence a Loeb. Detection of ultra-rare mutations by next-generation sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 109(36):14508–13, September 2012.

[97] Tilman Schneider-Poetsch, Jianhua Ju, Daniel E Eyler, Yongjun Dang, Shridhar Bhat, William C Merrick, Rachel Green, Ben Shen, and Jun O Liu. Inhibition of eukaryotic translation elongation by cycloheximide and lactimidomycin. *Nature chemical biology*, 6(3):209–217, March 2010.

[98] Premal Shah, Yang Ding, Malwina Niemczyk, Grzegorz Kudla, and Joshua B Plotkin. Rate-limiting steps in yeast protein translation. *Cell*, 153(7):1589–601, June 2013.

[99] P M Sharp, M Averof, a T Lloyd, G Matassi, and J F Peden. DNA sequence evolution: the sounds of silence. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 349(1329):241–7, September 1995.

[100] PM Sharp and WH Li. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic acids research*, 1987.

[101] J C Shen, W M Rideout, and P a Jones. The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. *Nucleic acids research*, 22(6):972–976, 1994.

[102] Katsuyuki Shiroguchi, Tony Z Jia, Peter a Sims, and X Sunney Xie. Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proceedings of the National Academy of Sciences of the United States of America*, 109(4):1347–52, January 2012.

[103] Mikhail Shugay, Olga V Britanova, Ekaterina M Merzlyak, Maria a Turchaninova, Ilgar Z Mamedov, Timur R Tuganbaev, Dmitriy a Bolotin, Dmitry B Staroverov, Ekaterina V Putintseva, Karla Plevova, Carsten Linnemann, Dmitriy Shagin, Sarka Pospisilova, Sergey Lukyanov, Ton N Schumacher, and Dmitriy M Chudakov. Towards error-free profiling of immune repertoires. *Nature methods*, 11(6):653–5, 2014.

[104] T. Smith and M. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197, 1981.

[105] Wendy Weijia Soon, Manoj Hariharan, and Michael P Snyder. High-throughput sequencing for biology and medicine. *Molecular Systems Biology*, 9(640):1–14, 2013.

[106] Michael Stadler and Andrew Fire. Wobble base-pairing slows in vivo translation elongation in metazoans. *RNA*, page 2073, 2011.

[107] Thomas a Steitz. A structural understanding of the dynamic ribosome machine. *Nature reviews. Molecular cell biology*, 9(3):242–253, 2008.

[108] Leonid Teytelman and DM Thurtle. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proceedings of the ...*, pages 2–7, 2013.

[109] Cole Trapnell and SL Salzberg. How to map billions of short reads onto genomes. *Nature biotechnology*, 27(5):455–457, 2009.

[110] Kevin J Travers, Chen-Shan Chin, David R Rank, John S Eid, and Stephen W Turner. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic acids research*, 38(15):e159, August 2010.

[111] Tamir Tuller, Asaf Carmi, Kalin Vestsigian, Sivan Navon, Yuval Dorfan, John Zaborske, Tao Pan, Orna Dahan, Itay Furman, and Yitzhak Pilpel. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, 141(2):344–54, April 2010.

[112] Stéfan van der Walt, S Chris Colbert, and Gaël Varoquaux. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering*, 13(2), 2011.

[113] S Varenne, J Buc, R Lloubes, and C Lazdunski. Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains. *Journal of molecular biology*, 180(3):549–76, December 1984.

[114] Michael Waskom, Kyle Meyer, Paul Hobson, Yaroslav Halchenko, Miikka Koskinen, Alistair Miles, Daniel Wehner, Olga Botvinnik, Tobias Megies, Cynddl, Erik Ziegler, Tal Yarkoni, Yury V. Zaytsev, Luis Pedro Coelho, John B. Cole, Tom Augspurger, Diego0020, Travis Hoppe, Skipper Seabold, Phillip Cloud, Stephan Hoyer, Adel Qalieh, and Dan Allan. seaborn: v0.5.0 (November 2014). November 2014.

[115] David E Weinberg, Premal Shah, Stephen W Eichhorn, Jeffrey A Hussmann, Joshua B Plotkin, and David P Bartel. Improved ribosome-footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation. 2015.

[116] Richard a White, Paul C Blainey, H Christina Fan, and Stephen R Quake. Digital PCR provides sensitive and absolute calibration for high throughput sequencing. *BMC genomics*, 10:116, January 2009.

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

[118] Richard Williams, Sergio G Peisajovich, Oliver J Miller, Shlomo Magdassi, Dan S Tawfik, and Andrew D Griffiths. Amplification of complex gene libraries by emulsion PCR. *Nature methods*, 3(7):545–550, 2006.

[119] Andrew S Wolf and Elizabeth J Grayhack. Asc1, homolog of human RACK1, prevents frameshifting in yeast by ribosomes stalled at CGA codon repeats. *RNA*, pages 1–11, 2015.

[120] Frank Wright. The 'effective number of codons' used in a gene. *Gene*, 87:23–29, 1990.

[121] Boris Zinshteyn and Wendy V Gilbert. Loss of a conserved tRNA anticodon modification perturbs cellular signaling. *PLoS genetics*, 9(8):e1003675, August 2013.