CHAPTER 3

High-Resolution Mapping of the *Escherichia coli* Chromosome Reveals Positions of High and Low Transcription

3.1 Contribution details

This work was reproduced from its published form, with permission, from Scholz et al. [197]. I am third author on this publication and initially became involved in it through the development of a command-line, sliding window, tool that enabled the primary author, Scott Scholz, to interrogate the primary data and create many of the figures in the final manuscript. Throughout this data analysis process, I had many conversations with Scott, Rucheng Diao (the second author), and Peter Freddolino about, not only technical issues with the data analysis itself, but also with biological questions and directions the data analysis could go. Once the initial draft was written by Scott, I provided editing and critical revisions of the manuscript. Through the review process, I contributed additional analyses and wrote manuscript text on the discovery of the periodicity of the transcriptional propensity signal and the correlation of the signal with Hi-C data from Lioy et al. [198], which I reanalyzed for this paper. Since publication of this manuscript, I have been involved in discussions about follow-up experiments and this work has heavily influenced the future direction of my research on the role of nucleoid associated proteins in regulating bacterial transcription.

3.2 Abstract

Recent studies on targeted gene integrations in bacteria have demonstrated that chromosomal location can substantially affect a gene's expression level. However, these studies have only provided information on a small number of sites. To measure position effects on transcriptional propensity at high resolution across the genome, we built and analyzed a library of over 144,000 genome-integrated, standardized reporters in a single mixed population of *Escherichia coli*. We observed

more than 20-fold variations in transcriptional propensity across the genome when the length of the chromosome was binned into broad 4 kbp regions; greater variability was observed over smaller regions. Our data reveal peaks of high transcriptional propensity centered on ribosomal RNA operons and core metabolic genes, while prophages and mobile genetic elements were enriched in less transcribable regions. In total, our work supports the hypothesis that *E. coli* has evolved gene-independent mechanisms for regulating expression from specific regions of its genome.

3.3 Introduction

The bacterial nucleoid is a dense structure composed of DNA, RNA, and proteins and excludes some other abundant cellular machinery, such as ribosomes, from its interior [199–201]. Several studies have demonstrated that packing of the nucleoid is non-random and condition dependent. For example, chromosome conformation capture (3C) studies in multiple bacterial species have revealed segments of DNA that preferentially self-interact and have been called chromosome interaction domains [198, 202–204] . During exponential growth, RNAPs are also organized into tight foci on the nucleoid surface, actively transcribing the ribosomal RNA operons (*rrn*) [205], most of which appear spatially co-localized [206]. Despite the specific localization of DNA and RNAP, previous findings based on site-specific integrations have suggested that gene expression from different genomic loci is roughly equivalent, except for the effect of gene dosage, which decreases from the origin of replication to the terminus during exponential growth [92, 207, 208]. Higher gene dosage near the origin is a result of multiple replication initiation events before terminus replication and cell division [209]; historically, the bacterial chromosome has otherwise been considered generally accessible structurally and for transcription, without detectable interference from chromosomal structure [208, 210].

By measuring GFP fluorescence from a terminator-flanked reporter integrated into several sites, Block et al. [211] observed that gene expression variation from the origin to the terminus corresponded to expected growth-rate-dependent gene dosage changes, consistent with the expectations outlined above. More recently, however, the dogma of uniform expression capability across the genome has been challenged by several lines of evidence. Using a similar approach to Block et al. [211], Bryant et al. [93] demonstrated widely varying expression from a GFP reporter in *E. coli* that did not correlate with genome copy number. Some of the lowest expressing sites were in transcriptionally silent extended protein occupancy domains (tsEPODs) [78], which are regions of high protein occupancy on the genome that appear to correlate with low transcript levels. In some cases, the reporter gene expression could be increased by replacing the tsEPOD with the reporter gene instead of integrating within it [93]. For some reporters outside of tsEPODs, expression interference from neighboring genes drove down reporter expression, depending on the relative gene orientation. Gene expression interference between neighboring genes has also been studied in more detail on plasmids within *E. coli* cells [212]. In that study, some of the gene expression interference observed between neighboring genes could be attributed to competition for negative DNA supercoiling and was gene orientation specific. DNA gyrases and topoisomerases maintain negative supercoiling, which compacts the nucleoid and is important for gene expression [213]. Brambilla and Sclavi [214] have also tracked expression of a reporter under a promoter known to be bound by the nucleoid protein H-NS from 9 different sites over the *E. coli* growth period and observed different site-specific expression levels depending on the growth phase.

Despite the specific observations described above, a systematic understanding of the effects of chromosomal position itself on gene expression has so far eluded the field. Previous studies on position-dependent expression variation have been limited to a small number of integration sites, which was appropriate for mechanistic studies into the effects of specific genomic features, but could not reveal the full range of position-dependent effects on transcription. DNA supercoiling, protein occupancy, transcriptional interference, and binding of promoters and genes by various nucleoid-associated proteins (NAPs) are examples of genomic features that affect expression of large portions of genes in the bacterial genome. Extensive work has been conducted to characterize the effects of a number of these factors for expression of specific genes. However, genomic features vary simultaneously across the genome, potentially leading to combinatorial effects on gene expression [215, 216]. Specific loci may have unique features affecting transcription, which could only be identified by high-resolution mapping of position-dependent expression variation.

Here, we employ Tn5 transposase to perform massively parallel integration of a standardized, barcoded reporter construct, allowing us to obtain an empirical map of gene-independent transcriptional propensity—that is, the amount of RNA produced per unit of DNA from a given reporter—across the bacterial genome (Figure 3.1). High-resolution transcriptional propensity comparisons with genomic features can reveal both strong and weak correlations with high statistical power. To test the effect of genome position on gene expression, and not native gene regulation, we designed a reporter construct with strong bi-directional terminators [217] and its own inducible promoter (Figure 3.1A). Each reporter construct is tagged with a unique barcode identifier, which allows simultaneous tracking of gene expression from thousands of integrations. Using a modified transposon footprinting procedure, unique barcodes were paired with integration location, allowing barcodes to serve as a proxy for the overall abundance of RNA or DNA at each integration address. The s70-dependent TetO1 promoter drives expression of mNeonGreen (mNG) followed by a 15 base barcode on the 3' UTR of the RNA upon induction by anhydrotetracycline (aTc) [218]. The reporter used here was designed to be relatively small in size and has an intermediate transcription rate [219] in order to minimize the effect of the reporter on the local genome structure [202]. The inclusion of an open reading frame in our construct ensures that the transcribed RNA will be sub-

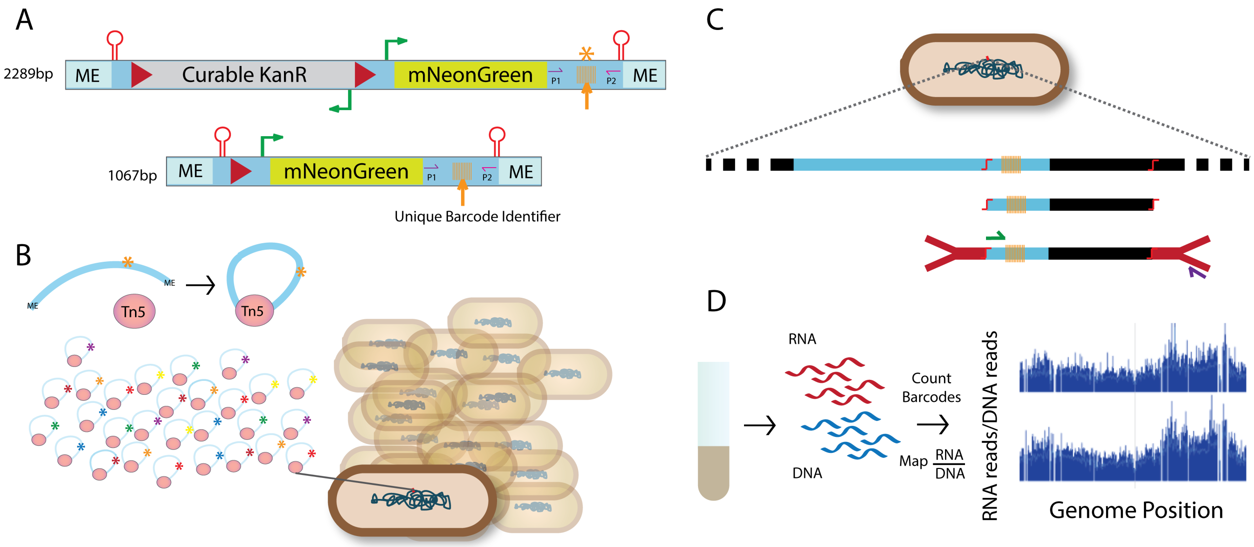


Figure 3.1: Library Construction and Data Acquisition for Position-Dependent Transcriptional Propensity Mapping. (A) mNeonGreen (mNG) reporter is controlled by the TetO1 promoter. The orange arrow indicates the position of the 15 bp barcode that is transcribed with mNG. The construct is flanked by strong bi-directional terminators and mosaic ends (ME), which are recognized by Tn5 transposase. P1 and P2 indicate sites used for light amplification in preparation for barcode sequencing. Construct size and features are shown before and after curing of a kanamycin resistance marker (KanR). (B) To produce the reporter library, randomly barcoded reporter constructs in complex with Tn5 are electroporated into cells and randomly integrated into the *E. coli* genome in parallel. (C) Transposon footprinting pairs barcode sequence (orange) with integration location on the genome (black). 4 bp recognition restriction enzymes cut upstream of the barcode and randomly in the downstream genomic DNA. After ligation of the Y-linker (red), construct-containing DNA fragments are specifically amplified and sequenced. Footprinting need only be done once for a given library to identify the insertion location corresponding to each barcode. (D) To measure transcriptional propensities, the reporter library is grown to an optical density (OD) at 600 nm of 0.2. Total RNA and DNA are extracted. After nucleic acid processing (Figure 3.2), the RNA:DNA ratio for each barcode is mapped to their corresponding genomic locations.

ject to typical post-transcriptional phenomena (e.g., co-transcriptional translation and subsequent protection by ribosomes). In keeping with efforts to minimize reporter size, the selection marker is an FRT (flippase recognition target)-flanked kanamycin resistance cassette and was removed by Flp recombinase before the full-scale profiling procedure (Figure 3.1A).

3.4 Results

Tn5 transposition was used to integrate the barcoded reporter in a massively parallel fashion into the *E. coli* genome. We mapped 144,672 unique reporter barcodes to 98,034 unique genomic integration sites, corresponding to an average of one unique location every 47 bp. As integration rate was not uniform across the genome, resolution varies depending on the region (Figures 3.4C and 3.3E). Neighboring integrations have high similarity in raw RNA barcode produced per unit DNA barcode (which we refer to as transcriptional propensity), indicating that reporter transcription is

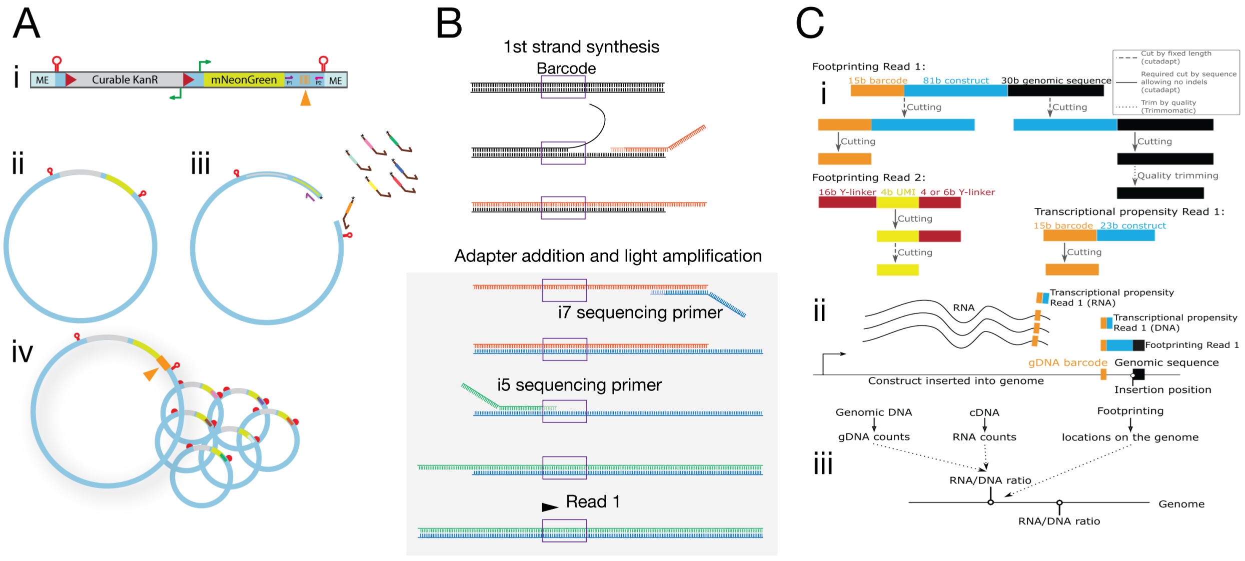


Figure 3.2: Plasmid barcoding, nucleic acid processing, and sequencing analysis workflow. (A) Barcoding of pSAS31 for generation of barcoded reporter integration construct. (i) Diagram of the mature barcoded integration construct. The orange arrow indicates the position of the random barcode. (ii) Representation of key reporter features on the pSAS31 plasmid; the plasmid backbone (pale blue) is un-annotated for clarity. (iii) After digestion with AscI, pSAS31 is amplified with primers that introduce a random 15 bp barcode (brown primer). (iv) After digestion of the PCR product from (iii) by AcsI, the plasmid is recircularized and transformed for selection. All plasmids in the library differ by only the barcode sequence. The orange arrow indicates the position of the random barcode. The barcoded integration construct is liberated by PvuII digestion of the plasmid library and used for transposome generation. (B) Nucleic acid processing for sequencing of RNA barcodes. The barcode is indicated in the purple box. First strand synthesis introduces a site for the i7 NEB sequencing primer to bind (P11). Through a low number of cycles of PCR the library is lightly amplified and has adapters added for sequencing. (C) Sequencing data analysis workflow. (i) Structures and processing of raw sequencing reads for footprinting and barcode sequencing runs. Both Read 1 and 2 in paired-end sequencing for footprinting are used. Read 1 in footprinting data are processed to retrieve barcode sequence (orange) and genomic sequences (black) at the insertion position by removing the uninformative parts. Footprinting Read 2 is used to obtain the unique molecular identifier (UMI, yellow) to each barcode-position combination in the same set of reads. (ii) Sources of transcriptional propensity and footprint reads relative to the construct inserted into genome. (iii) Use of information from sequencing runs to map transcriptional propensity to positions on genome. The cDNA barcode counts, genomic DNA barcode counts, and genomic DNA footprinting yield the transcript levels, DNA- based normalization, and insertion locations, respectively, needed for transcriptional propensity calculation.

dependent on integration location (Figures 3.4A and 3.4B). After smoothing the raw transcriptional propensity by taking the median value for reporters in a 500 bp window around each integration, the highly correlated replicates (Figure 3.4D, Spearman $\rho = 0.915$) were averaged to produce the high-resolution transcriptional propensity map (Figure 3.4E). The transcriptional propensity map for all analyses includes only sites where at least three independent integrations were measured within a 500 bp window. The transcriptional propensity signal is reported as a median of signal for all integration events within a 500 bp window centered on each integration in all calculations in order to minimize noise potentially arising from a single barcode (see Supplemental Table S7 for all transcriptional propensity and count values). Several other potentially confounding features, such as barcode-specific GC content and reporter-integration-specific growth rate changes, do not have systematic effects on this transcriptional propensity signal (Figure 3.5; Section 3.6.20). N.B. the barcode abundance measurements used are for the reporter barcodes only; RNA from native transcripts is not sequenced in our experiments. We also note that in principle, any potential genome-position-dependent effects on RNA stability would be part of the transcriptional propensity signal. Although there is a weak positive correlation between the degradation rate of RNA from neighboring operons [220], RNA abundance is well correlated with transcription rate of native genes in *E. coli* [221], whereas RNA stability is generally not predictive of overall transcript levels in *E. coli* [222], hence our use of the term transcriptional propensity (as opposed to RNA abundance propensity).

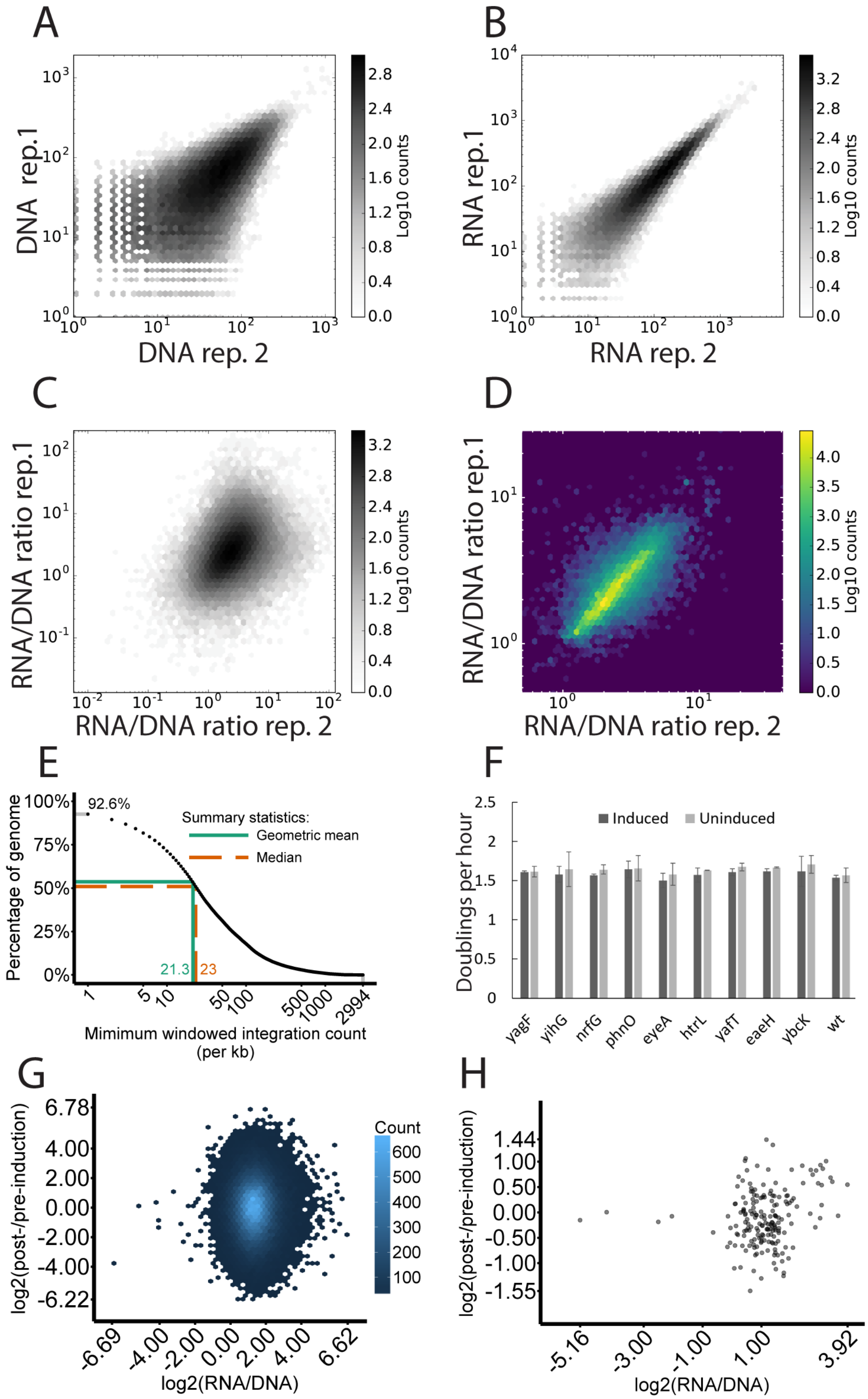

A
DNA rep.1
DNA rep. 2
10^0
10^1
10^2
10^3
Log10 counts
0.0
0.4
0.8
1.2
1.6
2.0
2.4
2.8
B
RNA rep.1
RNA rep. 2
10^0
10^1
10^2
10^3
10^4
Log10 counts
0.0
0.4
0.8
1.2
1.6
2.0
2.4
2.8
3.2
C
RNA/DNA ratio rep.1
RNA/DNA ratio rep. 2
10^{-2}
10^{-1}
10^0
10^1
10^2
Log10 counts
0.0
0.4
0.8
1.2
1.6
2.0
2.4
2.8
3.2
D
RNA/DNA ratio rep.1
RNA/DNA ratio rep. 2
10^0
10^1
Log10 counts
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
E
Percentage of genome
100%
75%
50%
25%
0%
92.6%
Summary statistics:
Geometric mean
Median
21.3
23
1
5
10
50
100
500
1000
2994
Mimimum windowed integration count
(per kb)
F
Doublings per hour
0
0.5
1
1.5
2
2.5
Induced
Uninduced
yagF
yihG
nrfG
phnO
eyeA
htrL
yafT
eaeH
ybcK
wt
G
log2(post-/pre-induction)
6.78
4.00
2.00
0.00
-2.00
-4.00
-6.22
-6.69
-4.00
-2.00
0.00
2.00
4.00
6.62
log2(RNA/DNA)
Count
600
500
400
300
200
100
H
log2(post-/pre-induction)
1.44
1.00
0.50
0.00
-0.50
-1.00
-1.55
-5.16
-3.00
-1.00
1.00
3.92
log2(RNA/DNA)

Figure 3.3 *(previous page)*: Reporter library properties. (A) Correlation between DNA barcode counts from replicate 1 and replicate 2 (Spearman $\rho = 0.72$). (B) Correlation between RNA barcodes from replicate 1 and replicate 2 (Spearman $\rho = 0.95$). (C) Correlation between unsmoothed RNA/DNA ratio replicate 1 and replicate 2 (Spearman $\rho = 0.4$). (D) Correlation between replicates as in Figure 3.4D after independent median windowing over 500 bp centered on each integration site (Spearman $\rho = 0.91$); this statistic corresponds to the data discussed in the main text. (E) Integration density across the genome as the percentages of genomic sites with a minimum of certain counts of integrations within the 1 kb windows centered at each genomic position. 92.6% of genomic sites had at least 1 integration within the 1 kb window. The orange dashed line indicates the value of the median of integration counts of all genomic sites (1 integration per 43.5 base pairs). The cyan-green solid line indicates the percentile and value of the geometric mean of integration counts. (F) Doublings per hour for strains with targeted reporter integrations downstream of the indicated gene with and without induction by aTc from cells grown in microplate (Corresponding dosage-transformed transcriptional propensity from these sites range from a maximum of 41.25 for *yagF* to a minimum of 2.74 for *ybcK*). The sites are ordered from left to right by highest to lowest transcriptional propensity. (G) Correlation of $\log_2$ ratios of barcodes genomic DNA abundances post- and pre- aTc induction and growth with $\log_2$ ratios of RNA/DNA ratios (Spearman $\rho = 0.01$). Barcodes were filtered to include only barcodes with over 10 counts in both replicates. (H) Correlation of $\log_2$ ratios of barcodes genomic DNA abundances post- and pre- growth with aTc induction with $\log_2$ ratios of RNA/DNA ratios. Barcodes were filtered to include only barcodes with over 100 counts in both replicates (Spearman $\rho = 0.14$). Note that as described in the text, the Spearman correlations drop in magnitude to below 0.03 upon application of the window-averaging used in processing our transcriptional propensity signals, demonstrating a lack of any meaningful impact due to clonal variations in growth rate.

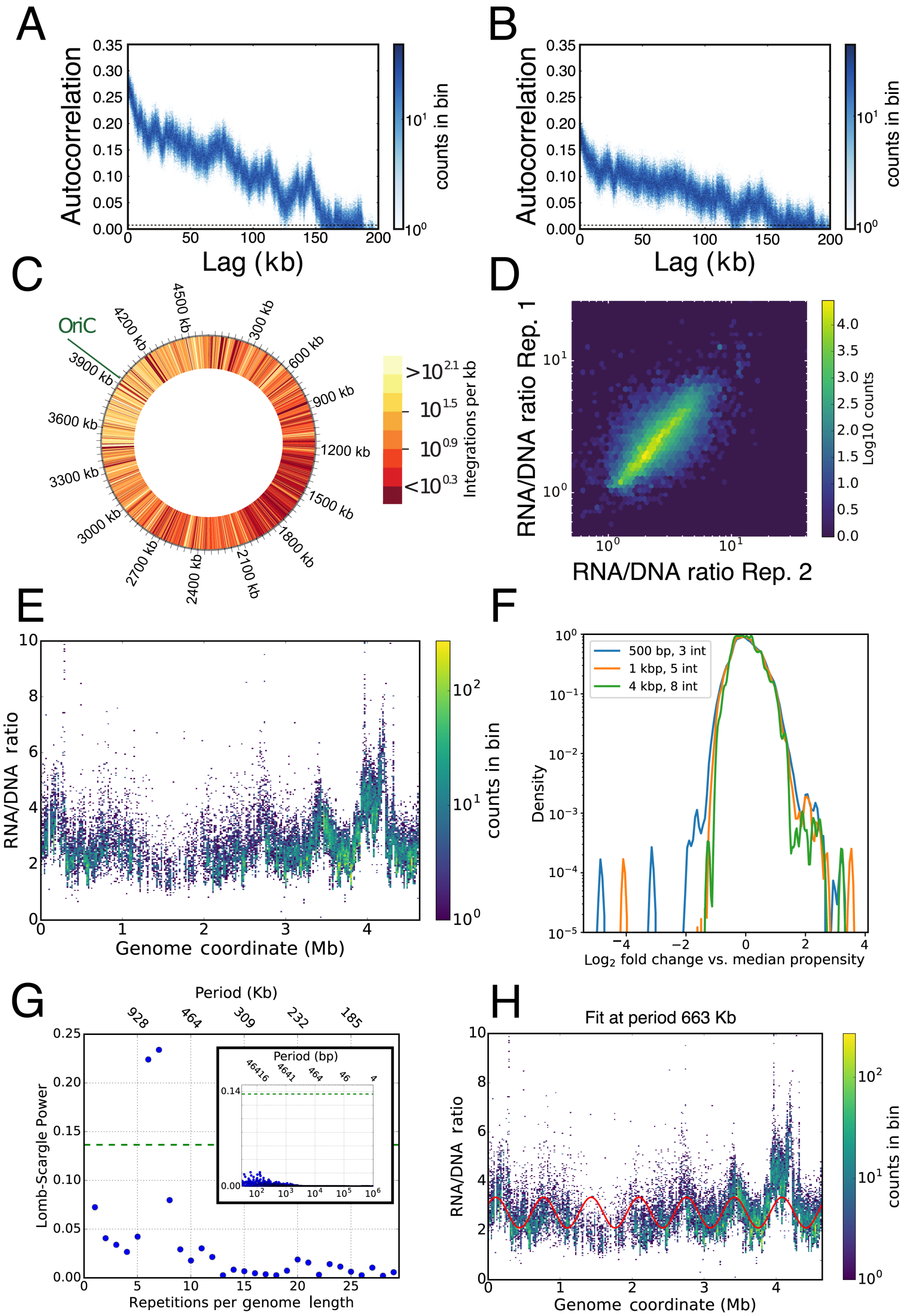

A
Autocorrelation
Lag (kb)
counts in bin
B
Autocorrelation
Lag (kb)
counts in bin
C
OriC
300 kb
600 kb
900 kb
1200 kb
1500 kb
1800 kb
2100 kb
2400 kb
2700 kb
3000 kb
3300 kb
3600 kb
3900 kb
4200 kb
4500 kb
>10^2.1
10^1.5
10^0.9
<10^0.3
Integrations per kb
D
RNA/DNA ratio Rep. 1
RNA/DNA ratio Rep. 2
Log10 counts
E
RNA/DNA ratio
Genome coordinate (Mb)
counts in bin
F
500 bp, 3 int
1 kbp, 5 int
4 kbp, 8 int
Density
Log2 fold change vs. median propensity
G
Period (Kb)
928
464
309
232
185
Lomb-Scargle Power
Repetitions per genome length
Period (bp)
46416
4641
464
46
4
H
Fit at period 663 Kb
RNA/DNA ratio
Genome coordinate (Mb)
counts in bin

Figure 3.4 *(previous page)*: Genome-Position-Dependent Transcriptional Propensity from Tn5-Based Integration of a Barcoded Reporter Is Nonrandom (A) Autocorrelation of raw RNA/DNA ratio values for replicate 1. Lag represents base pair distance. The dashed line represents the 99% autocorrelation confidence interval for a white noise process, thus representing the level of autocorrelation that would be observed in the absence of a true signal. (B) Autocorrelation of raw RNA/DNA ratio values for replicate 2. (C) Reporter integration count within 1 kb windows throughout the genome. (D) Correlation between replicates for calculated transcriptional propensity from 500 bp rolling median windows (Spearman $\rho = 0.91$). (E) Median transcriptional propensity (over 500 bp median rolling windows with at least three unique barcodes) mapped to specific integration locations on the E. coli genome. The color indicates the number of unique transposon insertions in the same bin of RNA/DNA ratio. All values used to generate these plots can be found in Supplemental Table S7. (F) Shown are kernel density estimates of the distributions of $\log_2$-fold deviations of transcriptional propensities (smoothed by taking the median value from different window sizes around each site) versus the global median of transcriptional propensity from the same sample. "Int" indicates the minimum number of integrations required to generate a median-smoothed value for each window size. The blue curve corresponds to the smoothing used throughout the remainder of the text. (G) Spectral analysis of the observed transcriptional propensity signal (averaged across biological replicates) using the Lomb-Scargle periodogram method [223, 224]. Only periods that represent an integer divisor of the total genome length were analyzed. The green dotted line represents an overall $<1\%$ false discovery rate as determined by a permutation test using blocks of 2,200 adjacent collection bins ($\sim$100,000 bp; see Section 3.6.16 for details). Inset; as in main figure, zoomed out to show periods $<$160,000 bp. (H) Sinusoidal function fit (red line) of the period with the highest Lomb-Scargle power (663,093.14 bp, 7 repetitions per genome length) to the transcriptional propensity.

3.4.1 Transcriptional Propensity Is Highly Variable across the *E. coli* Genome

Transcriptional propensity variation appears roughly periodic at the whole-genome scale (Figure 3.4E). Several sharp troughs are also apparent, independent of the overall waveform. Transcriptional propensities are not a result of gene dosage resulting from high Ori-Ter ratios during exponential phase growth or from differing representation of a library member because all transcriptional propensities are reported as RNA:DNA ratios. The distributions of transcriptional propensity values observed using different windowing sizes are plotted in Figure 3.4F. Substantial position-dependent variation is present throughout the genome. A smooth, roughly log-normal population is observed spanning a 16-fold range of propensities (with 99% of values contained within a central 4.2-fold range); furthermore, many genomic regions are apparent in the tails of the distribution that represent dramatically activated or silenced sites. There is a >250-fold propensity difference between the highest and lowest 500 bp windows (using the median of no fewer than three sites within each window to avoid undue impact of individual outliers). Even considered over broader regions, an overall >195-fold range persists using a window size of 1 kbp (requiring no fewer than 5 reporters) and a >22-fold range of transcriptional propensity difference between the highest and lowest 4 kbp regions of the chromosome (median of integration sites within a 4 kb window with at least 8 reporters, see Figure 3.4F for transcriptional propensity variation considered over different median-windowing sizes). A trade-off of course exists in expanding the window size used in the analysis above, as larger windows will be less subject to statistical fluctuations, but also will likely miss biologically meaningful local variations in transcriptional propensity and instead provide a

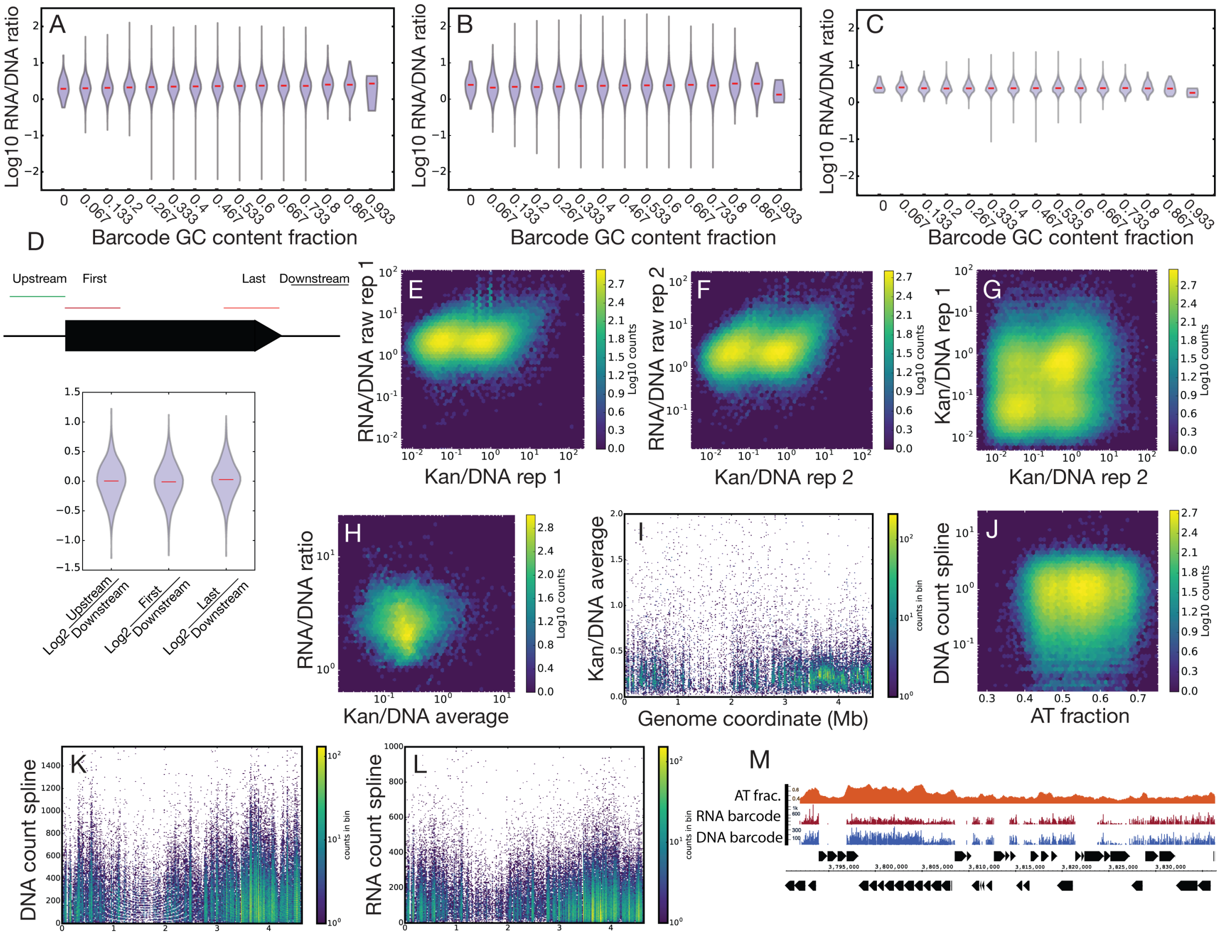


Figure 3.5: Barcode GC content and other barcode-specific properties have little to no effect on transcriptional propensity. Violin plot of raw transcriptional propensity from replicate 1 (A, Spearman $\rho = 0.15$) and replicate 2 (B, Spearman $\rho = 0.13$) from each possible barcode GC content fraction. (C) Correlation of the 500 bp median windowed and replicate-averaged transcriptional propensity (used for correlations with all other genomic features) with barcode GC content (Spearman $\rho = 0.01$). (D) Median transcriptional propensity within different 500 bp windows of each gene (cartoon of 500 bp windows on a generic gene indicated above) are divided by the downstream median value (which does not result in a knockout) and $\log_2$ transformed. In general, the values from the upstream, first and last windows are very similar to the downstream window, with median $\log_2$ fold changes of of 0.004, -0.01, and 0.028, respectively (Integrations within the first window are likely to cause loss of function). (E)-(I) Retention rate of kanR does not affect measurement of genome-wide transcriptional propensity. (E) Correlation between the raw replicate 1 transcriptional propensity and replicate 1 kanR-associated barcode per DNA barcode (Spearman $\rho = 0.14$). (F) Correlation between the raw replicate 2 transcriptional propensity and replicate 2 kanR- associated barcode per DNA barcode (Spearman $\rho = 0.19$). (G) Correlation between replicates for the kanR-associated barcodes per DNA barcode (Spearman $\rho = 0.25$). (H) Correlation between the standard transcriptional propensity values (Y-axis) and the average of the kanR-associated barcodes per DNA barcode replicates after taking the median value from each site and integrations in the surrounding 500 bp (these data are processed identically to the transcriptional propensity values) (Spearman $\rho = -0.05$). (I) Windowed average KanR per DNA barcode (as in D) mapped onto the *E. coli* genome. (J)-(L) DNA barcode count is not correlated with genomic AT content: (J) Insubstantial correlation between spline corrected barcode genomic DNA counts with AT content fraction from 500 bp around each integration site (Spearman $\rho = 0.05$) (K) Spline corrected genomic DNA barcode counts mapped to the chromosome. (L) Spline corrected RNA barcode counts mapped to the chromosome. (M) Zoom-in to a genomic region showing AT fraction and both RNA barcode counts (red) and DNA barcode counts (blue).

regional average across large chunks of the chromosome.

The superficially apparent periodicity in the transcriptional propensity is supported by spectral analysis via Lomb-Scargle periodograms [223, 224]. As shown in Figure 3.4G, strong spectral lines are apparent at both 663 kb and 773 kb periods, which would correspond to 7 and 6 repetitions, respectively, throughout one genome length, suggesting that this length scale may be characteristic of a key unit of functional and/or spatial organization in the *E. coli* chromosome. Modeling the transcriptional propensity with a sinusoidal function at a period of 663 kb shows a good fit (Figure 3.4H). These period lengths are roughly consistent with the size of macrodomains observed in recent 3C-sequencing experiments [198]. We also note that the absence of a $\sim$10 kb component in our periodogram, which might be expected based on experiments measuring the propagation of supercoiling relaxation upon DNA damage [225], may arise simply because of a lack of periodicity in the $\sim$10 kb domain organization.

3.4.2 Ribosomal RNA Operons Are Centered in Broad Transcriptional Propensity Peaks

Several genomic features are readily apparent as having substantial correlations with regions of high transcriptional propensity, as shown in Figure 3.6A. The seven *rrn* operons in the *E. coli* genome are located within the major peaks of transcriptional propensity, although a single major peak (near 1 Mb) occurs without an *rrn* operon. Thus, either the *rrn* operons have been selected to be contained in regions of exceptional transcriptional propensity or they contain some feature that itself enhances transcriptional propensity in their surroundings. By subtracting a LOWESS (locally weighted scatterplot smoothing) [226] smoothing on transcriptional propensity with distance from the nearest *rrn* operon (Figure 3.6C), from the overall transcriptional propensity signal, the major waveform pattern is mostly eliminated, while local peaks and troughs are still apparent (Figure 3.6D); thus, several additional features must contribute locally to both position-dependent activation and silencing. Another feature that could contribute to the transcriptional propensity signal is the structural organization of the genome at both long- and short-length scales. In order to explore the relationship between long-length scale organization and transcriptional propensity, we examined the interaction between our transcriptional propensity data and the macrodomain boundaries identified in Lioy et al. [198], shown in Figure 3.6A (lower panel). Working at the level of each macrodomain, we grouped the transcriptional propensity signal into 10 equally spaced bins according to length-normalized position within the corresponding macrodomain. Transcriptional propensity is higher at macrodomain boundaries near an rrn operon, but this effect is not present at macrodomain boundaries that do not have a nearby *rrn* operon (Figure 3.7C). On shorter length scales, such as the scale of the chromosomal-interacting domains (CIDs) observed in 3C-sequencing experiments [198] or topologically separated domains observed in isolated chromosomal DNA [227], we see a limited relationship between the transcriptional propensity and the position within a defined CID (Figure 3.7A). In fact, similar to the macrodomain boundaries, the transcriptional propensity tends to be higher at CID boundaries that coincide with a ribosomal RNA operon, but this effect is not present at CID boundaries that do not have a nearby *rrn* operon. Together, these results suggest that measures of the structural organization of the genome alone are not predictive of transcriptional propensity and other features must also contribute to the transcribability of any particular region.

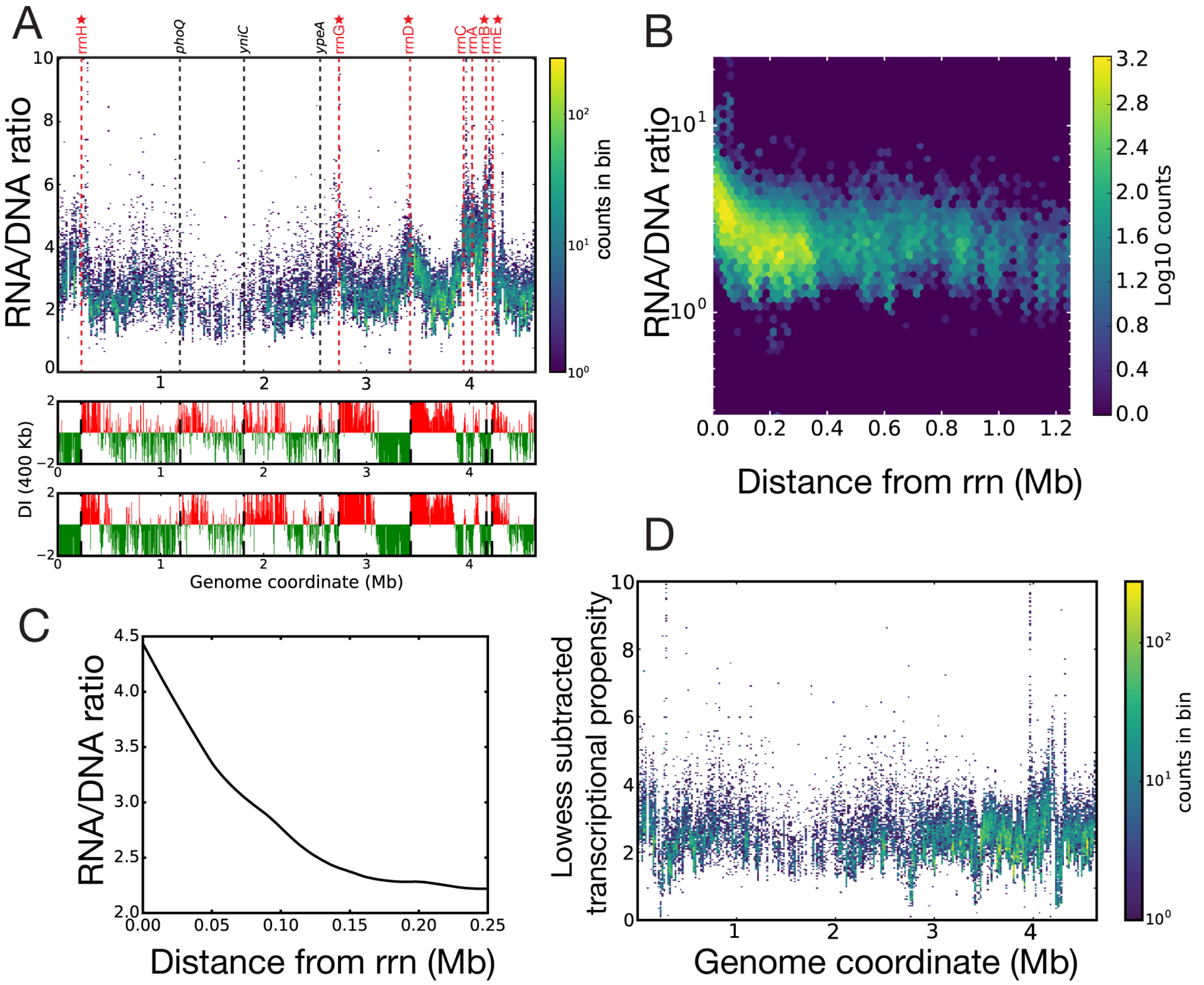


Figure 3.6: Transcriptional Propensity Peaks Correspond to Ribosomal RNA Operon and Macrodomain Boundaries. (A) (Top) Transcriptional propensity plot with macrodomain boundaries, as determined from Lioy et al. [198]. Red dashed lines indicate ribosomal RNA operons. Ribosomal RNAs labeled with a star indicate rRNA operons near macrodomain boundaries. Black dashed lines indicate macrodomain boundaries that are not near a ribosomal RNA operon, and gene names for the overlapping or nearest gene to these boundaries are indicated above. (Bottom) Directionality index (DI; see Section 3.6.17) determined at 400 kb scale for each of the two biological replicates taken from exponentially growing *E. coli* cells at 37°C in LB media with macrodomain boundaries indicated with black dashed lines, obtained by re-analysis of data from Lioy et al. [198]. For details on determination of macrodomain boundaries and accession of Lioy et al. [198] data, see Section 3.6.17. (B) Correlation of transcriptional propensity and distance from the nearest rrn operon (Spearman $\rho = -0.56$). (C) LOWESS fit of transcriptional propensity with rrn distance (fitting using a smoothing parameter of 0.33). (D) Transcriptional propensity signal with values from LOWESS regression subtracted from the overall signal in (A).

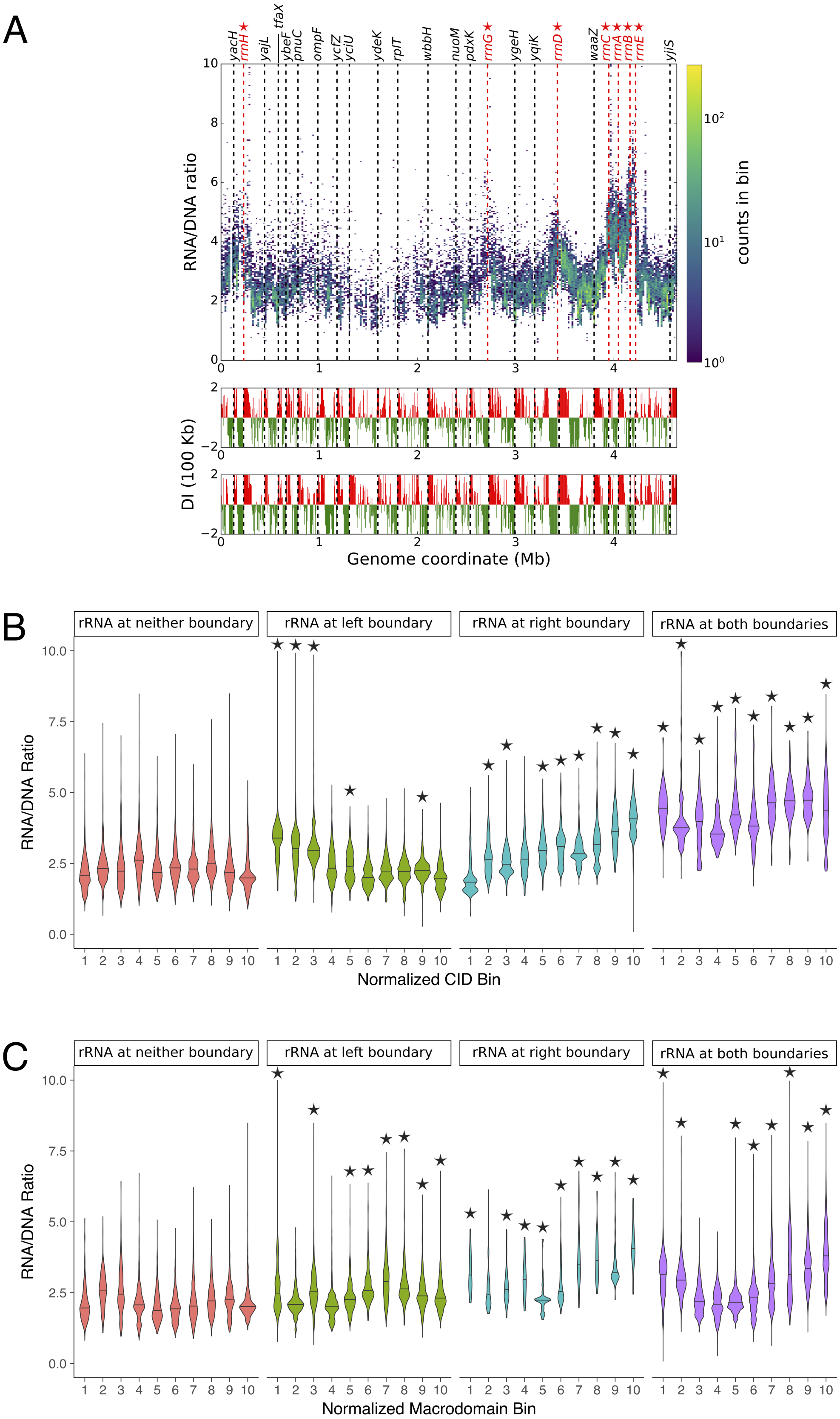

A
yacH
rrnH★
yajL
tfaX
ybeF
pnuC
ompF
ycfZ
yciU
ydeK
rplT
wbbH
nuoM
pdxK
rrnG★
ygeH
yqiK
rrnD★
waaZ
rrnC★
rrnA★
rrnB★
rrnE★
yjiS
RNA/DNA ratio
0
2
4
6
8
10
0
1
2
3
4
10^0
10^1
10^2
counts in bin
DI (100 Kb)
2
−2
Genome coordinate (Mb)
B
rRNA at neither boundary
rRNA at left boundary
rRNA at right boundary
rRNA at both boundaries
RNA/DNA Ratio
0.0
2.5
5.0
7.5
10.0
1 2 3 4 5 6 7 8 9 10
Normalized CID Bin
C
rRNA at neither boundary
rRNA at left boundary
rRNA at right boundary
rRNA at both boundaries
RNA/DNA Ratio
0.0
2.5
5.0
7.5
10.0
1 2 3 4 5 6 7 8 9 10
Normalized Macrodomain Bin

Figure 3.7 *(previous page)*: Relationship between Transcriptional Propensity and Chromosome Interacting Domains. (A) (top) Transcriptional propensity plot (this study) compared with CID boundaries [198], plotted as dashed black lines and labeled with an overlapping or nearby gene. Red dashed lines indicate one of seven ribosomal RNA operons. Stars above ribosomal RNA operon names indicate close proximity to a CID boundary. (bottom) Directionality index determined at 100 Kb scale for each of two biological replicates taken from exponentially growing *E. coli* cells at 37 °C in LB media with CID boundaries labeled as in the top panel. For details on determination of CID boundaries and accession of Lioy et al. [198] data see Section 3.6.17. (B) Violin plots of transcriptional propensity aggregated into ten bins evenly discretized over the length of each CID and conditioned on the presence or absence of rRNA operons at either or both boundaries for a given CID (i.e. bin 1 is aggregated data from the left-most bin of each CID and bin 10 is aggregated data from the right-most bin of each CID, where left and right are defined relative to standard genome coordinates). Lines on each violin indicate the 50th percentile estimate as determined by the kernel density for each distribution. Stars indicate bins whose medians are significantly higher (FDR corrected p-value < 0.01) than the corresponding "rRNA at neither boundary" baseline bin as determined by a one-sided permutation test where class labels were shuffled 1000 times. (C) Same as (B) but for macrodomains determined from directionality index analysis at 400 Kb scale as in (Figure 3.6A).

3.4.3 Binding of the NAPs H-NS and Fis Is Strongly Correlated with Transcriptional Propensity

We next examined the correlation of transcriptional propensity with several characterized genomic features using rolling-window medians over 500 bp for each dataset. We observed the strongest effects for binding of the nucleoid proteins Fis and H-NS, as well as global protein occupancy measured via *in vivo* protein occupancy display (IPOD; Figures 3.8, 3.9A, and 3.9B; see Table 3.1 for all Spearman correlations). Despite the fact that the abundant NAP Fis is not expected to bind the reporter construct itself, transcriptional propensity is highly positively correlated with Fis binding levels at genomic integration sites (Spearman $\rho = 0.50$, Figure 3.8A). Conversely, transcriptional propensity is strongly negatively correlated with H-NS binding (Spearman $\rho = -0.58$, Figure 3.8B), consistent with the previously described gene silencing role for H-NS (Kahramanoglou et al., 2011). Transcriptional propensity is also negatively correlated with overall protein occupancy, strongly supporting reporter silencing observed by Bryant et al. [93] when integrated within tsEPODs (Figure 3.9A) [78].

RNA abundance from native genes displays only a weak positive correlation with transcriptional propensity (Spearman $\rho = 0.24$, Figure 3.9D). However, when larger rolling median windows (50 kb) are used for RNA abundance from native genes, correlation with transcriptional propensity is much higher (Spearman $\rho = 0.51$, Figure 3.9E). These results show that while highly expressed genes are more frequently located in high transcriptional propensity regions, the regulatory logic governing expression of individual genes is dominant over the underlying transcriptional propensity of a given region.

Binding of other NAPs (HU, LRP, and SeqA) was not well correlated with transcriptional propensity nor was RNAP binding to active promoters (Figure 3.9). We likewise found no sub-

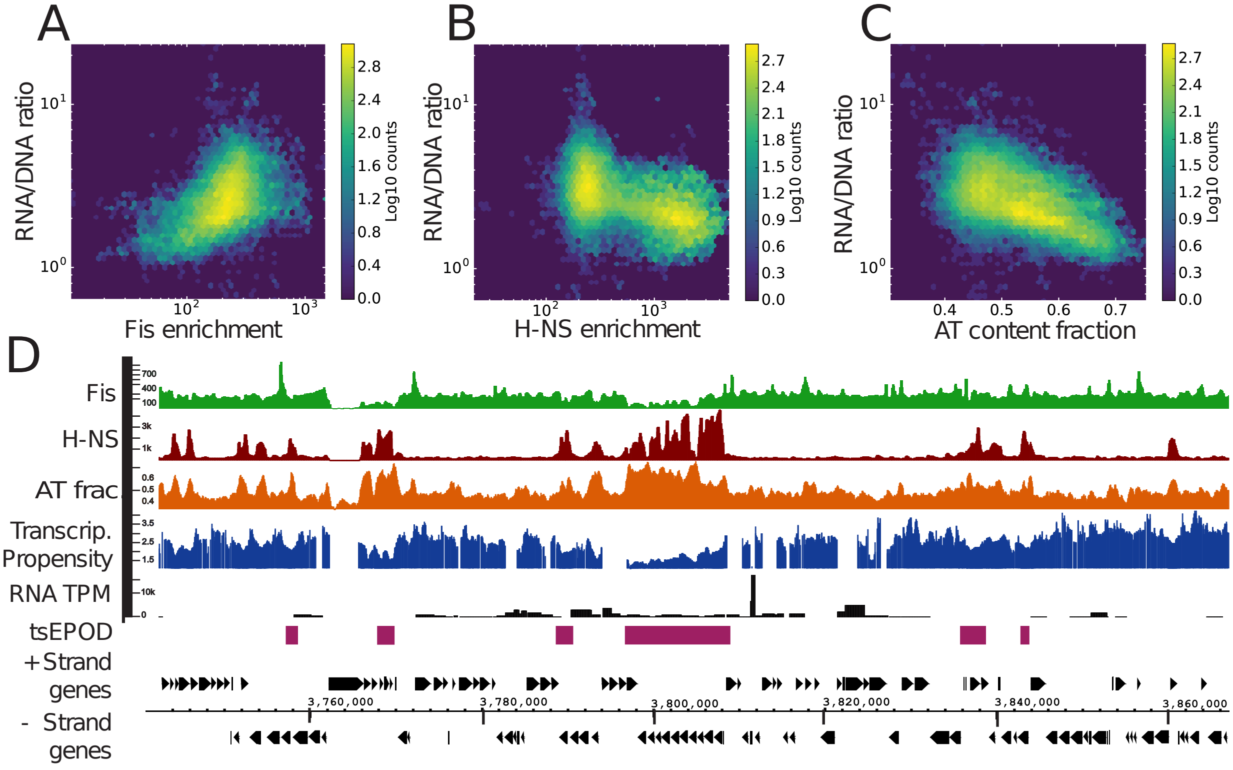


Figure 3.8: Correlation of Transcriptional Propensity with Binding of Abundant NAPs and Nucleotide Content. (A) Correlation of transcriptional propensity with Fis binding (500 bp rolling median, Spearman $\rho = 0.50$). (B) Correlation of transcriptional propensity with enrichment by H-NS binding (500 bp rolling median, Spearman $\rho = -0.58$). (C) Correlation of transcriptional propensity with AT content (500 bp rolling mean, Spearman $\rho = -0.59$). (D) Genome browser view of a large tsEPOD (2.79-2.81 Mb) and surrounding genomic context. Tracks from top to bottom for Fis binding, H-NS binding, AT content (AT frac.), and transcriptional propensity (Transcrip. Propensity), transcripts per million (TPM) RNA from native genes, and tsEPOD ranges. Strand-specific gene annotations are indicated below the data tracks [48].

stantial correlation of transcriptional propensity with a measure of DNA supercoiling density or with reporter location with respect to genes encoding proteins recognized by the signal recognition particle (Figure 3.9) [228, 229]. In contrast, mean adenine and thymine (AT) content in a 500 bp window around insertion locations was strongly negatively correlated with transcriptional propensity (Figure 3.8B). AT content is also highly correlated with H-NS and protein occupancy binding (it is notable that both H-NS and Fis have consensus motifs with high AT content [48], although the Fis consensus sequence is bookended by G and C nucleotides).

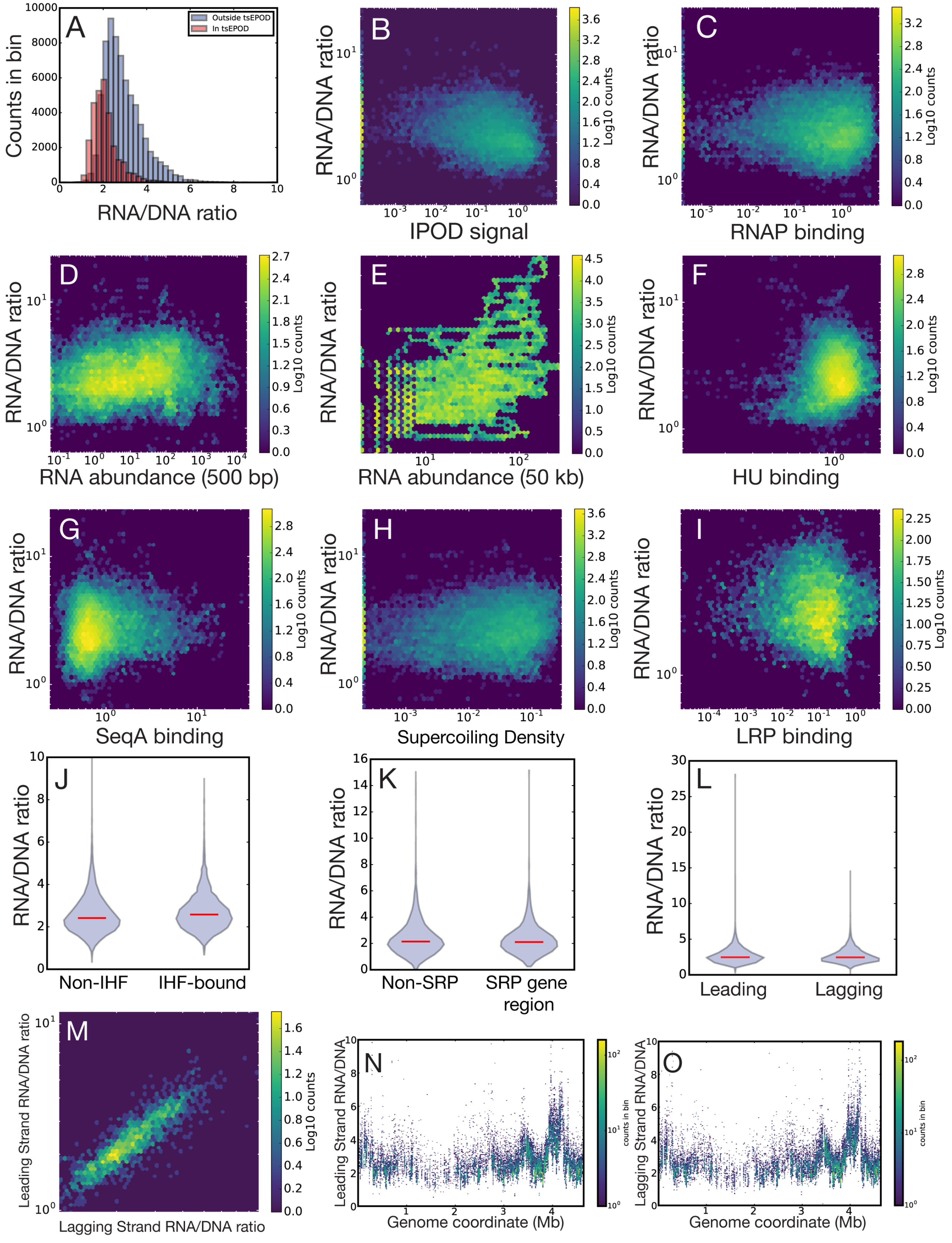

A
Outside tsEPOD
In tsEPOD
Counts in bin
10000
8000
6000
4000
2000
0
RNA/DNA ratio
0
2
4
6
8
10
B
RNA/DNA ratio
IPOD signal
Log10 counts
C
RNAP binding
D
RNA abundance (500 bp)
E
RNA abundance (50 kb)
F
HU binding
G
SeqA binding
H
Supercoiling Density
I
LRP binding
J
Non-IHF
IHF-bound
K
Non-SRP
SRP gene region
L
Leading
Lagging
M
Leading Strand RNA/DNA ratio
Lagging Strand RNA/DNA ratio
N
Leading Strand RNA/DNA
Genome coordinate (Mb)
counts in bin
O
Lagging Strand RNA/DNA
Genome coordinate (Mb)
counts in bin

Figure 3.9 *(previous page)*: Correlation of transcriptional propensity a variety of genome features. Unless otherwise noted, continuous datasets are all 500 bp median windowed. Replicates were quantile normalized and averaged before performing correlation analysis. (A) Histogram plot of transcriptional propensity within tsEPODs (median = 1.97) and all other sites (median = 2.74; the observed difference between the tsEPOD set and other regions is -0.77; 95% CI for difference in medians is -0.91 to -0.62 based on a circular block bootstrap). (B) Correlation of transcriptional propensity with protein occupancy from Vora et al. [78] (Spearman $\rho = -0.34$). (C) Correlation of transcriptional propensity with RNAP binding (Spearman $\rho = -0.22$). D) Correlation of transcriptional propensity with *E. coli* native RNA abundance over 500bp rolling median window (Spearman $\rho = 0.24$) data from Kroner et al. [138]. (E) Correlation of transcriptional propensity with *E. coli* native RNA abundance over 50 kb rolling median window (Spearman $\rho = 0.51$). (F) Correlation of transcriptional propensity with HU binding (Spearman $\rho = 0.13$) [230]. (G) Correlation of transcriptional propensity with SeqA binding (Spearman $\rho = 0.14$) [231]. (H) Correlation of transcriptional propensity with supercoiling density (Spearman $\rho = 0.096$) [228]. (I) Correlation of transcriptional propensity with LRP binding (Spearman $\rho = -0.06$). (J) Violin plot of Non-IHF (Median=2.45) and IHF bound sites (Median=2.65, the observed difference is -0.20; 95% CI for the difference in medians is -0.06 to -0.32 based on a circular block bootstrap, showing a small but statistically significant decrease of transcriptional propensity in IHF-bound sites) [230]. (K) Violin plot of transcriptional propensity for genomic regions around non- SRP genes (Median=2.11) and around SRP genes (Median=2.14, the observed difference is -0.03; 95% CI for the difference in medians is -0.12 to 0.12, demonstrating no meaningful difference between the sets) [229]. (L) Violin plot of transcriptional propensity from the leading (Median=2.47) and lagging strands (Median=2.46, the observed difference between the median of each strand is 0.01; 95% CI for the difference in medians is -0.02 to +0.05 based on a circular block bootstrap), demonstrating the presence of no meaningful difference between the strands. (M) Correlation of transcriptional propensity from the leading and lagging strand for sites that contain integrations in both orientations (Spearman $\rho = 0.89$). (N) Transcriptional propensity map for the leading strand. (O) Transcriptional propensity map for the lagging strand.

3.4.4 Transcriptional Propensity Does Not Show Substantial Strand Specificity and Is Insulated from Native Transcription

We also examined the correlation of neighboring RNA abundance in all orientations relative to the reporter on transcriptional propensity (Table 3.2). These data indicate that neighboring transcription has no more than a tiny impact on transcriptional propensity in our experimental setup, likely due to the strong bidirectional terminators flanking our insertion construct. Since the correlations of native RNA abundances with transcriptional propensity in the tandem orientation with respect to the reporter are very similar regardless of which is upstream (Spearman correlation between reporter and adjacent RNA abundance is 0.16 with the reporter upstream, 0.15 with the reporter downstream), insulation by the strong upstream transcriptional terminator of the reporter is also validated. Transcriptional propensity from reporters on the leading and lagging strands display the same overall waveform pattern, and transcriptional propensities at the same positions are highly correlated with each other (Spearman $\rho = 0.89$, Figures 3.9L-3.9O).

Genomic feature	Propensity ρ	Propensity p-value	Density Spearman ρ	Density p-value	citation
H-NS binding	-0.58	<0.0001	0.51	0.0002	Kahramanoglou et al. [84]
AT-content	-0.59	<0.0001	0.37	<0.0001	Blattner et al. [232]
Distance from *rrn*	-0.56	0.0098	0.12	<0.0001	Blattner et al. [232]
Fis binding	0.5	<0.0001	-0.09	<0.0001	Kahramanoglou et al. [84]
IPOD	-0.34	<0.0001	0.04	0.0017	Vora et al. [78]
RNA abundance	0.24	0.0002	-0.34	<0.0001	Kroner et al. [138]
HU binding	0.13	<0.0001	0.18	<0.0001	Prieto et al. [230]
SeqA binding	0.14	0.0345	0.01	0.4114	Joshi et al. [231]
Supercoiling density	0.01	0.3494	0.04	0.0042	Lal et al. [228]
RNAP binding	-0.22	<0.0001	0.12	<0.0001	Data from Tom Goss (Manuscript in preparation)
LRP binding	-0.06	0.0565	-0.06	0.0021	Kroner et al. [138]
density/propensity	-0.49	<0.0001	-0.49	<0.0001	This study

Table 3.1: Correlation of genomic features with transcriptional propensity and integration density. Correlations are all from 500 bp median windowing of each dataset. Replicates were quantile normalized and averaged before performing correlation analysis. The RNAP binding signal was obtained following a ChIP-seq procedure derived from the ChIP-chip protocol in Mooney et al. [233] and using the same antibody, but on cells that were treated briefly with rifampicin immediately prior to crosslinking in order to immobilize RNAP at active promoters.

Relative orientation	Leading strand ρ	Lagging strand ρ
Divergent	0.11	0.14
Convergent	0.11	0.13
Tandem upstream	0.16	0.14
Tandem downstream	0.16	0.15
Codirectional	0.17	0.14
Opposite	0.12	0.16
Unstranded*	0.24	0.24

Table 3.2: Correlation of transcriptional propensity with directional RNA abundance. Given are the Spearman correlation coefficients (ρ) of transcriptional propensity (from reporters on the header-indicated strands, with respect to replication) with the corresponding stranded RNA (variable strand depending on the indicated orientation) mean over a 500 bp window from wild type cells grown under equivalent conditions (RNA-seq data from Kroner et al. [138]). Divergent, convergent and tandem orientations include data from transcriptional propensity and RNA seq from adjacent, non-overlapping 500 bp windows. Codirectional, Opposite and Unstranded orientations include values from 500 bp around the same coordinate (*Note that the Unstranded measurement includes RNA-seq data from both strands). Transcriptional propensity is slightly more correlated with RNA abundance in the tandem orientations than in convergent or divergent orientations, and the correlations are virtually identical for upstream vs. downstream RNA.

3.4.5 Sequence Composition and Nucleoid Protein Occupancy Make Unique Contributions to Transcriptional Propensity

Given the numerous correlations between transcriptional propensity and other genomic features observed above, it is useful to consider how much independent information is contributed by the various features that we have noted, and to what extent transcriptional propensity can be predicted solely on the basis of those features. We applied lasso regression to obtain regularized models that predict transcriptional propensities based on a minimal number of useful features. The input feature set included a total of 96 characteristics including sequence composition, protein occupancy data, and ribosomal RNA positioning (see Section 3.6.19 for details). During lasso regression, a regularization parameter (lambda) is gradually scaled from higher to lower values; as it does so, the penalty associated with having non- zero coefficients for various features falls, and thus more of the features contribute to the model. As seen in the fits in Figures 3.10 and 3.11, the simplest justifiable model (based on 5-fold cross validation) incorporates six features (given here in the order in which they appear during the regression): proximity to ribosomal RNA, AT content, H-NS occupancy, Fis occupancy, HU occupancy, and total protein occupancy (measured via IPOD).

Consistent with the results in Figures 3.8 and 3.9, proximity to ribosomal RNA, Fis occupancy, and HU occupancy seem to characterize regions with high transcriptional propensity, whereas high AT content, H-NS occupancy, and overall protein occupancy are associated with lower transcriptional propensity. The fact that all six of the features discussed here have non-zero coefficients in the penalized lasso regression model, even when accounting for uncertainty under cross-validation, demonstrates that each bears significant information content, rather than one of them (e.g., AT content) acting as an underlying basis for others (e.g., H-NS occupancy, which does show a preference for high AT regions [84]). The resulting linear model explains 69.9% of the variance in transcriptional propensity using the most parsimonious parameterization and 72.3% of the variance using the parameterization that minimizes the mean squared error (see Figures 3.10 and 3.11; Table 3.3). Thus, while the predictions are not perfect, a substantial fraction of the observed fluctuations in transcriptional propensity may be predicted on the basis of simple chromosomal features and nucleoid-associated protein occupancy. Note that in less conservative parameterizations, many additional features are incorporated into the model, including transcript levels, initiating (rifampicin-treated) RNAP occupancy, and density of motif matches for several other transcription factors (Figure 3.11); however, in light of the observed uncertainties upon cross-validation, inclusion of features beyond those shown in Figure 3.10 cannot be strongly justified.

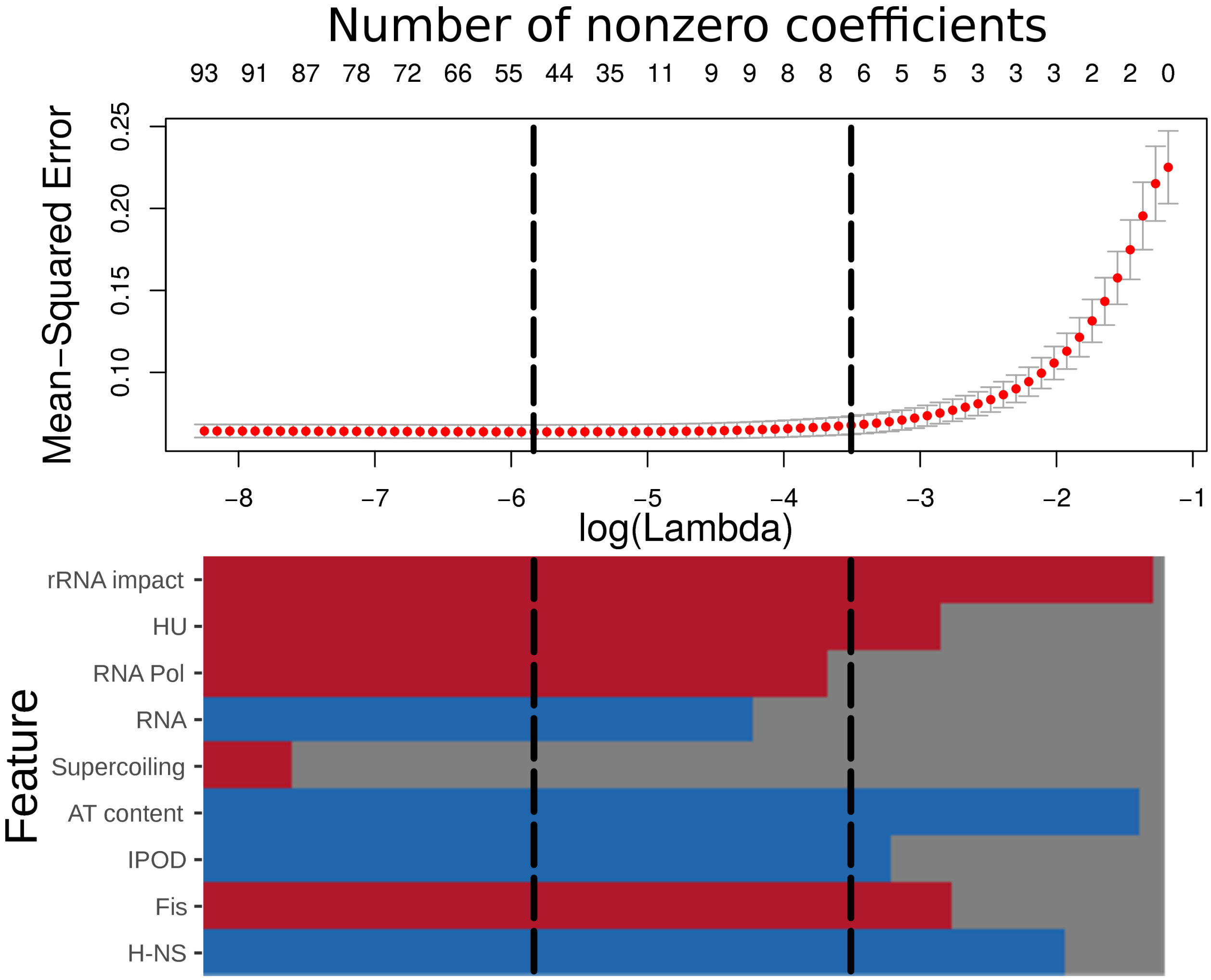


Figure 3.10: Contributions of Known Genomic Features to Prediction of Transcriptional Propensity. The top panel shows the mean-squared error (MSE) upon 5-fold cross-validation for lasso regression as a function of the regularization parameter lambda (see Section 3.6.19 for details); dashed lines occur at the point of lowest MSE (left) and at the point with the highest value of lambda that is within 1 standard error of the lowest MSE (right). The bottom panel shows the signs of coefficients for several key genomic features as a function of lambda (gray, zero; blue, negative; and red, positive). Fitted values for these coefficients at the two points shown by the dashed lines are given in Table 3.3, and a similar plot showing all features used in the fitted model given as Figure 3.11. "rRNA impact" refers to the LOWESS-fitted signal assignable to proximity to the nearest ribosomal RNA, as shown in Figure 3.6C.

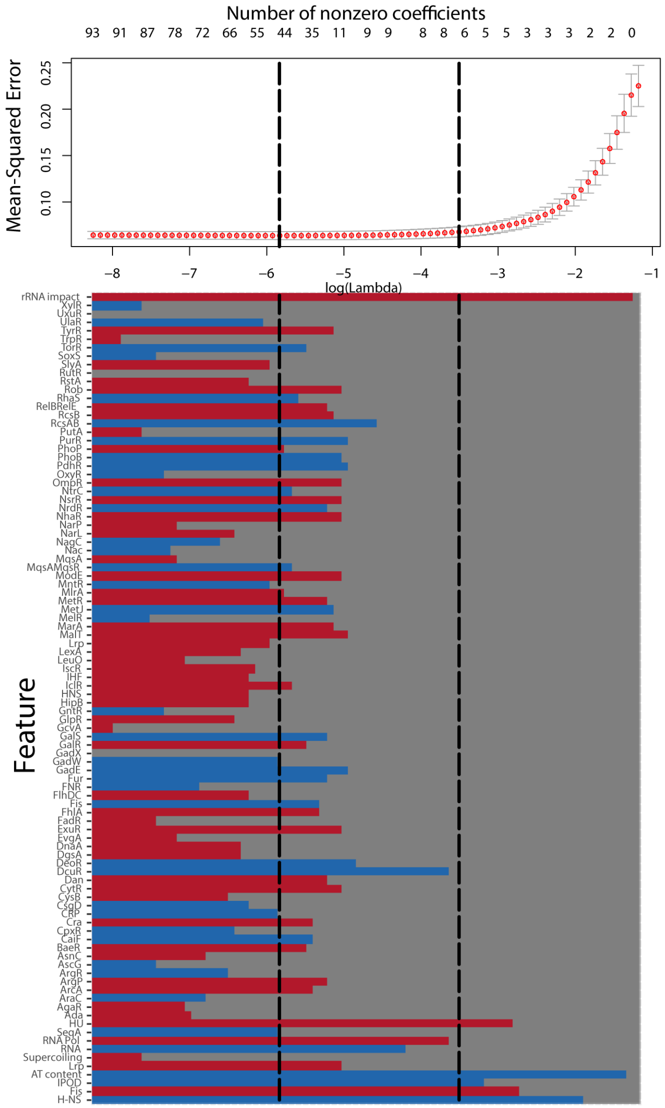


Figure 3.11: Contributions of all available features to prediction of transcriptional propensities. As in Figure 3.10, the top panel shows the mean-squared error (MSE) upon five-fold cross-validation for lasso regression as a function of the regularization parameter lambda, and the bottom panel shows the sign of the coefficient for each feature in the fitted model at that lambda (grey: zero, blue: negative; red: positive). Dashed lines occur at the point of lowest MSE (left), and at the point with the highest value of lambda that is within 1 standard error of the lowest MSE (right).

Feature	Lowest MSE	Most Parsimonious
rRNA impact	8.155826e-01	7.479827e-01
H-NS occupancy	-8.435171e-05	-7.538871e-05
Fis occupancy	1.644675e-04	1.066400e-04
Protein occupancy (IPOD)	-3.604436e-02	-1.209995e-02
AT content	-2.728446e+00	-2.356038e+00
HU occupancy	2.123566e-01	1.303875e-01
RNA Pol occupancy	3.256524e-02	–
RNA abundance	-5.131865e-05	–

Table 3.3: Fitted coefficients for the key parameters in LASSO models. Parameters correspond to features shown in Fig. 3.10 at the points of lowest mean-squared error and for the most regularized model within one standard error of the MSE (the latter should be considered the most parsimonious justifiable model); these points correspond to the dashed lines in Fig. 3.10. Note that all coefficients are on the scales of the features themselves, and thus are not of directly comparable magnitude to each other. "rRNA impact" refers to the LOWESS-fitted signal assignable to proximity to the nearest ribosomal RNA, as shown in Fig. 3.6C.

3.4.6 Location of Specific Classes of Genes Are Informative of Transcriptional Propensity

To obtain a global picture of the biological logic dictating the organization of genes into high and low transcriptional propensity regions, we used the iPAGE software package [155] to identify Gene Ontology [234, 235] terms (GO terms) that are informative about the transcriptional propensity at each gene location (Figure 3.12). As expected, large ribosomal subunit genes are informative of high transcriptional propensity (GO:0022625). Genes in pathways for enterobacterial common antigen biosynthesis or organic phosphonate catabolism (GO:0009246, GO:0019700) are clustered together in high transcriptional propensity peaks. Cellular amino acid biosynthetic process (GO:0008652), which has 105 genes in *E. coli*, is also associated with high transcriptional propensity. We also identify intracellular protein transmembrane transport (GO:0065002) as being enriched in regions of high transcriptional propensity. However, we note that there is no difference in local transcriptional propensity between genes that encode products that are recognized by the SRP machinery and all other genes (Figure 3.9K). Genes involved in cytolysis and DNA integration (GO:00019835, GO:0015074) are significantly enriched in regions of the lowest transcriptional propensity. They are also both composed predominantly of prophage genes, possibly reflecting selection for such genes to be in broadly silencing genomic contexts. Genes for lipopolysaccharide core region biosynthesis (GO:0009244) and O-antigen biosynthesis (GO:0009243) are also in regions of low transcriptional propensity, possibly because several of the GO term member genes are clustered in large tsEPODs between 3.795 and 3.810 Mb and between 2.102 and 2.115 Mb. It is important to note that iPAGE automatically filters out GO terms that do not convey substantial conditional mutual information above an already-present set; the absence of terms such as additional ribosomal protein GO terms simply arises because of this redundancy filtering.

3.4.7 RNA Abundance from Targeted Reporter Integrations Reveals a Range of Expression Compared to Native Gene Expression

In order to further test the effect of chromosomal position on reporter transcription, we used lambda Red recombineering to perform targeted integrations of the reporter construct into several different sites representing a range of transcriptional propensities. Spline-smoothed genomic DNA sequencing counts from cells grown under the same conditions (Figure 3.13A) were used to transform the transcriptional propensity map of RNA per DNA into a measure of reporter RNA counts per cell (Figure 3.13B). These dosage-corrected values can be used to identify the highest and lowest transcriptional propensity regions for heterologous gene expression (Table 3.4). We then used RT-qPCR to quantify reporter RNA from four targeted reporter integration strains relative to a set of native reference transcripts. To provide a representative range of transcriptional propensities for comparison with native promoters while avoiding extremes, sites from within EPODs (*ybcK* and *eaeH*) and from relatively high propensity regions (*phnO* and *yihG*) were selected to represent the higher- and lower-middle distribution of transcriptional propensity variation (compare to Figure 3.4F). Reporter transcription from targeted integrations was in good quantitative agreement with dosage-transformed transcriptional propensity (Figure 3.13C). Additionally, by measuring the RNA abundance per unit DNA from three native genes in each of the reporter integration strains, and comparing with insertions spanning a range of intermediate transcriptional propensities (1.5–8), we could determine the transcription from the targeted integration reporters relative to native gene expression (Figure 3.13D). These results show that RNA abundance per DNA from the reporter construct is in the 80–86th percentile when compared to native genes (Figure 3.13E), indicating a moderately strong (but far from overpowering) promoter and thus confirming the physiological relevance of our reporter.

Table 3.4 lists broad regions with the most extreme transcriptional propensity per cell (after transformation by genomic DNA copy number estimates), which may be useful information for heterologous gene expression from genomic sites.

3.4.8 Reporter Integration by Tn5 Is Biased toward Low Transcriptional Propensity Regions

The correlation between reporter integration density and known genomic features was also tested (Figure 3.14). H-NS binding had a very strong positive correlation with integration density (Figure

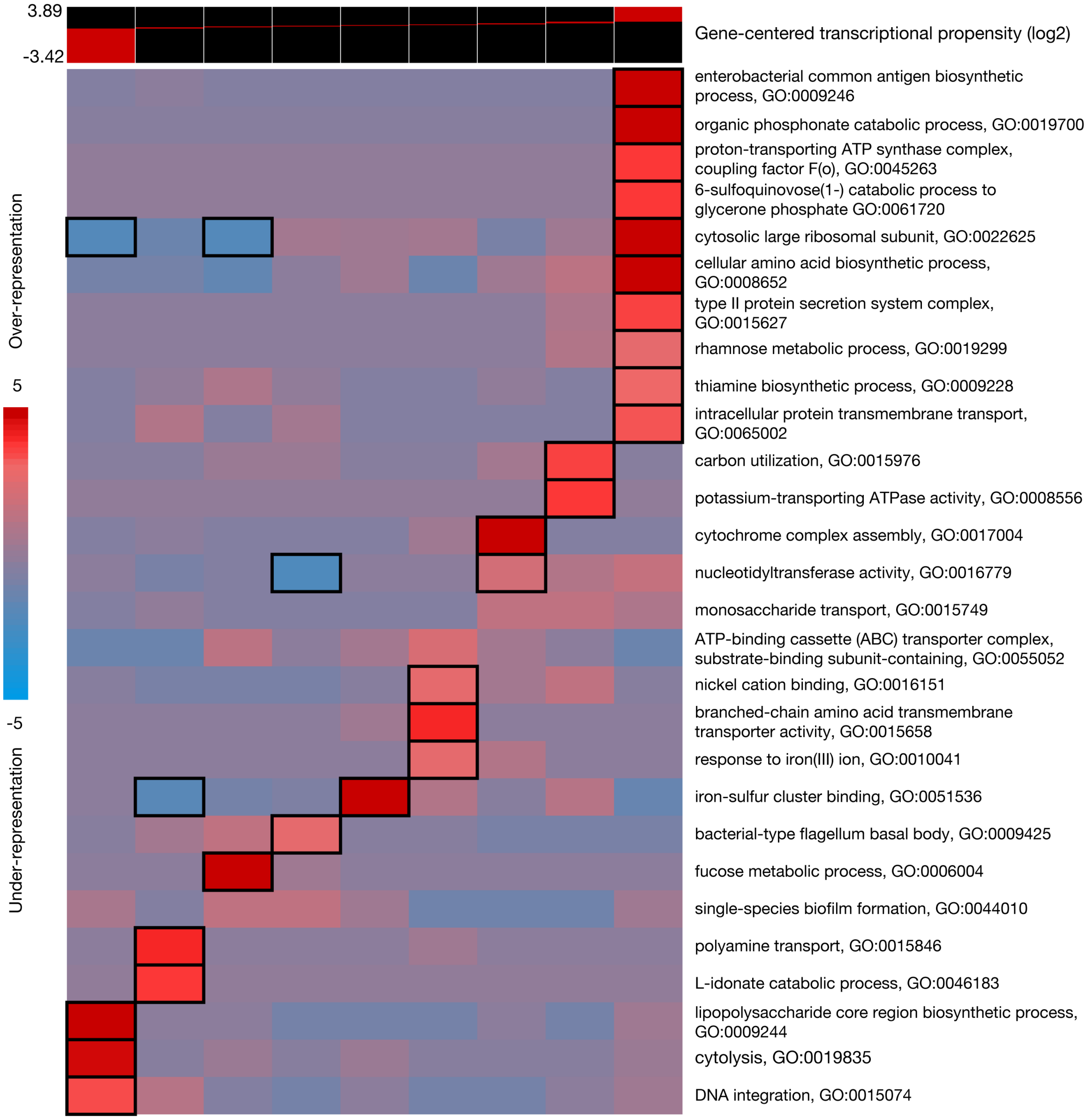


Figure 3.12: Over-Representation of GO Terms in Transcriptional Propensity Bins. Pathways identified by iPAGE analysis [155] as having significant mutual information with transcriptional propensity, and their over-representation within specific transcriptional propensity bins. Gene-specific metrics for transcriptional propensity are calculated as the $\log_2$ median values of replicate-averaged RNA/DNA ratios within the regions of the genes and all locations within 2.5 kb up- and downstream. iPAGE then discretizes the gene-specific propensity metrics into nine equally populated bins, as shown in the upper panel, where the range of the propensity metric within each bin (red boxes) is shown as a proportion of the overall range (background, black boxes). The leftmost bins contain genes within the lowest transcriptional propensity regions, with the rightmost bins containing the highest. Enrichment of GO terms within each bin is identified as the absolute value of $\log_{10}$ of enrichment p value, with sign set such that under-represented GO terms are negative (blue end of left scale) and the over-represented are positive (red end of scale). The heatmap of the sign-adjusted enrichment shows the over- (as red tiles in heatmap) and under-representation (as blue tiles) of GO terms in different transcriptional propensity levels that are visualized as separated bins on the horizontal axis of the heatmap. Tiles with bold borders indicate significant individual enrichments ($p < 0.05$ after Bonferroni correction across the row); note, however, that all displayed GO terms have significant mutual information with the transcriptional propensity profile (as assessed by the default series of tests used by iPAGE).

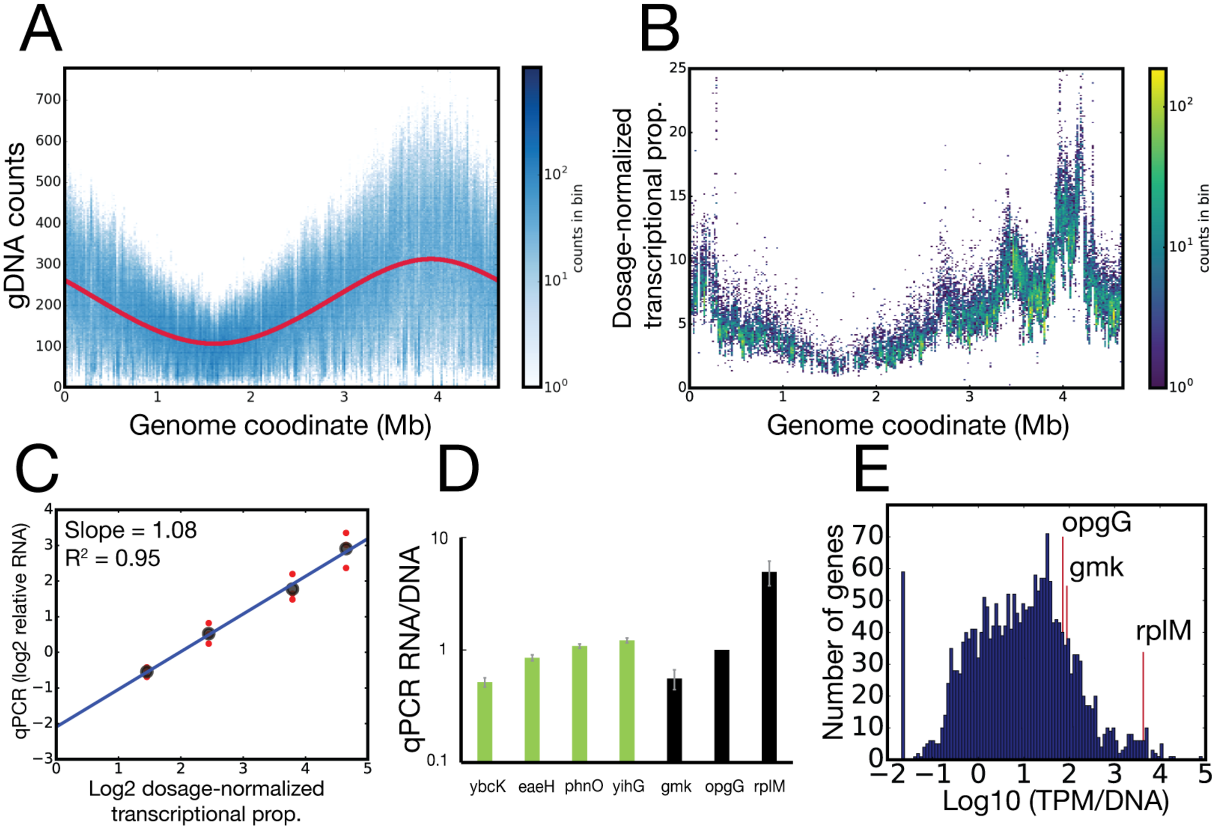


Figure 3.13: Expression from Targeted Reporter Integrations Indicate Transcription Level Relative to Native Genes. (A) Genomic DNA counts (blue) were used to generate spline-smoothed values (red) to estimate DNA dosage for cells grown under the same conditions as the reporter library. (B) Transcriptional propensity (as in Figure 3.4E) transformed by the DNA dosage spline line in (A). The map here reflects the transcriptional propensity per cell instead of per unit DNA (see Section 3.6.18). (C) qPCR measurement of RNA from targeted reporter integrations from strains grown under the same conditions as the reporter library compared to dosage-transformed transcriptional propensity. All values were normalized by opgG signal [236]. (D) qPCR measurements of RNA per DNA for mNeonGreen at four targeted integration strains (green) and for three native genes (black). DNA and DNA values are all relative to opgG signal within each replicate. Error bars for RNA:DNA ratio signal represent the standard deviation of three biological replicates. (E) Histogram of RNA abundance (as estimated from TPM using RNA sequencing [138] per DNA abundance, as estimated from DNA copy number as in (A), for each annotated gene in *E. coli*. The three native genes that were measured by qPCR in (D) are indicated.

Peaks								
Average	start	end	length	center	max	min	Intergenic center	max coord
29.4	281404	288692	7288	285048	52.9	20.9	285168	285070
18	4189573	4204803	15230	4197188	24.6	13.9	4199462	4200392
15.2	4128474	4165334	36860	4146904	22.1	10.14	4158225	4144823
13.9	4003592	4060559	56967	4032075.5	24.5	6.2	4042031	4041973
14.6	3949128	3980248	31120	3964688	24	10.2	3965702	3965302
15.3	4313204	4318232	5028	4315718	23.8	9.55	4314301	4315234
10.4	3387507	3439571	52064	3413539	18.1	5.5	3390503	3408955
Troughs								
Average	start	end	length	center	max	min	Intergenic center	max coord
2.2	1624903	1728263	103360	1676583	5.3	1	1629437	1629261
2.2	2093588	2115488	21900	2104538	3.7	1.28	2103389	2106137
3.3	2469179	2498548	29369	2483863.5	7.1	1.84	2483493	2469547
3.7	2986238	2996317	10079	2991277.5	5.1	2.7	2989806	2994033
5	3796920	3807766	10846	3802343	7	3.4	2798230	3797322
4.3	310546	325753	15207	318149.5	7.1	2.6	314320	325132

Table 3.4: Ranges for the highest and lowest transcriptional propensity regions. The average transcriptional propensity after transformation by genomic DNA dosage (as in Fig 3.13B), peak range coordinates and length (bp) are indicated, followed by the center coordinate. The maximum and minimum transcriptional propensity value are indicated. Intergenic center is a site within the indicated range where integration is not expected to result in a gene knockout. Finally the coordinate for the maximum or minimum transcriptional propensity within the range is indicated. The median and minimum values over the entire genome are 6.38 and 0.24, respectively.

3.14F). In addition, RNA abundance from native transcription showed the strongest negative correlation with integration density. Consistent with these two observations, transcriptional propensity itself was also negatively correlated with integration density. Although integration density is generally high, these results indicate that the resolution in low transcriptional propensity regions is generally better than the resolution at high transcriptional propensity regions and also suggest that the same biological mechanisms responsible for shaping low transcriptional propensity regions also tend to occur in an environment more permissive for transposon insertion. It is important to note that as the integration densities arise from libraries that have undergone growth and antibiotic selection, some bias may arise from exclusion of essential genes or those genes that cause severe growth phenotypes upon transposon insertion.

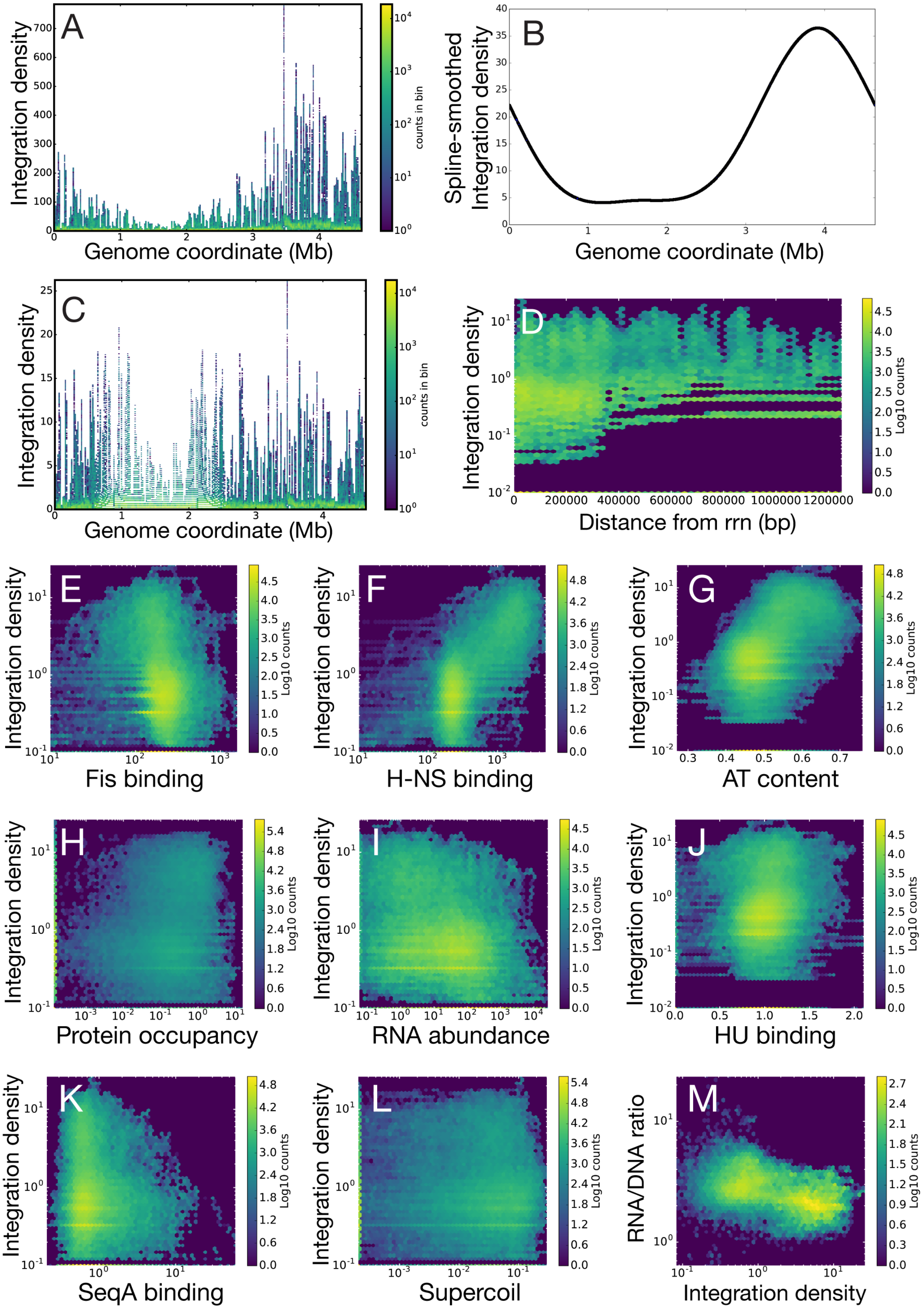

A
Integration density
0
100
200
300
400
500
600
700
Genome coordinate (Mb)
0
1
2
3
4
counts in bin
10^0
10^1
10^2
10^3
10^4
B
Spline-smoothed
Integration density
0
5
10
15
20
25
30
35
40
Genome coordinate (Mb)
0
1
2
3
4
C
Integration density
0
5
10
15
20
25
Genome coordinate (Mb)
0
1
2
3
4
counts in bin
10^0
10^1
10^2
10^3
10^4
D
Integration density
10^{-2}
10^{-1}
10^0
10^1
Distance from rrn (bp)
0
200000
400000
600000
800000
1000000
1200000
Log10 counts
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
E
Integration density
10^{-1}
10^0
10^1
Fis binding
10^1
10^2
10^3
Log10 counts
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
F
Integration density
10^{-1}
10^0
10^1
H-NS binding
10^1
10^2
10^3
Log10 counts
0.0
0.6
1.2
1.8
2.4
3.0
3.6
4.2
4.8
G
Integration density
10^{-2}
10^{-1}
10^0
10^1
AT content
0.3
0.4
0.5
0.6
0.7
Log10 counts
0.0
0.6
1.2
1.8
2.4
3.0
3.6
4.2
4.8
H
Integration density
10^{-1}
10^0
10^1
Protein occupancy
10^{-3}
10^{-2}
10^{-1}
10^0
10^1
Log10 counts
0.0
0.6
1.2
1.8
2.4
3.0
3.6
4.2
4.8
5.4
I
Integration density
10^{-1}
10^0
10^1
RNA abundance
10^{-1}
10^0
10^1
10^2
10^3
10^4
Log10 counts
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
J
Integration density
10^{-2}
10^{-1}
10^0
10^1
HU binding
0.0
0.5
1.0
1.5
2.0
Log10 counts
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
K
Integration density
10^{-1}
10^0
10^1
SeqA binding
10^0
10^1
Log10 counts
0.0
0.6
1.2
1.8
2.4
3.0
3.6
4.2
4.8
L
Integration density
10^{-1}
10^0
10^1
Supercoil
10^{-3}
10^{-2}
10^{-1}
Log10 counts
0.0
0.6
1.2
1.8
2.4
3.0
3.6
4.2
4.8
5.4
M
RNA/DNA ratio
10^0
10^1
Integration density
10^{-1}
10^0
10^1
Log10 counts
0.0
0.3
0.6
0.9
1.2
1.5
1.8
2.1
2.4
2.7

Figure 3.14 *(previous page)*: Correlations of reporter integration density with genomic features. Reporter integration by Tn5 is non-random. (A) Genome map of reporter integration density over a 500bp rolling window. (B) Spline smoothing of integrations density which reflects the integration density variation expected due to only gene dosage effects from exponentially growing cells; all smoothing splines used here had four knots located at the origin and equally spaced other points around the chromosome. (C) Integration genome map (as in A) divided by the spline smoothing values. (E)-(M) Correlation of spline-corrected integration values (as in C) with several genomic features. Spearman ρ values are in Table 3.1.

3.5 Discussion

Random integration of barcoded reporters in the E. coli genome has allowed us to map transcriptional propensity at an unprecedented resolution across the genome. Previously, a reporter has been integrated into 27,000 sites in parallel in mouse embryonic stem cells using piggyBac transposition [237]. The average resolution of one integration per 100 kbp revealed a stronger association of low transcription with lamina-associated domains than with repressive H3K9me2 histone modification. To our knowledge, as many as 38 sites have previously been tested in a single study for position-dependent expression variation in bacteria, which, because of the small 4.22 Mb *Bacillus subtilis* genome size, is a similar resolution to the mouse genome study described above [238]. Here, we used Tn5 transposition to integrate and track 144,000 barcoded reporters into the 4.6 Mb *E. coli* genome, to produce a map with an average resolution of one integration per 47 bp, the highest resolution gene-independent expression map for any species to date that we are aware of. This integration density uniquely allowed testing of reporter transcription from multiple sites within genomic neighborhoods with rare and distinct features (e.g., ribosomal RNA operon regions and extreme nucleotide content).

Considered over the entire genome, the fold-change between the highest and lowest transcriptional propensity locations is 272-fold, which is on the order of the fold-change from different sites reported from reporter fluorescence at a small number of integrations [93]. These calculated propensities represent the value for a rolling median over 500 bp windows that required at least three independent integrations, thus avoiding strong influences of single outliers. It is also important to note that most of the genome shows intermediate levels of transcriptional propensity, with 99% of sites contained within a 4.2-fold range centered upon the median (using our standard window averaging). The full range of observed propensities arises because of substantially higher values at biologically important sites such as *rrn* operons and dramatically lower values in silenced regions such as some EPODs; we consider the biological meanings of both extremes in detail below.

3.5.1 Ribosomal RNA Operons Occur in Broad Regions of High Transcriptional Propensity

Several large peaks of transcriptional propensity across the genome are centered on *rrn* operons (Figure 3.6A). The *rrn* operons are the most highly transcribed genes in the *E. coli* genome, with an average of one RNAP molecule per 85 bp, compared to one every 1020 kb for the rest of the genome [239, 240]. An *rrn* encoded on a plasmid can also physically relocate RNAP away from the nucleoid, which causes a decrease in growth rate [205], suggesting that the *rrn* genes themselves affect RNAP localization. With the exception of *rrnC* (which appears to have its physical location controlled by its proximity to the origin), the *rrn* operons also co-localize in the cell [206]. Regardless, we find that *rrnC* is also within a transcriptional propensity peak. Together, these findings suggest a model in which the very high concentrations of RNAP involved in active transcription of *rrn* occur in regions of increased transcriptional propensity, although we cannot yet determine whether the local propensity is a consequence of *rrn* transcription or has evolved to facilitate it. In general, highly transcribed native *E. coli* genes are more frequently located in *rrn*-proximal regions with high transcriptional propensity (Figure 3.9E), suggesting that both gene-specific regulation and genome organization evolve for specific expression outcomes.

3.5.2 Fis and H-NS Are Markers for Activation and Repression, Respectively

Transcriptional propensity is highly correlated with Fis binding and anticorrelated with H-NS binding, which are by far the two strongest correlations for protein binding (Figure 3.8) [84]. As two of the top five most abundant NAPs during exponential phase growth, they bind to and affect expression of many genes both directly and indirectly [71, 76, 84, 241]. In general, genes bound by H-NS are directly repressed, shown by increased expression of bound genes in the *hns* knockout strains. Fis-bound genes are typically more highly expressed. However, only 15% of Fis-bound genes are differentially expressed in a *fis* knockout strain during mid- exponential phase [84]. Our reporter is essentially identical at every integration site. Therefore, any mechanistic effects of H-NS or Fis on transcriptional propensity must occur directly on chromosomal integration neighborhood or region, rather than due to specific binding to the promoter of the reporter itself.

It is conceivable that Fis activates reporter transcription through binding to local regions around integration sites in a similar manner to rrn transcriptional activation [88, 89, 242]. Alternatively, regions of high Fis density may be activated through other mechanisms, such as spatial localization to regions of high RNAP availability, or effects on supercoiling state and DNA conformation that promote transcriptional initiation. Consistent with these models, Fis binding is also anticorrelated with distance from the nearest *rrn* (Spearman $\rho = -0.52$).

The negative correlation of transcriptional propensity with identified binding sites of the NAP H-NS from chromatin immunoprecipitation sequencing (ChIP-seq) experiments is consistent with a silencing role for H-NS (Figure 3.8B) [84]. H-NS can also oligomerize along DNA in a process dependent on non-specific electrostatic interactions with DNA [243]. Therefore, H-NS, and likely other proteins such as StpA and Hha [87], may oligomerize and bridge from silenced genomic regions into small integrated reporters and silence their transcription [81, 244]. The size, promoter strength, AT content, and other features of the reporter itself may play an important role in determining the particular transcriptional outcome at different sites because of the conflict between H-NS and RNAP, as has been suggested [82]; we reiterate, however, that for comparison among the barcoded reporters used in our study, we showed that variations in AT content of the barcode itself has no impact on observed transcriptional propensities (Figures 3.5A3.5C). We also find that reporter integration density is most highly and positively correlated to H-NS binding and low transcriptional propensity (both of which partially overlap with genomic AT content). Furthermore, integration density is anticorrelated with RNA abundance from native genes (Spearman $\rho = -0.34$). These results were surprising because the opposite occurs in eukaryotes, where low gene expression is well correlated with heterochromatin that is inaccessible to Tn5 transposon insertion, a fact used to great effect in ATAC-seq assays [245]. Although this is only a single observation for the present study, it suggests a model in which foreign DNA may more readily integrate into H-NS-bound sites, thereby increasing the likelihood that integrated foreign DNA is silenced, as has been previously proposed [246–248]. Such a mechanism would also be consistent with the enrichment of prophages and mobile elements that we observed in low transcriptional propensity regions (Figure 3.12). As opposed to horizontally transferred genes, which are often bound by and silenced by H-NS and are generally AT-rich in E. coli and closely related species [79, 80], the integration construct has 53.7% GC content. Additionally, barcode GC content has no correlation with genomic GC content in 500 bp surrounding each integration site (Spearman $\rho = 0.001$), indicating that small changes in overall reporter GC content do not affect integration site. The exact mechanism by which H-NS influences integration and expression of foreign DNA remains an important but challenging subject for ongoing studies because of partial functional redundancies and suppressive potential of other NAPs [249, 250]. We also observed very low expression from reporters integrated into tsEPODs (Figure 3.9A), supporting and greatly expanding on previous functional tests from a few sites [78, 93]. We note that there may also be sites that were silenced to the extent that integrations could not be selected for in kanamycin-containing media and were thereby lost from the library.

3.5.3 Transcriptional Propensity Has Minor Differences Depending on the Reporter Strand

In general, we observe no more than minor effects of neighboring transcription on transcriptional propensity (Table 3.2). These results may indicate that DNA supercoiling regulation is highly efficient on the chromosome, ameliorating supercoiling-mediated transcription conflicts. Additionally, the expression level for genes in various orientations used in previous studies is likely high compared to the global transcriptional activity considered for this analysis (note the relative RNA abundances in Figures 3.13D and 3.13E) [93, 212]. Similarly, we did not observe a bias in transcriptional propensity depending on whether reporters were encoded on the leading or lagging strands, indicating that replication conflicts generally do not impose a major effect on transcriptional propensity. It is possible that global strand differences would be detectable in cells that are deficient in R-loop resolution, as has been reported for reporters and native genes in RNase HIII mutant *B. subtilis* cells [251], or in the presence of higher levels of transcription through our integrated reporter.

3.5.4 Functional Classes of Genes Are Enriched at Specific Transcriptional Propensity Levels

Clustering of genes involved in the same pathway is a hallmark of bacterial genome organization [252–254]. By definition, clustered genes will end up within the same transcriptional propensity region. For example, the large operon encoding genes for organic phosphonate catabolism is entirely contained in a region of very high transcriptional propensity. However, there are other classes of genes that are not clustered, which are nonetheless significantly enriched at specific transcriptional propensity levels. For example, the GO term for cellular amino acid biosynthetic process (GO:0008652) is composed of over 100 genes, which are scattered throughout the *E. coli* genome in operons and as single genes but are significantly enriched in high transcriptional propensity regions. For genes within a specific pathway, however, clustering for co-regulation or as a result of horizontal transfer also allows genes within a common pathway to reside in the same transcriptional propensity neighborhood, which may be another evolutionary strategy by which genes in the same pathway are expressed at optimal levels. Perhaps gene clustering within a transcriptional propensity region could be considered another method of co-regulation.

In considering the implications of our results, it is important to bear in mind that all experiments described here were performed on cells growing in rich media during early exponential phase. In all likelihood, growth-phase-dependent changes in NAP occupancy [255], as well as (potentially) local regulation of transcriptional propensity across changing physiological conditions may substantially alter the positional effects of transcription. Future mapping of transcriptional propensity under different growth conditions will be particularly interesting in light of the enrichment of specific gene classes involved in rapid growth that we found in transcriptional propensity levels observed during exponential growth in rich defined media. It is also important to consider the properties of the reporter construct itself. Although design and analysis choices were made to optimize the collection of detectable signals while simultaneously minimizing the effect of the reporter on the underlying biology, there is a large diversity of gene organization in the *E. coli* genome, which may be differently affected by position depending on the physiological condition of the cell. To that end, future studies may elucidate how different gene architectures are affected by position for each cell in a population, as opposed to the population averages reported here. Our findings also provide a roadmap for how chromosomal positioning can be utilized to add another layer of regulatory tuning to control expression of chromosomally integrated heterologous pathways and potentially will enable the design of dedicated integration platforms to target particular expression levels (see Table 3.4). Future investigation into condition-dependent changes in transcriptional propensities at different genomic regions will be essential to realizing the full potential of this regulatory tool for synthetic biology applications.

Taken together, our results reveal the presence of regional variations in the transcriptional propensity of an identical construct integrated into different regions of the *E. coli* chromosome. Both extremes of transcriptional propensity appear to have functional significance: ribosomal RNA operons and important biosynthetic operons are disproportionately located in regions of high transcriptional propensity, whereas mobile genetic elements and prophages are located in regions of low transcriptional propensity. We have also elucidated several mechanistic details determining transcriptional propensity: regions of low transcriptional propensity are characterized by high levels of H-NS occupancy, high overall protein occupancy, and high AT content, whereas regions of high transcriptional propensity are characterized instead by higher binding of the nucleoid-associated proteins Fis and HU. The fact that high local levels of one nucleoid protein or another in adjacent regions of the chromosome can so profoundly impact the transcription of a uniform reporter suggests a functional compartmentalization in the bacterial chromosome akin to the distinction between euchromatin and heterochromatin in eukaryotes, where active and silenced genes are characterized by the binding state and epigenetic marks of histone proteins [256] and the three-dimensional structure of the chromosome itself [257]. We expect that future work will more fully explore both additional molecular details giving rise to these distinctions in bacteria and determine the role played by position-dependent transcriptional propensity in gene regulation and evolution.

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and Virus Strains		
E. coli K12 MG1655	CGSC	CGSC#:7740
DH5$_{\alpha}$™	Invitrogen	18265017
MG1655 Z1 malE	Addgene	a gift from Keith Tyo (Addgene plasmid) # 65915
Strain BL21(DE3)/pCP20	CGSC	CGSC#:14177 Cherepanov and Wackernagel [258]
Chemicals, Peptides, and Recombinant Proteins		
CviAII	NEB	Cat# 0640
CviQI	NEB	Cat# R0639
T4 DNA ligase	Invitrogen	Cat# 15224017
iTaq™Universal SYBR®Green Supermix	Biorad	1725120
EZ-Tn5™Transposase	Lucigen	TNP92110
Q5 Hot Start polymerase	NEB	Cat# M0493
10× ACGU	Teknova	Cat# M2103
5× Supplement EZ	Teknova	Cat# M2008
Protoscript II	NEB	Cat# M0368
AscI	NEB	Cat# R0558
PvuII	NEB	Cat# R0151
EvaGreen	Biotium	Cat# 31000
RNAProtect Bacteria Reagent	Qiagen	Cat No./ID: 76506
Critical Commercial Assays		
AxyPrep™Mag PCR Clean-Up Kit	Axygen	MAGPCRCL50
NEBNext®Multiplex Oligos for Illumina® (96 Index Primers)	NEB	E6609S
Deposited Data		
Raw sequence files for cDNA barcodes, genomic DNA barcodes, KanR-associated barcodes and transposon footprinting	Sequence Read Archive (SRA)	SRP149841
Experimental Models: Organisms/Strains		
MG1655 Z1 with proD-mCherry	This Study	ecSAS17
lambda red integration of yihG-pSAS31 integration fragment & CURED KanR	This Study	ecSAS20
lambda red integration of yafT-pSAS31 integration fragment & CURED of KanR	This Study	ecSAS21
lambda red integration of eaeH-pSAS31 integration fragment & CURED of KanR	This Study	ecSAS22
lambda red integration of htrL-pSAS31 integration fragment & CURED of KanR	This Study	ecSAS23
lambda red integration of ybcK-pSAS31 integration fragment & CURED of KanR	This Study	ecSAS33
lambda red integration of in_yagF-pSAS31 integration fragment & CURED of KanR	This Study	ecSAS34
lambda red integration of eyeA-pSAS31 integration fragment & CURED of KanR	This Study	ecSAS35
lambda red integration of nrfG-pSAS31 integration fragment & CURED of KanR	This Study	ecSAS36
lambda red integration of phnO-pSAS31 integration fragment & CURED of KanR	This Study	ecSAS37
Oligonucleotides		
Supplementary Table File S6: primer list	This Study	
Recombinant DNA		
pSIM5		Gift from Prof. Don Court
pBT1-proD-mCherry	Addgene	Gift from Michael Lynch (Addgene plasmid # 65823)
PCNS-mNeonGreen	Allele Biotech	N/A
pBAD-Flp	This Study	N/A
pSAS31	This Study	N/A
Software and Algorithms		
Autocorrelation code	Shweta Ramdas	https://github.com/shwetaramdas/autocorrelation
cutadapt, version 1.8.1	Martin [177]	https://cutadapt.readthedocs.io/en/stable/
Trimmomatic, version 0.33	Bolger et al. [178]	http://www.usadellab.org/cms/?page=trimmomatic
Bowtie2, version 2.1.0	Langmead et al. [259]	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
iPAGE	Goodarzi et al. [155]	https://tavazoielab.c2b2.columbia.edu/iPAGE/
Data analysis code	This Study	https://github.com/freddolino-lab/2018_genomeProfiling

Table 3.5: Key resources used in this study

3.6 Methods

3.6.1 Experimental Model and Subject Details

MG1655 (CSGC 7740) was obtained from the Coli Genetic Stock Center (CGSC, Yale University) [232]. We used P1vir transduction to introduce the Z1 cassette from MG1655 Z1 malE, a gift from Keith Tyo (Addgene plasmid # 65915), into MG1655. The TetR repressor itself is integrated at the *attB* site (genomic coordinates 807,328-807,342); as this is a region of high transcriptional propensity, the variations that we observe cannot be attributed to simple proximity to the site of repressor production. The MG1655 Z1 strain was then transformed with the lambda red plasmid pSIM5 (gift from Prof. Don Court). We then used the primers BT1promCh F and BT1promCh R to amplify the mCherry and ampicillin resistance cassette from pBT1-proD-mCherry, a gift from Michael Lynch (Addgene plasmid # 65823). The mCherry cassette was then integrated into a site directly downstream from *yihG* using lambda red recombination to produce ecSAS17 (MG1655 Z1 *mCherry*$^+$ *AmpR*). We confirmed the mCherry integration by genotyping and the transduction of the Z1 cassette by observing TetR-mediated repression of mNG compared to a blank MG1655 strain. ecSAS17 was then transformed with the pBAD-Flp plasmid (see below) to provide the starting strain for library generation.

3.6.2 Reporter Construct Design

The mNeonGreen (mNG) coding sequence was obtained through license from Allele Biotech [260]. We put mNG under control of the TetO1 promoter and the B0030 ribosome binding site, which is predicted to have 30-fold lower translation initiation rate than the highest rate of a native gene in *E. coli* [219, 261]. Upstream of the mNG cassette, an FRT-flanked kanamycin resistance cassette amplified from a Keio collection strain was introduced in the divergent orientation relative to mNG [262]. Directly downstream of the mNG coding sequence, we introduced an Illumina i5 adapter primer complement sequence and an AscI recognition site for later barcoding of the integration construct. The reporter and antibiotic cassettes are flanked by the strong bidirectional terminators L3S2P21 and ECK120026481 [217]. Finally the entire cassette is flanked by mosaic ends (MEs) to allow for binding to Tn5 transposase. The ME-flanked construct was modified to remove two PvuII restriction sites in order to allow for PvuII digestion of the plasmid pSAS31 and release the integration construct for Tn5 transposase binding in vitro. The full annotated pSAS31 sequence, with the exception of the mNeonGreen CDS, can be found in Supplementary Data File S1.

3.6.3 Large-Scale Plasmid Barcoding

pSAS31 was digested with the restriction enzyme AscI (NEB Cat#R0558). Primers were used to introduce the barcode and amplify the entire plasmid by PCR (Figure 3.2). The reverse primer includes one base that is either an A or T directly 5' of the annealing sequence. Fourteen hand-mixed random nucleotides followed by an AscI site are directly 5' of the A/T (Integrated DNA Technologies) (Supplementary Table File S6). Six mg of resulting fragment was digested by DpnI (NEB Cat# R0176) and AscI. The digested DNA was column purified and then ligated in a 2.4 mL reaction with 40 ml T4 ligase overnight at 14°C (Invitrogen Cat# 15224017). The reaction was quenched with 40 ml of 0.5 M EDTA. We then scaled up the Hanahan procedure to transform chemically competent cells with the ligated plasmid [263]. Cells were recovered in SOC for 1 h at 37°C before removing an aliquot for transformation efficiency counts and adding 50 mg/mL kanamycin for 8h liquid selection at 37°C. Cells were then pelleted for 7 min at 4600 $\times g$ and snap frozen in liquid nitrogen. To obtain the plasmid, snap-frozen cells were resuspended in lysis buffer for plasmid miniprep. By colony counts we estimate that 48.55 million cells were uniquely transformed with a barcoded plasmid, with a transformation efficiency of approximately one in 4,500 cells.

3.6.4 pBAD-Flp Plasmid Construction

Upon initial attempts at library construction, the pCP20 plasmid (Cherepanov and Wackernagel, 1995) caused over 90% of cells with an FRT-flanked kanamycin resistance cassette to lose resistance even at the non-inducing 28°C temperature, presumably due to leaky expression of Flp recombinase (data not shown). Since leaky expression of Flp recombinase from the pCP20 plasmid appeared to be severely reducing transposon integration efficiency, probably due the removal of the KanR cassette soon after integration and prior to liquid-phase selection, we replaced the PR temperature sensitive promoter on pCP20 with the arabinose-inducible promoter pBAD and repressor *araC* gene. The modified pBAD-Flp plasmid did not cause detectable loss of the KanR cassette under repressing conditions, yet still allowed efficient excision upon arabinose induction (data not shown). The full sequence of pBAD-Flp can be found in Supplemental Data File S2.

3.6.5 Tn5 Integration of Barcoded Reporter Constructs

To generate stable transposomes for electroporation into our target strain, we utilized the Epicentre EZ-Tn5 custom transposome construction kit following the manufacturer's instructions. In brief, barcoded pSAS31 plasmid was digested with PvuII (NEB Cat#R0151) for 1 h at 37°C and fragments were separated on a 0.8% agarose gel. The band corresponding to the integration fragment size was excised from the gel and purified (Qiagen Cat# 28706). Two ml of 200 ng/ml fragment was then incubated with 2 ml Tn5 transposase and 1 ml glycerol, according to the manufacturer's instructions. After 30 min incubation at room-temperature, the mixture was stored at -20°C. Electrocompetent cells were prepared using ecSAS17 with 34 mg/L chloramphenicol included in the growth medium in order to maintain the pBAD-Flp recombinase plasmid. One ml of the Tn5-DNA complex was mixed with 50 ml of fresh electrocompetent cells. Four separate electroporations were carried out in 2 cm electroporation cuvettes at 2500kV and immediately resuspended in 1mL of 30°C SOC medium. Each reaction was pooled into SOC medium including 34 mg/mL chloramphenicol and incubated at 30°C for 1.5 h. An aliquot for plating on selective plates to assess integrant counts was removed from the recovery medium before adding 50 mg/mL Kanamycin. Liquid selection proceeded for 16 hrs at 30°C. After liquid selection, all cells were pelleted at 4600 $\times g$ for 7 min. Cells were then resuspended in 30 mL 15% glycerol, pipetted into 30 1 mL aliquots and snap frozen in a dry-ice ethanol bath before storage of the transposon library at -80°C (the entire transformation procedure is adapted from Girgis et al. [264]). According to colony forming unit counts from plating after recovery, 609,000 cells were uniquely transformed and maintained pBAD-Flp, as indicated by resistance to kanamycin and chloramphenicol, corresponding to approximately one in 5,600 cells integrated with a reporter; thus, the odds of dual integration in a single cell are exceedingly small, and we have never observed such an event in transposon footprinting experiments performed on single colonies (data not shown).

3.6.6 Pairing Integration Site with Barcode via Transposon Footprinting

Cells from one aliquot of the transposon library were recovered in 5 mL SOC for 30 min at 30°C with shaking. Genomic DNA was isolated from the library using the Qiagen Blood and Tissue kit for Gram negative bacteria. 1 mg of the resulting DNA was digested separately with each of CviAII (NEB Cat#0640) and CviQI (NEB Cat#R0639) restriction enzymes (each has a different 4 bp cut site but leaves compatible overhangs; the use of both enzymes prevents inability to identify footprints in the rare event when a restriction site is close to the transposon insertion). An annealed Y-linker (final concentration of 10 pM of each CviQI-YTA3 with CviQI-YTA5 or CviAII-YAT3 with CviAII-YAT5, Supplementary Table File S6) that complements the overhangs was ligated to the digested DNA fragments with T4 DNA ligase (Invitrogen Cat# 15224017) for 10 min. The reaction was quenched with 1 ml 0.5 M EDTA. The DNA from the ligation mix was purified with Axygen AxyPrep Mag PCR cleanup beads at a 0.9:1 bead to DNA ratio to remove unligated Y-linker. The resulting DNA was amplified by PCR using the primers that bind within the transposon and on the Y-linker to amplify transposon-genomic DNA specific fragments (P9, P10, Supplementary Table File S6). NEBnext dual index primers (NEB Cat#E7600) were then used to add sequencing adapters by PCR with Q5 Hot Start polymerase (NEB Cat#M0493). Sequencing preparation was completed in parallel for the CviQI and CviAII cleaved samples, and they were then combined computationally during postprocessing by concatenating the resulting reads.

3.6.7 Full-Scale Genome Profiling Procedure

For each biological replicate, a single aliquot of the cryopreserved transposon library was scraped into 1 mL of M9-EZrich medium (NH_4Cl 1 g/L, KH_2PO_4 3 g/L, NaCl 0.5 g/L, Na_2HPO_4 6 g/L, $MgSO_4$ 240.7 mg/L, ferric citrate 2.45 mg/L, $CaCl_2$ 111 ng/L, 200 mL/L 5x Supplement EZ (Teknova cat # M2008), 1 mL/L micronutrient solution (Neidhardt et al., 1974)) and diluted into 50 mL of M9-EZrich with 1% Arabinose + 0.4% glycerol + 34 mg/mL chloramphenicol in a baffled 125mL flask to achieve OD_{600} (optical density at 600 nm) of 0.0031. Micronutrient solution is composed of 0.3 mM Ammonium heptamolybdate, 400 mM boric acid, 30 mM cobalt (ii) chloride, 10 mM copper (ii) sulfate, 80 mM manganese (ii) chloride, 10 mM zinc sulfate [154]. The flask was incubated at 30°C for 8 h with shaking at 225 rpm to allow Flp recombinase to excise the kanamycin resistance cassette. Cells were then pelleted at 4600 $\times g$ for 7 min and resuspended in 25 mL PBS. In parallel, an aliquot of the culture was diluted and plated on LB-kanamycin and LB plates to determine the fraction cell that permanently lost kanamycin resistance (>92.5%). Assessed by qPCR, there was 0.1% and 1.4% kanR relative to the amount of kanR remaining in the library where Flp recombinase was not induced, for replicates one and two, respectively (mNG primers P34 and P35 were used to normalize library DNA concentrations with and without Flp induction). See below for analysis and interpretation regarding the low rate of kanR retention.

After the pre-growth and kanR excision described above, cells were pelleted again and resuspended in 10 mL M9RDM. Cells were then diluted into 100 mL of M9RDM (Glucose 4 g/L, NH_4Cl 1 g/L, KH_2PO_4 3 g/L, NaCl 0.5 g/L, Na2HPO4 6 g/L, $MgSO_4$ 240.7 mg/L, ferric citrate 2.45 mg/L, CaCl2 111 ng/L, 200 mL/L 5x Supplement EZ, 100 mL/L 10X ACGU (Teknova cat # M2103), 1 mL/L micro- nutrient solution). Then a final concentration of 100 mg/L anhydrotetracycline (aTc) was added to 0.0031 OD_{600} cells. The culture was incubated at 37°C until an OD_{600} of 0.2 was reached (about 3.5 h) to allow induction of the transposon-born reporter construct. The entire flask was then immediately transferred to an ice-slurry bath. Three aliquots of 5 mL were then pelleted at 6600 $\times g$ for 3 min and snap-frozen in a dry-ice ethanol bath to allow harvest of genomic DNA. In parallel, three additional aliquots of 5 mL of the culture was rapidly mixed with 10 mL RNAProtect Bacteria Reagent (Qiagen) and frozen according to the manufacturer's instructions to allow harvest of RNA from matched samples of the growing library. All samples were then stored at -80°C.

3.6.8 Nucleic Acid Processing and Sequencing

Genomic DNA (gDNA) from harvested samples was extracted following the Qiagen Blood and Tissue kit instructions. 1 mg of gDNA was then digested for 1 hour with CviQI. The resulting DNA was purified with PCR cleanup kit and eluted into 0.1x TE. The DNA was then amplified with primers P9 and P11 flanking the barcode for eight cycles using Q5 polymerase, resulting in a 186 bp fragment (Figure 3.2; Supplementary Table File S6). The DNA from the PCR mix was purified with Axygen AxyPrep Mag PCR cleanup beads at a 1.8:1 bead to DNA ratio to remove unincorporated primers.

RNA from the exponentially growing cells was extracted following the Qiagen RNeasy Bacterial RNA protect protocol including on-column DNaseI treatment. 1 mg of the resulting RNA and a single reverse primer (P11) were used for first strand synthesis with the NEB Protoscript II First Strand cDNA kit using the manufacturer's instructions, and the resulting cDNA was stored at -20°C. No-polymerase controls (-RT) were included. 20 ml of the gDNA or 5 ml of cDNA reaction mixture was used for a 50 ml minimal-cycle PCR amplification using NEB Q5 hotstart polymerase, following the manufacturer's instructions with the following modifications: NEB i5 or i7 primers were used to add Illumina adapter sequences. EvaGreen dsDNA dye to a final 1x concentration was added to each reaction. 10 ml of each reaction (including -RT controls) were then monitored for qPCR fluorescence signal during PCR amplification. The remaining 40 ml of each reaction was then amplified with the number of PCR cycles corresponding to 25% of the maximum fluorescence observed in the 10 ml qPCR pilot reaction. We verified that the cycle threshold for the -RT cDNA controls were at least 7 cycles greater than the standard cDNA samples (indicating background from DNA contamination of less than 1%). Each 40 ml PCR reaction was then purified with 90 ml of Axygen MAG-S1 beads and eluted in 0.1x TE. The purified, prepared DNA library was was submitted to the University of Michigan sequencing core for sequencing on an Illumina NextSeq 550.

3.6.9 Construction and Testing of Targeted Reporter Integrations

The ecSAS17 strain was transformed with the lambda Red plasmid pSIM5 (gift from Prof. Don Court). The reporter construct was amplified with primers with 37-40 bp 5' flanks to introduce homology domains for each integration site (Supplementary Table File S6). Each purified reporter DNA fragment and digested with DpnI to remove the pSAS31 template plasmid. Integration constructs were then electroporated into the ecSAS17 + pSIM5 strain and plated on LB-kanamycin agar. Integration strain candidate colonies were streaked out and grown overnight at 37°C. A single colony from each candidate streak was then grown overnight at 37°C in LB broth in order to eliminate the temperature-sensitive pSIM5 plasmid. The resulting strain was then transformed with pCP20 and selected on ampicillin at 30°C in order to cure the KanR cassette. Single colonies were streaked out on LB ampicillin plates and grown overnight at 30°C. The resulting single colonies were then spotted onto both LB agar and LB-kanamycin agar in order to confirm loss of kanR. The strains showing sensitivity to kanamycin were then checked for integration using primers from each side of the chromosomal integration site and primers P34 and P35 within mNG (Supplementary Table File S6). PCR reactions that produced bands of the expected size were purified and sent to the University of Michigan Sequencing Core for Sanger sequencing and confirmed for the integration site and sequence. Confirmed strains were grown in LB broth at 37°C overnight and cryopreserved indefinitely.

For qPCR analysis of mNG transcript level in targeted integration strains, cells were grown overnight in LB broth at 37°C. The strains were then diluted 1:100 into M9RDM and grown for two h and 37°C. After the pre-growth, the cells were diluted to a final concentration of 0.0031 OD_{600} cells in fresh M9RDM including 100 mg/L aTc. The culture was incubated at 37°C until an OD_{600} of 0.2 was reached (about 3.5 h) to allow induction of the reporter construct. The entire flask was then immediately transferred to an ice-slurry bath. Two aliquots of 2 mL were then pelleted at 4600 $\times g$ for 6 min at 4°C and snap-frozen in a dry-ice ethanol bath for later harvest of genomic DNA. In parallel, two additional aliquots of 650 ml of the culture was rapidly mixed with 1.3 mL RNAProtect Bacteria Reagent (Qiagen) and frozen according to the manufacturer's instructions to allow harvest of RNA from matched samples of the growing library. All samples were then stored at -80°C. The procedure was performed in its entirety on three separate days. Purified RNA was converted to cDNA using the standard protocol for NEB Protoscript II using random hexamers. cDNA and genomic DNA was quantified by qPCR with primers P46-P51 using iTaq™Universal SYBR®Green Supermix. Cycle thresholds for *opgG* were used to normalize loading of DNA and cDNA for all other primer sets (Heng et al., 2011).

For growth rate analysis, the same pre-growth procedure as above was performed. Next, cells from the pregrowth were diluted to 0.0031 OD_{600} in fresh M9RDM with and without aTc and 150 ml was added into a clear-bottom, black-walled microplate in triplicate. Each culture was covered with 100ml of sterile mineral oil. The microplate was shaken at 37°C and monitored every 10 min for OD_{600} in order to derive the doublings per hour. The procedure was completed in its entirety on three separate days.

3.6.10 Footprinting Positions of Insertions on the Genome

Sequencing results returned from Illumina NextSeq 550 had sequencing depths of 33,197,291 and 39,030,663 for RNA barcodes, 13,260,924 and 14,763,447 for DNA barcodes, and 5,604,268 and 5,622,546 paired-end reads for kanR retention samples, for each replicate. Barcodes from inserted reporter constructs in the genome are included at the 3' end of the mNG transcript. The footprinting process sequences both the barcodes and genomic sequences after the insertion site obtained from fragmented genomic DNA. Barcodes and genomic region sequences were extracted from the obtained Read 1 sequences, using Cutadapt 1.8.1 [177] to remove a fixed length of leading or trailing sequences and to remove construct sequences. Only barcodes and genomic region sequences from reads with an identifiable construct sequence were extracted (Figure 3.2).

The extracted barcodes from the two sequencing runs for both CviAII and CviQI (four samples in total) were pooled into a single barcode read pool for further analysis, and the genomic region sequences were similarly treated. Pooled barcodes were filtered to remove any barcodes with any base of quality score below 30. The filtering survival rate for the barcodes was 65.96% (68,933,499 out of 104,513,169 reads). In parallel with barcode filtering, pooled genomic region sequences were trimmed by quality using Trimmomatic 0.33 [178], removing trailing bases with quality scores below 3, any sliding window of 4 bases that had average quality score below 15, and keeping reads with a minimum length of 20 bases. Quality trimming survival rate for the pooled genomic region sequences was 74.74% (78,114,467 out of 104,513,169). Only reads which passed both filtering steps noted above were included in alignments to the genome. For alignment, the reference was built using sequences from MG1655 genome (U00096.3), pBAD-Flp, pSAS31 , and pBT1-proD-mCherry sequences. The alignment of the extracted genomic sequences to the reference was performed using Bowtie2 (2.1.0), under the "very sensitive" preset. Alignment rate was 58.17% (45,437,982 out of 78,114,467). The query read names, 5' aligned positions, and strandedness information were extracted to match the transcriptional propensity data.

The Y-linkers used in footprinting incorporated a random 4 bp unique molecular identifier (UMI), which was then observed in Read 2 of footprinting data. Cutadapt was used to cut construct sequences as anchored 5' adapters, allowing no indels and discarding uncut sequences. The trailing sequences of construct were removed from the remainder sequences to retrieve UMI sequences. UMIs, barcode sequences, and insertion locations were matched to identify the corresponding insertion location and UMI count for each barcode, keeping only entries with all three types of information.

Tables of barcodes, UMIs, and genome positions were first deduplicated to keep only unique records of every combinations of all three source of information. Barcodes with multiple insertion positions were removed. The barcode-position pairs were then supplemented by the counts of unique corresponding UMIs. As a result, each barcode sequence was mapped to its unique insertion position on genome and its unique UMI count number. Combined, these data allowed mapping of transcriptional propensity for each barcode onto the *E. coli* genome. In total, 355,314 barcodes were mapped to *E. coli* genome (excluding sequences derived from plasmids), of which 184,575 were supported by at least 2 different UMIs; only the latter category of locations were included in downstream analysis.

For each genomic position identified above, we used the number of integration sites falling into a 1-kb window (500 bases on each side) to describe the integration density across genome (Figure 3.3E). We investigated the percentage of genomic positions with at least a certain number of integration sites within the window, for all positions on genome. Integration sites included all integrations identified and mapped in footprinting that were on the genome. The geometric mean and median were calculated including sites with no integrations in the windows (that is, across the entire genome).

3.6.11 Quantitation and Mapping of Transcriptional Propensity to Integration Sites

To retrieve the barcode sequences from RNA or DNA sequencing of the barcodes themselves, all Read 1 sequences containing the barcodes were processed using Cutadapt, with part of the construct and primer sequence removed as anchored 3' adapters (Figure 3.2). The construct cutting process allowed no indels and discarded any reads that were not cut. The counts of barcodes were measures of abundances of barcodes in RNA (as cDNA libraries) and DNA (as gDNA libraries) samples. Ambiguous barcodes were removed.

The barcode abundances were mapped to insertion positions by barcode sequences, keeping only barcodes that had at least one count in both the RNA and DNA samples of both replicates, and had a mapped location on the genome from the footprinting data. The barcodes were further filtered to require at least two different UMIs. A total of 140,292 barcodes, mapping to 98,034 locations, passed all filters on the merge, and were thus included in the transcriptional propensity calculations described in the text.

3.6.12 Quantitation of Knock-out and Growth Rate Effects of Insertions

"Knock-out effects" refer to the situation where the insertions of barcodes in or near genes could disrupt that gene's function. To identify potential knock-out effects of genes, for each gene annotation (NCBI annotation for U00096.3), four statistics were calculated: median RNA:DNA ratio within the windows of upstream 500 base pairs, downstream 500 base pairs, first 500 base pairs, and last 500 base pairs of a gene. All calculation of medians required a minimum of 10 insertions in the window. For genes with length between 500 and 1000 base pairs, overlapping windowing within the gene was allowed. Genes with length less than 500 base pairs were filtered out.

In addition to gene-specific knockout effects, we also considered the more general possibility that insertions with a strong impact on growth rate might affect the observed transcriptional propensity, due to alteration of the dilution rate of the transcript of interest. We have observed in low-throughput experiments that the impact of our reporter insertions on growth rate are very low, even for insertion locations with large differences in transcriptional propensity (Figure 3.3F), thus immediately arguing against growth rate effects as a major confounder of our observations. Nevertheless, we also assessed impacts of clonal growth rate on transcriptional propensity observations using the relative abundance of DNA barcodes, regardless of their position relative to genes or the transcription level from any reporter (Figures 3.3G and 3.3H). To evaluate the potential effect of cell growth on transcriptional propensity, the DNA abundance ratio between after induction and growth (post-) and before (pre-) were calculated. Briefly, in the pre-/post-induction experiment, the reporter library was grown under the same conditions as in the "Full-scale genome profiling procedure" and genomic DNA was collected before induction (pre-) and at after induction and growth to OD_{600} 0.2 (post-), processed and sequenced using the same methods as in Section 3.6.8. To examine reproducibility between the pre-/post- induction experiment and the experiment for transcriptional propensity profiling, we visualized and quantified the correlation of counts of barcodes in common between two sets of experiments. To reduce the effect of noise, we performed filterings of a minimum requirement of 10 or 100 counts for each barcode in each sample (Figures 3.3F and 3.3H for a minimum requirement of 10 counts).

For the growth rate experiment, we examined the reproducibility when different levels of filterings were applied, by examining how well the barcode counts from two replicates agreed with each other. More specifically, after filtering by a minimum requirement of 10 or 100 counts, we calculated the Spearman correlation between replicates as post-growth counts ($\rho = 0.17$ when filtered by 10 and $\rho = 0.90$ when filtered by 100), pre-induction counts ($\rho = 0.30$ when filtered by 10 and $\rho = 0.89$ when filtered by 100), and ratios of post- growth barcode counts over pre-growth counts (0.02 when filtered by 10 and 0.37 when filtered by 100). The correlations were generally low, suggesting a low signal-to-noise ratio (consistent with our low-throughput observations that the insertions present in the library generally had no effect on growth rate, possibly because clones containing detrimental insertions were already selected out).

To directly test whether relative growth rate had an effect on RNA/DNA ratios or transcriptional propensity, we applied the minimum count filter, then calculated the Spearman correlations of the $\log_2$(RNA/DNA) or $\log_2$(transcriptional propensity), respectively, with $\log_2$(post-/pre-induction). We required barcodes evaluated for growth-rate effects to be detected in all four samples in both experiments: growth-rate experiment with two replicates and transcriptional propensity profiling experiment with two replicates. With a requirement of minimum of 10 barcode counts from growth rate experiment replicates, 64,003 barcodes passed the filter; with a requirement of a minimum of 100 counts, 185 barcodes passed the filter. As noted below, in neither case was any substantive correlation with transcriptional propensity observed.

3.6.13 Profiling Genome-wide HU Binding Landscape

HU binding data was obtained from [230], accessed via Accession Number SRP008538 in NCBI SRA database. HupA (SRR353962) and HupB (SRR353967) single-end ChIP-seq data were downloaded as sra data files, converted into single fastq files via sra-toolkit (2.8.2-1). Quality controls were performed using FastQC (version 0.10.1). Cutadapt was used in preset Illumina sequencing cutting mode to remove adapter sequences from reads. Trimmomatic was used to remove reads and read ends of low quality. The trimming parameters were Phred+33 scores, leading quality score 3, trailing 3 and sliding windows of length 4 and quality score 15, and a minimum read length of 20 bases, in single-end mode. Alignments were performed using Bowtie2, in very sensitive preset mode, to *E. coli* genome version U00096.3. For each position on genome, the coverages were defined as the counts of aligned templates of reads that spanned the position, calculated using an in-house Python script. The resulting coverages were divided by a spline-smoothed version of the same data to correct for origin-to-terminus effect of circular genome of *E. coli* (see below for details).

3.6.14 Estimation of Gene-Level Transcriptional Propensity for Functional Analysis

For each gene annotation, the $\log_2$ median value of RNA/DNA ratio within the region of the gene's open reading frame plus a flanking region of 2500 bases on each side of the gene were calculated as the gene-level transcriptional propensity for functional analyses. iPAGE analysis was performed using nine uniformly populated bins with dependency of GO terms (Figure 3.12). Genes were also categorized by whether or not the coding products were recognized by SRP, according to the list of gene names provided in [229]. Gene names in the list were mapped to b numbers based on gene annotation for genome version U00096.3. For gene names that corresponded to multiple b numbers, the b number with the gene name as its primary name was prioritized over b numbers with the gene name as synonyms. For genes with no matching gene name but multiple synonyms, the smallest b number was used.

3.6.15 Autocorrelation

To compute the autocorrelation in transcriptional propensity across the genome, we estimated the Spearman correlation coefficient between pairs of loci separated by different base pair lags (ranging from 1 to 200,000 base pairs). For each base pair distance, we first created two lists representing expression levels at pairs of loci separated by a distance equal to the lag. We then computed the Spearman correlation coefficient between these two lists. If a given locus had multiple raw transcriptional propensity values (without median windowing), we took the median for all reporters at that coordinate. To compute the null distribution for the autocorrelation we used a white noise process with N samples where N is the total number of unique insertion locations (98,034); the algorithm described here was implemented in Matlab [265].

3.6.16 Discovery of Periodic Signals in Transcriptional Propensity

To detect periodic signals in the transcriptional propensity signal we used the astropy implementation [266] of the Lomb-Scargle algorithm [223, 224] to perform spectral analysis for signals at all possible frequencies that could repeat between 1 and 100000 times across the *E. coli* genome. Due to the circular nature of a bacterial genome, any periodic signal must repeat at some integer divisor of the full length of the bacterial chromosome and thus we restrict our analysis to only frequencies that are possible under this constraint (n.b. each period need not, however, consist of an integer number of base pairs). The Lomb-Scargle algorithm was designed for unevenly sampled linear time series data and cannot natively handle data coming from a strictly periodic series as is the case for our data. Therefore, in order to better detect low frequency, high period signals over the circular chromosome, we ran the Lomb-Scargle algorithm on the linear transcriptional propensity signal repeated, end-to-end, five times to simulate the circular genome. For all Lomb-Scargle calculations, the spectral power was normalized using the standard normalization based on data residuals around a constant reference model as described in the astropy documentation. In order to assess the statistical significance of the spectral power for the periods we observed we repeated the Lomb-Scargle analysis on 1000 permuted transcriptional propensity signals generated by shuffling blocks of 2200 adjacent collection bins (∼100,000 bp) of the original transcriptional propensity signal, and repeating the same shuffled signal end-over-end five times. Periods discovered in the original transcriptional propensity with a Lomb-Scargle power higher than the period with the highest Lomb-Scargle power in all 1000 permuted signals represent periodic signals discovered at a false discovery rate of $< 1\%$.

3.6.17 Calculation of Macrodomain and Chromosome Interacting Domain (CID) Boundaries

Macrodomain and chromosome interacting domains were calculated from data published with Lioy et al. [198] using GCC contact matrices obtained from *E. coli* cells in exponential phase growth at 37°C in LB media (GEO data sets GSM2870426 and GSM2870427). Processed count matrices taken directly from GEO and were normalized using code from https://github.com/koszul-lab/E_coli_analysis. Directionality indices were calculated as described in Lioy et al. on normalized matrices at both 100 Kb (CIDs) and 400 Kb (macrodomains) scales. Significant boundaries were determined by choosing locations where the value of the directionality index t-test transitioned to a value of +2 or greater after previously obtaining a value of -2 or less upstream. Final boundaries were chosen by taking the average of boundaries found within 25 kB of each other between both replicates. Boundaries found only in one replicate were not considered, and final boundaries were converted from the U00096.2 gene coordinates to the U00096.3 gene coordinates used in this study. Boundaries were labeled with the either the first gene overlapping or the closest gene to the boundary as found in the GeneProductSet dataset in RegulonDB version 9.4 and sorted by start coordinate in the annotation file.

3.6.18 Transcriptional Propensity, Integration Density and Experimental Data Processing

RNA barcode counts were divided by DNA barcode counts separately for each replicate to generate raw transcriptional propensity values. Each replicate was then smoothed by a rolling median window over 500 bp for all windows with at least 3 reporters. Smoothed transcriptional propensity values for all integration sites were retained. The replicates (Figure 3.4D) were then quantile normalized and averaged to generate the transcriptional propensity values used in this study, unless otherwise noted. All external experimental data sets (see Table 3.1) were subjected to the same smoothing and averaging of replicates described above. For each correlation python and Matplotlib were used to generate the hexbin plots, histograms, violin plots and Spearman statistics. All spline normalization was carried out using a smoothing B-spline with four knots, located at equidistant points along the chromosome including one at oriC to provide a low-pass filter responding primarily to DNA abundance.

Integration density was calculated by summing reporter integration over a rolling 500 bp window. Since the reporter was integrated during exponential growth phase, integration density was expected to be higher around the origin of replication (Figure 3.4C). In order to eliminate density variation arising from gene dosage effects from all correlation analysis, we performed a B-spline smoothing of integration rates over the length of the chromosome (Figure 3.14). The raw integration density data was divided by the spline values to generate the gene-dosage corrected integration density values that were used for all correlation analysis (Figures 3.14A-3.14C; Table 3.1).

In order to approximate the transcriptional propensity per cell instead of per DNA copy number (as in Figure 3.4E), we multiplied transcriptional propensity by genomic DNA copy number during exponential phase for cells grown under the same conditions (Data from Goss, manuscript in preparation). Specifically, B-spline smoothing was used on read depth of total genomic DNA (Figure 3.13A). Transcriptional propensity was then multiplied by the spline values to generate the dosage transformed transcriptional values shown in Figure 3.13B and in Table 3.4.

3.6.19 Modeling of Transcriptional Propensity Based on Chromosomal Features

To obtain a minimal set of informative parameters for use in predicting transcriptional propensity, we performed lasso regression [267] using the R package glmnet. Table 3.1 shows correlation statistics and data source for each experimental data set. We fitted the regression under five-fold cross validation, using a blocked strategy with each group consisting of four ∼230 kilobase regions of contiguous locations, in order to account for the correlation structure inherent to the data itself. We only used data points for which all features were available, and thus the cross validation regions all contain consistent numbers of locations, but not necessarily precisely the same size of genomic region.

3.6.20 Elimination of Potentially Confounding Features

As described here, we considered and subsequently eliminated several possible sources of systematic bias in our transcriptional propensity measurements. For the effects considered here, we observed in some cases small effects at the level of individual barcodes, but all such effects were eliminated upon applying the window averaging used in our actual transcriptional propensity statistic (500 bp rolling median requiring at least three independent barcodes), demonstrating a lack of meaningful contribution from any of the factors noted here to the reported signal.

3.6.20.1 Barcode Sequences

We observed very low correlation of transcriptional propensity with GC content of the barcode (Spearman $\rho = 0.13$ and 0.15 for each replicate, respectively, considered at the level of individual barcodes), and essentially all detectable bias was eliminated by the median windowing that we applied in our analysis (Spearman $\rho = 0.013$ for the 500 bp moving median signal) (Figures 3.5A-3.5C), eliminating any impact on our final analysis.

3.6.20.2 Transposon-Based Knockout Effects

In principle, it is possible that the signal that we observed could be altered by the effects of gene disruptions caused by reporter integration. In practice, however, we observed that reporters integrated within the beginning of a gene coding sequence (thus knocking out its activity) were nearly identical to reporters integrated directly downstream of the CDS for the vast majority of genes (Figure 3.5D; Supplementary Table File S5), and thus knockout effects appear to play little role in our observations. We could identify only five genes where transcriptional propensity from reporters within a gene differed from the surrounding neighborhood (by at least 1.7 fold) that could be potentially attributed to gene knockout effects instead of H-NS peak location (*rep*, *bioH*, *dtpC*, *yfdL*, and *ftsN* - see Supplementary Table File S5). The largest effect was a 2.4 fold-change and these represent exceptions rather than the rule. Most likely, integrations in genes that would globally affect transcriptional propensity also result in a competitive disadvantage during growth and were therefore lost from the library, as were integrations in essential genes. However, there may be some loci where reporter integration causes minor growth defects and results in the appearance of a slightly elevated transcriptional propensity over specific loci as a result of the decreased growth rate. Based on the threshold used for our identification of knockout effects above, we would expect most of these cases to have an effect of less than 1.7-fold. Potential cases that were not already identified with the knockout analysis (Figure 3.5D) do not explain the genome-scale signal variation visible in Figure 3.1E, which varies over 10 - 100 kb.

3.6.20.3 Effects of Clone-Dependent Growth Rates

We also considered the possibility that differences in growth rate might impact transcriptional propensity measurements by altering the effective stability of the transcript (through altering its dilution rate). Some evidence against this possibility arose from our consideration of targeted insertions (Figure 3.3F), where we saw no growth rate effects from any insertion locations tested, despite those locations showing a 15-fold range in transcriptional propensities. To provide a more comprehensive test of possible effects of growth rate, we estimated the relative growth rates from many transposon-inserted reporters by measuring the abundance of each genomic DNA barcode at the start of the assay period compared to the end of the assay period. We found low correlations of growth-rate dependent genomic DNA abundance of barcodes with either raw RNA/DNA ratios and transcriptional propensity (see Table 3.1 for correlation coefficients), which suggested that the differences in relative growth rates did not have a substantial effect on our measures of transcriptional propensity. We also noticed that the signal-to-noise ratio in these assays was generally low, likely due to a relative rarity of insertions that caused strong changes in growth rate. The Spearman correlation of the observed replicate-averaged genomic DNA abundance ratios of barcodes after and before growth with raw transcriptional propensity as RNA/DNA ratios before median windowing, was 0.01 using a threshold of 10 counts (Figure 3.3G) and 0.14 using a threshold of 100 counts (Figure 3.3H), demonstrating that any effect of growth rate on transcriptional propensity is exceedingly weak, and in particular is not needed even for very low or high propensities (Figure 3.3H). Note well that the correlations stated here are for individual barcodes, rather than the window-averaged statistics used for transcriptional propensities. Indeed, the Spearman correlations of the window-averaged, growth-dependent changes in DNA barcode abundance ratios with transcriptional propensity (500 bp median windowing) at each genomic position that had insertions of filtered barcodes was -0.02 with a 10 count filtering threshold, and 0.02 with a 100 count filtering threshold, demonstrating a complete lack of meaningful correlation with the key statistic used in our work.

3.6.20.4 kanR Excision Efficiency

Although the rate of kanR retention after the excision step is generally very low (0.1-1.4% percent of total mNG signal as assessed by qPCR), there was a mild correlation of the rate of kanR retention between replicates for different reporters (Figure 3.5G), and between the rate of kanR retention and the transcriptional propensity at individual integration sites (Figures 3.5E and 3.5F). However, unlike the transcriptional propensity signal, the fluctuations in kanR retention were not highly correlated between nearby sites, and lose all correlation with the transcriptional propensity signal upon median windowing (Figures 3.5H and 3.5I). The combination of this lack of overall correlation, and the very low absolute rates of retention of the kanR marker (see above), leads us to conclude that any site specific variations in kanR excision efficiency have no meaningful effect on our overall transcriptional propensity profiles (indeed, it may well be that the site level correlation that is observed is caused by the variation in propensity/ accessibility per se, rather than by retention of the marker).

3.6.20.5 Effects of Neighboring Sequence Context

The transcriptional propensity signal was not a result of bias in DNA barcode amplification that could be present due to variations in neighboring genomic AT content, as the two had no meaningful correlation (Figure 3.5J). As expected from the large and diverse reporter library strain, counts for DNA barcodes and RNA barcodes vary substantially across the genome and at the local scale (Figures 3.5K-3.5M). Taken together, these figures indicate that genomic nucleotide content has a strong effect at the level of transcription (Figure 3.8C), but is not a result of bias introduced by light PCR amplification of genomic DNA barcodes.

3.6.21 Data and Software Availability

Source code implementing the autocorrelation analysis: https://github.com/shwetaramdas/autocorrelation.

Source code implementing all other statistical analysis and modeling: https://github.com/freddolino-lab/2018_genomeProfiling

We provide a comprehensive table of transcriptional propensity data (Supplemental Table S7), which contains the following columns (with description): t1_cDNA (replicate 1 raw RNA barcode counts), t1_gDNA (replicate 1 raw DNA barcode counts) ,t2_cDNA (replicate 2 raw DNA barcode counts), t2_gDNA (replicate 2 raw DNA barcode counts), pos (U00096.3 coordinates), strand (relative to U00096.3 reference), gc_fraction (Fraction of GC in barcode), raw_propensity1 (replicate 1 RNA/DNA barcode counts), raw_propensity2(replicate 2 RNA/DNA barcode counts), 500_med_win_propensity1(replicate 1 median of RNA/DNA barcode counts in 500 bp window around each integration when at least 3 integrations are included), 500_med_win_propensity2(replicate 2 median of RNA/DNA barcode counts in 500 bp window around each integration when at least 3 integrations are included), Avrg_500_med_win_propensity (average of replicates after median windowing and quantile normalization).

We provide the raw sequence files for the cDNA barcodes, genomic DNA barcodes, KanR-associated barcodes and transposon footprinting reads. The accession number for all sequencing data is Sequence Read Archive: SRP149841.

In addition, three large supplementary tables are provided as separate files.

3.7 Acknowledgements

We are grateful to all members of the Freddolino and Lin labs and the University of Michigan Microbiology Super Group for many helpful comments and suggestions. We thank Shweta Ramdas for implementation of the autocorrelation analysis. This work was supported by NIH R00-GM097033 (to P.L.F.), R03-AI130610 (to P.L.F.), and R35-GM128637 (to P.L.F.). S.A.S. was additionally supported by the NIH Cellular and Molecular Biology Training Grant T32-GM007315. M.B.W. is supported by an NSF Graduate Research Fellowship DGE 1256260.

3.7.1 Author Contributions

Conceptualization, S.A.S., P.L.F., and X.N.L.; Methodology, S.A.S. and P.L.F.; Investigation, S.A.S. and R.D.; Data Curation, R.D. and M.B.W.; Writing - Original Draft, S.A.S., P.L.F., and R.D.; Writing - Review & Editing, S.A.S., P.L.F., M.B.W., R.D., and X.N.L.; Funding Acquisition, X.N.L. and P.L.F.; Resources, M.B.W. and E.M.F.

CHAPTER 4

Global analysis of RNA metabolism using bio-orthogonal labeling coupled with next-generation RNA sequencing

4.1 Contribution details

This work was reproduced from its published form, with permission, from Wolfe et al. [268]. As the primary author, I performed the literature review, performed the analyses, and wrote the manuscript. Peter assisted with the creation of some key figures. Discussions between him and I resulted in the section on the role of spike-ins. Discussions with and edits from both Peter Freddolino and Aaron Goldstrohm helped vastly improve the content and direction of the review article. Their mentorship made this review possible.

4.2 Abstract

Many open questions in RNA biology relate to the kinetics of gene expression and the impact of RNA binding regulatory factors on processing or decay rates of particular transcripts. Steady state measurements of RNA abundance obtained from RNA-seq approaches are not able to separate the effects of transcription from those of RNA decay in the overall abundance of any given transcript, instead only giving information on the (presumed steady-state) abundances of transcripts. Through the combination of metabolic labeling and high-throughput sequencing, several groups have been able to measure both transcription rates and decay rates of the entire transcriptome of an organism in a single experiment. This review focuses on the methodology used to specifically measure RNA decay at a global level. By comparing and contrasting approaches and describing the experimental protocols in a modular manner, we intend to provide both experienced and new researchers to the field the ability to combine aspects of various protocols to fit the unique needs of biological questions not addressed by current methods.

4.3 Introduction

Gene expression is modulated at multiple stages including transcription and processing of nascent transcripts, regulation of translation efficiency and intracellular localization, and control of the rate of RNA degradation. This chapter focuses on advanced methods to measure mRNA decay rates on a transcriptome-wide basis. Multiple RNA decay pathways that degrade RNAs have been discovered and specific regulatory factors that control the rates of RNA decay have been reported [123, 269, 270] . These include short regulatory RNAs (siRNA and microRNAs) [271] and a plethora of RNA binding proteins [126]. The key challenge now is to determine the impact of each of these factors on the transcriptome using facile quantitative approaches. Early approaches to measuring mRNA decay involved shutting off transcription and measuring RNA abundance over time using either northern blots, dot blots, or radioactively labeled RNA [272–276]. However, concerns over the impact on the underlying biology for cells undergoing transcription shutoff have led to the development of various methods to metabolically label RNA to measure RNA decay in a less intrusive manner. By incorporating a chemically modified nucleobase into the cellular pool of ribonucleotide triphosphates (NTPs), RNAs can be labeled without disrupting gene expression, thereby minimally perturbing the underlying biology. Additionally, the indiscriminate nature of metabolic labeling combined with label-based purification methods and modern RNA sequencing allows for transcriptome-wide determinations of both transcription rates and RNA decay in a single experiment. Here we review both historical and recent advances in methods using metabolic labeling to quantitatively measure RNA decay in living cells. We take a modular approach, by describing individual aspects of the methods that have been developed in such a way that each step can be mixed and matched with later steps, so that unique experimental designs can be developed to answer challenging biological questions using an optimal combination of approaches.

4.4 Metabolic Labels

The cornerstone of most modern sequencing-based workflows for measuring RNA decay is the use of metabolic labeling, via the incorporation of nucleotide analogs into RNA, which are used to separate or distinguish the labeled RNA from the rest of the cellular RNA pool. Here we review the development and characteristics of several of the most frequently used metabolic labels in modern practice.

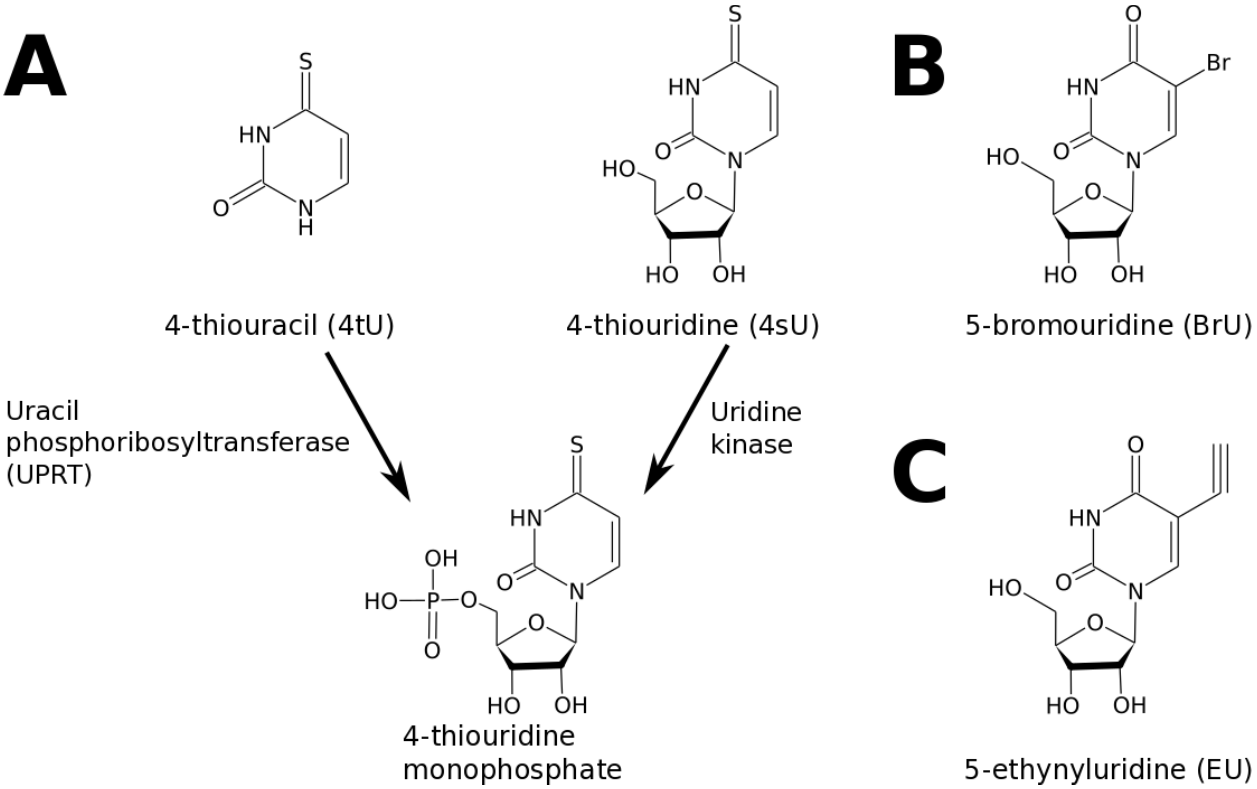


Figure 4.1: Structures and inclusion chemistries of common RNA metabolic labels. **A)** 4-thiouracil variants and pathways for incorporation into nucleotide metabolism; once the nucleotide monophosphate is formed, the resulting compound is readily incorporated into cellular RNA. **B)** Structure of 5-bromouridine, which can be assimilated through the uridine kinase pathway as on the right side of panel **A**. **C)** Structure of the click chemistry substrate 5-ethynyluridine, again typically incorporated into the cellular nucleotide pool via uridine kinase activity.

4.4.1 Thiol-containing uracil analogs

A variety of different modified uracil labels have been used to measure both mRNA decay and mRNA synthesis. The most basic requirements for such a label are that it be cell permeable, readily incorporated into RNA, minimally perturb cellular physiology, and permit either the purification or specific detection of RNA molecules containing the label. Several commonly used metabolic labels meeting these criteria are shown in Figure 4.1. The most widely used label is 4-thiouridine in either its nucleoside (4sU) or nucleobase (4tU) forms (Figure 4.1A). Both 4tU and 4sU are readily taken up by yeast [277, 278], archaea [279], and higher eukaryotes including human cells [280–282]. In contrast to other thiol-modified nucleotides, incorporation of 4sU at concentrations of up to 100 μM in cell culture does not have a discernible impact on the synthesis of RNA or protein degradation rates indicating limited perturbation of transcription and translation following incorporation of the label [283]. In contrast, 6-thioguanine (6sG) and related compounds are still at times used for metabolic labeling of RNA [284], but as 6sG has been shown to perturb both transcription and translation [283], 6sG is of substantially less utility for the long term labeling required for RNA stability experiments. Although long term culture (48 hr) in the presence of 4sU has been associated with a decrease in cell viability [285], short term labeling (10 hrs) of up to 4 mM 4tU does not appear to have a discernible impact on cell growth in yeast [278]. However, in vitro translation assays have revealed that 4sU-containing mRNAs can decrease ribosomal elongation processivity and increase downstream initiation rates [286]. For organisms such as *S. cerevisiae* and *E. coli* that express a functional Uracil Phosphoribosyltransferase (UPRT), 4-thiouracil can be used in place of 4-thiouridine as it is readily converted to 4sU as needed by the cells. However, in both mouse and human cells incorporation of 4tU into cellular RNA does not readily occur and coexpression of UPRT from another organism is needed in order to incorporate 4tU into nascent RNA [277, 287]. Expression of the well-characterized *Toxoplasma gondii* UPRT has been used to successively label RNA with 4tU in human foreskin fibroblasts [277] and, subsequently, in a variety of other cell types [288, 289]. The requirement for UPRT activity in labeling with 4tU has also led to the development of “TU-tagging”, a method to selectively label mRNAs in only one cell type in the context of a mixed population of cells. By expressing UPRT only in the cell type of interest, one can determine both the identity and mRNA decay rates of the mRNAs from that cell type [290]. Additionally, 4tU [Sigma Aldrich Cat. No. T4509] is substantially cheaper than 4sU [Sigma Aldrich Cat. No. 440736] and is more economical to use for organisms that already have robust endogenous UPRT activity (please note that throughout this review we indicate product numbers merely as examples, and not as a reflection of endorsement of any particular product or manufacturer). In whatever form it is introduced, RNA-incorporated 4sU readily crosslinks to both RNA and protein upon exposure to 365 nm UV light, a feature that is taken advantage of for the analysis of RNA-protein interactions but should be minimized in the analysis of mRNA decay

[291, 292]

4.4.2 Halogen-containing uracil analogs

Incorporation of 5-bromodeoxyuridine (BrdU) into cellular DNA was first described in the 1950s [293, 294] and the development of an anti-BrdU antibody allowed for visualization of DNA within a living cell [295]. Some BrdU antibodies cross-react with 5-bromouridine (BrU) and labeling with BrU can be used to selectively purify BrU labeled RNA with an anti-BrdU antibody [296]. Like 4sU, BrU (Figure 4.1B) is readily taken up by mammalian cells [297] and BrU does not appear to have the same general toxicity effects that 4sU has under long exposure [285], making it an attractive reagent to use for measuring mRNA decay over a longer time course. However, in vitro translation assays have revealed that BrU containing mRNAs have a modest negative impact on both ribosomal elongation and initiation, but not as large in magnitude as the effects seen from 4sU [286]. Additionally, BrU [Sigma Aldrich Cat. No. 850187] is comparable in price to 4tU and does not require UPRT activity for incorporation into mammalian cellular RNA.

4.4.3 Alkyne-containing uracil analogs

First described as a labeling reagent for fixed cells, 5-Ethynyluridine (EU; Figure 4.1C) is a uracil derivative capable of performing “click” chemistry (reviewed [298]) both in vivo and in vitro [299]. Like BrU, 4sU and 4tU, EU is rapidly taken up into the cellular pool of NTPs and incorporated into transcribed RNAs. Similar to 4sU and 4tU, short-term labeling with EU does not appear to have negative effects on cellular health, but longer incubation times do negatively impact growth rates [285]. Although EU could be used for high throughput determinations of RNA synthesis and decay rates, most studies have been primarily focused on targeted measurements of select RNAs through the use of qRT-PCR [300, 301]. Recent development of 5-Ethynylcytosine (EC) [302] in conjunction with expression of cytidine deaminase and UPRT in *Drosophila* has led to the development of “EC-tagging”, a method to purify cell-type specific RNAs with higher specificity than “TU-tagging” with 4tU as described above [303]: EU is generated in situ by target cells through the combined activities of ectopically expressed cytidine deaminase (to generate 5-ethynyluracil) and UPRT (to generate EU, analogous to the reaction in Figure 4.1A). Additionally, EU has recently been used to determine the nascent-RNA “interactome” through a combination of EU labeling and UV crosslinking coupled with RNA-seq and proteome analysis, indicating that EU labeling can be successively used with high throughput methods [304]. EU is significantly more costly than 4sU, 4tU or BrU, and can be purchased either stand alone [Invitrogen Cat. No. E10345] or in the Click-iT Nascent RNA Capture Kit [Invitrogen Cat. No. C10365] along with buffers and protocols for its use.

4.4.4 Impact of exogenous labels on RNA decay rates

A growing body of evidence has suggested a role for RNA modifications in the post-transcriptional control of gene expression, including control of RNA processing, binding of RNA binding proteins, changes in secondary structure, and stop-codon readthrough (reviewed in [305]). Although none of the labels introduced above perfectly match the modifications that have been found naturally in eukaryotic cells, in principle, these exogenous labels could still disrupt RNA decay rates through similar mechanisms. While this represents an important consideration, to the best of our knowledge there have been no targeted experiments designed to test the impact of any of the labels above on RNA decay rates themselves. Genome-wide comparisons between 4sU labeling and transcriptional shutoff experiments in yeast have shown that RNA decay rates determined from transcriptional shutoff experiments have greater agreement with one another than they do with RNA decay rates determined using metabolic labeling [306]. However, Sun et al. also show that decay rates determined from transcriptional shutoff experiments correlate well with genome-wide measurements of mRNA decay made using metabolically labeled RNA in cells displaying a transcriptional shutoff phenotype. They further show, by using measurements of metabolically labeled RNA, that RNA decay in cells under osmotic stress or heat shock also correlate well with RNA decay rates determined in transcriptional shutoff experiments, suggesting that perturbations of RNA abundances from cellular responses to transcriptional shutoff may mimic stress responses and confound measurements of RNA decay [306]. On the other hand, comparisons of RNA decay rates determined from seperate labs using different experimental strategies with the same metabolic label do not correlate well with one another, suggesting that there may be sources of experimental error in labeling experiments that are poorly understood [306]. One possible source of error could be attributed to differences in normalization between RNA abundance measurements. For example, Lugowski et al. report better replicate to replicate correlation, as well as better agreement to transcriptional shutoff experiments and metabolic labeling experiments from other labs, using an internal normalization method (normalize to introns) as opposed to an external method (normalize to spike-in) [307]. Further discussion on the impact of normalization methods on measurements of RNA decay can be found in Section 4.9. As it stands, it is unclear if there is a single major source of discrepancy that results in disagreement between measurements of RNA decay between different labs and experimental approaches.

4.5 Selection and purification of labeled RNAs

In the majority of metabolic labeling experiments at present, labeled RNA molecules are physically isolated from the total RNA pool prior to analysis (one notable exception, SLAM-Seq, is described

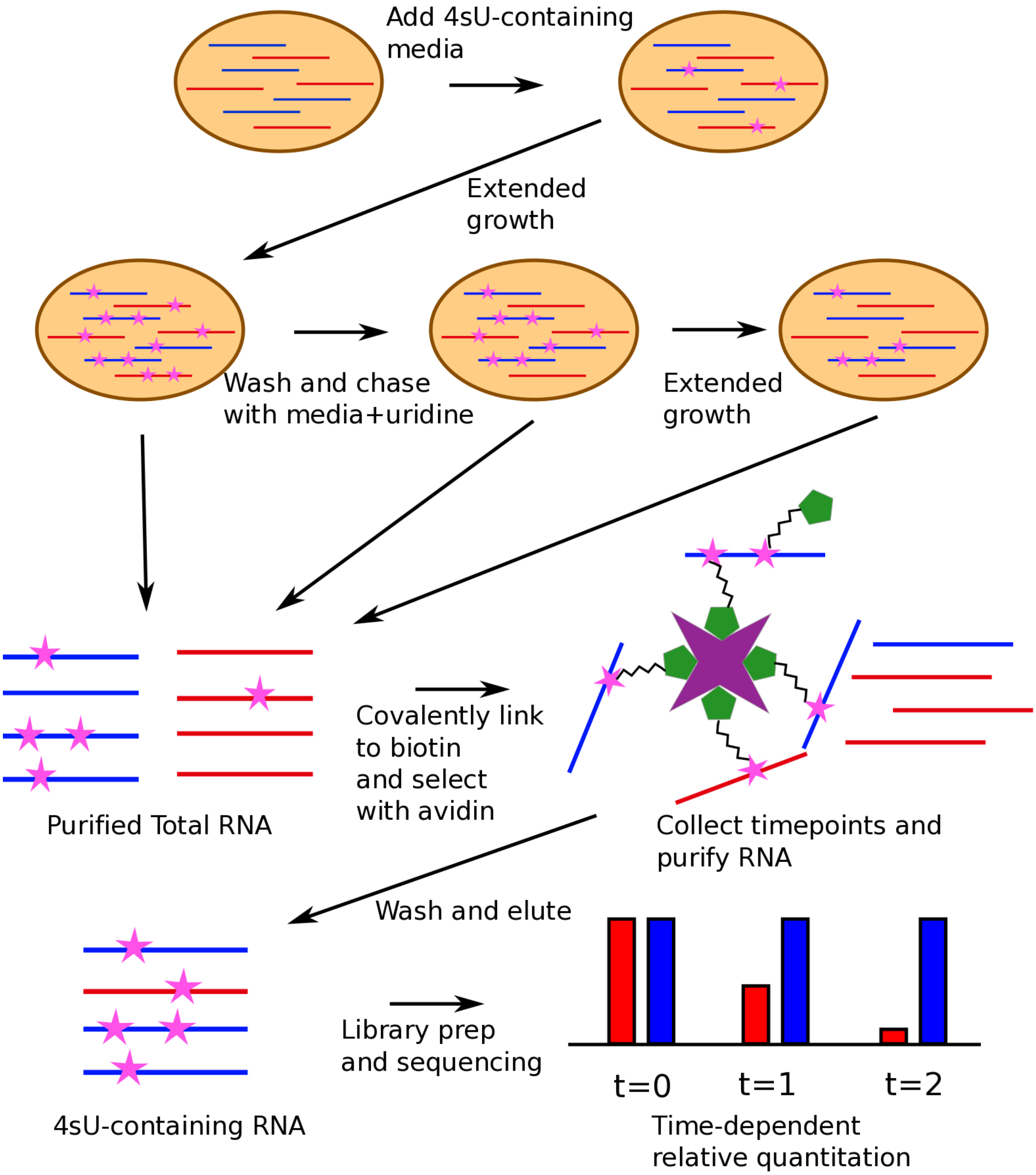

Add 4sU-containing media
Extended growth
Wash and chase with media+uridine
Extended growth
Purified Total RNA
Covalently link to biotin and select with avidin
Collect timepoints and purify RNA
Wash and elute
4sU-containing RNA
Library prep and sequencing
t=0
t=1
t=2
Time-dependent relative quantitation

Figure 4.2: Workflow of a 4sU chase experiment to measure the stabilities of different RNA species. Shown is a hypothetical cell containing two types of transcript (blue and red), with similar equilibrium levels but differing stabilities. Cells are grown in media with 4sU added to label transcripts, and then washed and chased with media containing unlabeled uridine, with samples harvested for RNA extraction at two or more time points during the pulse/chase. 4sU-containing transcripts are then covalently linked biotin and purified using streptavidin, and the enriched RNA prepared for sequenced using standard methods. Note that the RNA purification and 4sU enrichment steps are performed separately for each time point.

below). After purification of total RNA from the cell lysate, the newly labeled RNAs must be separated and purified using methods specific for the incorporated label. Each label discussed above uses different chemistry for selection, but the general principle is the same: select the label with as high affinity as possible thereby minimizing the amount of starting material needed and maximizing capture specificity. Once purified, the labeled RNAs are then quantified using standard RNA-seq methods (Figure 4.2) [282, 308–310].

4.5.1 HDPD-biotin

For 4sU and 4tU, purification is performed by chemically linking the labeled RNA to biotin and using the well-studied affinity between biotin and streptavidin to purify the RNA-biotin complex [311]. 4sU labeled RNA can be covalently linked to biotin by taking advantage of the thiol-containing uridine and forming a disulfide bond to modified biotin molecules. The most commonly used modification to biotin is N-[6-(Biotinamido]hexyl]-3'-(2'-pyridyldithio)-propionamide (HPDP-biotin) [277, 278, 280, 282] and HPDP-biotin is readily available in the form of the EZ-link HPDP-Biotin kit [Thermo Scientific Cat. No. 21341]. The covalent link between 4sU and HPDP-Biotin is completely reversible and elution is performed through the reduction of disulfide bonds with a reducing agent such as DTT, which results in RNA without covalently bound adducts as input into downstream sequencing.

4.5.2 MTS-biotin

While the HDPD-biotin based procedure described above has been widely used, the formation of a disulfide bond between 4sU and HPDP-biotin is inefficient; disulfide exchange reactions between 4sU and HPDP-biotin indicate that less than 20% of free 4sU is converted to 4sU-HPDP-biotin in reactions as long as 120 minutes. Recent developments using methylthiosulfonate-biotin (MTS-biotin) have indicated greater than 95% conversion of free 4sU to 4sU-MTS-biotin in as little as five minutes, indicating a fast and efficient reaction resulting in capture of labeled RNA without the need for as much starting material [310]. The MTS-biotin purification protocol has been used to study miRNA turnover [312], response to viral infection [313], and transcription rates in yeast [314], but it has not enjoyed as much widespread use as HPDP-biotin, possibly because of MTS-biotin's relatively recent introduction as a viable alternative to HPDP-biotin. Additionally, MTSEA-biotin [Biotium Cat. No. 90064] is less costly than HPDP-biotin, making it a more economical alternative.

4.5.3 Anti-BrdU antibody

Unlike 4sU, BrU does not have a chemical group that can be easily used to create reversible crosslinks with modified biotin. Thus, purification of BrU-containing RNAs must proceed with non-covalent interactions mediated through well-established anti-BrdU antibodies (which frequently also bind BrU). Many commercially available Anti-BrdU antibodies have been used for the quantification of mRNA synthesis or decay through BrU labels: mouse anti-BrdU [Roche 11170376001] [297], BrdU Antibody (IIB5) [Santa Cruz sc-32323] for GRO-Seq [315], Anti-BrdU mAb 2B1 [MBL International Corporation, cat. No. MI-11-3] for BRIC-seq [316], and mouse anti-BrdU [BD Pharmingen, 555627] for Bru-Seq and BruChase-Seq [308]. Imamchi et al. [316] indicated that they have tried multiple anti-BrdU antibodies and the reported 2B1 antibody resulted in the highest yields, but to the best of our knowledge, no extensive comparison of antibody purification efficiencies has been published.

4.5.4 Click chemistry

As with 4sU, purifying RNAs labeled with EU usually relies on a covalent linkage with biotin and selection using streptavidin beads, in this case using the bio-orthogonal copper-catalyzed azide-alkyne cycloaddition reaction typical of modern ‘click’ chemistry. Most uses of EU to purify RNA follow the Click-iT Nascent RNA Capture Kit protocol, which involves the use of PEG4 carboxamide-6-azidohexanyl biotin (azide-biotin) with a copper (I) catalyst (generated in situ in the reaction by reduction of copper (II)) to covalently link the EU to biotin [300, 301, 317]. Unlike 4sU, this covalent bond is not easily reversed and generation of cDNA libraries for sequencing or qRT-PCR for direct quantification has to be done while linked to the streptavidin-beads [318]. It is not clear what effect, if any, this has on the error rate of the reverse transcriptase.

It may be possible to take advantage of the ability of very low-salt solutions to cause surprisingly rapid dissociation of the streptavidin-biotin interaction [319] prior to quantitation or sequencing library preparation. To our knowledge, this strategy has not been employed to date in the published literature.

4.5.5 Purification-free detection through enhanced T→C mutation rates

The use of 4sU-containing RNAs for cDNA synthesis results in the reverse transcriptase misincorporating a guanine residue opposite the 4sU at a low level that is exacerbated when cross-linked to protein [292]. Substituting iodoacetamide (IAA) in place of cross-linked protein allows for non-specific enhancement of T→C conversion rates in the reverse transcriptase reaction for all 4sU sites in a library through disulfide bond formation between the IAA and 4sU. T→C mutation rates increase from 10% without IAA to 94% with IAA. SLAM-Seq takes advantage of this increase in mutation rates to quantify mRNA synthesis and decay rates without a purification step. By labeling with 4sU and treating with IAA before library preparation, SLAM-Seq can differentiate labeled RNA from unlabeled RNA strictly through quantification of T→C mutation rates of the final library. Removal of a purification step vastly decreases the amount of input RNA needed and greatly simplifies the mRNA decay protocol [320].

4.5.6 Impact of pulldown efficiency and label incorporation rates on experimental measurements

Two additional parameters that could introduce noise into measurements of RNA decay using metabolic labeling include the incorporation rate of the label into newly synthesized RNA, and the efficiency of pulling down labeled RNA from the total purified RNA. We are unaware of any systematic characterization of the differences in label incorporation between the different metabolic labels discussed in Section 4.4. In some sense, differences in incorporations rates, so long as they are consistent across timepoints, are of no consequence in the experimental designs discussed below since quantification of RNA abundance is either relative to the total amount of RNA pulled down or is normalized by sequencing both the unlabeled and labeled RNAs for each time point (see Section 4.9 for details). However, incorporation rates may be a crucial parameter for measurements of either fast-decaying or slow-decaying RNAs, as they may limit detection. In such cases, optimization of the amount of label added to the cells, incubation times with label, and/or choice of time points may allow for detection of difficult transcripts. As with incorporation rates, a systematic comparison of pulldown efficiencies between different labels and selection strategies has also not been performed. In a typical RNA decay experiment, differences in pulldown efficiency within a single experiment will be controlled for through the use of spike-ins or internal normalization (as discussed in Section 4.9), thereby largely eliminating pulldown efficiencies as a major source of experimental error as long as saturation is not reached. However, improvements in pulldown efficiency can result in less needed biological material for a given experiment. Furthermore, many of the computational methods used to analyze RNA decay experiments operate under the implicit assumption that the sequenced pool of labeled RNA contains no contaminating unlabeled RNA, which may not be accurate to actual experimental conditions but will be closer approximated with better pulldown efficiencies. As discussed above, some improvements have been made to biotin based pulldown strategies for experiments using 4sU as a label through changing the identity of the chemical crosslinker [310]. Additionally, the use of mutation rates induced by the metabolic label removes the need for a pulldown step but introduces a separate source of experimental error related to modification efficiencies of the label itself and misincorporation rates of the reverse

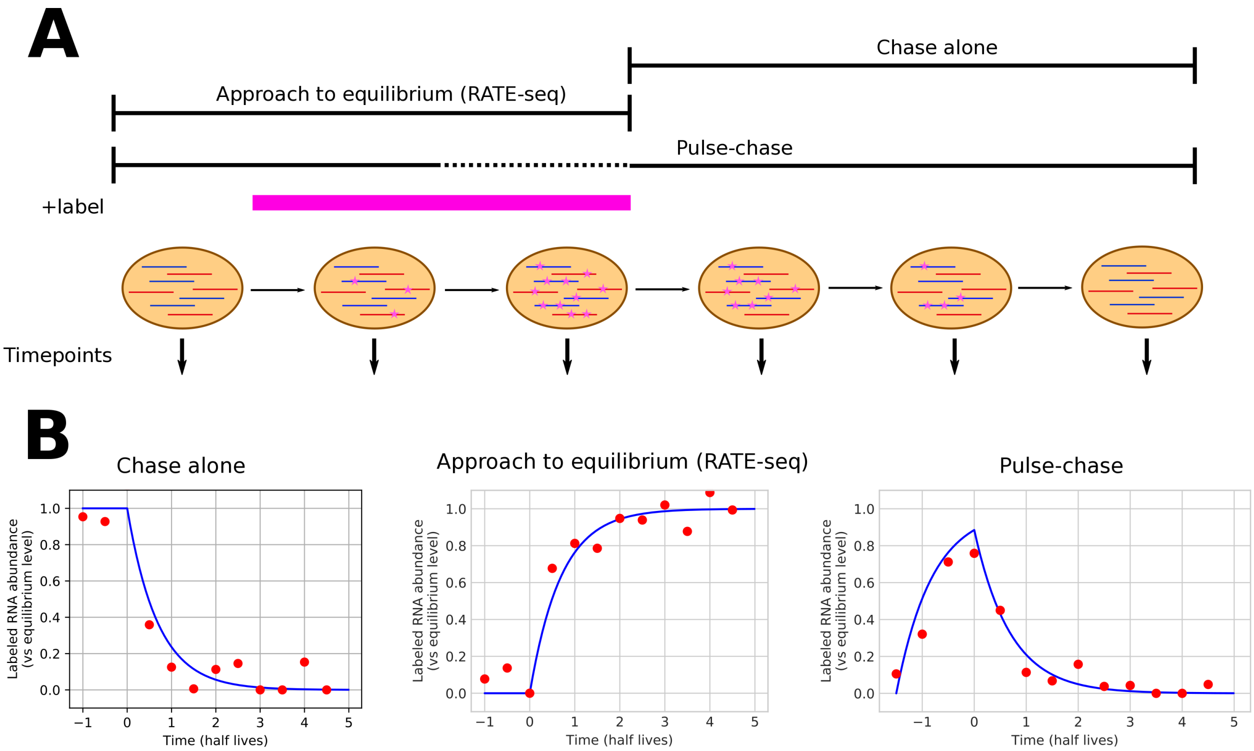


Figure 4.3: Overview of different metabolic labeling time strategies, as discussed in detail in the text. **A)** Schematic of the timing of labeling and sample harvest for three different methods; n.b. labeling in a pulse-chase experiment is typically too short for equilibrium levels to be reached. The pink bar (+label) indicates the time period during which labeled nucleotide is present. **B)** Expected abundance curves (blue) and hypothetical experimental data (red) for the fractional abundance of labeled transcript for any particular RNA under each experimental procedure shown in panel **A**. Time is relative to a zero point at the time of labeled nucleotide removal/washout (chase-alone and pulse-chase) or addition (RATE-seq).

transcriptase [320].

4.6 Experimental Design for measuring RNA decay

With a label and purification method in hand, an experimental design must be chosen that maximizes the amount of information to be gained per unit cost. Different considerations must be made if both synthesis rates and decay rates are to be determined. Additionally, it is critical to decide whether precise RNA half-lives are to be measured or if end-point abundance estimations are sufficient for the biological question of interest. A comparison of different experimental designs frequently used for the determination of RNA decay is shown in Figure 4.3.

4.6.1 Chase alone

To determine RNA decay alone, cells can be grown for an extended period of time, often 24 hours, in the presence of a label. At time zero, the growth media is replaced with identical media containing the same concentration of unlabeled uridine and the labeled RNAs are tracked via purification and sequencing. If determining RNA half-lives, several time points are taken and used for fitting a single exponential decay model [316]. For a more coarse-grained determination of decay, a single time point can be taken after the switch to unlabeled media and compared to a sample taken at time zero. There are major trade offs to consider between these two approaches. By taking only two time points, one drastically cuts down on the costs of sequencing and the labor to prepare the samples. This can be particularly useful when comparing the difference in RNA decay between two biological conditions where the exact half-life is not as useful as as the relative change in decay between the two conditions is. On the other hand, taking several time points allows one to capture both short-lived and long-lived transcripts that may be missed with a single time point. In cultured mammalian cells, the average mRNA half-life is 7–9 hours [282, 321, 322] and it is critical to choose time points that capture the decay of mRNA transcripts of interest. Furthermore, many time points are needed to accurately fit the exponential models used for half-life determination. Thus, selection of the duration and number of time points to be analyzed typically needs to be optimized (left side of Figure 4.3B).

4.6.2 Approach to equilibrium

The converse of the chase-alone experimental methodology, approach to equilibrium, allows for RNA decay rates to be determined from measuring time points after the addition of the labeled uridine to the media. Although cells harvested after a short incubation time can be used to measure transcription rates [323], taking several time points over an extended time course in the presence of the labeled uridine can allow for the mRNA decay rates to be determined instead. The biological motivation behind approach to equilibrium is the concern that labeled nucleotides can be recycled within a cell leading to an ineffective chase with unlabeled nucleotides [324]. To see the quantitative motivation for the approach to equilibrium method, it is useful to consider the overall dynamics of a given transcript. Assuming a constant rate of transcription, the concentration of any particular RNA species, X, will generally follow the equation

$$[X]' = \tau - \delta[X] - \gamma[X] \tag{4.1}$$

Here, τ represents the rate of transcription under the condition of interest, δ is the decay rate of the RNA (typically the quantity of interest), and γ is a dilution term dependent on the growth rate of the cell (if not explicitly accounted for, dilution effects will be incorporated into the inferred value of δ, which for the slow-growing cells of higher eukaryotes is typically a negligible correction)[325]. If one considers the labeled form of an RNA of interest as a separate species, $X*$, then Eq. 4.1 will likewise be followed for the labeled species, except that the synthesis rate will be proportional to τ when the label is present, and equal to zero when the label is not. As the steady state level is defined by the point at which the synthesis and decay rates are perfectly balanced, the steady state concentration requires $[X]' = 0$, or $\tau = (\delta + \gamma)[X]_{eq}$. From this equation it immediately follows that knowledge of any two of the equilibrium concentration, overall decay rate, and synthesis rate are sufficient to specify the third.

By growing cells in a constant amount of label, the fraction of each RNA that is labeled will increase at a rate that is determined only by its degradation rate and the growth rate of the cells until it reaches a steady state level [105]. By measuring time points along this increase, one can capture the decay rate of any given RNA molecule [309], as the equilibrium value will be known from a very late time point and a curve fit can then reveal the decay parameters (see Figure 4.5B and Figure 4.3B (middle)). However, approach to equilibrium requires cells to grow in the presence of the label for an extended period of time, which may be problematic for labels that have demonstrated toxicity under longer exposure, such as 4sU.

4.6.3 Pulse-chase

It is often advantageous to determine both the synthesis and decay rates of an RNA molecule within the bounds of a single experiment. By incubating with a short "pulse" of label and "chasing" with unlabeled media one can both minimize exposure of the cells to the label and determine both synthesis and decay rates separately [277, 282, 308, 326, 327]. Through taking time points at the initial addition of the label, the switch to unlabeled media, and throughout the "chase" period, the lifespan of all nascent labeled RNAs can be tracked (Figure 4.3B (right)). Pulse-chase methods have the advantage of subjecting the cells to short-exposures of the label thereby mitigating any potential toxicity.

4.7 Quantification of RNA abundance

Although specialized DNA microarrays have been used previously [280, 321, 322], global analysis of RNA decay is more recently measured through the use of high throughput sequencing and well-established bioinformatics tools are used to analyze the resulting sequencing reads. Library preparation for RNA sequencing experiments is available through several commercial kits or custom methods that are specific to the experiment of interest. As a general rule, paired-end and

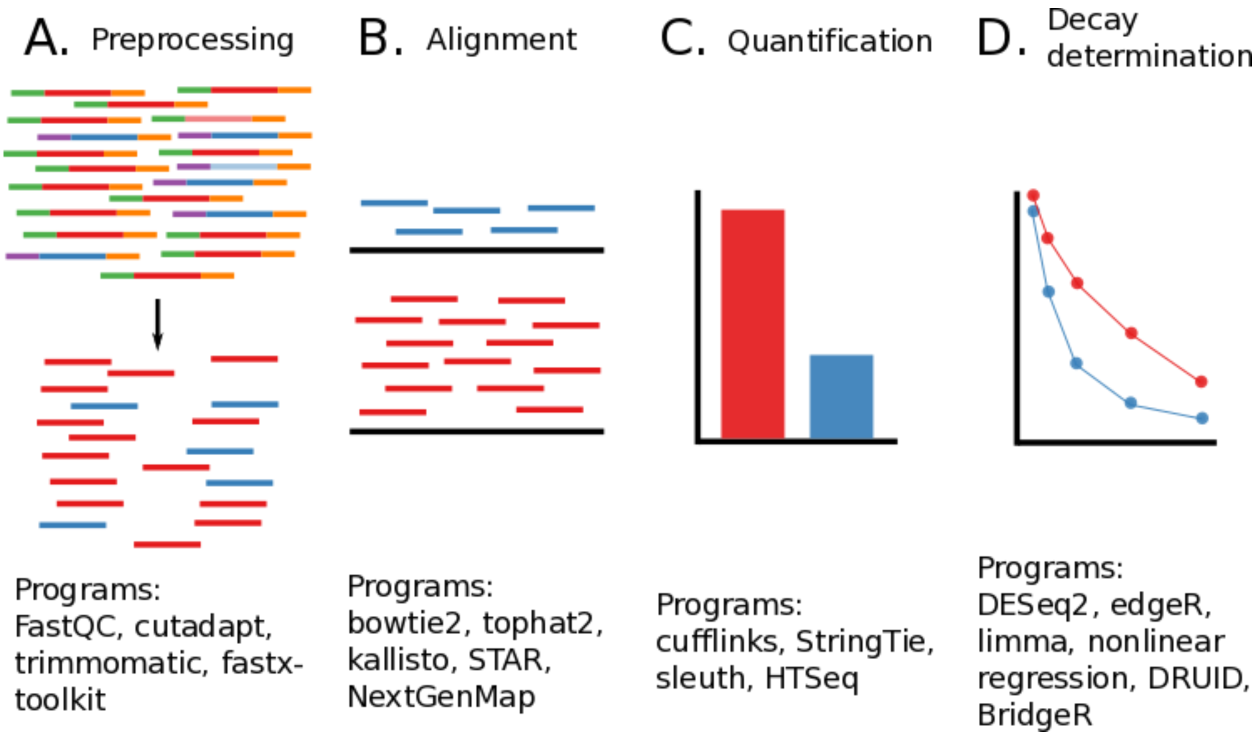

A. Preprocessing
B. Alignment
C. Quantification
D. Decay determination
Programs:
FastQC, cutadapt, trimmomatic, fastx-toolkit
Programs:
bowtie2, tophat2, kallisto, STAR, NextGenMap
Programs:
cufflinks, StringTie, sleuth, HTSeq
Programs:
DESeq2, edgeR, limma, nonlinear regression, DRUID, BridgeR

Figure 4.4: Overall diagram of data analysis steps needed to process high throughput sequencing reads from RNA decay experiments. Widely used example software packages are noted underneath each step. **A)** Preprocessing and quality control, here adapters and low quality reads are removed from analysis. **B)** Alignment of reads to a reference genome or transcriptome. Several key considerations are highlighted in the text below. **C)** Quantification of each transcript or feature of interest. Several different programs can be used to convert alignment information into a measure of RNA abundance that is comparable between experiments. **D)** Modeling of RNA decay. Many different models can be used to determine the decay rates of each transcript of interest.

stranded sequencing is preferred, particularly for organisms that perform splicing or have transcripts regulated by antisense RNAs. Additionally, several strategies exist to remove highly abundant ribosomal RNAs (rRNA) from samples prior to library preparation, including rRNA depletion with custom oligos or selection of poly-adenylated mRNAs. Because poly(A) metabolism plays an important role in mRNA decay pathways, it is advisable to avoid poly(A) selection when analyzing mRNA decay kinetics. [269, 270] After sequencing, several data processing steps must occur to take the raw sequencing reads to a measurement of RNA abundance. Many of the programs and tools written for the analysis of high throughput sequencing data are driven by a text interface, so it is expected that users have some familiarity with the Unix command line. Many institutions have workshops designed to teach new users both familiarity and comfort with the command line, and readers who feel uncomfortable working with command line programs can find help both online and locally. For most applications, the RNA sequencing reads obtained from the methods above can be treated like data from any other RNA-seq experiment. Typically, sequencing reads are stored in the fastq file format where both sequence and base-calling quality information can be stored. Here we will briefly outline the set of steps needed to analyze RNA-sequencing data for RNA decay experiments with extra commentary on possible locations in the analysis that may differ for RNA decay-type experiments as compared to standard RNA-seq workflows. For more information on best practices concerning RNA-seq data we point the reader to recent reviews in the literature [328, 329].

4.7.1 Adapter removal and quality control

As with any sequencing analysis, standard quality control must be employed. Several steps must be taken to remove adapters needed for Illumina sequencing as well as reads containing low confidence base calls. For the removal of adapters and low quality sequences, several programs exist including cutadapt [177], fastx toolkit [http://hannonlab.cshl.edu/fastx_toolkit/], and trimmomatic [178]. Several key statistics about the quality of the sequencing reads can be calculated both before and after adapter and quality trimming using FastQC [179] (Figure 4.4A). Next the reads must be aligned to a reference transcriptome which is available from either NCBI or the UCSC Genome Browser for most model organisms. Several different aligners have been developed for processing RNA sequencing reads including bowtie2 [182], tophat2 [330], STAR [331], kallisto [190], and many others. A comparison of the most commonly used aligners indicates tradeoffs between each tool and the specific aligner used will depend on the question being asked [332]. However, if one is using the SLAM-seq methodology that is dependent on T→C mutations then it is recommended to use the T→C mutation aware aligner NextGenMap [333] with special settings designed to weaken the penalty for mismatches resulting from a T→C mutation event [320] (Figure 4.4B).

4.7.2 Reference-based alignment, transcriptome assembly or pseudoalignment?

Several additional considerations need to be made when choosing both the aligner and the downstream quantification software for processing the data from a high-throughput RNA decay experiment. For single-celled organisms such as bacteria or archaea where a high quality reference transcriptome is known for the organism and that organism does not process RNAs through splicing, a simple aligner such as bowtie2 will perform well. However, most higher eukaryotes do process RNAs through splicing and thus splice-aware aligners, such as hisat2 [334] and STAR, are recommended. Under some biological conditions, novel transcripts are may be expected and have not yet been characterized and logged in the reference transcriptome of the organism under study. Here, downstream software will be needed to infer the presence of novel transcripts and assemble a transcriptome either *de novo* or through assistance of an existing reference transcriptome. However, many experiments are not designed to look for new transcripts and are instead concerned with the abundance of well-characterized transcripts annotated in a reference transcriptome. Pseudoaligners such as kallisto [190] and salmon [335] are designed to deal efficiently with this latter case. Rather than do a full alignment, pseudoaligners allow for RNA quantification without needed to fully align the reads to the reference transcriptome. Pseudoaligners have the advantage of being substantially faster than traditional alignment methods, but will not be able to detect any novel transcripts and are wholly reliant on the quality of the reference transcriptome. Unlike the pseudoaligners, most major aligners will output a sequence alignment map (SAM) file or its binary equivalent (BAM) that contains several details of where a particular sequence aligned and the quality of that alignment. Key statistics and simple manipulations of this file format can be obtained using samtools [183]. After alignment, downstream tools are needed to convert the sequence alignment information into some form of quantification of RNA abundance. The most commonly used software suite that performs this quantification is cufflinks [336] however, StringTie has shown better performance than cufflinks and is currently recommended as a replacement [337]. Both cufflinks and StringTie (as well as other related tools) perform novel transcript discovery and transcriptome assembly, which is useful under conditions where new transcripts are expected and informative but is not always necessary. If transcriptome assembly is not needed due to the existence of an already annotated, high quality reference transcriptome, or if the investigators biological question is not concerned with novel transcripts, then a simple feature level quantification can be obtained using HTSeq [338] instead (Figure 4.4C).

4.7.3 Gene level or exon level?

Another key consideration when quantifying data from RNA decay experiments is to determine whether to quantify at the gene level (where all reads for a gene are pooled together regardless of transcript isoforms) or exon level (where each exon is quantified separately). Most reports for determining RNA decay have focused on gene level quantification, but exon level information may be needed if one is tracking decay of specific transcript isoforms.

4.7.4 Count level or TPM?

When considering differences between two experimental conditions, another major consideration to make is how to quantify the amount of change in RNA decay between the two conditions. Without proper statistical analysis, differences in sequencing depth, the efficiency of labeled RNA recovery, and biological variability between replicates can confound any true biological difference that is being measured. Fragments Per Kilobase per Million (FPKM) or Reads Per Kilobase per Million (RPKM) are two measures that were designed to correct for both sequencing depth and transcript length bias between different samples and genes (or exons). However, the Transcripts Per Million (TPM) unit has superseded RPKM and FPKM as the preferred value for reporting RNA expression, since TPM values can more accurately be directly compared between experiments [339]. TPM is commonly reported as measure of relative RNA abundance under a particular experimental condition for a feature of interest, but more sophisticated statistical models have arisen that better account for the biological variability seen in the quantification of RNA-seq data. The use of negative binomial models based on count-level data instead of FPKM or TPM for each feature of interest allow for better estimation of biological variability and thus more accurate and reproducible results. Negative binomial models are implemented in all of the major differential expression packages currently used in RNA-seq analysis and are applicable to RNA decay analysis. Some of the key differential expression software packages include DESeq2 [340], edgeR [341], limma [342], cufflinks [336], and StringTie [337]. These packages will take count-level data for each feature (at the gene or exon level) of interest and use negative binomial-based statistical models to properly account for variability between conditions. Additionally, the kallisto pseudo-aligner has a downstream package, sleuth [191] designed specifically for use with kallisto, and uses the same general principles as the packages mentioned above (Figure 4.4D).

4.8 Modeling RNA Decay

The ultimate goal for most RNA decay experiments is to quantitatively measure the kinetics of RNA abundance over time. For some research questions, a measure of relative changes in RNA decay between two conditions or two transcripts may suffice. However, for another subset of research questions, the determination of a quantitative rate constant with meaningful units is the object of interest. A careful consideration of normalization procedures for measurements of RNA abundance using high-throughput sequencing techniques is essential for this latter class of experiments (and still useful for the former class), as discussed in Section 4.9. However, a discussion of the theory that underlies models used for the determination of RNA decay as applied to perfect measurements of RNA abundance and discussed below is, nevertheless, instructive.

4.8.1 Single exponential decay

Guided by historical transcription shutoff experiments, most chase experimental designs use a single exponential equation to determine RNA decay half-lives. A single exponential model assumes that RNA decays at a rate proportional to its instantaneous concentration over the measurement time of the experiment:

$$\frac{A_i(t)}{A_i(t_0)} = e^{(-\alpha_i t)} \tag{4.2}$$

Where $\frac{A_i(t)}{A_i(t_0)}$ is the relative abundance for labeled RNA i at time t as compared to time t_0: the initial time point taken when the labeled RNA has reached equilibrium. Here α_i represents the constant decay rate for RNA i. Note that the exponential form for RNA abundance is obtained directly from integration of Eq. 4.1 with the production term set to zero and growth term omitted. Thus the half-life of the RNA can be determined by fitting the data with the following equation (Figure 4.5A):

$$T_{\frac{1}{2}} = \frac{ln(2)}{\alpha_i} \tag{4.3}$$

It is important to note that both the approach given here, and the more sophisticated variations below, work under the assumption that the fitted parameters (*e.g.*, decay rate) do not vary throughout the experimental time course. An additional modification for the half-life determination to account for dilution due to cell growth has also been suggested by several groups. [278, 309]:

$$T_{\frac{1}{2}} = \frac{ln(2)}{\alpha_i - k_{growth}} \tag{4.4}$$

Where k_{growth} is the same for all RNAs and is determined by the growth rate of the culture; again this equation arises directly from the presumed time-dependent change in RNA abundance stated in Eq. 4.1. Note that in the context of Eq. 4.4 the "half life" so calculated yields a half life for the individual RNA molecules themselves, rather than the bulk half life that would be observed for a

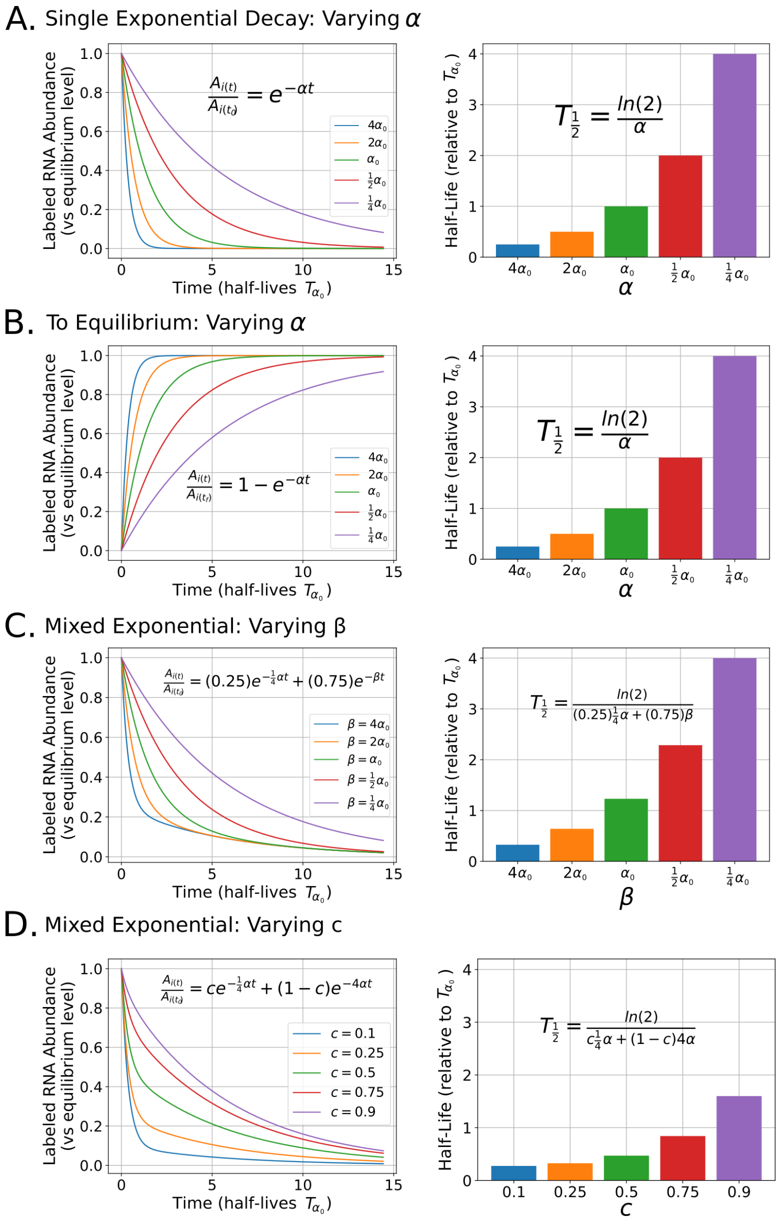


Figure 4.5: Impact of varying parameters in the equations used for modeling RNA decay: For each panel, the left graph represents the shapes of several arbitrary exponential curves after perturbing a single parameter in each model. Here, the y-axis represents labeled RNA abundance relative to the equilibrium level of labeled RNA in label-containing media. The x-axis represents time in the number of half-lives for a single exponential curve with a decay rate of α_0, indicated with T_{α_0}. The right graph in each panel represents half-lives calculated from each of the curves from the corresponding left graph in the same panel and relative to T_{α_0}. In each graph, the equation used to either model the decay or determine the half-life is displayed. **A)** The effect of varying α on relative RNA abundance (left) and half-life (right) when modeling RNA decay with a single exponential. **B)** The effect of varying α when modeling RNA decay in a to equilibrium experimental design. **C)** The effect of varying β with a fixed c and a fixed α when modeling RNA decay with a two component mixed exponential. **D)** The effect of varying c with a fixed α and β when modeling RNA decay with a two component mixed exponential.

population of molecules (the latter ought to include dilution effects, while the former should not).

4.8.2 Mixed exponential decay

Imamachi et al. [316] have noted that a subset of RNAs do not decay in a manner that is easily described by a single exponential and have suggested fitting the data with a model that considers a mixed population of RNAs with different decay rates:

$$\frac{A_i(t)}{A_i(t_i)} = (c)e^{-\alpha_i t} + (1-c)e^{-\beta_i t} \tag{4.5}$$

Where c indicates a weight for one subpopulation vs. the other subpopulation and β_i is the decay rate for a second population for a particular RNA. In principle, even more complex functional forms could be considered, such as adding additional exponential terms or using a stretched exponential, which might better account for data where multiple subpopulations decayed on different timescales. Precisely such a situation might easily emerge if multiple different subpopulations of cells were present in the measurements, or if gene-level quantification was in use but multiple transcript isoforms existed with differing stabilities. Using more complicated models can be prone to overfitting and appropriate model selection criteria [343] must be made when choosing between models with more or fewer parameters (Figure 4.5C-D).

4.8.3 Approach to equilibrium

For approach to equilibrium experimental designs, several assumptions and considerations must be made to properly model the RNA half-lives. Neymotin et al. determine RNA half-lives by considering the decay of unlabeled RNAs and also taking into account the cell growth rates [309]. They ultimately model the abundance of any given labeled RNA at time t as the following:

$$\frac{A_i(t)}{A_i(t_f)} = (1 - e^{-(\alpha_i + k_{growth})(t-t_d)}) \tag{4.6}$$

Where t_f is the final time point where the labeled RNA has reached steady state levels (at the end of the time course) and t_d is the time between the addition of the label and the first measurement of labeled RNA. Both the overall α, which is equal to the $\alpha_i - k_{growth}$, and the Y_{eq} for each RNA can then be estimated from the experimental data, here assuming the t_d is fixed for all RNAs based on experimental measurements for when RNA first appears in after label selection (Figure 4.5B). Half-lives can then be calculated as above using the growth rate-corrected half-life formula above (Eq. 4.4). DRUID, an automated pipeline for approach-to-equilibrium experiments, has been developed to deal to help analyze data from this type of experimental design without the need to have complicated spike-ins or sophisticated ways to deal with normalization [307].

4.8.4 Pulse-chase considerations

Pulse-chase experimental designs have the advantage of allowing the experimenter to separately determine both transcription rates and decay rates from a single experiment. By comparing labeled RNA abundances to unlabeled RNAs (or labeled RNAs between two different experimental conditions) at the beginning of the chase, one can have a general idea of nascent transcription rates [320] or condition-specific effects on transcription [308]. By taking several time points throughout the chase part of the experiment, one can use the single exponential equations described above to fit half-lives of each RNA of interest. Alternatively, one can take a single time point after a chase and measure differences between labeled RNA abundances in two different experimental conditions or between the start and end of the chase to determine relative decay rates without determining the half-life of the RNAs. It is worth noting that many of the methods mentioned above have focused on following a single species: the labeled RNAs. Whether through following the decay of the labeled species over time (chase and pulse-chase), or through measuring the decay of the unlabeled species over time indirectly by measuring the approach to equilibrium of the labeled species, these methods allow for accurate determinations of mRNA decay rates. However, valuable information that may also be gained by sequencing both the labeled and unlabeled pools. More sophisticated methods that take into account both transcription rates, RNA decay rates and measurement of both the pool of unlabeled RNAs and the pool of labeled RNAs have also been reported throughout the literature [281, 282, 344] and can give deeper insight into the full kinetics of individual mRNA transcripts but are outside the scope of this particular review.

4.8.5 Half-lives vs. differential abundance

The exponential equations above are typically fit using non-linear least squares methods to determine α_i by minimizing the squared sum of the errors between the model and the data for each RNA. Although half-lives can be determined with as few as three time points, it has been recommended to use at least 5 time points [345] in order to accurately determine half-lives. The Akimitsu lab has developed a custom R package [https://github.com/AkimitsuLab/BridgeR] for determining the difference in mRNA half-lives between two different conditions of interest. However, determining the RNA half lives for many different replicates and experimental conditions can be incredibly costly due to the amount of sequencing samples needed in order to properly fit the exponential equations. Instead of determining full half-lives for every RNA of interest, one could consider capturing an initial and final time point and using differential expression software to measure the impact of a particular condition on the relative abundance of RNA in the final time point compared to the initial time point. Simple models designed to measure the condition-specific effects on RNA abundance can be specified easily in differential expression analysis software such as DEseq2 [340], which at least permits determination of whether or not the decay of a particular transcript changes between a pair of conditions.

4.9 Normalization and the use of spike-ins for estimation of labeled RNA abundance

It is important to note that high-throughput sequencing reactions only give relative abundance measurements of RNA. Any comparison of two separate RNA sequencing reactions will thus require some sort of normalization in order to put RNA abundance estimations on the same numerical scale relative to one another. The most commonly reported normalization schemes for RNA-seq type experiments include RPKM and TPM, which act to normalize the count data obtained from a typical RNA seq workflow to both the length of the genomic feature of interest as well as the sequencing depth for that particular sample as discussed above. As most metabolic labeling experiments described here involve a pulldown step, however, the normalization provided by TPM-type measurements is insufficient, because the resulting abundance measurements are still only known relative to the total set of labeled RNA. Comparison of different time points, essential for calculation of RNA stability, is thus impossible without some sort of normalization that allows for proper scaling of the observed abundances relative to the total (and not only labeled) RNA present in the sample.

4.9.1 Rationale for the use of spike-ins

To more clearly demonstrate the necessity and utility of a constant reference value for normalization of RNA abundance, we must consider what is actually being measured when one performs an RNA decay experiment where only the labeled RNA is sequenced. Let us represent the abundance of labeled RNA for any given transcript i as $X_{i,L}(t)$ and the corresponding abundance of unlabeled RNA in the same experiment as $X_{i,U}(t)$. We can then consider the entire abundance of labeled RNA for all genes at any given time point t to be:

$$\Gamma(t) = \sum_j X_{j,L}(t) \tag{4.7}$$

Likewise, the entire abundance of unlabeled RNA for all genes can be represented as:

$$\beta(t) = \sum_j X_{j,U}(t) \tag{4.8}$$

Thus the total RNA abundance A is simply:

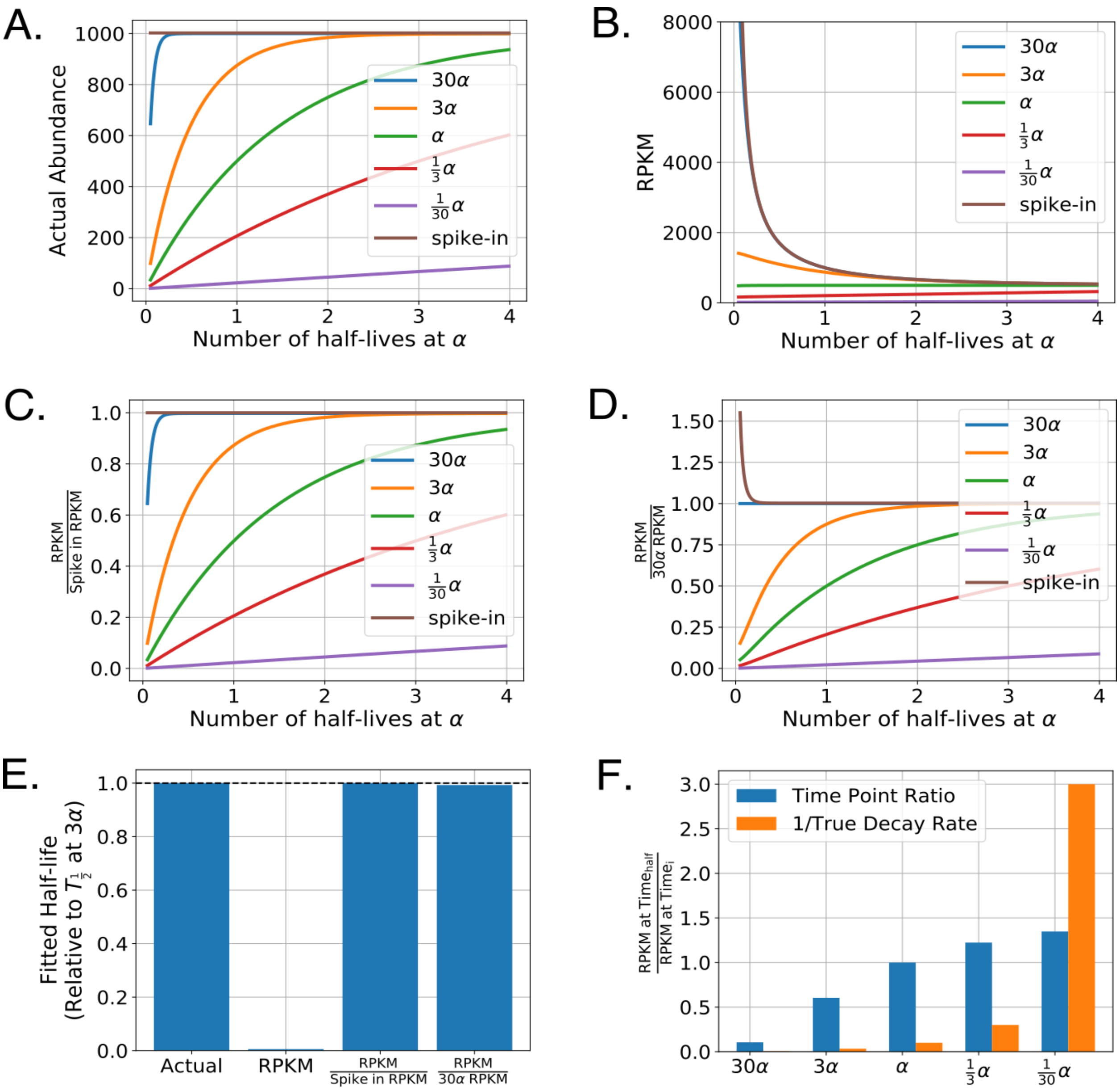


Figure 4.6: Impact of common normalization procedures on the determination of RNA half-lives. **A)** Simulated labeled mRNA counts for several transcripts decaying at the indicated rates in an approach to equilibrium experiment. For this simulation, bulk RNA (not plotted) decayed at a rate of α and represented 99% of the total RNA sample. Spike-ins were added at 0.5% of the total RNA (that is, sum of labeled and unlabeled). **B)** Raw RPKM values for each transcript and spike-in RNA. For simplicity, each simulated time point was sequenced to the same depth of 5,000,000 reads and each transcript and spike-in RNA was considered to be the same exact length. Time is indicated in number of half-lives of the bulk RNA, which decays at a rate of α. **C)** As in **B** but RPKM values are normalized to the spike-in RPKM values for each sample. **D)** As in **B**-**C** but RPKM values are normalized to a transcript that decays at a rate of 30α. **E)** Calculated half-lives for the transcript with a decay rate of α. Each half-life was determined by fitting the the approach to equilibrium equation indicated in Figure 4.5B using non-linear least squares on five evenly spaced time points from the indicated simulated traces in panels **A**-**D**. **F)** Relative mRNA decay as determined by the change in raw RPKM from two time points. $Time_{half}$ was chosen to be the time point at exactly one half-life for the bulk RNA. For comparison, orange bars represent the inverse decay rate for each of the indicated transcripts.

$$A(t) = \Gamma(t) + \beta(t) \tag{4.9}$$

And what is actually measured for any given gene in an RNA decay experiment when only the labeled RNA pool is sequenced is:

$$R_i(t) = \frac{X_{i,L}(t)}{\Gamma(t)} \tag{4.10}$$

Where $R_i(t)$ is the relative abundance of RNA i in the total labeled RNA pool at time point t. Since both $\Gamma(t)$ and $X_{i,L}$ are changing throughout the course of an experiment, attempting to fit the RNA decay equations we have described here to raw RPKM measurements is not physically meaningful. However, if one is able to change the variable quantity $\Gamma(t)$ in the denominator of Eq. 4.10 into something that is known to be constant throughout the experiment, then $R(t)$ can be transformed into a reliable estimator of RNA abundance on an arbitrary scale. One approach to add a constant to any RNA decay experiment is to add a labeled spike-in RNA at a known ratio $1/d$ of labeled spike-in to total RNA and normalize RPKM values to that of the measured RPKM of the spike-in. Thus, the spike-in will be added at some constant value $S(t)$ that is a function of $A(t)$:

$$S(t) = \frac{A(t)}{d} \tag{4.11}$$

Now we can modify Eq. 4.10 to include a known constant amount of spiked-in label S:

$$R_i(t) = \frac{X_{i,L}(t)}{\Gamma(t) + S(t)} \tag{4.12}$$

Likewise, we can represent the relative abundance of the spike in by $R_s(t)$:

$$R_s(t) = \frac{S(t)}{\Gamma(t) + S(t)} \tag{4.13}$$

By normalizing the fractional abundance of labeled RNA in the total labeled pool with the spike-in $R_i(t)$ (Eq. 4.12) to the relative abundance of the spike-in $R_s(t)$ (Eq. 4.13) we can see that the denominator $\Gamma(t) + S(t)$ will cancel, resulting in a spike-in normalized estimation of labeled RNA abundance, $N_i(t)$.

$$N_i(t) = \frac{X_{i,L}(t)}{S(t)} = \frac{d \cdot X_{i,L}(t)}{A(t)} \tag{4.14}$$

Furthermore, upon substitution of $S(t)$ with Eq. 4.11 we can see that $N_i(t)$ in Eq. 4.14 is a reliable estimator for the fractional abundance of labeled RNA i in the total RNA $A(t)$ rather than just the labeled RNA pool $\Gamma(t)$, clearly demonstrating the need for normalization to some source of constant labeled RNA when determining RNA decay rates and half-lives. It is important to note, that any error in the fraction of added spike-in $1/d$ will add additional noise to the normalized RNA abundance estimate described by Eq. 4.14.

To further illustrate the impact of normalization on the determination of RNA decay rates, we simulated an approach to equilibrium experiment where the bulk RNA representing 99% of the sequences decayed at a rate of α. We then considered several different RNA transcripts each at the same steady state level of overall abundance but at several different multiples of the overall bulk RNA decay rate. The actual abundances of labeled RNA from this simulation can be seen in Figure 4.6A. We then added in a spike-in RNA at 0.5% of the total RNA for each time point and determined what the resulting RPKM values for each of these transcripts would be under a scenario where labeled RNA is pulled-down with perfect efficiency (Figure 4.6B). In this simulation, it is evident that the raw RPKM values do not represent the actual RNA abundances. Note that the spike-in RNA rapidly decays in RPKM abundance throughout the time course even though it is added at a constant amount relative to the total RNA. This is to be expected, as in early time points the spiked-in RNA represents the only labeled RNA species in the reaction. As more labeled RNA is created in the cells, the relative fraction of spike-in RNA drops precipitously. However, if we normalize the RPKM traces to the spike-in RPKM (Figure 4.6C) the actual RNA abundances are exactly reproduced, demonstrating both the utility and necessity of a constant reference value in RNA decay experiments. Therefore, it should not be surprising that many groups advocate for the use of labeled RNA spike-ins when determining RNA half-lives across a variety of model organisms [285, 307, 309, 310, 316, 346].

4.9.2 Practical use of labeled spike-ins for RNA decay experiments

For standard quantification of RNA in RNA-seq experiments, a set of agreed upon standards have been adopted and maintained by the External RNA Controls Consortium (ERCC) [347–349]. Furthermore, spike-ins are seeing widespread use throughout most high-throughput sequencing technologies (reviewed in [350]). However, unlike RNA-seq, no agreed upon set of labeled RNA spike-in standards have been established for RNA decay experiments and the ERCC collection is not available in labeled form. Instead, each lab has developed their own set of standards to use as spike-ins for their system. Tani et al. have established the use of the exogenous luciferase RNA, in vitro transcribed with a known quantity of label, and added to the total purified RNA directly before label selection [285, 316]. Russo et al. use an expensive synthetic labeled positive control that is not reliant upon the efficiency of labeling within an in vitro transcription reaction [346]. Neymotin et al. used a combination of three spike-ins with different lengths from a different organism but with matched GC content to their organism of interest [309]. Likewise, Duffy et al.

also use a mix of RNAs from a different organism as a spike in [310]. Finally, Lugowski et al. use two sets of spike-ins, a labeled spike-in of whole genome reads from one organism and an unlabeled spike-in of whole genome reads from a second organism, where both spike-in species originate from organisms that are sufficiently different from the organism of interest [307]. There are several advantages and disadvantages to each of the approaches used above. Spike-ins labeled by in vitro transcription are much cheaper than buying synthetic spike-ins but are also sensitive to variations in the in vitro transcription reaction itself. To mitigate this effect experiments using in vitro transcription to create labeled spike-in RNAs should use RNAs from the same transcription reaction for all samples that are to be compared. From the ERCC experiments, it is evident that sequence bias can have a major impact on measurements from high throughput sequencing experiments [351]. Thus, the use of a single spike-in may not be sufficient for precise measurements of mRNA half-lives. The use of whole-genome labeled RNAs from a non-target organism may help alleviate some of these concerns since a variety of length distributions and sequence compositions are present from those samples, but mismatches between sequence bias in two different organisms can add additional source of noise to the experiment. Additionally, any spike-in is particularly subject to pipetting errors as any mis-quantification of the precise amount of spike-in added to a reaction will add a considerable amount of noise to the quantification procedure, as the spike-in provides the sole normalizing factor for recovering proper decay rates (Eq. 4.14).

4.9.3 Spike-in free approaches

Despite the clear utility of a spike-in in estimating RNA abundance, several groups have found additional ways to accurately estimate RNA abundance in RNA decay experiments without using a spike-in RNA [280, 282, 307, 308, 320]. Both Dolken et al. and Schwanhausser et al. use a procedure in which they determine the abundance of both the labeled and unlabeled RNA species by sequencing both the selected labeled RNAs and the unlabeled RNAs found in the unbound fraction, which allows them to determine absolute RNA abundance and decay rates for each transcript, albeit at greater cost than a typical RNA decay experiment [280, 282]. Similarly, Herzog et al. have the ability to measure both labeled and unlabeled pools of RNA abundance with a single sequencing reaction since their method relies on the determination of T to C mutations to determine labeled RNAs and they are able to internally normalize to the total abundance of RNA through this method [320]. Lugowski et al. developed an entirely new pipeline (DRUID) that uses rapidly decaying RNA introns as a constant internal normalization in approach to equilibrium experiments, which they found to be superior to the spike-in based normalization that they attempted in parallel [307]. To illustrate how the DRUID procedure works, we simulated normalization to a rapidly decaying transcript in an approach to equilibrium experiment (Figure 4.6D). Here it is evident that the rapidly decaying transcript approaches a constant labeled value quickly within the experimental procedure and can be used, instead of a constant spike-in, to normalize the RPKM abundances and recover a true estimator of RNA abundance. We also demonstrate that RNA half-lives determined using the DRUID approach vs. the spike-in approach are able to easily recover the true half-life for a transcript in our simulations (Figure 4.6E). Lugowski et al. directly compare their DRUID approach to a spike-in approach and find that half-lives determined from the DRUID approach have higher replicate-replicate agreement and also outperformed both spike-in normalization and transcription shutoff experiments when compared against a benchmark dataset [307], possibly due to the pipetting error inherent in the use of spike-ins. In theory, a similar approach could be used for pulse-chase and chase-alone experimental set-ups. However, instead of normalizing to a highly unstable transcript, one would need to normalize to an extremely stable transcript (after sufficient labeling time) as has been suggested by some groups [281, 316]. All such internal-reference approaches provide a potentially simpler workflow than spike-in based methods, and avoid concerns such as pipetting and RNA quantitation errors, but necessitate the identification of extremely unstable or stable pieces of RNA that can be relied upon to have far longer or shorter half lives than any transcripts of biological interest.

As a simpler alternative, Paulsen et al. 2014 suggest measuring the labeled RNA species at just two time points, one time point after a short labeling period, and a second time point chosen at the average half-life of RNA in the organism of interest [308]. A comparison between these two time points can then be made using differential expression software to get a semi-quantitative view of RNA decay at a much lower cost. To illustrate this approach we compared the RPKMs of two different time points in our simulation and compared these ratios with the true decay rates of the transcripts (Figure 4.6F). Here, it is clear that the rank ordering of the transcript stabilities is preserved, but no interpretation can be given as far as the magnitude change between each of the transcripts, and any attempts to fit a decay rate using such data would fail even if many timepoints were collected. However, the true utility of this approach can be seen when comparing these relative measurements of RNA decay between two different experimental conditions. We further illustrate this approach with a case study in Section 4.10.

4.10 Interpretation and follow-up

A careful consideration of each aspect of a successful RNA decay experiment can be best described through a sample case study. Consider the scenario where one wants to identify the set of mRNA targets for which RNA decay is primarily mediated by a particular RNA binding protein of interest. To determine possible targets, mRNA decay is measured transcriptome-wide in both mock-treated cells and cells in which the RNA binding protein of interest is knocked down with a silencing RNA.

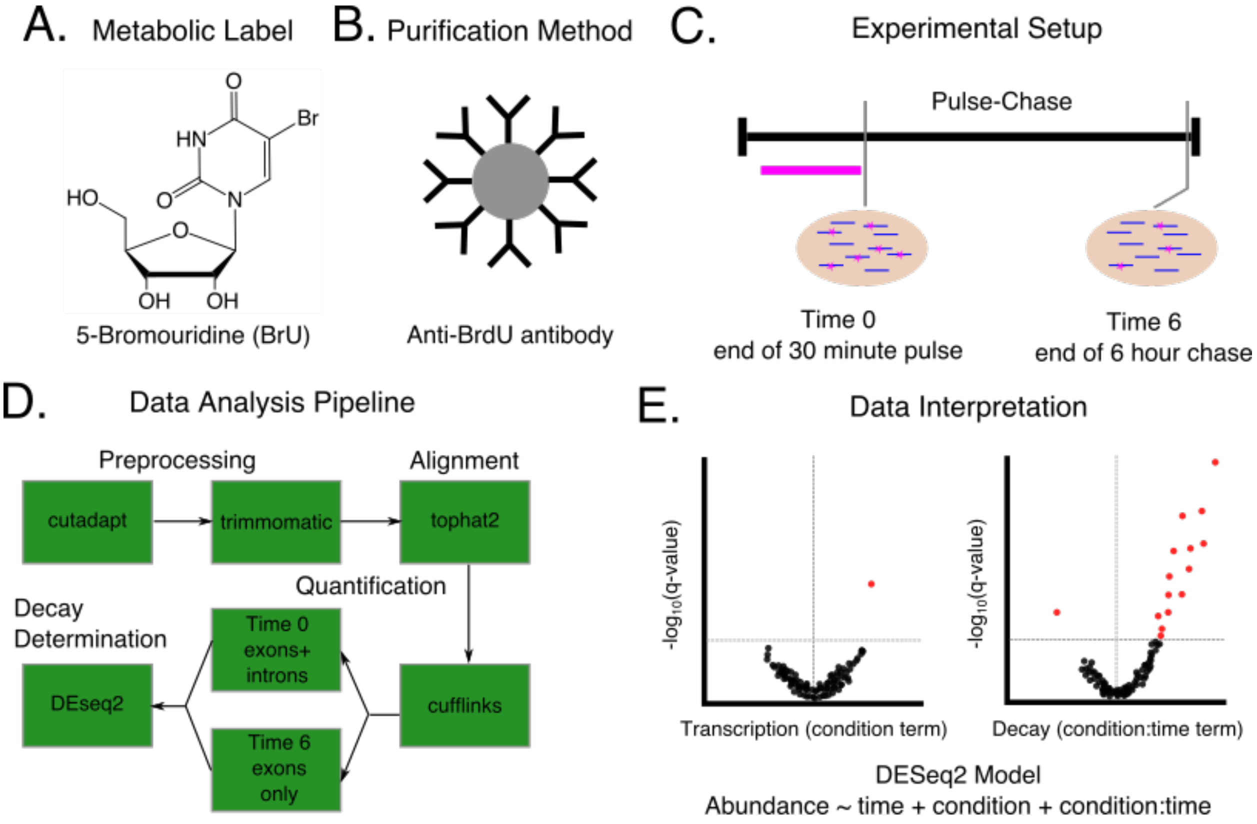


Figure 4.7: A hypothetical experimental design for determining changes in RNA decay between a WT and RBP knockdown condition. **A)-C)**. The choices to be made in designing the experiment including which metabolic label to use (**A**), how to purify labeled RNA (**B**), and the timing of label introduction and sample harvest (**C**). Note that this experimental design is done in parallel for knockdown and control cells. **D)** A sample data analysis pipeline to be used to analyze sequencing results from the experiment described in **A**. Choices must be made at the preprocessing, alignment, quantification, and decay determination stages as indicated in Figure 4.4. **E)** Hypothetical volcano plots to visualize the results from the experiment in **A**-**C** as analyzed by the pipeline in **D**. A generalized linear model is used with DEseq2 to determine knockdown specific changes in both transcription and decay. Red dots indicate significant genes as measured by a FDR corrected p-value < 0.1. The vertical gray line in each plot indicates a $\log_2$ fold change of zero. For this hypothetical experiment, few genes had a significant change in transcription under the knockdown condition (left plot), but many genes were stabilized in the knockdown condition (right plot), suggesting that experiment identified several genes that can be considered putative targets for the RBP of interest and represent good candidates for targeted experimental follow-ups.

Possible targets will include RNAs that have differential mRNA decay in the knockout genotype compared to the wild-type cells. For this case study, we select an experimental procedure designed to minimize both the cost and cellular manipulations needed to conduct the experiment. Given these constraints, BrU is chosen as the labeling reagent for its low toxicity, cost, and the avoidance of any requirement to incorporate a functional UPRT into the human cell line used for the experiment (Figure 4.7A). With BrU, an anti-BrdU antibody with known cross-reactivity to BrU is chosen as the selection reagent (Figure 4.7B). For this particular experiment we are not interested in the exact half-lives of expressed RNAs, but rather the effect of the RNA-binding protein of interest on mRNA decay. Since the RNA binding protein of interest is hypothesized to be involved only in post-transcriptional regulation and not in transcriptional regulation, we are also interested in differentiating transcriptional effects from stability effects. Thus, the pulse-chase experimental design is chosen in order to be able to determine effects on both processes. In this case, we take a single time point at the start of the chase after 30 minutes of labeling and take a second time point at the end of the chase several hours later (Figure 4.7C). The end of the chase was chosen to coincide with the average mRNA half-life in cultured mammalian cells [282, 321, 322]. To assess biological reproducibility, three replicates for each time point and genotype are performed and analyzed. Three replicates were chosen to be consistent with long RNA-seq ENCODE guidelines which suggest that at least two biological replicates should be used to assess biological reproducibility [352]. Furthermore, the ENCODE ChIP-Seq guidelines suggest that more than two replicates are not absolutely necessary as experiments with RNA pol II indicated that more than two replicates did not increase the number of sites discovered [62, 353]. Since RNA decay experiments have aspects in common with both ChIP-Seq (with an immunoprecipitation step) and RNA-seq (with quantification of RNA abundance), elements of both recommendations are likely applicable here. After RNA quantification, replicate agreement among a single time point can be assessed using rank-based statistics, such as Spearman correlation coefficients. However, correlation coefficients between samples at different timepoints are not meaningful as the RNA abundances are expected to decay at different rates throughout the experiment. Major disagreements between replicates at the same time point can indicate a need for more replicates to better assess variability or a need to repeat the experiment and obtain higher quality samples. It is important to note that this experimental design disfavors detection of regulation of mRNAs with very short or very long half lives. After preparing stranded paired-end libraries for each sample and sending them for sequencing, we perform quality control and clean-up of the sequencing reads using a combination of FastQC, cutadapt, and trimmomatic. Since we want to differentiate transcription effects from decay effects of the RNA binding protein on the transcriptome, we choose to use the splice-aware aligner tophat2 and associated analysis suite cufflinks to assign read counts at both the exon and gene level. We follow the recommendation of Paulsen et al. [308] and use full gene level counts (including exons and introns) at the early time point to measure nascent RNA abundance and use the sum of all possible exons (but not introns) at the late time point to measure mature RNA abundance (Figure 4.7D). We then take this count data and use a simple model to determine changes in transcription and stability resulting from the RNA binding protein knock down with DEseq2:

$$A \sim time + condition + condition : time \tag{4.15}$$

Where A is the abundance of any particular transcript, $time$ is a binary term for the time point (start or end of the chase), $condition$ is a binary term for which condition the RNA is in (knockdown or control) and $condition : time$ is an interaction term between the time and knockdown information. Here, the magnitude and direction of the $condition$ term is interpreted as the knockdown effect on RNA abundance after 30 minutes transcription during the pulse. The magnitude and direction of the interaction term $condition : time$ is interpreted as the knockdown effect on the change in RNA abundance from the start to the end of the chase. After false discovery rate correction using the Benjamini-Hochberg procedure [354], several high confidence targets for the RNA binding protein of interest can be identified and followed up with targeted experiments (Figure 4.7E).

4.11 Concluding Remarks

This review provides a general overview of the decisions to be made when planning experiments to globally analyze RNA decay using metabolic labeling coupled with high throughput sequencing. We hope this article will serve as a resource for new and experienced researchers in the field. For additional information, we refer readers to several recent reviews that provide more depth on each topic presented above [345, 355, 356]. With the advent of low cost high-throughput sequencing, measurements of RNA decay at a global scale are broadly achievable. Metabolic labeling of RNA has allowed for the measurement of both transcription rates and decay rates with minimal perturbation of the underlying biology. Recent advances in chemistry have allowed for enhanced selection of labeled RNAs from the pool of total RNA in the cell [310] or removal of the need to select the labeled species from the pool of RNA altogether [320], greatly reducing the amount of starting material needed for these experiments and lowering the overall cost. Additionally, new experimental approaches using the metabolic labels and methods described here have allowed for novel insights into RNA biology including the identification of RNA binding proteins involved in nascent transcription [304], the impact of a single RNA binding protein in amyotrophic lateral sclerosis [327], and the discovery of antisense RNAs expressed during herpes infection [313], to name a few. Future applications can include analysis of RNA metabolism during development, differentiation, the course of the cell cycle, and in response to external cues, stresses and infections. Many open questions in RNA biology involve the kinetics of RNA abundance and the effect of various players on RNA synthesis and degradation, rather than the steady-state abundance of RNA alone. The combination of metabolic labeling with high-throughput sequencing has allowed researchers to address these questions at a global level and will prove to be a valuable asset in the RNA biologist's toolkit.

4.12 Acknowledgments

This work was supported in part by the National Institute of General Medical Sciences, National Institutes of Health grant R35 GM128637 to P.L.F. and grant R01 GM105707 to A.C.G. Additionally M.B.W. is supported by an National Science Foundation Graduate Research Fellowship DGE1256260.

CHAPTER 5

Metabolic labeling reveals global modulation of mRNA stability by the human Pumilio proteins

5.1 Contribution Details

This work is currently being prepared for submission and has not yet been published. I am the primary author of this manuscript and I have performed all analyses presented below. I also wrote this manuscript in its entirety, with editing from Lindsay Cannon, Peter Freddolino, and Aaron Goldstrohm. The work presented below, particularly the experimental side, involved many labs as listed in the author contributions at the end of the chapter. The original experiments were initiated in Aaron Goldstrohm's lab and I performed all analyses under both Peter Freddolino's and Aaron Goldstrohm's mentorship and guidance.

5.2 Abstract

The human members of the PUF family of proteins, PUM1 and PUM2, are RNA binding proteins that post-transcriptionally regulate gene expression through binding to a PUM recognition element (PRE) in the 3'UTR of target mRNAs, promoting RNA decay. Recent RNA-seq experiments in PUM1/2 knockdown conditions have identified hundreds of known and new human PUM targets through measurement of changes in steady state RNA levels. However, steady-state RNA levels do not allow for measurement of changes in RNA stability between conditions and do not allow for the differentiation between the contributions of changes in initial transcription rates and changes in RNA decay. Here, we identify hundreds of human PUM1/2 targets that have changes in RNA stability under PUM1/2 knockdown. We separate the contributions of changes in initial transcription rate and RNA decay and find that human PUM proteins primarily modulate RNA abundance through changing RNA decay. In addition, we find that the sequence preferences of all possible 8mers are largely similar between PUM1 and PUM2 through the use of high throughput *in vitro*

RNA binding assays, suggesting that PUM1 and PUM2 recognize similar targets. We identify an ideal PRE "rulebook" by finding key features around PREs, including local AU content, location of a PRE within a 3'UTR, clustering of PREs, and number of miRNA sites near a PRE, that help differentiate functional PREs from non-functional ones as measured by our decay dataset. Consistent with previously identified functional roles of mammalian PUMs, we find that human PUM1 and PUM2 modulate the decay of genes related to signaling cascades and neuronal function. Finally, we use conditional random forest models to predict functional regulation of RNA targets by the human PUM proteins and find that, although we are able to predict changes in steady state RNA levels with accuracy, there is still substantial room for improvement in predicting PUM-mediated gene regulation.

5.3 Introduction

The control of gene expression at the post-transcriptional level is critical for diverse biological processes including the proper organismal development in eukaryotes. Diverse regulators act to control the mRNA stability of transcripts through the recognition of key sequence elements in the 3'UTR of target transcripts [123, 357] and a large fraction of the human RNA binding proteins (RBPs) surveyed thus far bind to mRNAs [126]. The PUF (Pumilio and FBF [fem-3 binding factor]) family of proteins are a set of RNA binding proteins (RBPs) with a similar C-terminal high homology domain (PUM-HD) that results in sequence-specific binding in target RNAs. The founding member of the family, *Drosophila* Pum, together with the Nos protein, is needed for correct body patterning in the developing fly embryo [127, 358]. Patterning is accomplished by location-specific repression of the *hunchback* mRNA through sequence-specific recognition of a nanos response element (NRE) in the 3'UTR [128]. In humans, there are two members of the PUF family, PUM1 and PUM2, which share 75% overall sequence identity and 91% sequence identity in the PUM-HD. In addition, human PUM1 and PUM2 share the 78% and 79% sequence identity in the PUM-HD to DmPum, respectively [359]. Human PUM1 and PUM2 are expressed across tissues and their expression is highly overlapping [359] suggesting that they likely act redundantly in human cells. Functionally, mammalian PUM proteins have been implicated in spermatogenesis [360, 361], neuronal development and function[362–367], immune function [368, 369], and cancer [370–373]. In humans, PUM1 missense and deletion mutants lead to Adult-onset Ataxia (Pumilio1-related cerebellar ataxia, [PRCA]) and loss of one copy leads to developmental delay and seizures (Pumilio1-associated developmental disability, ataxia, and seizure; [PADDAS]) [374]. Structurally, the human PUM-HD consists of 8 helical repeats containing specific amino acids that both intercalate and form Watson-Crick base pairs with target RNA, resulting in exquisite specificity for a UGUA-NAUA consensus sequence motif or Pum Recognition Element (PRE) [132, 375]. Recognition by the PUM-HD is modular and specificity for a given base can be changed through mutation of a set of three key amino acids in a single repeat [132, 376]. Furthermore, sequence specificity by PUM-HD across species can be predicted from the identity of these three key amino acids across the helical repeats in any given PUM-HD [377]. High-throughput measurements of PUM1 and PUM2 binding sites *in vivo* have confirmed this high specificity for a PRE and have identified a diverse set of PUM targets in human cell lines, including those involved in regulating neuronal function and signaling cascades [135, 292, 378, 379]. Thus, sequence-specific recognition of the PRE is a key aspect of target recognition for the PUM proteins.

Targeted experiments have indicated that human PUM1 and PUM2 are able to repress expression of a luciferase reporter through recognition of PREs in the reporter gene's 3'UTR, likely through recruitment of the CCR4-NOT complex and subsequent degradation of the mRNA [131]. Additionally, similar targeted assays have shown that repression by the human PUM2 PUM-HD alone requires the polyA binding protein PABPC1 [380]. However, PUM-mediated repression is not the only type of gene regulation by human Pumilio proteins. Recently, expression of a key regulator of hematopoietic stem cell differentiation FOXP1 was shown to be activated by human PUM1/2 binding to the 3'UTR [372]. Furthermore, measurements of changes in global steady-state RNA abundance between wild-type (WT) and PUM1/2 knockdown conditions have identified hundreds of RNAs that either increase or decrease in abundance upon PUM1/2 knockdown. Targeted experiments have confirmed activation of key targets by human PUMs through the use of a reporter gene-target 3'UTR fusion construct [134].

Key questions about PUM-mediated gene regulation remain. There are on the order of 10,000 PRE sites across the full set of annotated human 3'UTRs, but only roughly 1000 genes change in steady state RNA levels under PUM1/2 knockdown [134]. Additionally, models using a simple count of PREs in the 3'UTR of a transcript do not completely capture the complexity of PUM-mediated gene regulation [134]. The identification of additional sequence features that discriminate functional PREs from apparently non-functional PREs will improve the understanding of PUM-mediated gene regulation. Furthermore, as the measurement of steady-state RNA levels do not allow for differentiation between the individual contributions of initial transcription rates and RNA decay, we instead set out to measure changes in RNA stability under PUM1/2 knockdown conditions. This has allowed us to determine RNA targets that display PUM-mediated changes specifically in RNA decay and facilitated our understanding of functional PRE sites. Through the use high-throughput sequencing methodologies, we demonstrate the human PUM1/2 modulate the RNA abundance of mRNA targets primarily through controlling mRNA decay and not initial transcription rates. We demonstrate, through high-throughput *in vitro* binding assays, that PUM1 and PUM2 RBDs have highly similar preferences for the same sets of sequences. We find that PUM1/2 control the mRNA decay of transcripts involved in signaling pathways, neuronal development, and transcriptional control. In addition, we identify a key set of contextual features around PREs that help predict PUM-mediated regulation including proximity to the 3'end of a transcript and the AU content around PRE sites. Taken together, our study illuminates key contributors to determining functional PRE sites and represents a rich resource for interrogating the control of mRNA decay by the PUM RBPs.

5.4 Results

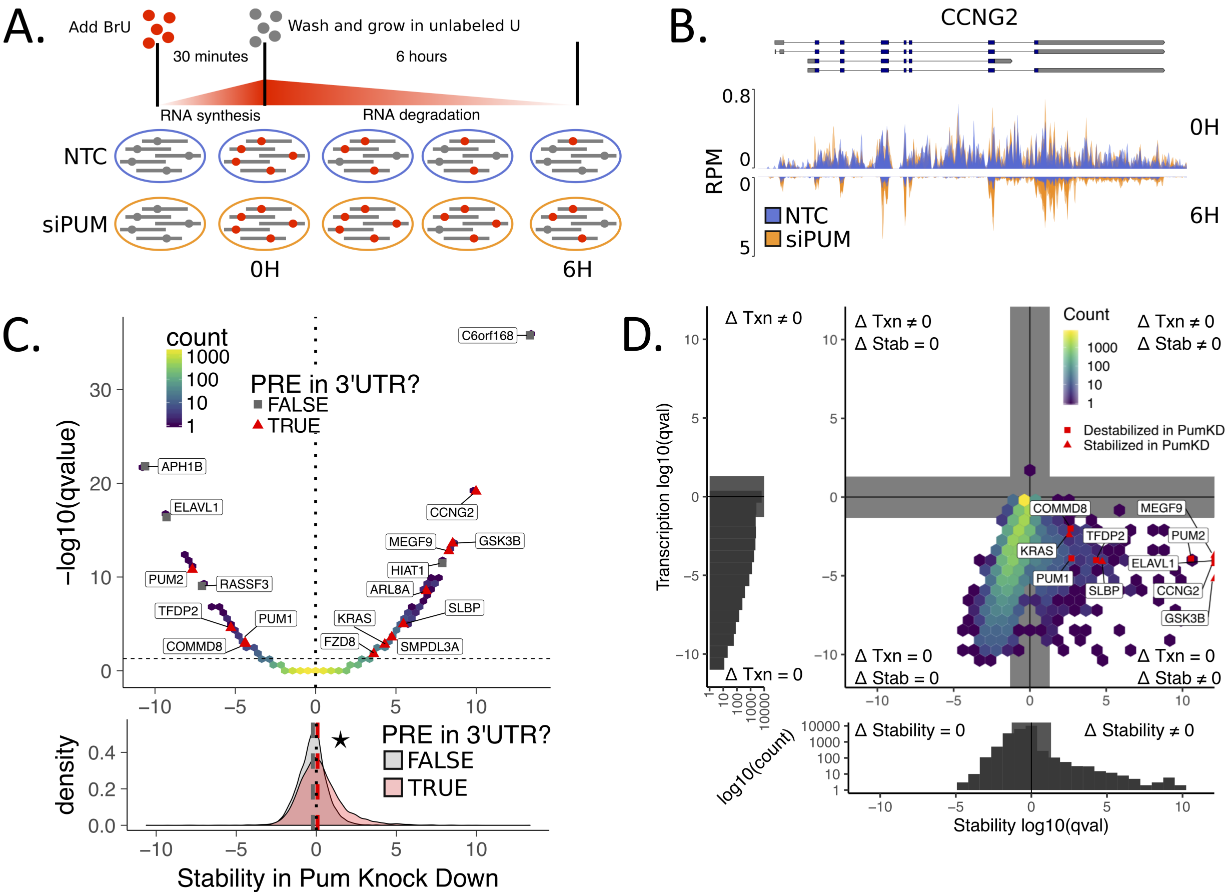
5.4.1 Metabolic labeling with BruSeq reveals Pum-mediated effects on mRNA stability

In order to measure the effect of the human PUM1 and PUM2 proteins on mRNA stability at a transcriptome-wide scale, we employed the Bru-Seq and BruChase-Seq methodology [308]. In brief, Bru-Seq involves the metabolic labeling of RNA using 5-bromouridine (BrU), which is readily taken up by the cells and incorporated into the nascent NTP pool [297]. After incubation with BrU over a short time period, newly synthesized and labeled RNAs are selectively pulled out of isolated total RNA using an anti-BrdU antibody and sequenced. Labeled RNA abundance is then tracked over time by continuing to grow the cells in the absence of BrU and isolating BrU labeled RNA at additional time points. For this study, two time points were chosen: (1) a zero hour time point taken at the transition to unlabeled media after 30 minutes of incubation in BrU-containing media and (2) at six hours, a time point chosen to coincide with the average mRNA half-life in cultured mammalian cells [282, 321, 322]. The experiment was performed in the presence of a mix of siRNAs targeting both *PUM1* and *PUM2* mRNAs (siPUM) or in the presence of scrambled non-targeting control siRNAs (NTC) (Figure 5.1A). It is important to note that the use of two time points does not allow for determination of full decay rate constants for each transcript, but it does allow for measurements of relative changes in mRNA stability between the two conditions [268].

In Figure 5.1B the read coverage for cyclin G2 (*CCNG2*), a cyclin-dependent kinase involved in the cell cycle, is shown at the 0 hr and 6 hr time points for the NTC (blue) and siPUM conditions (orange). At the 0 hr time point, read coverage resulting from initial transcription for four distinct replicates in each condition can be seen (Read coverage includes immature RNAs that still contain introns) (Figure 5.1B top). At the six hour time point, only mature RNA remains, with read coverage primarily observed at exons and no longer prevalent in the intronic regions (Figure 5.1B bottom). Here, silencing of both PUM1 and PUM2 clearly increases RNA abundance relative to the non-targeting control at the 6 hr time point, but does not appear to impact initial transcription as seen at the 0 hr time point.

To quantify the effect of silencing PUM1 and PUM2 on changes in relative labeled RNA abundance between the 0 and 6 hour time points, we used DEseq2 [340] to model the count of reads observed from each gene using a generalized linear model that considers the effects of time, condition, and the interaction between time and condition (see Methods for details). We interpret the term associated with the interaction between condition and time to be the PUM-mediated effect on stability – where a positive value indicates that an RNA was stabilized in the PUM knockdown condition and a negative value indicates that an RNA was de-stabilized in the PUM knockdown condition. Likewise, we interpret the condition term as the PUM-mediated effect on initial transcription rates. Figure 5.1C displays an overview of PUM-mediated effects on stability as a volcano plot, with 12,165 genes represented in a two-dimensional histogram. A selection of both previously-identified [134, 378] and newly discovered PUM activated and PUM repressed genes are represented as individual points with shape and color indicating whether a PRE, as defined by a match to the PUM1 position weight matrix (PWM) described below, was found in any annotated 3'UTR for that gene. Using an FDR-corrected p-value threshold of 0.05 and a fold-change cutoff of 1.75 we found 44 genes were statistically significantly de-stabilized (56 with no fold-change cutoff) and 200 genes were statistically significantly stabilized in the PUM knockdown condition (252 with no fold-change cutoff). Of these genes, 30 were also identified as having lower abundance under PUM knockdown in the Bohn et al. [134] RNA-seq data set (37 with no fold-change cutoff). Likewise, 95 were also identified as having higher abundance under PUM knockdown in the Bohn et al. [134] RNA-seq data set (106 with no fold-change cutoff). As expected, in our data both *PUM1* and *PUM2* were destabilized in the PUM knockdown condition relative to the WT condition. Additionally, we found that genes with a PRE in their 3'UTR were, on average, more stabilized in the PUM knockdown condition than those without a PRE in their 3'UTR (Figure 5.1C bottom).

Using our statistical methodology, we separated the impact of silencing PUM on stability from its impact on initial transcription rates. For each term, we tested for statistically significant changes under a null model centered around 0. In addition, we tested for a statistically significant lack of change by considering a null model centered around the boundary of a defined region of practical equivalence spanning from $-log_2(1.75)$ to $log_2(1.75)$(see Methods for details); such a test is important because failure to reject the null hypothesis cannot, by itself, be taken as evidence favoring the alternative. In total, four statistical tests were run for each gene: a test for change and a test for no change for both transcription and stability. For each axis, the smaller of the two FDR-corrected p-values (i.e. test for change vs. test for no change) was chosen as the coordinate for that term, which enabled classification of each gene into one of four quadrants: 1. Genes that change in both stability and transcription (Figure5.1D, upper right quadrant), 2. genes that change only in stability (Figure5.1D, lower right quadrant), 3. genes that change only in transcription (Figure 5.1D, upper left quadrant) and 4. genes that change in neither (Figure 5.1D, lower left quadrant). Thus, using



this methodology, we identified 213 genes with a statistically significant change in stability (Figure5.1D lower right quadrant). We were also able to identify a set of 2,834 genes with evidence for no change in stability under our experimental conditions (Figure5.1D lower left quadrant). Additionally, we show only one gene, *ETV1*, with a statistically significant change in transcription and 11,527 genes with statistically significant lack of change in transcription. Taken together and consistent with the Pumilio proteins' role in post-transcriptional regulation, these results suggest that PUMs primarily regulate gene expression at the level of RNA decay and not transcriptional initiation. Furthermore, this analysis allows us to divide the genes into those in which Pumilio knockdown has an EFFECT on RNA stability and those in which there is evidence for NOEFFECT on RNA stability, a stronger statement than simply failing to reject the null hypothesis that a change was occurring. The words EFFECT and NOEFFECT will be used to refer to these gene classes throughout the rest of the paper.

Figure 5.1 *(previous page)*: BruSeq allows for determination of Pum-mediated effects on RNA stability. A) Experimental design for measuring Pum-mediated effects on RNA stability. HEK293 cells were incubated for 30 minutes in the presence of BrU prior to time 0. Cells were then washed and cultured in media containing unlabeled uridine for six hours. At times 0 and 6 hours, a portion of cells were harvested and BrU labeled RNA was isolated for sequencing. Changes in relative RNA abundance between the 0 and 6 hour time points were compared between cells grown in the presence of silencing RNA targeting *PUM1* and *PUM2* (siPUM) and a non-targeting control siRNA (NTC). B) Read coverage traces for CCNG2 as measured in reads per million (RPM). Traces are shown for siPUM (orange) and NTC (blue) conditions at both 0H (top) and 6H (inverted bottom) time points. Four replicates for each combination of siRNA and time point are overlaid. Known isoforms for *CCNG2* are represented above. C) (Top) Volcano hexbin plot displaying global changes in RNA stability under Pum knockdown conditions. Stability in Pum knockdown is represented by a normalized interaction term between time and condition (see methods for details). No change in stability is represented with a dotted line at 0. Statistical significance at an FDR corrected p-value < 0.05 is represented with a horizontal dashed line. A combination of genes known to be regulated by Pum and genes newly identified in this study are labeled. Red triangles indicate genes that have a PRE in any annotated 3'UTR as determined by a match to the Pum1 motif we identified using SEQRS (Figure5.2A). Gray squares indicate genes that did not have a PRE in their 3' UTR. Unlabeled genes are binned into a two-dimensional histogram to avoid overplotting. (Bottom) Marginal distribution of Stability in Pum knockdown for genes with a PRE in their 3'UTR (red) and genes without a PRE in their 3'UTR (gray). Median values for each distribution are plotted as a dashed line in the appropriate color. The star indicates a statistically significant difference in the median stability as measured by a two-sided permutation of shuffled labels (n =1000, p $<$ 0.001). D) Analysis of changes in transcription vs. changes in stability. Four separate statistical tests were calculated for each gene: 1. a test for statistically significant changes in RNA stability (Δ Stability $\neq 0$), 2. a test for statistically significant changes in transcription (Δ Txn $\neq 0$), 3. a test for no change in RNA stability (Δ Stability = 0), and 4. a test for no change in transcription (Δ Txn = 0). Genes are plotted as an (x,y)-coordinate where each coordinate represents the $\pm \log_{10}$(FDR corrected p-value) of the test with greater evidence ($\Delta \neq 0$, +log_{10}; or Δ = 0, -$\log_{10}$) for each axis (see methods for details). Representative genes displaying a range of stability effects are labeled. Red squares represent genes that were destabilized in Pum knockdown, whereas red triangles represent genes that were stabilized in Pum knockdown. All other genes were binned into a two dimensional histogram. Gray rectangles represented a statistical significance cutoff of q-value $>$ 0.05. (Left and Below) Marginal histograms for each axis are plotted with matching gray rectangles to represent the same statistical significance cutoff of q-value $>$ 0.05.

5.4.2 SEQRS shows conserved preference for canonical UGUANAUA PRE by Pumilio proteins

To determine the binding specificity of the human PUM1 and human PUM2 proteins, we applied *in vitro* selection and high-throughput sequencing of RNA and sequence specificity landscapes (SEQRS) to purified RBDs of each protein [381]. Similar to systematic evolution of ligands by exponential enrichment (SELEX) [382], SEQRS allows for the determination of an RNA binding protein's sequence specificity by selecting for RNAs that interact with the RBP out of a pool of random 20mers generated by T7 transcription of a synthesized DNA library. The RNA pulled-down from a previous round is reverse-transcribed into DNA to be used as the input for the next round of transcription and selection, allowing for exponential enrichment of preferred sequences for any RBP of interest. We applied five rounds of SEQRS to the PUM1 and PUM2 RBDs separately and quantified the abundance for each of the 65536 possible 8mers in the sequencing libraries for each round (including 8mers that would overlap with at least one base with the adjacent static adapter sequences see Methods for details). To obtain representative PWMs for each round of selection (Figure 5.2A,B (top)), we used the top enriched 8mer, UGUAAAUA, as a seed sequence to create a multinomial model from the abundance of every possible single mismatched 8mer to the seed sequence (see methods for details). This data analysis approach has yielded similar results to that of expectation-maximization algorithms such as MEME [383] and has been used successfully with SELEX experiments using DNA binding proteins [384, 385]. We also apply this same analysis pipeline to previously published SEQRS analysis of the *D. Melanogaster* Pumilio RBD [381] and find that it readily captures the *D. mel* Pum sequence preference for the canonical UGUANAUA PRE (Figure 5.2D (top)). When considering the enrichment of all 8mers relative to sequencing of the input pool, we see that 8mers within 1-2 mismatches of the UGUAAAUA seed sequence are highly enriched compared 8mers with more than 2 mismatches (Figure 5.2A,B,D (bottom)). However, variation in enrichment scores with higher numbers of mismatches suggests that sequences matching the canonical UGUAAAUA may not fully explain PUM binding specificity. Additionally, our SEQRS experiment suggests that the PUM2 RBD has much weaker enrichment for the canonical PUM PRE compared to PUM2 which is inconsistent with PUM2 sequence preferences obtained from *in vivo* transcriptome-wide experiments [135, 292]. This may indicate differences between *in vitro* and *in vivo* conditions that specifically impact PUM2 or may indicate that PUM2 RBD does not bind as efficiently to RNA as the full-length protein. Figure 5.2E shows a comparison of enrichment scores for all possible 8mers between PUM1 and PUM2 and indicates that PUM1 and PUM2 RBDs have overall similar and highly correlated enrichments for all 8mers, with PUM1 RBD having overall higher enrichment than PUM2. Unless otherwise indicated, the SEQRS round 5 PWM for PUM1 will be used to determine PREs throughout the text.

5.4.3 Features associated with PREs explain variability in Pum-mediated RNA stability

Determining what defines a functional binding site from a non-functional binding site as well as factors that control the magnitude of the regulation effect are a major questions for any RBP. Taken as a whole, RBPs tend to bind similar low sequence complexity motifs *in vitro* [386]. Additionally, probing of RBP binding *in vivo* at a transcriptome-wide scale, has indicated that the majority of predicted binding sites are not bound for some RBPs [387]. Targeted experiments with the Pumilio-family of proteins have established that mammalian Pumilio proteins recognize the UGUANAUA PRE in the 3'UTR of target genes [131, 135, 292, 365].

To determine sequence motifs *de novo* that have explanatory power for our RNA stability dataset, we used FIRE [388] to find motifs in the 3'UTR of transcripts that share high mutual information with our RNA stability dataset by taking the normalized interaction term (see methods

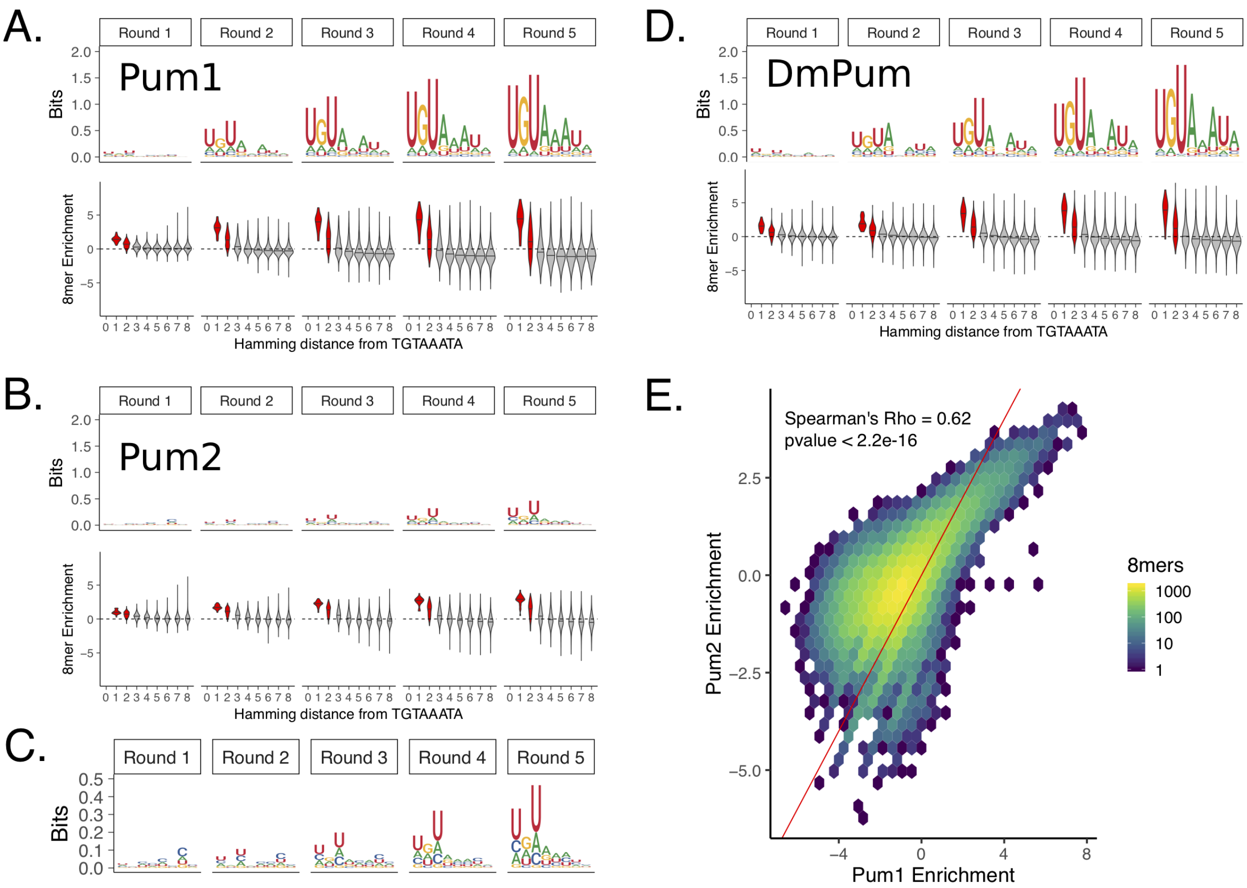


Figure 5.2: SEQRS analysis of Human Pum1 and Pum2 RBDs reveals preference for canonical the Pum Recognition Element. A) (Top) Position weight matrices representing 8mer sequence preferences for purified Human Pum1 RBD, as determined for each SEQRS round. (Bottom) 8mer enrichment, as measured by $\log_2$(Enrichment SEQRS round/ Enrichment no protein) (see methods for details) for each 8mer as binned by hamming distance from the canonical TGTAAATA Pum recognition element. Enrichment scores for 8mers within 2 mismatches are filled in red. B) Same as in A, but for Human Pum2 RBD. C) Closer view of Human Pum2 RBD PWMs. D) Same as in A, but for Drosophila Pum RBD. E) Correlation of 8mer enrichment between Human Pum1 and Human Pum2 RBDs. Enrichment for all possible 8mers are displayed in a two dimensional histogram.

for details) and discretizing it into ten bins, with an equal number of genes in each bin. Figure 5.3A shows that FIRE rediscovers the canonical UGUANAUA PRE using only the RNA stability data as input. Furthermore, the UGUANAUA PRE is enriched in transcripts that are highly stabilized under PUM knockdown conditions, suggesting that these transcripts are regulated by PUM through recognition of a UGUANAUA PRE in the 3'UTR of the transcript. However, this analysis does not provide direct evidence for PUM binding *in vivo* at PREs associated with transcripts.

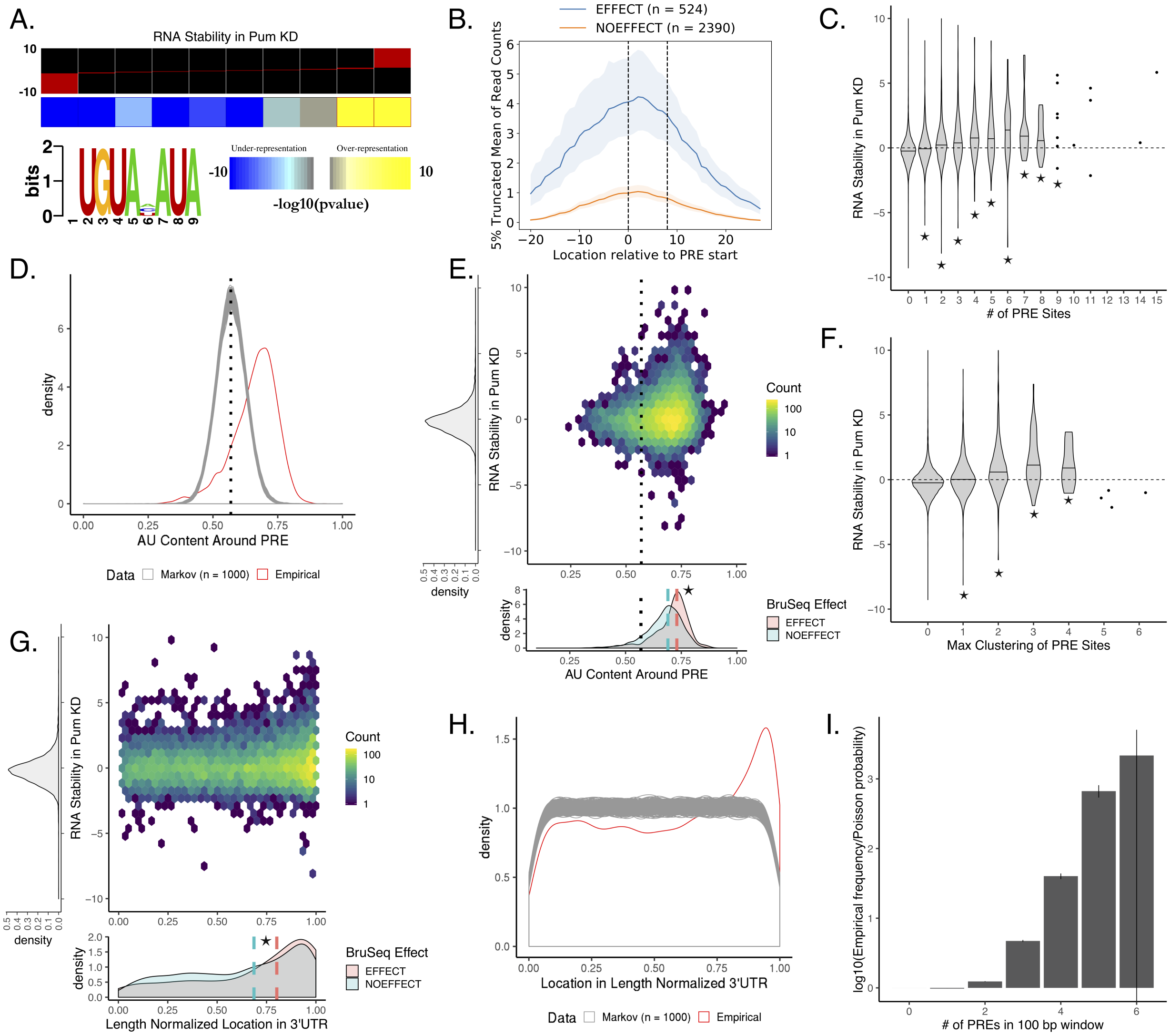
To determine whether there was evidence for PUM binding at PREs associated with a change in RNA stability, we used publicly available PAR-CLIP data for human PUM2 [292] to determine the amount of read coverage at PREs associated with transcripts that have a statistically significant change in RNA stability under PUM knockdown (EFFECT class, Figure5.1D) and compared it to transcripts with a statistically significant lack of change in RNA stability (NOEFFECT class, Figure5.1D). In Figure 5.3B, we report the average read coverage in a 40 bp window around PREs in the 3'UTR of transcripts associated with the EFFECT and NOEFFECT classes. We use a 5% truncated mean to remove the impact of extreme outliers on the average coverage reported. To estimate a 95% confidence interval on the average coverage (shaded region), we performed bootstrap replicates (n = 1,000) by sampling vectors of read coverage for individual PREs with replacement. Here, we clearly see that PREs in transcripts with a change in RNA stability have higher read coverage than those with no change in RNA stability. This is consistent with higher overall PUM binding at PREs associated with changes in RNA stability but, as the PAR-CLIP signal is not normalized to RNA abundance, the possibility that these transcripts were simply more abundant under the PAR-CLIP conditions cannot be definitively ruled out.

We have shown that a PRE in the 3'UTR is associated with a change in RNA stability under PUM knockdown and that PREs in transcripts with a change in RNA stability have more evidence for being bound by PUM *in vivo*. However, knowledge of the presence or absence of a PRE in the 3'UTR alone is not sufficient to predict the magnitude of PUM-mediated repression and a wide variation in the effect of knocking down human PUM1 and PUM2 on steady-state RNA levels has been observed in previous transcriptome-wide analysis [134]. Here, we demonstrate that a similar level of variation can be seen in measurements of RNA stability. Figure 5.3C, displays the overall distribution of RNA stability measurements for transcripts with increasing numbers of PREs in annotated 3'UTRs. An increase in the number of PREs is associated with an increase in RNA stability on average under PUM knockdown conditions compared to transcripts that do not have a PRE in their 3'UTR. However, wide variations in RNA stability can be seen for each category consistent with previous measurements of changes in steady state RNA levels under PUM knockdown [134]. Thus, we considered additional contextual features around PREs that could help to better explain the large variance in RNA stability we observe.

To explore the local sequence context around PREs, we trained a 3rd order Markov model on the full set of unique annotated hg19 3'UTRs that were greater than 3 basepairs long (29,380 3' UTRs). Using this Markov model, we simulated 29,380 3'UTRs that were the same length and shared similar sequence composition to the empirical set of true 3'UTRs. We then searched for matching PREs in the simulated set of 3'UTRs and calculated the AU content in a 100 bp window around these PREs. On average, we discovered 11264.18 matching PREs (standard deviation of 106.167) in simulated sets of 3'UTRs compared to the 12582 matching PREs in the emprical set of 3'UTRs. Figure 5.3D displays the local AU content for PREs found in 1000 simulated sets of 29,380 3'UTRs (gray), as compared to the empirical distribution of local AU content for PREs found in the actual set of human 3'UTRs (red). The dotted line represents the average AU content over all 3'UTRs. Here, the true set of PREs have higher local AU content than one would expect from chance, as represented by the Markov models (p-value < 0.001). In the Markov models, the local AU content for PREs is centered around the average AU content for all 3' UTRs, as would be expected if there was no selective pressure for PREs to occur in AU rich areas of 3'UTRs. This analysis is consistent with Jiang et al. [136] who also observed a preference for PREs to occur in AU rich areas as compared to shuffled PREs with preserved overall sequence content.

Furthermore, we observe that transcripts with PREs that have higher local AU content also have a larger measured change in RNA stability. Figure 5.3E displays a two-dimensional histogram of highest local AU content in 100 bp surrounding a PRE in a gene's 3' UTR and the RNA stability measurement associated with the transcript for that PRE. The y axis marginal kernel density plot displays the distribution of RNA stability for transcripts with no PRE in their 3'UTR. PREs with local AU content above the average AU content for all 3' UTRs (dotted line) appear to have a larger effect than those with lower local AU content. Additionally, PREs in transcripts that had a statistically significant stability effect in PUM knockdown had higher local AU content compared to PREs in transcripts with no change in stability ($p < 0.001$, Figure 5.3E bottom).

Using the Markov models described above, we also looked at the location of PREs within 3'UTRs. In Figure 5.3F, we observe that the empirical distribution of true PRE locations in length-normalized 3'UTRs appear enriched towards the 3' end of 3'UTRs (red) as compared to PREs found within 1000 simulated sets of 3'UTRs (gray). Again, this suggests a selective pressure for PRE sites to exist at the 3' end of 3'UTRs as compared to the uniform distribution of PREs found in simulated 3'UTRs with similar sequence properties. Like the AU content analysis, this analysis is also consistent with observations made by Jiang et al. [136] who saw an enrichment towards the 3' end for PRE locations in the full set of human 3'UTRs compared to a shuffled PRE motif with preserved overall sequence content. While these approaches are complementary, our approach allows for the exact identity of the PRE to remain intact thereby maintaining a PRE-centric assessment rather than one based solely on the general sequence content within the motif. Additionally, we observe that transcripts with a PRE towards the 3' end of the 3'UTR tend to have



a larger RNA stability effect (Figure 5.3G center) and PREs in transcripts that had a statistically significant change in stability in PUM knockdown were, on average, closer to the 3' end of the 3'UTR than those with no change in RNA stability ($p < 0.001$, Figure 5.3G bottom).

In Figure 5.3H, we discretized transcripts according to how many full PREs were clustered within a 100 bp window within the 3' UTR of that transcript. Similar to the association with the number of PREs (Figure 5.3), we find that having more PREs clustered together is associated, on average, with a higher stabilization effect under PUM knockdown conditions. We also find that PREs tend to cluster together more than one would expect by chance by determining the divergence from a simple Poisson model (Figure 5.3I, $p < 0.001$ for clusters 2-5; see Methods for details).

Figure 5.3 *(previous page)*: Features associated with a Pum Recognition Element explain some variability in Pum-mediated effect on decay. A) Results of motif inference using FIRE [388] on RNA stability data discretized into 10 equally populated bins. B) 5% Truncated average of Pum2 PAR-CLIP read coverage [292] over each PRE site in the 3'UTRs of genes with a statistically significant change in RNA stability (blue) compared to genes in which there was a statistically significant lack of change in stability (orange; see methods for details on no effect test). Shaded regions represent 1,000 bootstrap replicates within each group. Dashed lines indicate the PRE site. C) Violin plots representing the distributions of RNA stability for genes with 0 to 15 PRE sites within their 3'UTR. Stars represent statistical significance as measured by a Wilcoxon rank sum test using the 0 PRE case as the null distribution. D) Distribution of AU content in a 100 bp window around all unique PRE sites in the 3' UTRs of the human transcriptome. The empirical distribution (red) is compared to the distribution of AU content around PRE sites in 1,000 simulated sets of 3'UTRs the same size as the true set of 3'UTRs as simulated from a third order Markov model trained on the true 3'UTR sequences. The dotted line represents the average overall AU content of the entire set of 3'UTRs in the human transcriptome. E) Relationship of AU content in a 100 bp window around a PRE to RNA stability. (left) Marginal kernel density plot of RNA stability for genes with 0 PREs in their 3'UTRs. (right) 2D histogram of RNA stability and AU content around each PRE site for all genes with at least one PRE in the 3'UTR. Dotted line represents the average AU content over the entire set of 3' UTRs in the human transcriptome (bottom). Marginal kernel density plot of AU content around a PRE site split amongst genes with a statistically significant change in RNA stability (red) and genes with a statistically significant lack of change in stability (blue). Dotted black line represents the average AU content (right). Dashed lines represent the median AU content around a PRE for the effect (red) and no effect (blue) genes. The star represents a statistically significant difference in medians using a one-sided permutation test (n=1,000) of shuffled class labels. F) Violin plots representing the distributions of RNA stability for genes with 0 to 6 full PRE sites clustered within a 100 bp window. Stars represent statistical significance as measured by a Wilcoxon rank sum test using the 0 PRE case as the null distribution. G) Relationship of normalized location of PRE site in 3' UTR to RNA stability. Plots as in (D). H) Distribution of length normalized locations of PRE sites in the 3'UTRs of the human transcriptome. The empirical distribution (red) is compared to that of PRE sites found in 1,000 simulated sets of 3'UTRs calculated as in (G). I) Comparison of the empirical frequencies of PRE site clustering over all possible 100 bp windows in the full set of human 3'UTRs with at least 1 PRE in them to the probabilities expected from a poisson null distribution. Error bars represent 95% confidence intervals based on 1,000 bootstraps of the empirical distribution.

5.4.4 Pumilio proteins modulate the stability of genes involved in neural development and regulators of gene regulation

Mammalian Pumilio proteins have been shown to regulate a diverse set of genes, including those involved in signaling pathways, transcriptional regulation, and neurological functions [134, 361, 365, 366, 378]. Consistent with prior observations, we see changes in RNA stability for genes involved in these functions. For example, multiple epidermal growth factor-like-domains 9 (*MEGF9*) is a transmembrane protein that is highly expressed in the central and peripheral nervous system and its expression appears to be regulated over nervous system development in mice [390]. We see strong stabilization of the *MEGF9* transcript under PUM knockdown conditions (Figure 5.5A top). Furthermore, of the five PREs we identify in two unique 3'UTRs for *MEGF9*, we see the strongest PUM2 binding signal for the 3'-most PRE (Figure 5.5A bottom right). Additionally, we see that the 3'-most PRE has high local AU content compared to the overall distribution of PRE sites (Figure 5.5A bottom left). Taken together, these data implicate the PUM proteins as direct post-transcriptional regulators of *MEGF9*.

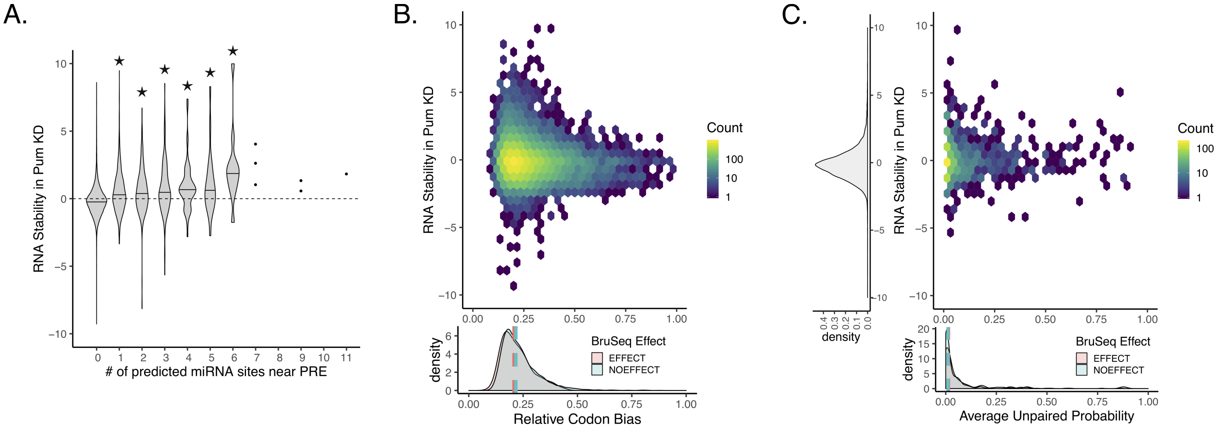


Figure 5.4: Additional features considered in determining PUM-mediated decay. A) Here the count of predicted conserved miRNA sites from conserved families that overlapped within 100 bp of a PRE was tallied for each gene. Stars indicate statistical significance from a Wilcoxon test compared to the 0 overlapping miRNA case. B) Interaction between codon usage bias as measured by Relative Codon Usage Bias [389] and PUM-mediated effect as measure in our BruSeq data. C) Interaction between the probability of a given PRE being unpaired in predicted RNA secondary structure. Only genes with a PRE with >0 probability of being unpaired where shown in the heatmap. All other genes are shown in the marginal y-axis density plot. See methods for details of secondary structure prediction.

Another transcript that is strongly stabilized under PUM knockdown conditions is Glycogen synthase kinase-3 B (*GSK3B*) (Figure 5.5B top). GSK3B is a serine-threonine kinase that is involved in the regulation of diverse cellular processes and its misregulation is associated with neurological disease [391, 392]. We identify four PREs in *GSK3B* 3'UTRs (Figure 5.5B below) with largely similar adjacent AU content (Figure 5.5B bottom left). We also find that the 3' most distal PRE has evidence for PUM2 binding consistent with the global trends we describe in Figure 5.3). Like *MEGF9*, this evidence suggests that PUM proteins are directly involved in destabilizing *GSK3B* transcripts.

We also see examples of RNAs that are destabilized when PUM is knocked down, suggesting that PUM may actually act to stabilize these transcripts under conditions containing WT levels of PUM expression. Transcription dimerization partner 2 (*TFDP2*) encodes a protein that cooperates with E2F transcription factors to regulate genes important for cell cycle progression and dysregulation of this system can lead to cancer [393]. PUM proteins have been previously shown to regulate another member of the E2F family by functionally cooperating to enhance the effect of miRNA-mediated regulation of *E2F3* expression [137]. Furthermore, regulation of *TFDP2* by the liver-specific miRNA *miR-122* has been shown to be important for preventing up-regulation of *c-Myc* in hepatic cells [394]. We observe that *TFDP2* is highly destabilized under PUM knockdown conditions (Figure 5.5C top). Additionally, we find that the *TFDP2* 3'UTR has a single PRE site toward the 3' end of the 3'UTR and has high adjacent AU content (Figure 5.5C bottom and lower left). However, there is limited evidence for PUM2 binding in PAR-CLIP data (Figure 5.5C lower right). One possible mechanism for PUM mediated activation of *TFDP2* is by acting to block regulation by miRNAs; however, the nearest conserved miRNA site of a conserved miRNA family to the PRE is over 100 bases away [124] and further evidence would be needed to establish this link.

Another example of a highly destabilized transcript under PUM knockdown conditions is the embryonic lethal abnormal vision 1 (*ELAVL1*) or HuR RNA binding protein (Figure 5.5D top). The *ELAVL1* RBP stabilizes RNA transcripts by binding to AU-rich elements in the 3'UTR of transcripts [395] and its dysregulation is associated with several different types of cancer [396]. We did not find any matching sequence to the PRE we have defined from our PUM1 SEQRS analysis in the 3'UTR of *ELAVL1*; however we do find two matching PREs using the Hafner et al. [292] motif from PAR-CLIP analysis of PUM2 (Figure 5.5D bottom). These motifs are spread evenly across the 3'UTR and have either below average or average local AU enrichment compared to other sites defined by the Hafner motif (Figure 5.5D lower left). Additionally, there is limited evidence for binding by PUM2 at either of the PREs in the *ELAVL1* 3'UTR (Figure 5.5D lower right). Taken together, this suggests that *ELAVL1* may be indirectly regulated by PUM.

To discover categories of genes that are globally associated with RNA stability in PUM knockdown, we applied iPAGE—a computational tool that uses mutual information to find informative Gene Ontology (GO) terms associated with discretized gene expression data [155]—to our stability dataset as represented by the normalized interaction term discretized into 5 equally populated bins. It is worth noting that this analysis will pick up pathways regulated both indirectly and directly by PUM out of the full set of annotated GO terms. Figure 5.6A displays the iPAGE results with several GO terms that are either significantly overrepresented (red-filled box) or underrepresented (blue-filled box) across the full range of stability data. We see several enriched GO term categories that are consistent with previous reports of changes in steady-state RNA levels under PUM knockdown in HEK293 cells [134] including categories related to guanyl-nucleotide exchange factor activity (GO:0005085), WNT signaling (GO:0030177), nucleosome (GO:0000786) and platelet-derived growth factor receptor signaling (GO:00048008).

For a finer grain view, we plotted the RNA stability measurement for each gene involved in selected GO terms as indicated by either blue (destabilized in PUM KD) or red (stabilized in PUM KD) text for that GO term in Figure 5.6A. In Figure 5.6B, we show two selected GO terms whose members tend to be de-stabilized upon PUM knockdown: nucleosome (GO:0000786, left) and myelin sheath (GO:0043209, right). For genes related to the nucleosome, we see a general destabilization under PUM Knockdown conditions. However, when comparing genes within this GO term that have a PRE in their 3'UTR to those that do not, we see that genes with a PRE in their 3'UTR have a median stability that is significantly higher than those without a PRE in their 3'UTR (p ¡ 0.001), suggesting that the destabilization of nucleosome genes under PUM knockdown conditions may be mediated indirectly. Some of these effects could be explained by perturbation of the stem-loop binding protein (SLBP), as SLBP is a protein involved in the proper maturation of replication-dependent histone mRNAs [397], and we observe that *SLBP* is significantly stabilized under PUM knockdown conditions (Figure 5.1C). Like the nucleosome GO term, we see a general de-stabilization of genes categorized into the myelin sheath GO term. A role for PUM in controlling the stability, either indirectly or directly, of genes involved in the myelin sheath is consistent with the previously identified role of Mammalian PUMs in neurogenesis and neurodegenerative diseases [362, 365, 366, 374]. However, we see no evidence for a difference in stability between genes that have a PRE in their 3'UTR compared to genes that do not have a PRE in their 3'UTR. Furthermore, the genes that have a statistically significant de-stabilization under PUM knockdown have no PRE in their 3'UTR, whereas the genes with a significant stabilization do, suggesting a complex role of PUM in modulating the stability of genes in this GO term, possibly arising mainly through indirect effects.

In Figure 5.6C, we report specific GO terms associated with genes that were stabilized under PUM knockdown. Each of these GO terms are involved in regulating gene expression. For instance, guanine nucleotide exchange factors (GEFs) activate Rho-family GTPases to regulate a diverse suite of cellular functions, including cell-cycle progression, the actin cytoskeleton, and transcription [398]. We find that genes associated with guanyl-nucleotide exchange factor activity (GO:0005085) are stabilized under PUM knockdown conditions. Furthermore, genes within this GO term that have a PRE in their 3'UTR are significantly stabilized compared to those with no PRE in their 3'UTR, suggesting that PUM directly acts to destabilize the mRNA transcripts of these genes under normal conditions. We see a similar pattern with genes involved in peptidyl-serine phosphorylation (GO:0018105), which includes a broad class of kinases, including those involved in neurological disease and inflammation [392, 399]. This same pattern also holds with genes involved in transcriptional repressor activity (GO:0001078), which includes proteins involved in regulating hematopoiesis and controlling neurological development [400–402]. Again, genes in these GO terms with a PRE in their 3'UTR are more stabilized under PUM knockdown than those with no PRE, suggesting that PUM has a direct role in regulating a subset of genes in each of these GO terms. Of particular interest is the mild enrichment of CCR4-NOT complex GO term (GO:0030014) in genes that were stabilized in PUM knockdown (Figure 5.6C far right). Almost every gene in this GO term was stabilized under PUM knockdown to some extent. Although the effect of a PRE site for genes in this category did not meet our threshold for statistical significance, several of the genes have a PRE in their 3'UTR and both genes with a statistically significant change in stability have a PRE in their 3'UTR. Human Pumilio proteins have been shown to interact with the CCR4-NOT complex and recruit the complex to target mRNAs for de-adenylation [131]. These data suggest that PUM could also be acting to directly inhibit CCR4-NOT expression and thus globally lower deadenylation rates, perhaps providing a feedback loop that further regulates PUM activity.

5.4.5 Conditional random forest models allow for prediction of Pum-mediated effects from sequence-specific features

A long standing question for any RBP is how to predict that RBPs effect on a given transcript. Previous models of PUM-mediated regulation were created using nonlinear regression based on the number of PREs in various locations across the transcript including the 5'UTR, CDS, and 3'UTR [134]. Here, we use a different approach, which allows us to include a larger feature set of possible predictors for PUM-mediated regulation. Using conditional random forest models [403], we classified genes into EFFECT and NOEFFECT classes, as determined in Figure 5.1D. We used four different definitions for a PRE, (Figure 5.7A) including the SEQRS motifs we defined for PUM1 and PUM2 in Figure 5.2A-B, the PUM2 motif determined from Hafner et al. [292], and a regular expression UGUA.AU[AU] defined from the PUM consensus sequence which has been used extensively to define PREs in previous publications [132, 134, 136]. We focused our analysis on PREs found in the 3'UTRs of target genes. For each definition of a PRE, we calculated several features based on our analysis in Figure 5.3, including AU content around a PRE, clustering of PREs, total count of PREs, a score for PRE match to the specific PRE definition, relative location of the PRE in the 3'UTR, number of miRNA sites near a PRE, and predicted secondary structure around a PRE. In addition to these features, we included motif matches for additional human RBPs, *in vivo* PUM binding data, predictions of secondary structure, and a measure of codon bias for the CDS of target genes (see Methods for details). As our data is highly unbalanced (199 EFFECT genes and 2535 NOEFFECT genes, after performing an inner join on all features) we trained 10 different machine learning models where the NOEFFECT class was randomly downsampled to match the number of EFFECT class genes in each model. Within each downsampled dataset, 5-fold cross validation was performed to assess performance.

To determine which features best help predict EFFECT genes from NOEFFECT genes, we used an AUC-based permutation variable importance measure [404], which indicates the average change in the area under the curve (AUC) of a receiver operator characteristic (ROC) plot across all trees with observations from both classes in the forest when the predictor of interest is permuted. Typically values of the AUC of a ROC curve span from 0.5 to 1.0 where 1.0 indicates perfect classification performance and 0.5 indicates random guessing of class distinctions. Since the AUC-based variable importance measure is calculated using the change in AUC when the predictor is permuted, the expected values are much smaller and fall between 0.0 and 0.06 in simulated cases with 65 predictors and variable numbers of observations from n=100 to n=1,000 [404]. Higher values indicate a larger drop in performance when that variable is permuted; thus,

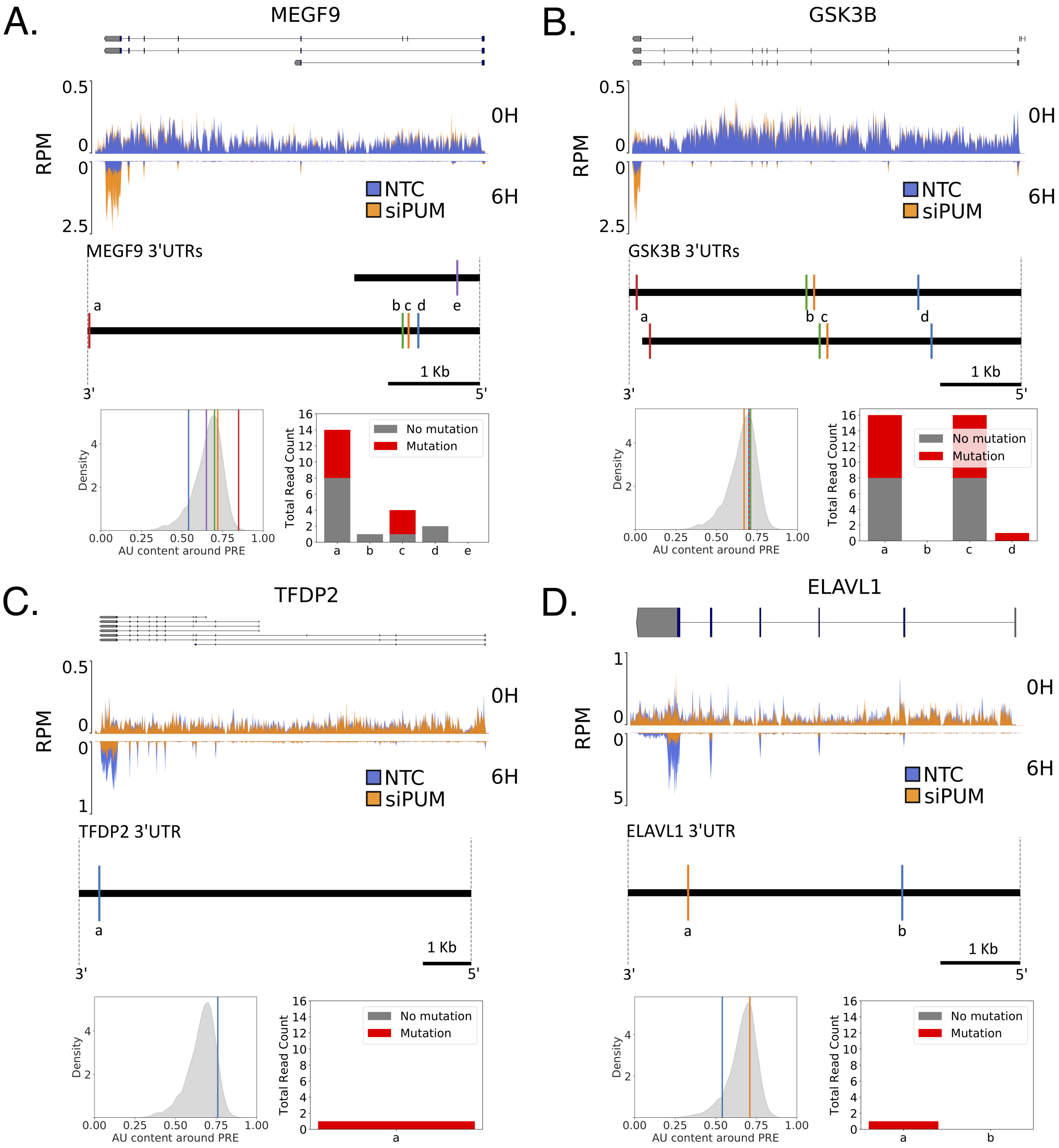


Figure 5.5: Pum-mediated effects on RNA stability under Pum knockdown include stabilization and destabilization. A) (top) Read coverage traces for *MEGF9* as measured in reads per million (RPM). Traces are shown for siPUM (orange) and NTC (blue) conditions at both 0H (upper track) and 6H (inverted lower track) time points. Four replicates for each combination of siRNA and time point are overlaid. Known isoforms for MEGF9 are represented above. (Below) Diagram of unique *MEGF9* 3'UTRs. Sites matching the PUM1 SEQRS motif are represented as vertical lines and labeled alphabetically from 3' to 5'. (Below left) AU content of a 100 bp window around each PRE labeled above in the overall distribution of surrounding AU content for all PUM1 SEQRS motif matches in the entire set of 3'UTRs. (Below right) PAR-CLIP read coverage [292] of 40 bp around each indicated PRE. Number of reads with a T→C mutation are shown in red, whereas the number reads with no T→C mutation are shown in gray. B) As in A), but for *GSK3B*. C) As in A), but for *TFDP2*. D) As in A), but for *ELAVL1* and PREs were determined using the Hafner et al. [292] PUM2 motif as no match was found with the SEQRS PUM1 motif.

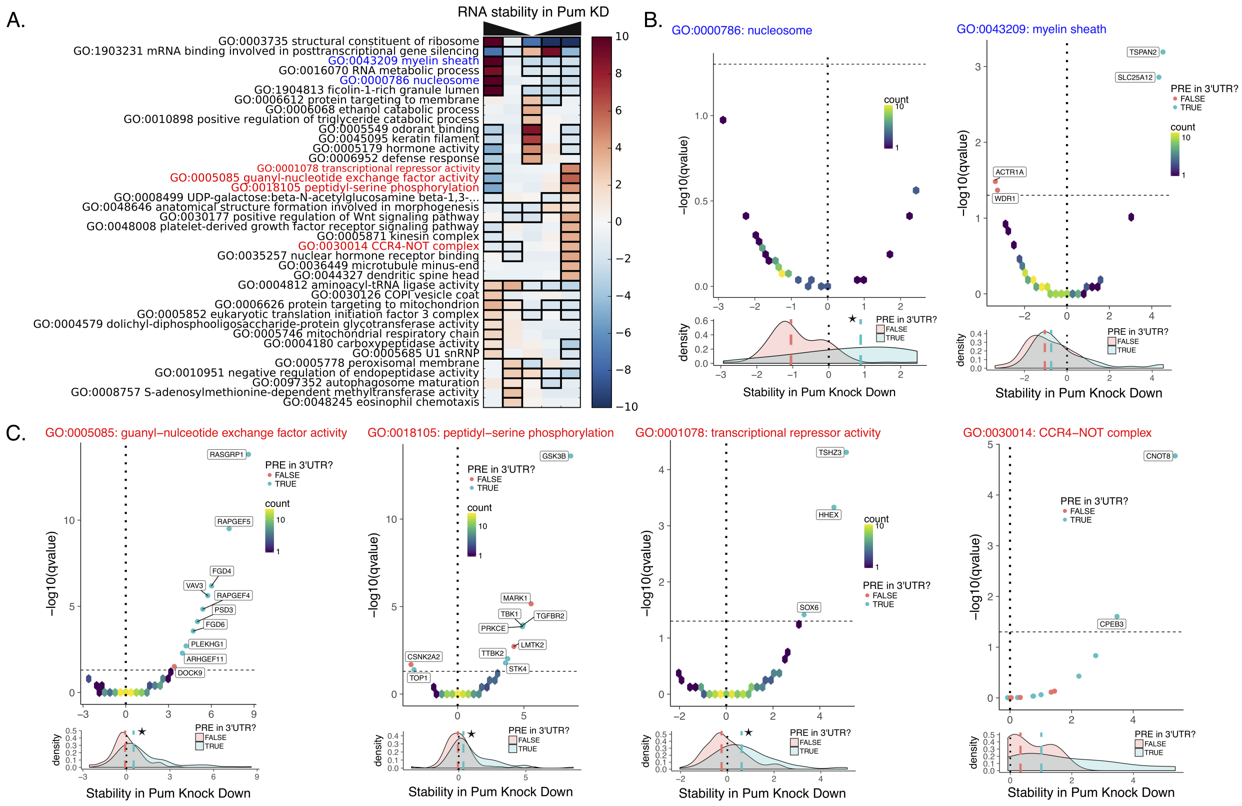


Figure 5.6: Gene ontology terms associated with Pum-mediated changes in RNA stability. A) Results of iPAGE analysis to find GO terms sharing mutual information with RNA stability discretized into 5 equally populated bins. Red bins indicate over representation of genes associated with the corresponding GO term. Blue bins indicate under representation of genes associated with the corresponding GO term. A black box indicates a statistically significant over or under representation with a p-value < 0.05 using a hypergeometric test [155]. B) Selected GO terms whose members are over represented in the RNAs that are destabilized under Pum knockdown, as labeled in blue in panel A. For each GO term, a volcano plot is shown for all genes within the GO term. Volcano plots are shown as two dimensional histograms for genes below a statistical significance threshold (q-value < 0.05) and as individual points for genes above the statistical significance threshold. Individual points are blue if a PRE can be found within any annotated 3'UTR for that gene and red otherwise. The dashed line represents the statistical significance threshold and the dotted line represents no change in RNA stability under Pum knockdown. Below each volcano plot is a marginal density plot for the RNA stability split into two categories: Genes with a PRE in any annotated 3'UTR (blue) and genes with no PRE in any annotated 3'UTR (red). Medians for each distribution are shown as dashed lines in the appropriate color. The black dotted line represents no change in RNA stability, as in the volcano plot above. A star represents a statistically significant ($p < 0.05$) difference in the medians as tested by a two-sided permutation test of shuffled group labels ($n = 1000$). C) As in (B), but for selected GO terms whose members are over represented in the RNAs that are stabilized under Pum knockdown, as labeled in red in panel A.

the variables can be ranked based on their unique contribution to the model, with higher values indicating a more important individual contribution. Figure 5.7B displays the top 20 variables ranked according to their average AUC-based variable performance across all 50 models (10 sets of downsampled models with 5-fold cross-validation each). Count based metrics enumerating the total number of PREs within the 3'UTR appear to be the most important variable for predicting a PUM-mediated effect in the Bru-Seq data. In addition, local AU content and PRE clustering appear to be substantial contributors to the models. To a lesser extent, the number of miRNA sites around a PRE, the location of the PRE in the 3'UTR, and the "Bound" status of the 3'UTR also appear to contribute meaningfully to our models. It is possible that each of these variables contain largely the same information (i.e., whether or not the 3'UTR has a PRE or not in it). To rule out this possibility, we created the same models, where we subset the data to only include genes with one of the four definitions of a PRE within their 3'UTR, thus only using information beyond a simple binary classification ("contains a PRE or not"). Each of these models also displayed substantial contributions for AU content, clustering, and total count in predicting PUM-mediated regulation, as measured by Bru-Seq (Figure 5.8A-D left panel). It is also noteworthy that the variable that contributes most meaningfully to our models is a simple count of the regular expression definition of a PRE, and not the more information-rich PWM definitions.

The high similarity in appearance between each of the definitions of a PRE we include here led us to explore how much redundant information is contained between each of the top 20 highest contributing features. To measure redundancy, we use an information theoretic definition based on discretization of each feature (see Methods for details). In Figure 5.7C, we display the redundancy between the top 20 features as a hierarchically clustered heatmap, where a value of 1.0 indicates that the features contain exactly the same information and a value of 0.0 indicates that the features share no information. Here, we can see that features that are defined around the same motif definition or feature-type tend to share information (as expected). However, despite their similarity in appearance, there are some differences in information content between the different motif definitions and different feature types, indicating that there is knowledge to be gained outside of a simple PRE count.

To assess the performance of our conditional random forest models we, considered several typical machine learning metrics including summary metrics (Accuracy, F1 measure, Matthews correlation coefficient [MCC], Area Under the Curve of a Precision-Recall Curve [AUC PRC], and AUC ROC), and metrics more focused on performance for positive or negative cases (Negative Predictive Value [NPV], Precision, Recall, Specificity). We considered each of these metrics for all 50 models (10 downsampled datasets with 5-fold cross-validation each) at a classification probability cutoff of 0.5. The full range of values obtained are displayed in Figure 5.7D. It is evident that the models are robust to both downsampling and cross validation and the performance hovers around 0.75 for each metric (and 0.5 for MCC), indicating balanced performance in predicting both positive and negative classes. These results are robust even in the case where we only use one PRE definition and only consider genes that contain a PRE in their 3'UTR (Figure 5.8A-D).

We also tested the performance of these models on the Bohn et al. [134] RNAseq dataset that was not used to train the models. We first trained 10 different conditional random forest models on randomly downsampled subsets of our BruSeq data to balance the EFFECT and NOEFFECT classes. Using these 10 models, we then tested the performance on the Bohn et al. [134] steady state RNA dataset and measured the performance as shown in Figure 5.7E. Here, the performance on the trained Bru-Seq data is reported as the five-fold cross-validation performance for each of the 10 downsampled models. As expected, the performance drops on the new dataset, particularly in its ability to correctly classify EFFECT genes. However, a single probability cutoff (here chosen to be 0.5) for classification does not show the full performance of these models. To observe the overall performance of the models, we display precision-recall curves on both the Bru-Seq data on which the model was trained and the RNA-seq data for each of the 10 different models (Figure 5.7F). Here, the baseline is defined separately for each dataset as the overall class balance between the positive and negative class. A perfect model tends toward the upper right of the graph, and a poor model follows the dotted baseline for that dataset. Despite the differences in technique and biological implications between RNA-seq and Bru-Seq in determining PUM-mediated gene regulation, we find that the models trained on Bru-Seq are able to perform adequately well in predicting PUM-mediated regulation in RNA-seq data. We see similar performance when considering a single definition for a PRE and only considering genes that have a least one PRE in their 3'UTR (Figure 5.8A-D). However, there is still substantial room for improvement in predicting PUM-mediated gene regulation and the features we have included here are not sufficient to fully describe PUM-mediated gene regulation in human cells.

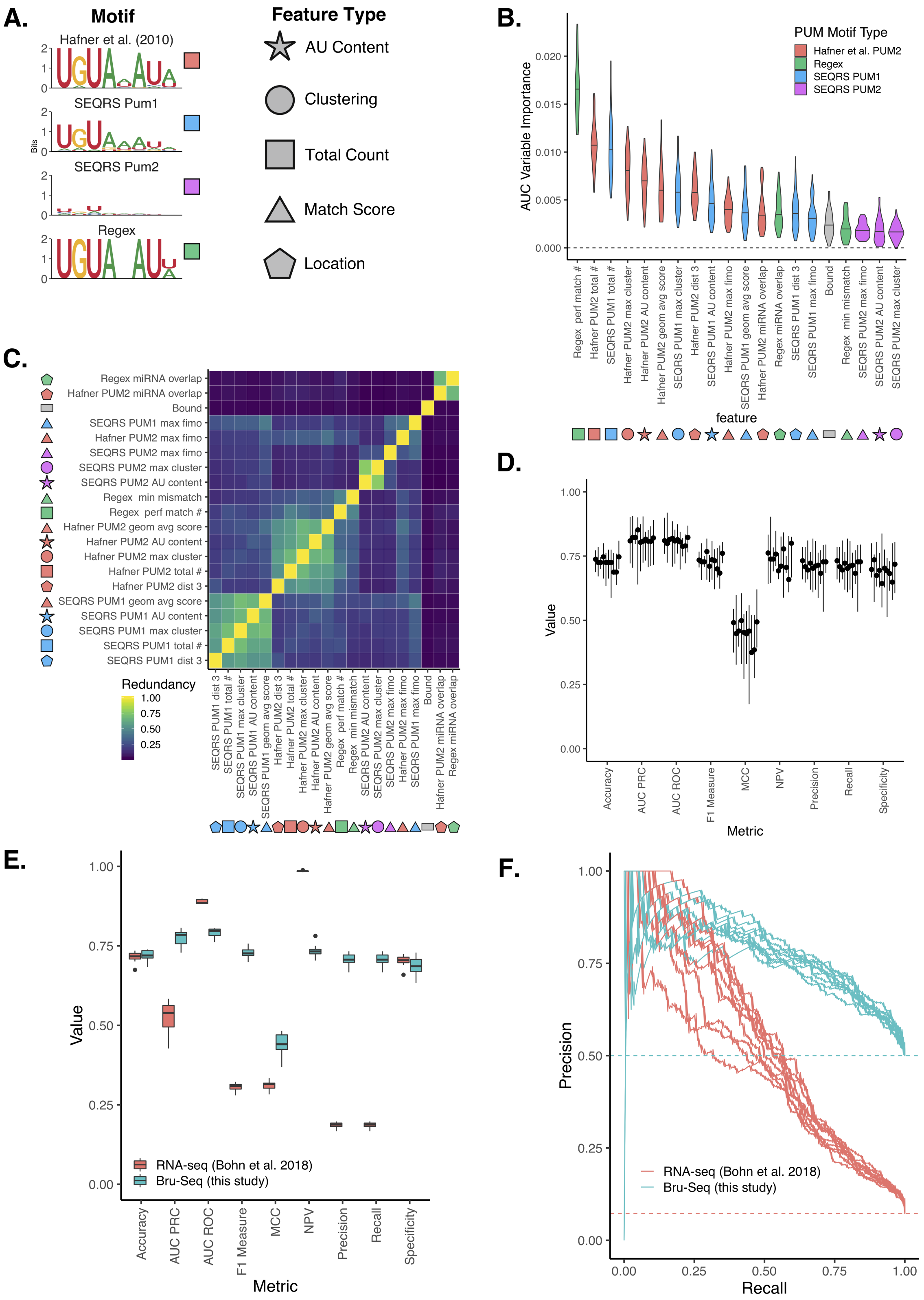

A.
Motif
Hafner et al. (2010)
SEQRS Pum1
SEQRS Pum2
Regex
Bits
Feature Type
AU Content
Clustering
Total Count
Match Score
Location
B.
AUC Variable Importance
PUM Motif Type
Hafner et al. PUM2
Regex
SEQRS PUM1
SEQRS PUM2
feature
Regex perf match #
Hafner PUM2 total #
SEQRS PUM1 total #
Hafner PUM2 max cluster
Hafner PUM2 AU content
Hafner PUM2 geom avg score
SEQRS PUM1 max cluster
Hafner PUM2 dist 3
SEQRS PUM1 AU content
Hafner PUM2 max fimo
SEQRS PUM1 geom avg score
Hafner PUM2 miRNA overlap
Regex miRNA overlap
SEQRS PUM1 dist 3
SEQRS PUM1 max fimo
Bound
Regex min mismatch
SEQRS PUM2 max fimo
SEQRS PUM2 AU content
SEQRS PUM2 max cluster
C.
Redundancy
Regex miRNA overlap
Hafner PUM2 miRNA overlap
Bound
SEQRS PUM1 max fimo
Hafner PUM2 max fimo
SEQRS PUM2 max fimo
SEQRS PUM2 max cluster
SEQRS PUM2 AU content
Regex min mismatch
Regex perf match #
Hafner PUM2 geom avg score
Hafner PUM2 AU content
Hafner PUM2 max cluster
Hafner PUM2 total #
Hafner PUM2 dist 3
SEQRS PUM1 geom avg score
SEQRS PUM1 AU content
SEQRS PUM1 max cluster
SEQRS PUM1 total #
SEQRS PUM1 dist 3
D.
Value
Metric
Accuracy
AUC PRC
AUC ROC
F1 Measure
MCC
NPV
Precision
Recall
Specificity
E.
Value
Metric
RNA-seq (Bohn et al. 2018)
Bru-Seq (this study)
F.
Precision
Recall
RNA-seq (Bohn et al. 2018)
Bru-Seq (this study)

Figure 5.7 *(previous page)*: Predicting Pum-mediated effect on decay using both sequence-based and experimental features. A) Motifs used to calculate features for machine learning. Shapes indicate the type of feature calculated, whereas colors indicate the motif used to calculate those features. Shapes filled in with the appropriate color are used to label features throughout the rest of the figure. B) Variable importance plot displaying the top twenty most important features, as determined by training a conditional random forest classifier on Pum decay data (see methods for details). Violin plots represent density from ten separate downsamplings of the majority class, each with five fold cross-validation. An AUC based variable importance measure is used as described in Janitza et al. [404]. C) Calculation of the redundancy in information between the top twenty most important variables, as determined in A. Redundancy is calculated in the information-theoretic sense (see Methods for details) where 1 is completely redundant information and 0 is no redundancy in information between the two variables. D) Cross-validation of conditional random forest classifier performance. Each boxplot represents a separate downsample of the majority, no pum-mediated effect class. Values for each boxplot represent the performance metric as calculated for each of five folds using a classification cutoff of 0.5. E) Performance of conditional random forest models on the steady state RNA data-set from [134]. Blue boxplots represent values from seperate downsamplings of the majority, no pum-mediated effect class used to train the model on the BruSeq data set. Red boxplots indicate values from testing each model on the Bohn et al. [134] steady-state RNA-seq data set. Metrics were calculated using a classification cutoff of 0.5. F) Precision Recall curves using the models in E. Each line represents one of ten conditional random forest models trained on separate down sampled sets of the entire BruSeq data set and tested on the steady state RNA data set.

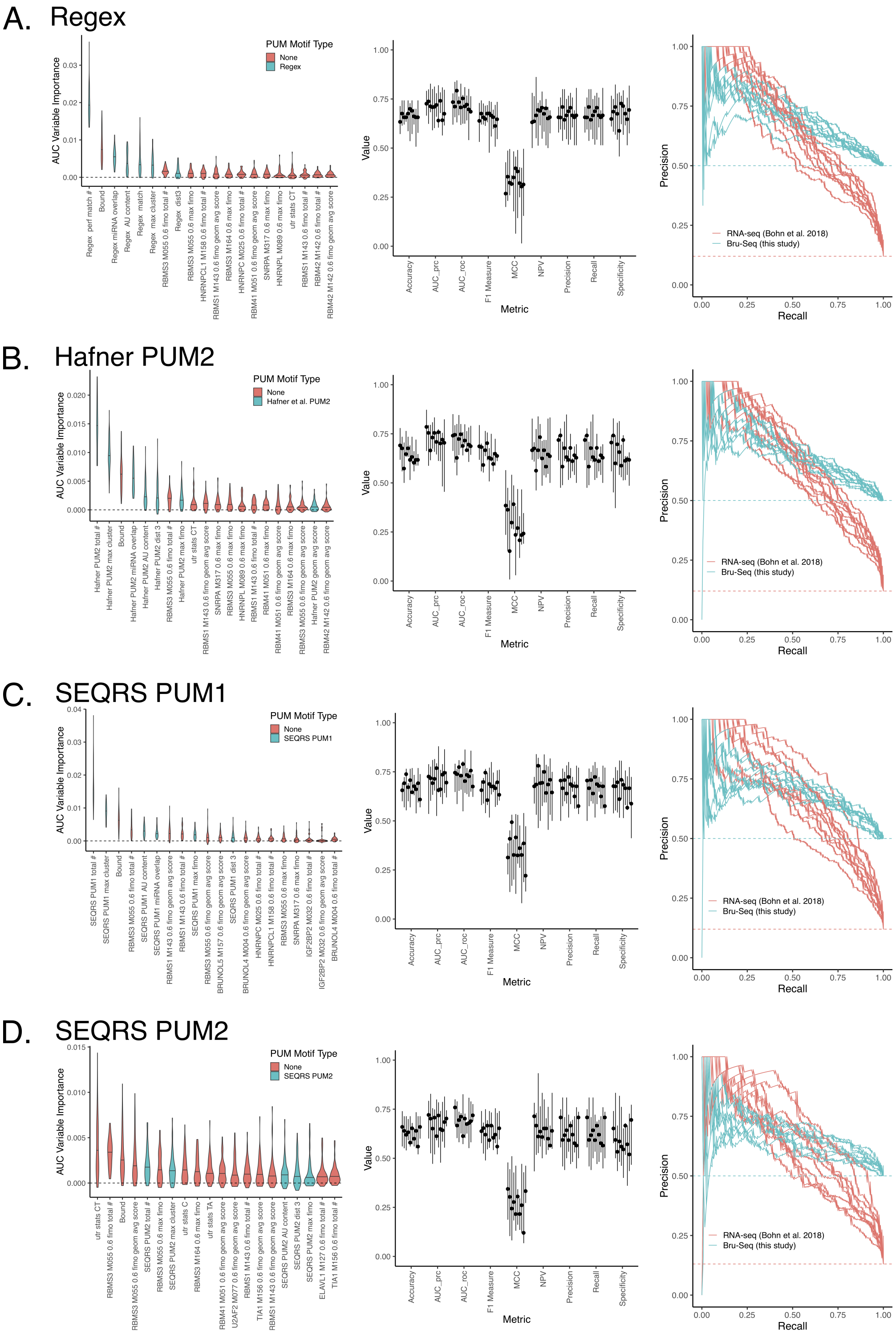

A.
Regex
PUM Motif Type
None
Regex
AUC Variable Importance
0.00
0.01
0.02
0.03
Regex perf match #
Bound
Regex miRNA overlap
Regex AU content
Regex match
Regex max cluster
RBMS3 M055 0.6 fimo total #
Regex dist3
RBMS3 M055 0.6 max fimo
HNRNPCL1 M158 0.6 fimo total #
RBMS1 M143 0.6 fimo geom avg score
RBMS3 M164 0.6 max fimo
HNRNPC M025 0.6 fimo total #
RBM41 M051 0.6 fimo geom avg score
SNRPA M317 0.6 max fimo
HNRNPL M089 0.6 max fimo
utr stats CT
RBMS1 M143 0.6 fimo total #
RBM42 M142 0.6 fimo total #
RBM42 M142 0.6 fimo geom avg score
Value
0.00
0.25
0.50
0.75
1.00
Accuracy
AUC_prc
AUC_roc
F1 Measure
MCC
NPV
Precision
Recall
Specificity
Metric
Precision
0.00
0.25
0.50
0.75
1.00
RNA-seq (Bohn et al. 2018)
Bru-Seq (this study)
0.00
0.25
0.50
0.75
1.00
Recall
B.
Hafner PUM2
PUM Motif Type
None
Hafner et al. PUM2
AUC Variable Importance
0.000
0.005
0.010
0.015
0.020
Hafner PUM2 total #
Hafner PUM2 max cluster
Bound
Hafner PUM2 miRNA overlap
Hafner PUM2 AU content
Hafner PUM2 dist 3
RBMS3 M055 0.6 fimo total #
Hafner PUM2 max fimo
utr stats CT
RBMS1 M143 0.6 fimo geom avg score
SNRPA M317 0.6 max fimo
RBMS3 M055 0.6 max fimo
HNRNPL M089 0.6 max fimo
RBMS1 M143 0.6 fimo total #
RBM41 M051 0.6 max fimo
RBM41 M051 0.6 fimo geom avg score
RBMS3 M164 0.6 max fimo
RBMS3 M055 0.6 fimo geom avg score
Hafner PUM2 geom avg score
RBM42 M142 0.6 fimo geom avg score
Value
0.00
0.25
0.50
0.75
1.00
Accuracy
AUC_prc
AUC_roc
F1 Measure
MCC
NPV
Precision
Recall
Specificity
Metric
Precision
0.00
0.25
0.50
0.75
1.00
RNA-seq (Bohn et al. 2018)
Bru-Seq (this study)
0.00
0.25
0.50
0.75
1.00
Recall
C.
SEQRS PUM1
PUM Motif Type
None
SEQRS PUM1
AUC Variable Importance
0.00
0.01
0.02
0.03
0.04
SEQRS PUM1 total #
SEQRS PUM1 max cluster
Bound
RBMS3 M055 0.6 fimo total #
SEQRS PUM1 AU content
SEQRS PUM1 miRNA overlap
RBMS1 M143 0.6 fimo geom avg score
RBMS1 M143 0.6 fimo total #
SEQRS PUM1 max fimo
RBMS3 M055 0.6 fimo geom avg score
BRUNOL5 M157 0.6 fimo geom avg score
SEQRS PUM1 dist 3
BRUNOL4 M004 0.6 fimo geom avg score
HNRNPC M025 0.6 fimo total #
HNRNPCL1 M158 0.6 fimo total #
RBMS3 M055 0.6 max fimo
SNRPA M317 0.6 max fimo
IGF2BP2 M032 0.6 fimo total #
IGF2BP2 M032 0.6 fimo geom avg score
BRUNOL4 M004 0.6 fimo total #
Value
0.00
0.25
0.50
0.75
1.00
Accuracy
AUC_prc
AUC_roc
F1 Measure
MCC
NPV
Precision
Recall
Specificity
Metric
Precision
0.00
0.25
0.50
0.75
1.00
RNA-seq (Bohn et al. 2018)
Bru-Seq (this study)
0.00
0.25
0.50
0.75
1.00
Recall
D.
SEQRS PUM2
PUM Motif Type
None
SEQRS PUM2
AUC Variable Importance
0.000
0.005
0.010
0.015
utr stats CT
RBMS3 M055 0.6 fimo total #
Bound
RBMS3 M055 0.6 fimo geom avg score
SEQRS PUM2 total #
RBMS3 M055 0.6 max fimo
SEQRS PUM2 max cluster
utr stats C
RBMS3 M164 0.6 max fimo
utr stats TA
RBM41 M051 0.6 fimo geom avg score
U2AF2 M077 0.6 fimo geom avg score
RBMS1 M143 0.6 fimo total #
TIA1 M156 0.6 fimo geom avg score
RBMS1 M143 0.6 fimo geom avg score
SEQRS PUM2 AU content
SEQRS PUM2 dist 3
SEQRS PUM2 max fimo
ELAVL1 M127 0.6 fimo total #
TIA1 M156 0.6 fimo total #
Value
0.00
0.25
0.50
0.75
1.00
Accuracy
AUC_prc
AUC_roc
F1 Measure
MCC
NPV
Precision
Recall
Specificity
Metric
Precision
0.00
0.25
0.50
0.75
1.00
RNA-seq (Bohn et al. 2018)
Bru-Seq (this study)
0.00
0.25
0.50
0.75
1.00
Recall

Figure 5.8 *(previous page)*: Predicting Pum-mediated effecf subset by motif. A) Conditional random forest models for the datasets considering only genes that had at least one match to the regex in a 3'UTR. PRE features only consider those around the regex. Panels are as in Figure 5.7B, D, and F. B) As in A), but for the Hafner et al. [292] PUM2 motif. C) As in A), but for the SEQRS PUM1 motif. D) As in A), but for the SEQRS PUM2 motif.

5.5 Discussion

Through the combination of our high-throughput probing of RNA decay and the mining of sequence information in the 3'UTRs of human transcripts, we were able to establish several general rules of PUM-mediated gene regulation in human embyronic stem cells.

5.5.1 Human PUM proteins control gene expression at the RNA level through mediating RNA stability

Previous studies have established that both PUM1 and PUM2 mediate the RNA stability of transcripts through recognition of a UGUANAUA PRE [131]. Transcriptome-wide measurements in PUM1 and PUM2 knockdown conditions have shown that hundreds of RNAs change in abundance, as measured using RNA-seq [134]. However, measurements of RNA abundance using RNA-seq only allow for determination of changes in steady-state RNA abundances and do not allow one to differentiate effects from changes in RNA stability versus changes in transcription rates. Through the use of metabolic labeling, we are able to differentiate between the effects of knocking down both PUM1 and PUM2 on initial transcription from the effects on RNA stability [308]. Our results indicate that perturbing the expression of human *PUM1* and *PUM2* has a widespread effect on the mRNA stability of many transcripts in HEK293 cells, but does not appear to perturb initial transcription rates in any meaningful way, as measured by our system. Rather than determine full decay rate constants for each transcript, which would have required the use of additional time points throughout the chase period of our experiment, we chose to determine relative changes in RNA stability using just two time points. The measurements obtained from these experiments cannot be interpreted on an absolute scale, but the rank order of stability measurements within the experiment is preserved, allowing us to determine the relative effects of PUM knockdown between any two genes [268]. Consistent with the changes in steady-state RNA levels determined under PUM knockdown conditions, we see transcripts that are both destabilized and stabilized. As expected, the number of genes that are stabilized under PUM knockdown is much higher than the number of genes that were destabilized, which is consistent with PUMs role in reducing the expression levels of target genes likely through the recruitment of the CCR4-NOT complex and subsequent destabilization of the transcript [131].

5.5.2 Rules for PUM-mediated activation are only partially clear

In contrast with the clear and robust effects of PUM on PUM-repressed transcripts, the mechanism for the rarer case of PUM-mediated stabilization remains unclear. Measurements using luminescent reporter assays have shown activation of a subset of predicted PUM-activated transcripts that is dependent on the presence of a PRE in the 3'UTR of the reporter [134]. Furthermore, direct binding of PUM1 or PUM2 to PREs present in the *FOXP1* 3'UTR has been reported to promote expression of the FOXP1 protein, an important regulator of the cell cycle in hematopoietic stem cells [372]. Conversely, when considering PAR-CLIP measurements of PUM2 occupancy at PREs for only the transcripts that were destabilized under PUM knockdown, we find inconclusive evidence for binding in targeted examples (Figure5.5C,D), and when considering the group as a whole separately from the stabilized transcripts (data not shown), mostly due to the low number of binding sites that can be considered. Furthermore, attempts to classify transcripts that were stabilized in PUM knockdown from those that were destabilized using random forest models with identical feature sets to those used in Figure 5.7 showed poor performance, possibly due to the small number of examples for transcripts that were destabilized under PUM knockdown. There is also the possibility that the destabilization of the transcripts under PUM knockdown are indirect effects mediated through another factor that PUM is directly regulating. We see at least one example of this in our data with the SLBP protein which is needed for the maturation of replication-dependent histones [397]. It is likely that *SLBP* is an example of a transcript directly regulated by PUM with a PRE in it's 3'UTR and *SLBP*s significant stabilization under PUM knockdown conditions. In addition, we observe that the set of genes in the nucleosome GO term are enriched in the set of genes that were destabilized under PUM knockdown. Taken together, this evidence suggests that PUM could be stabilizing these transcripts indirectly through the destabilization of the *SLBP* transcript and perturbation of SLBP-dependent processing of histone transcripts. Similar mechanisms may explain a substantial fraction of other "PUM-activated" targets, although direct stabilization by PUM in a subset of cases is also possible.

5.5.3 PUM1 and PUM2 have shared sequence preferences

The sequence preferences for both the full length PUM1 and PUM2 have been previously probed *in vivo* [135, 292, 379, 405] and the sequence preferences for the RNA binding domains of both PUM1 and PUM2 were probed *in vitro* [406, 407]. Each of these approaches and methodologies agree on a general preference for the UGUANAUA consensus motif for both PUM1 and PUM2, with subtle differences in the information content for the PWMs obtained from each technique, particularly at the 3' end of the PWM. Using SEQRS [381] on purified RNA binding domains for both PUM1 and PUM2, we find a strong preference for the UGUANAUA motif for PUM1 and, somewhat surprisingly, a much weaker preference for this motif for PUM2. However, when considering the enrichment of all possible 8mers, we see that the preferences for each RBD are highly correlated with a larger magnitude in enrichment for PUM1 RBD compared to PUM2 RBD. Our approach uses a random library of RNA sequences to determine RNA binding preferences and our analysis of PUM1 qualitatively agrees with previous *in vitro* approaches with randomized libraries [406]. However, using a curated library of sequences based on mutations from the consensus UGUANAUA motif, Jarmoskaite et al. [408] created a thermodynamic model for PUM2 binding that considers the effects of non-consecutive bases in target recognition, as opposed to our simpler model that only considers the frequency of occurrence of consecutive bases in a fully randomized library. Using this model, they show that RBDs from both PUM1 and PUM2 share nearly identical sequence preferences, which is in agreement with our strong correlation in enrichment between the two proteins. Thus, the weaker enrichment for the canonical PRE for PUM2 may be the result of an experimental artifact, or may come from the larger number of non-target sequences in our library compared to that of Jarmoskaite et al. [408].

When we considered the local sequence content and location of PREs, we found that PREs tend to be located towards the 3' end of the 3'UTR and have high local AU content. We are not the first to observe these properties, as Jiang et al. [136] also arrived to this conclusion by comparing the locations of shuffled PREs. However, we instead considered the locations of PREs in simulated sets of 3'UTRs that share similar trinucleotide content to that of the true set of 3'UTRs and this strengthens the claim that PREs are enriched in these areas more than one would expect by chance. Furthermore, we show that PREs in transcripts that had a significant change in RNA stability under PUM knockdown are closer to the 3' end of the 3'UTR and have higher AU content, suggesting a functional role for the location of PREs within the 3'UTR itself. The non-random selection of a PRE to occur towards the 3'end of the 3'UTR is consistent with a model where PUMs recruit the CCR4-NOT complex for de-adenylation of target sequences.

5.5.4 Human Pumilio proteins regulate genes involved in signaling pathways

When looking at the classes of genes that are stabilized under PUM knockdown, we find that many GO terms with evidence for direct repression by PUMs revolve around regulating signaling pathways mediated by proteins including kinases (GO:0018105), GEFs (GO:0005085), and receptor signaling (GO:0030177, GO:0048008). The role of mammalian Pumilio proteins in modulating signaling through controlling mRNA levels has been well established. In human testes, PUM2 is thought to interact with DAZL proteins to regulate germ-line development and many GTP-binding, receptor-associated, and GEF encoding-mRNAs are found among a list of targets that co-immunoprecipitate with both proteins [360]. Similarly, PUM1 has been shown to be important in mouse testis development through downregulation of many proteins involved in MAPK signaling and ultimate activation of p53 [361]. In fact, it has been argued that an ancestral function of the PUF family of proteins is to regulate the maintenance of stem cells and cells that behave in a stem cell-like manner through the down-regulation of kinases involved in critical signaling pathways [357]. Many studies looking at genes associated with PUM1 or PUM2 binding in mammalian cells tend to find similar sets of GO terms overlapping with PUM bound targets. Early RIP-Chip experiments with human PUM1 and PUM2 found that genes bound by both proteins belonged to GO terms associated with the Ras pathway, MAPK kinase cascade, PDGF signaling pathway, WNT signaling pathway, small GTPase-mediated signal transduction, and transcription factor activity, among others [135, 378]. More recent iCLIP experiments in mouse brains have found that mouse PUM1 and PUM2 bind genes associated with WNT signaling, regulation of MAP kinase activity, small GTPase-mediated signal transduction, and several categories related to neural development [365]. Similarly, changes in steady-state RNA abundance under both human PUM1 and human PUM2 knockdown identified several similar classes of genes including WNT signaling, GEF activity, NOTCH signaling, and PDF signaling [134]. Each of these categories is consistent with identified biological roles for mammalian PUMs. For example, mice lacking PUM1 and PUM2 have impaired learning and memory, as well as decreased neural stem cell proliferation and survival [365]. Further, human PUM1 haploinsufficiency is associated with developmental delay and ataxia [374]. Likewise, PUM2-deficient mice are more prone to chemically-induced seizures and have impaired nesting abilities [363], and mouse PUM2 regulates neuronal specification in cortical neurogenesis [366]. Our work shows that genes in these GO categories are modulated at the level of mRNA stability, likely through direct interaction of the human PUM proteins by recognition of PREs in the 3'UTR of transcripts.

In many ways, post-transcriptional regulation of proteins involved in signaling cascades is an ideal way to rapidly modulate those pathways. In contrast to the long distance between the site of regulation and the site of protein production involved in regulating a gene at the transcriptional level, post-transcriptional regulation allows for a dampening of expression levels directly where synthesis is occuring. Furthermore, gene regulation in the cytosol allows for the possibility of localized control of expression [409]. In fact, temporal and localized control of gene expression – important for proper development of the fly embryo – was exactly how the PUF family of proteins were initially discovered [358]. Given the emerging role for human PUM proteins in neuronal development and function, and the need for localized control of gene expression in neuronal tissue [410] it is conceivable that PUM proteins could be heavily involved in localized control of signaling pathways within the neuron.

5.5.5 Prediction of PUM-mediated regulation defines a set of general principles for an ideal PUM target site

Many attempts have been made to predict gene regulation by Pumilio proteins given sequence information about the possible targets. Previously, a biologically inspired model based strictly on the count of PREs within the 5'UTR, CDS, and 3'UTR was fit to steady state RNA levels [134]. In this model, the effects of having multiple PREs on a single transcript were found to be less than linear on the target response to PUM knockdown, which was interpreted to indicate that multiple PRE sites function to increase the odds of having a PUM bound and that a single PRE likely performs most of the functions needed for PUM-mediated regulation [134]. In this study we expanded the feature set of possible predictors for PUM-mediated activity and determine a set of rules that define an ideal functional PRE. Consistent with the Bohn et al. [134], we find that a simple count of PREs in the 3'UTR acts as the best predictor for PUM activity. However, surprisingly we find that the simple UGUA.AU[AU] regular expression outperforms more sophisticated PWM-based definitions from either *in vivo* and *in vitro* high throughput data. This may indicate that, although PUMs can bind PREs with mismatches from this consensus motif, the UGUANAUA may represent the "ideal" PRE for functional regulation. In fact, structural studies of human PUM1 and PUM2 have identified three different modes of binding between the nucleotide bases of the fifth base in the consensus motif and the amino acids of PUM repeats 4 and 5. [411] show that changes between these modes of binding do not alter PUM binding affinity, but could conceivably present different surfaces for effector proteins. Although our regular expression allows for any base at the fifth position, PUM repeats are modular [132] and it is conceivable that a similar mechanism could apply to other bases in the motif. Additionally this suggests that PUM binding to the UGUANAUA consensus motif could represent the ideal structure for PUMs interaction with effector molecules. We also find sequence features surrounding a PRE to be important in predicting PUM activity on a target. High AU content and position within the 3'UTR both appear to be important for predicting mammalian PUM regulation. Consistent with prior reports of cooperativity between PUM and miRNAs [135–137, 405], we find that a count of predicted miRNA sites near PREs helps predict PUM effect, with a higher number of miRNA sites near a PRE indicating a larger stabilization under PUM knockdown (Figure5.4A). It is possible that PUM could act to block or enhance miRNA function through direct interactions with the miRNA machinery or through local rearrangements of RNA secondary structure.

Secondary structure has been predicted to have an effect on many RBPs [386] and PUM has been shown to change secondary structure upon binding to facilitate miRNA interaction [137]. However, we found that *in silico* predictions of RNA secondary structure around PREs were not predictive of PUM function (Figure5.4C). Targeted regression models considering PRE count and structure only performed worse when structural information was added (data not shown). Recent studies have shown that structural probing experiments used in tandem with *in silico* folding algorithms vastly improve biological predictions based on structural information [412]. Similar methods may be needed to determine the role of secondary structure in PUM-mediated regulation. Alternatively, PUM proteins may be able to overcome RNA secondary structure in order to bind PREs; thus, secondary structure would have no bearing on PUM binding. There has also been a recent interest in the role of codon optimality in mRNA decay in human cells [413, 414]. Using relative codon bias as a measure of the rarity of codon usage in mRNAs [389], we find no evidence for differences in codon bias between targets undergoing PUM-mediated decay in our data set and those that do not (Figure 5.4B). Similarly, recent efforts have also identified m6A sites across the human transcriptome at single nucleotide resolution [415]; however, we find limited to no overlap between m6A sites and PREs (data not shown).

Despite our extensive efforts to predict PUM-mediated regulation using sequence information, there is still substantial room for improvement. Recent successes in Pumilio target prediction in *Drosophila* have come from characterizing binding partners of DmPum: Nos and Brat [416]. Nos binds together with DmPum to modulate the 5' sequence specificity of the Pum-Nos complex, thus introducing fine-tune control over Pum target recognition [417]. A recent study identified many new and previously known interacting partners for the human PUM1 and PUM2 proteins including DAZL, PABP, FMRP, miRISC, and members of the CCR4-NOT complex [418]. Like the Nos/DmPum example, these partners likely add an additional layer of information in the control of PUM-mediated gene regulation. Furthermore, the probing of RNA secondary structure *in vivo* may allow for better incorporation of secondary structural information into models of PUM-mediated regulation. Finally, we were unable to find determinants of PUM-mediated activation, an area that is rich for future targeted experiments.

5.6 Materials and Methods

5.6.1 Experimental methodology

5.6.1.1 SEQRS protein purification

Methods are reproduced here from Weidmann et al. [417]. Recombinant PUM1/2 were expressed in KRX *E. coli* cells (Promega) in 2xYT media with 25 μg/mL Kanamycin and 2mM $MgSO_4$ at 37°C to OD_{600} of 0.70.9, at which point protein expression was induced with 0.1% (w/v) rhamnose for 3hr. Cell pellets were washed with 50mM Tris-HCl, pH 8.0, 10% [w/v] sucrose and pelleted again. Pellets were suspended in 25mL of 50mM Tris-HCl pH 8.0, 0.5mM EDTA, 2mM $MgCl_2$, 150mM NaCl, 1mM DTT, 0.05% (v/v) Igepal CA-630, 1mM PMSF, 10 μg/ml aprotinin, 10 μg/ml pepstatin, and 10 μg/ml leupeptin. To lyse cells, lysozyme was added to a final concentration of 0.5 mg/mL and cells were incubated at 4°C for 30min with gentle rocking. $MgCl_2$ was increased to 7mM and DNase I (Roche) was added to 10 μg/mL, followed by incubation for 20 min. Lysates were cleared at 50,000×g for 30min at 4°C. Halo-tag containing proteins were purified using Magnetic HaloLink Resin (Promega) at 4°C. Beads were washed 3 times with 50mM Tris-HCl pH 8.0, 0.5mM EDTA, 2mM MgCl2, 1M NaCl, 1mM DTT, 0.5% [v/v] Igepal CA-630) and 3 times with Elution Buffer (50mM Tris-HCl, pH 7.6, 150mM NaCl, 1mM DTT, 20% [v/v] glycerol).

To confirm protein expression, beads were resuspended in Elution Buffer with 30 U of AcTEV protease (Invitrogen), cleavage proceeded for 24hr at 4°C, and beads were removed by centrifugation through a micro-spin column (Bio-Rad).

SEQRS was conducted as described in Campbell et al. [377] with minor modifications on the following samples: PUM1 RBD, PUM2 RBD

Magnetic Halolink beads (Promega) were used and the Pum test proteins remained covalently bound via N-terminal Halotag to the beads.

The initial RNA library was transcribed from 1μg of input dsDNA using the AmpliScribe T7-Flash Transcription Kit (Epicentre). 200 ng of DNase treated RNA library was added to 100 nM of Halo-tagged proteins immobilized onto magnetic resin (Promega). The volume of each binding reaction was 100μl in SEQRS buffer containing 200 ng yeast tRNA competitor and 0.1 units of RNase inhibitor (Promega). The samples were incubated for 30min at 22°C prior to magnetic capture of the protein-RNA complex. The binding reaction was aspirated and the beads were washed four times with 200μl of ice cold SEQRS buffer. After the final wash step, resin was suspended in elution buffer (1mM Tris pH 8.0) containing 10 pmol of the reverse transcription primer. Samples were heated to 65°C for 10min and then cooled on ice. A 5μl aliquot of the sample was added to a 10μl ImProm-II reverse transcription reaction (Promega). The ssDNA product was used as a template for 25 cycles of PCR using a 50μl GoTaq reaction (Promega).

5.6.1.2 Bru-Seq experimental procedure

Bru-Seq was conducted as described in Paulsen et al. [308] in HEK293 cells grown in the presence of siPUM1/2 or siNTC. Resulting cDNA libraries were sequenced using an Illumina HiSeq 2000 via the University of Michigan Sequencing core.

5.6.2 BruSeq Computational analysis

5.6.2.1 Modeling PUM-mediated RNA decay

Sequencing reads were aligned and processed according to Paulsen et al. [308] up to obtaining read counts for exons and introns for each gene and sample. Our experimental design resulted in four different replicates of siNTC (WT) and siPUM1/2 (PUMKD) conditions with two different time points each: t_{0hr} and t_{6hr}. For the t_{0hr} time points, read counts from both exons and introns were pooled for each gene. For the t_{6hr} time points, only read counts from exons were used. Read abundance was modeled using DESeq2 [340]. As described in Love et al. [340], DESeq2 models read count abundance K for gene i in sample j using the generalized linear model described below:

$$K_{ij} \sim NB(\mu_{ij}, \alpha_i) \tag{5.1}$$

Where α_i is a gene-specific dispersion parameter for gene i and μ_{ij} is defined by the following:

$$\mu_{ij} = s_j q_{ij} \tag{5.2}$$

Here, s_j is a sample specific size factor used to put read count abundances on the same scale between samples. Finally, $q_{i,j}$ is defined according to our design matrix:

$$\log_2(q_{i,j}) = \beta_0 + \beta_c c + \beta_t t + \beta_{tc} tc \tag{5.3}$$

Where, c is an indicator variable that is 0 when the sample is in condition WT and 1 when the sample is in condition PUMKD. Likewise, t is an indicator variable that is 0 when sample is in the 0 hour time point and 1 when the sample is in the 6 hour time point. We interpret the β_{tc} term to represent changes in RNA stability resulting specifically from the PUM KD condition. Similarly we interpret the β_c term to represent changes in initial transcription rates between the two conditions. Throughout the text, unless otherwise noted, we report β_{tc} normalized by the reported standard error for the coefficient, which amounts to the Wald statistic computed for that term by DESeq2. Thus, the Wald statistic for the interaction term is denoted as "RNA stability in PUM KD" throughout the text and is a unitless quantity.

5.6.2.2 Analysis of transcriptional vs. stability effects

To test for significant changes in transcription or stability, the Wald test statistic for the appropriate term – β_c for transcription and β_{tc} for stability – was calculated as described above. The Wald statistic was compared to a zero centered Normal distribution and a two-tailed p-value was calculated using statistical programming language R's pnorm function (n.b. this is virtually equivalent to the p-values calculated by the DEseq2 package for contrasts [340]). To test for a statistically significant lack of change in transcription or stability, the Wald statistic for the appropriate term was compared to a Normal distribution centered at the nearest boundary of a region of practical equivalence (ROPE) and a two-tailed p-value was calculated using R's pnorm function. The ROPE was defined as $-\log_2(1.75)$ – $\log_2(1.75)$ and was chosen to be within the range of fold expression change of a RnLuc reporter gene with between one and three PREs in its minimal 3'UTR [134]. Each p-value was FDR-corrected using the Benjamini-Hochberg procedure [354] and, for each term, the smaller of the two FDR-corrected p-values was reported. In order for a gene to be classified in the EFFECT, class the following conditions had to be met: 1. its change in stability q-value had to be smaller than its no change in stability q-value, 2. Its change in stability q-value had to pass a cutoff of 0.05 for statistical significance and 3. The original $\log_2$ fold-change value had to be outside the defined ROPE. In contrast, in order for a gene to be classified in the NOEFFECT class the following conditions had to be met: 1. it was not classified as an EFFECT gene, 2. its no change in stability q-value had to be smaller than its change in stability q-value, 3. its no change in stability q-value had to pass a cutoff of 0.05 for statistical significance and 4. The original $\log_2$ fold-change value had to be within the defined ROPE. Genes not passing the criteria for either the EFFECT or NOEFFECT groups are those for which we lack sufficient information to make any strong statement on the effects of PUM knockdown.

5.6.3 SEQRS Computational analysis

The raw sequencing data was processed according to Weidmann et al. [417] to contain the sequences within only the 20mer variable region. The 20mer variable regions of each read where then broken into all possible 8mer sequences using a sliding window, and raw counts for all possible 8mer abundances for each sequencing round for each protein were calculated using custom python scripts. Prior to 8mer abundance estimates, the adapter sequences were computationally added at the 5' and 3' end to account for any 8mers resulting from a combination of the invariant adapter sequences and the variable region. Only 8mers that had at least one base in the variable region were considered.

To determine position-weight matrices that best represented selection by the protein of interest for that round, we followed the approach of Jolma et al. [385] in the analysis of DNA binding proteins using SELEX. Briefly, a seed sequence is determined from the most abundant N-mer within that round. From this seed sequence, the abundance of each base at a given position was tallied when all other positions match the seed sequence. The PWM frequencies were determined by dividing each column of the resulting count matrix by its column sum. Unlike Jolma et al. [385] we do not include the correction for non-specific carryover of DNA from the previous cycle. The enrichment of a particular 8mer was calculated with the following equation:

$$E = \log_2 \left(\frac{\frac{c_{s,i}}{\sum_{i=1}^{N_s} c_{s,i}}}{\frac{c_{b,i}}{\sum_{i=1}^{N_b} c_{b,i}}} \right) \tag{5.4}$$

Where $c_{s,i}$ represents the count for 8mer i in sample s and $c_{b,i}$ represents the count for 8mer i in blank round where the input sequences were sampled. The DmPum data and corresponding blank sample was accessed from Weidmann et al. [417] and only the first five rounds were considered.

5.6.4 GO term analysis and iPAGE

GO term analysis was performed using the integrative pathway analysis of gene expression (iPAGE) software package [155]. Genes were discretized by the interaction term Wald test statistic into five-equally populated bins and iPAGE was run with default settings.

5.6.5 Determination of matching PREs

The full set of 3'UTRs for hg19 genome was downloaded using the TxDb.Hsapiens.UCSC.hg19.-knownGene, BSgenome.Hsapiens.UCSC.hg19, and GenomicFeatures R packages. Matches to a given PWM across all 3'UTRs were determined using the FIMO package with a uniform background using default cutoffs for reporting matches [419]. For PRE-centric figures, such as the heatmaps and violin plots in Figure 5.3 and Figure 5.4, each unique 3'UTR isoform is matched to its corresponding "RNA stability in PUM KD" value by gene name, and each feature's value is reported as the given summary statistic over a given 3'UTR isoform for that feature, as described in the section below (i.e., for AU content, the value reported is the maximum AU content around any given PRE within that 3'UTR isoform).

For *de novo* discovery of informative motifs in our Bru-seq dataset, we applied the finding informative regulatory elements (FIRE) software with default settings to each unique 3'UTR isoform matched to its "RNA stability in PUM KD" value and discretized into ten equally populated bins.

To calculate the location and AU content of PREs in randomly generated sets of the 3'UTRs, a third order Markov model was trained on the empirical set of unique 3'UTR isoforms from the hg19 genome. One thousand randomly simulated sets of 3'UTRs – the same length as the empirical set of 3'UTRs – was generated then generated using custom python scripts. For each of the thousand simulated sets of 3'UTRs, the fifth round SEQRS PUM1 (Figure 5.2A) was used to search for matches using FIMO as described above. Here each individual PRE was considered in the calculation of the kernel density plots shown in Figure 5.3.

To determine the PAR-CLIP read coverage at identified PRE sites in the set of known unique 3'UTR isoforms, raw reads were downloaded from SRA with accession numbers SRR048967 and SRR048968. Raw fastq files were processed with trimmomatic [178] and cutadapt [177] to remove low quality reads and illumina adapters. Processed reads were aligned to the hg19 genome using the STAR aligner with default parameters [331]. Read coverage and T to C mutations were determined for reads within 20 bp of each PRE in each unique 3'UTR isoform for both EFFECT and NOEFFECT genes, individually, using custom python scripts. Coverage over all PREs was aligned and the bottom and top 5% of read coverage at each position was removed from the average calculation. Error bars were determined by sampling with replacement read coverage from individual PREs in each group separately.

5.6.6 Determination of PRE clustering

To determine whether the PREs cluster together more than would be expected by chance, we determined the ratio of the empirically observed frequency of PUM sites within all possible 100 bp windows of 3'UTRs with a least 1 PRE in them to a Poisson model with the rate parameter, λ, set to the average count of PREs within all 100 bp windows. 95% confidence intervals were determined by sampling the empirical distribution of PRE counts within all windows with replacement.

5.6.7 Predicting PUM-mediated regulation using conditional random forest models

In order to predict the PUM-mediated regulation on a given transcript, we used conditional random forest models from the cforest function from the party R package [420–422]. Binary classification models were trained using default settings with no parameter tuning on the Bru-Seq EFFECT and NOEFFECT classes and a permutation-based AUC variable importance metric was calculated for each individual model [404]. Due to the large class imbalance, ten separate datasets were generated from the full dataset, where the majority NOEFFECT class was randomly downsampled to match the EFFECT class. Within each of the ten datasets, five-fold cross validation was performed to assess performance and detect overtraining. Final models were generated using the ten downsampled datasets without cross-validation and performance was tested on the RNA-seq dataset from Bohn et al. [134]. Precision-recall plots were calculated using the PRROC package based on the methodology of Davis and Goadrich [423].

5.6.7.1 Calculation of features associated with a PWM

For each of the features described in this section, the feature was first calculated individually for each unique 3'UTR isoform. Values for each isoform were combined by taking the average of the value for that feature and isoform weighted by the number of isoforms that shared that unique 3'UTR in the full set of annotated 3'UTRs in the hg19 genome. For features ending in "fimo_best_bygene_max_fimo", the maximum FIMO match score for each unique 3'UTR isoform for that PWM was calculated by setting the p-value cutoff threshold in FIMO to 1.1, thereby allowing FIMO to consider every possible match for a given sequence. The maximum match score for each sequence was reported for each unique 3'UTR isoform. For features ending in “fimo_best_bygene_total_num”, the total number of matching sites for a given unique 3’UTR isoform was calculated as described above in the “Determination of matching PREs” section. For each sequence, the geometric average of FIMO scores for each matching PRE was calculated and reported in the “fimo_bygene_geom_avg_score”. The maximum match score, geometric average match score, and total match number was calculated for the SEQRS PUM1 round 5 PWM, SEQRS PUM2 round 5 PWM, Hafner et al. [292] PUM2 PWM, and each of the PWMs for human RBPs found in the CISBP-RNA database [424].

For PREs, the shortest distance to the 3’UTR for any given PRE is converted to normalized coordinates (i.e., 0.0 is the 5’ end and 1.0 is the 3’ end) and reported in the “fimo_best_bygene_dist_3”. For “fimo_bygene_at_content” the largest percentage AT content in a 100 bp window surrounding any PRE within a given sequence was reported. Similarly for “fimo_bygene_max_cluster”, the maximum number of full PRE sites within a sliding of 100 bp was calculated. For both of these features, windows were truncated at the 3’ and 5’ ends of the sequence.

5.6.7.2 Calculation of *in silico* basepairing probabilities for PREs

For each identified PRE, the probability of the given PRE being base-paired within predicted secondary structure was calculated using RNAfold [425] by calculating the ensemble free energy of an unconstrained sequence F_u of 50 bp flanking each side of a given PRE and the ensemble free energy of a constrained sequence where no base within the PRE is allowed to form a base pair F_c. The probability of the PRE being constrained from base-pairing can be calculated using:

$$P_c = \exp\left(\frac{(F_u - F_c)}{RT}\right) \tag{5.5}$$

Where T is the temperature (set to physiological temperature, 310.15K), and R is the gas constant (set to 0.00198 kcal K^{-1} mol^{-1}). Thus the probability of any given PRE being unpaired is P_c. We define two features associated with P_c for each PRE in a given 3’UTR isoform. “_avgprob_unpaired” is the average P_c of all the PREs within a given 3’UTR and “_maxprob_unpaired” is the maximum P_c of all the PREs within a given 3’UTR. Values for each isoform were combined into gene level estimates, as described above.

5.6.7.3 Calculation of information redundancy between features

In order to calculate the information redundancy between features, each feature was discretized into ten equally populated bins. The redundancy between feature 1 (F_1) and feature 2 (F_2) was calculated with the following equation:

$$R = \frac{2 \times I(F_1; F_2)}{(H(F_1) + H(F_2))} \tag{5.6}$$

Where H is the entropy of a given vector X of discrete values, as defined below:

$$H(X) = -\sum_{x \in X} P(x) \log_2(P(x)) \tag{5.7}$$

And the mutual information $I(X; Y)$ of vectors X and Y of discrete values is defined as:

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} P(x, y) \log \left(\frac{P(x, y)}{P(x)P(y)} \right) \tag{5.8}$$

5.6.7.4 Determination of EFFECT and NOEFFECT classes for RNA-seq data

RNA-seq data was obtained from Bohn et al. [134] and a gene was only considered if the FPKM for both the PUM1/2 knockdown condition and the siNTC condition were greater than 5. Genes that passed this cutoff and that were considered to have statistically significant differential expression in the original analysis were considered EFFECT genes. Genes that passed the cutoff and were not considered to have statistically significant differential expression were considered NOEFFECT genes.

5.7 Acknowledgments

This work was supported in part by the National Institute of General Medical Sciences, National Institutes of Health grant R35 GM128637 to P.L.F. and grant R01 GM105707 to A.C.G. Additionally M.B.W. is supported by an National Science Foundation Graduate Research Fellowship DGE1256260. Additional funds come from NIH funding awarded to Mats Ljungman and Zak Campbell.

5.7.1 Author contributions

Lead author: Mike Wolfe (Bioinformatics/Computational analysis, writing, creating all figures); BruChase: Trista Schagat (RNAi, RNA labeling and purification), Aaron Goldstrohm (Funding and Concept), Michelle Paulsen (BrU RNA Seq), Brian Magnuson (initial BruChase data analysis), Mats Ljungman (Funding and concept); PUM Protein Purification for SEQRS: Daeyoon Park, rotation student in Aaron Goldstrohm Lab, Chi Zhang and Zak Campbell (SEQRS and data analysis, funding)

CHAPTER 6

Conclusion: Structure and context in regulatory control

Through the use of high-throughput sequencing technology and computational analysis, I have characterized the regulatory network of a transcriptional regulator (Chapter 2), gained insight into promoter-independent mechanisms of transcriptional control (Chapter 3), made recommendations for measuring RNA decay (Chapter 4), and used experimental measurements to predict mRNA decay as mediated by a post-transcriptional regulator (Chapter 5). Each of these projects has advanced the understanding of mechanisms of gene regulation; however, each has also left unanswered questions. Of particular interest is the importance of contextual information in each of these systems, whether that be the existence of adjacent and condition-specific binding sites for Lrp, the binding signal from key architectural proteins in controlling location-dependent expression, or the relevance of sequence context outside the primary binding site for PUM. Here, I summarize the key findings for each project. I also suggest additional hypotheses that are generated by this work that could be tested by future experiments, some of which are suggested below.

6.1 Lrp at the interface between global regulator and architectural protein

In their 2002 review article, Martıńez-Antonio and Collado-Vides [49] refined a set of criteria to establish the identity of global regulators in bacterial species. Even then, Lrp was identified as a global regulator, but the evidence we have accumulated has solidified Lrp's place among the global regulators in *E. coli*. We have shown that Lrp regulates, either directly or indirectly, a third of the transcripts in *E. coli*. In addition, we have found that Lrp's regulon changes in a condition-specific manner and its effects are attenuated in conditions with high concentrations of branched chain amino acids. We have also found that it regulates more traditional transcription factors, such as LrhA and CysB, and tends to coregulate with promoters that are bound by the σ^{54}-RNAP holoenzyme. Each of these findings match the critera for a bon-a-fide global regulator. However, unexpectedly, we also found that changes in the types of genes that Lrp regulates do not seem to only depend on the presence of branched chain amino acids, but also seem largely influenced by growth stage. This is also reflected in Lrp's sequence specificity, which shifts from a general preference for high A/T content in early growth phases to slightly more sequence specific binding in late growth phase regardless of the concentration of branched chain amino acids in the media.

In addition, the presence of adjacent, condition-specific peaks in our data is intriguing. It is enticing to suggest that these peaks are the result of two octamers of Lrp binding on either side of a promoter and looping out the intervening DNA, similar to the mechanism behind the Lac repressor [43]. However, with Lrp's abundant binding across the genome, as identified in our data and in others [49, 58], it is possible that Lrp could mediate interactions between disparate regions of the nucleoid through the formation of hexadecamers from octamers at distal sites. This would allow regions of the DNA that are separated by a large distance in linear space to be localized in three dimensional space. Thus, one basis for regulation by Lrp could be the condition dependent formation of localized DNA, not dissimilar to the model of localized "transcription factories" that have been proposed to be involved in transcriptional regulation [426]. In fact, changes in genomic structure preceding changes in steady-state RNA levels have been shown to occur in the differentiation of eukaryotic stem cells [427], suggesting that the changes in genomic structure are critical for gene regulation in some instances. Furthermore, the "poised" Lrp binding sites we identify may be octamers that are in position to mediate these contacts but are not in the correct conformation to form higher order Lrp oligomers. An ideal experiment to probe Lrp-dependent interactions between regions of the DNA would be through the use of Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) [428], using our highly specific monoclonal antibody to identify contact between distant regions of the genome that may be mediated through Lrp. A pilot experiment could be done using qPCR on ligated fragments with a known Lrp binding site under conditions where it is differentially regulated.

Additional questions abound for Lrp. Originally three different modes of Lrp mediated activation or repression regulation were introduced by Cho et al. [58]: 1. independent (no Leucine dependence), 2. concerted (Leucine enhances Lrp's effect), 3. reciprocal (Leucine diminishes Lrp's effect). Analysis of our data to categorize Lrp-mediated gene regulation into these groups by Christine Ziegler of the Freddolino lab has indicated that the vast majority of the regulation seen in our data falls under the reciprocal mode. Furthermore, studies by other groups have indicated the Lrp activity is modulated by more than just Leucine [57]. Both of these observations are consistent with a role for Lrp as a general nutrient sensor; however, the mechanistic details of how Lrp modulates gene expression of its targets, and particularly, how Lrp activates some targets and represses others, is still poorly understood. My attempts to use machine learning models with sequence-based features to try to differentiate between activated and repressed genes performed poorly, indicating that we likely do not have the correct or sufficient set of features to predict gene regulation by Lrp. Furthermore, mechanistic details of gene regulation by Lrp are lacking. What conformational changes occur when Lrp is bound by a co-regulator? What protein partners, if any, does Lrp interact with? What is the structural basis for Lrp's changes in sequence specificity over growth phases? Each of these questions are actively being pursued by members of the Freddolino lab.

6.2 The role of bacterial nucleoid structure in transcriptional regulation

Our analysis of the effects of position on the steady state RNA levels of an identical reporter represent the highest resolution look at position-effects on gene expression to date. We identify a critical role for two nucleoid associated proteins, Fis and H-NS, in identifying regions that facilitate and inhibit gene expression, respectively. Of particular interest is the association between the proximity to a ribosomal RNA and higher transcriptional propensity. Previous studies have demonstrated that ribosomal RNA operons cluster together within the bacterial nucleoid [206]. Thus, several models for the effect of proximity to the *rrn* operons on transcriptional activation immediately come to mind:

1. The reporters' co-localization with the ribosomal RNA operons in an area of assumed high RNAP concentration could result in higher spontaneous initiation rates for said reporter.

2. Transcriptional read through at the end of the highly transcribed *rrn* operons could result in higher transcription of the reporter downstream of the *rrn* operon.

3. Negative supercoiling upstream of an *rrn* operon, induced by the action of RNAP elongation at the ribosomal RNAs could facilate initiation of RNAP at the reporter.

Models 2 and 3 can be ruled out as the *sole* effect on expression of the reporter from a single experiment. Using the targeted integration of the reporter construct to replace a single *rrn* operon, higher protein expression is observed (through measurement of fluorescence) than a targeted integration upstream or downstream of the *rrn* operon (Scott Scholz, personal communication). This suggests that effects coming from read-through (model 2) or the propogation of supercoiling (model 3) are likely not playing a large role in the associations we are seeing, as the *rrn* operon is no longer there to mediate those effects in this experiment. However, this does not rule out model 1 as a possible mediator of the proximity effect to *rrn*. If model 1 is responsible for the effect that is seen, then the resulting co-localization for the *rrn* operons must occur in an *rrn* sequence-independent manner. Thus, an alternative interpretation of this experiment could be that the *rrn* operons themselves exist in a region of the bacterial chromosome that is naturally permissive to gene expression. Both interpretations bring up questions about the actual structural nature of the bacterial chromosome itself and its impact on transcription.

Unlike the regularly structured *C. crescentus* genome [429], the *E. coli* genome structure has been difficult to capture. The most recent structure of the *E. coli* genome [198] differs substantially both from earlier reports [430], and from data obtained from experiments in the Freddolino lab. It is possible that this experiment performed in *E. coli* may be more sensitive to the exact experimental protocols used than other bacterial species (Peter Freddolino & Grace Kroner, personal communication). My analysis of the overlap between the Lioy et al. [198] data and our transcriptional propensity map has suggested that only structural boundaries associated with a *rrn* operon seem to be correlated with changes in transcriptional propensity. However, regions of the chromosome that are tied up in high densities of protein occupancy (tsEPODS, Vora et al. [78]) tend to be highly repressed, consistent with H-NS occupancy at those sites. The interaction between local compaction and higher order chromosomal structure in *E. coli* remains to be elucidated. Additionally, the role of DNA supercoiling is gaining wide appreciation as a factor in controlling transcription [431]. Another NAP, the Histone-like protein from *E. coli* strain U93 (HU), is thought to be involved in constraining negative supercoils in bacterial cells [432]. We found limited evidence for a correlation between HU ChIP-seq signal and our transcriptional propensity signal; however, the ChIP-seq signal is highly non-specific and may not reflect the biological function of the HU protein [230]. HU is highly conserved [433], highly expressed [76], and bound non-specifically across the nucleoid [77]. Although it does seem to impact the expression of genes involved in dealing with stress conditions [434], the details of how HU regulates bacterial transcription are poorly understood. Given its NAP status and putative function in constraining supercoils, it is possible that HU is involved at the confluence of chromatin structure, DNA supercoiling, and RNAP recruitment. Building off the intellectual foundation I have created in my graduate work, my post-doctoral work will focus on this mysterious protein and its role in transcriptional control. Taken together, the full nature of the interaction between chromatin structure, NAPs, and their impact on transcription suggests a bacterial heterochromatin/euchromatin analogue and will be a fruitful area for further inquiry.

6.3 RNA secondary structure and post-transcriptional regulation

The Pumilio family of proteins are perhaps the most well-studied family of RNA binding proteins involved in post-transcriptional regulation. Our work characterizing the effects of the human members of the family—PUM1 and PUM2—on mRNA decay shows their widespread impact on modulating mRNA decay, particularly for transcripts involved in neuronal development and protein-mediated signaling pathways. Despite the decent performance using sequence-based features to predict regulation by the Pumilio proteins, we still cannot fully predict the magnitude of the impact resulting from Pumilio-mediated regulation. Our analyses also indicated that the use of *in silico* models of RNA secondary structure were not useful in predicting PUM-mediated repression. This is consistent with recent analyses of secondary structural predictions when compared to experimentally determined accessibility determined through mid-throughput structural probing assays. Mustoe et al. [412] used selective 2'-hydroxyl acylation analyzed by primer extension and sequencing (SHAPE-seq) to determine the secondary structure of roughly 200 transcripts in *E. coli* and found that the SHAPE-informed structures—particularly when performed within cells rather than in cell extracts —vastly outperformed purely *in silico* structures in predicting translational efficiency from the structure of ribosome binding sites using a kinetic model. Additionally, secondary structure has been shown to play a critical role in the binding of splicing factors and can be used to engineer alternative splicing by creating mutants to increase secondary structure and block binding [387]. In the case of PUM proteins, we find that both local AU content and adjacent miRNA sites are important in helping to predict the effect of PUM-mediated gene expression. Previous studies have also highlighted the relationship between miRNAs and Pumilio proteins, including specific examples of PUM binding serving to rearrange secondary structure to modulate miRNA-mediated repression [135–137]. Thus, the ability to accurately determine RNA secondary structure *in vivo* and its role in PUM-mediated repression represents an area for rich exploration.

Consistent with previous high-throughput identifications of targets of PUM-mediated gene expression, we also find a subset of genes that are activated by PUM [134]. Recently, human PUMs have been implicated in the activation of FOXP1, a critical driver of hematopoeitic stem cell growth [372]. This, together with measurements of PUM-mediated activation of reporter genes with the 3'UTRs of targets identified in a high throughput screen [134], have demonstrated that, under some cases, PUM proteins can activate transcripts. The mechanisms behind this activation remain completely unknown. My attempts to separate PUM-mediated activation from PUM-mediated repression using machine-learning methods with sequence-based features were not successful and the PREs associated with PUM-mediated stabilization in our study appeared to be less bound than the transcripts that were destabilized in a PUM-specific manner. However, we lacked statistical power to say this with certainty due to the low number of transcripts with PUM-mediated stabilization. Work with *Drosophila* Pum has indicated that the N-terminal domains of Pum can mediate repression of a reporter transcript independent of the RNA-binding domain when tethered to a reporter transcript [435]. Furthermore, large protein-protein interaction screens have identified many putative partners for the human Pumilio proteins, including those involved in mediating the RNA-decay process [418]. Taken together, the interaction of PUM with protein partners, perhaps through the N-terminal domain, may determine whether PUM-mediated activation or PUM-mediated repression occurs for a particular transcript. However, interactions between the N-terminal domain and protein partners, if any, for the Human PUM proteins are poorly understood.

The use of reporter systems, where the sequence content of the 3'UTR can be carefully controlled, enables targeted probing of the particular features we identified as important for PUM-mediated control of mRNA decay. Future experiments may use such systems to interrogate the contributions of some of the contextual features we identify here, such as AU content and PRE location. Furthermore, the concentration of putative co-regulators could be easily manipulated in such a system, thereby allowing for the determination of their functional effects in PUM-mediated gene regulation. These type of experiments may be particularly useful when considering the interaction between PUMs and miRNAs. Functional assays coupled with measurements of RNA secondary structure may illuminate the effects of PUM on miRNA regulation or vice-versa. It seems likely that PUM could be acting to either occlude miRNA binding sites or allow for secondary structural rearrangements that facilitate regulation by miRNAs.

6.4 The interface between protein and nucleic acids

Central to the discussion of gene regulation has been the interaction between proteins and nucleic acids. In addition, determining the locations of any given regulator's binding on the nucleic acid sequence is critical to understanding and predicting that regulator's effect on gene expression. Traditionally and throughout this dissertation, position weight matrices (PWMs) have been used to model a protein's affinity for a particular nucleic acid sequence [436]. However, PWMs treat the Watson-Crick identity of each base in a particular sequence as an independent and additive measure for sequence preference, which can limit their utility. More recently, biophysically-motivated models have had greater success in predicting protein binding to DNA, as have those that consider higher order interactions between each base, but for many DNA binding proteins, there is still substantial room for improvement in predicting and understanding the determinants of protein binding [437]. Recent studies have shown that some DNA binding proteins prefer more general structural elements, such as a narrow minor groove width, rather than a simple identification of the Watson-Crick face of the bases themselves [438]. Furthermore, an algorithm has been developed to predict general DNA shape features from sequence information alone [439]. Application of this algorithm in a machine learning context has led to improvements in predicting the location of ChIP-Seq peaks for many DNA binding proteins when supplementing sequence information from a PWM with shape information around a PWM for a given protein [440]. Recently, I have developed an algorithm that uses structural features to predict transcription factor binding sites and preliminary analyses have indicated that these structural "motifs" outperform traditional PWMs in some cases. Similarly, algorithms have been developed to predict secondary structure motifs for RNA binding proteins [441]. Key to each of the systems I have presented in my dissertation is the prediction of a particular regulators' binding site and its effect once bound. Algorithms that take into account the structure and context of these regulators either at a local level (such as the local secondary structure of DNA or RNA) or the global level (such as the nucleoid structure and associated stretchs of high protein occupancy) will improve our ability to model and predict biological systems.

In fact, the interaction between protein and nucleic acids is so fundamental to biology that it may have been the origin of life. Crucial to the center of information for all biological systems is the ribosome, the translation apparatus that allows for the nucleic acid language to be converted into the amino acid language. Work by Bowman et al. [442] has indicated that ribosomal RNA structure has slowly accumulated on top of a common rRNA core over the history of evolution all the way back to the last common ancestor. Thus, the ribosome represents a molecule fossil and the common core is a frozen snapshot of the beginning of life. Bowman et al. [442] also argue that protein and nucleic acid co-evolved together through a chemical evolution that preceded traditional Darwinian evolution. Lending support to this model are experiments by Li et al. [443] that showed that enzymes encoded by the same DNA sequence, but on opposite strands, are capable of replicating the enzymatic reactions of the two different classes of tRNA synthetases (the enzymes responsible for charging the correct tRNA with the correct amino acid). They also argue that these early synthetases conduct spontaneous protein synthesis quickly enough to be a minimal working system for the nucleic acid and protein codes to simultaneously co-evolve. Another paper by the same group also argues that the properties of the amino acids can be predicted based off sequence elements in the acceptor stem or anticodon, thereby suggesting an early system involving tRNA acceptor stems together with their minimal working synthetases as a substrate for developing the genetic code [444]. Taken together, this work suggests that the interactions between proteins and nucleic acids have always formed the fundamental basis for life, forever intertwined in an intricate molecular dance of ever-increasing complexity.

BIBLIOGRAPHY

[1] Siddhartha Mukherjee. *The Gene: An Intimate History*. Scribner, 1230 Avenue of the Americas New York, NY 10020, 2016. ISBN 978-1-4767-3350-0.

[2] Aristotle, W. D. (William David) Ross, and J. A. (John Alexander) Smith. *The Works of Aristotle Translated into English*. Oxford, Clarendon Press, 1912.

[3] Gregor Mendel. Experiments in Plant Hybridization (1865). *Electronic Scholarly Publishing Project*, page 41, 1996.

[4] Thomas Hunt Morgan. *The Mechanism of Mendelian Heredity*. H. Holt, 1915.

[5] Fred Griffith. The Significance of Pneumococcal Types. *Epidemiology & Infection*, 27(2): 113–159, January 1928. ISSN 0022-1724. doi: 10.1017/S0022172400031879.

[6] Oswald T. Avery, Colin M. MacLeod, and Maclyn McCarty. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types: Induction of Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii. *Journal of Experimental Medicine*, 79(2):137–158, February 1944. ISSN 0022-1007, 1540-9538. doi: 10.1084/jem.79.2.137.

[7] A. D. Hershey and Martha Chase. Independent Functions of Viral Protein and Nucleic Acid in Growth of Bacteriophage. *The Journal of General Physiology*, 36(1):39–56, September 1952. ISSN 0022-1295, 1540-7748. doi: 10.1085/jgp.36.1.39.

[8] J. D. Watson and F. H. C. Crick. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, 171(4356):737, April 1953. ISSN 1476-4687. doi: 10.1038/171737a0.

[9] Matthew Meselson and Franklin W. Stahl. The replication of DNA in Escherichia coli. *Proceedings of the National Academy of Sciences*, 44(7):671–682, July 1958. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.44.7.671.

[10] J. D. Watson and F. H. C. Crick. Genetical Implications of the Structure of Deoxyribonucleic Acid. *JAMA*, 269(15):1967–1969, April 1993. ISSN 0098-7484. doi: 10.1001/jama.1993.03500150079031.

[11] G. W. Beadle and E. L. Tatum. Genetic Control of Biochemical Reactions in Neurospora. *Proceedings of the National Academy of Sciences*, 27(11):499–506, November 1941. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.27.11.499.

[12] Sydney Brenner, Francois Jacob, and Matthew Meselson. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature*, 190:576–581, May 1960.

[13] Francois Gros, H. Hiatt, Walter Gilbert, C. G. Kurland, R. W. Risebrough, and J. D. Watson. Unstable Ribonucleic Acid Revealed by Pulse Labelling of Escherichia Coli. *Nature*, 190 (4776):581, May 1961. ISSN 1476-4687. doi: 10.1038/190581a0.

[14] F. H. C. Crick, J. S. Griffith, and L. E. Orgel. Codes Without Commas. *Proceedings of the National Academy of Sciences*, 43(5):416–421, May 1957. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.43.5.416.

[15] S. Brenner. On the Impossibility of All Overlapping Triplet Codes in Information Transfer from Nucleic Acid to Proteins. *Proceedings of the National Academy of Sciences*, 43(8): 687–694, August 1957. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.43.8.687.

[16] F. H. C. Crick, Leslie Barnett, S. Brenner, and R. J. Watts-Tobin. General Nature of the Genetic Code for Proteins. *Nature*, 192(4809):1227, December 1961. ISSN 1476-4687. doi: 10.1038/1921227a0.

[17] Marshall W. Nirenberg and J. Heinrich Matthaei. The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. *Proceedings of the National Academy of Sciences*, 47(10):1588–1602, October 1961. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.47.10.1588.

[18] C. Thomas Caskey and Philip Leder. The RNA code: Nature's Rosetta Stone. *Proceedings of the National Academy of Sciences*, 111(16):5758–5759, April 2014. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1404819111.

[19] Peter Lengyel, Joseph F. Speyer, and Severo Ochoa. Synthetic Polynucleotides and the Amino Acid Code. *Proceedings of the National Academy of Sciences*, 47(12):1936–1942, December 1961. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.47.12.1936.

[20] Marshall Nirenberg and Philip Leder. RNA Codewords and Protein Synthesis: The Effect of Trinucleotides upon the Binding of sRNA to Ribosomes. *Science*, 145(3639):1399–1407, September 1964. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.145.3639.1399.

[21] C. T. Caskey, R. Tompkins, E. Scolnick, T. Caryk, and M. Nirenberg. Sequential Translation of Trinucleotide Codons for the Initiation and Termination of Protein Synthesis. *Science*, 162(3849):135–138, October 1968. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.162.3849.135.

[22] F. H. Crick. On protein synthesis. *Symposia of the Society for Experimental Biology*, 12: 138–163, 1958. ISSN 0081-1386.

[23] Francois Chapeville, Fritz Lipmann, Günter von Ehrenstein, Bernard Weisblum, William J. Ray, and Seymour Benzer. On the Role of Soluble Ribonucleic Acid in Coding for Amino Acids. *Proceedings of the National Academy of Sciences*, 48(6):1086–1092, June 1962. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.48.6.1086.

[24] Robert W. Holley, George A. Everett, James T. Madison, and Ada Zamir. Nucleotide Sequences in the Yeast Alanine Transfer Ribonucleic Acid. *Journal of Biological Chemistry*, 240(5):2122–2128, January 1965. ISSN 0021-9258, 1083-351X.

[25] Robert W. Holley, Jean Apgar, George A. Everett, James T. Madison, Mark Marquisee, Susan H. Merrill, John Robert Penswick, and Ada Zamir. Structure of a Ribonucleic Acid. *Science*, 147(3664):1462–1465, March 1965. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.147.3664.1462.

[26] Richard Giegé and Gilbert Eriani. Transfer RNA Recognition and Aminoacylation by Synthetases. In *eLS*. American Cancer Society, 2014. ISBN 978-0-470-01590-2. doi: 10.1002/9780470015902.a0000531.pub3.

[27] P. Schimmel, R. Giegé, D. Moras, and S. Yokoyama. An operational RNA code for amino acids and possible relationship to genetic code. *Proceedings of the National Academy of Sciences*, 90(19):8763–8768, October 1993. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.90.19.8763.

[28] Marina V. Rodnina, Kirill B. Gromadski, Ute Kothe, and Hans-Joachim Wieden. Recognition and selection of tRNA in translation. *FEBS Letters*, 579(4):938–942, 2005. ISSN 1873-3468. doi: 10.1016/j.febslet.2004.11.048.

[29] Douglas Hofstadter. *Gödel, Escher Bach: An Eternal Golden Braid*. Basic Books, New York, twentieth-anniversary edition, April 1979.

[30] Jacques Monod. From Enzymatic Adaptation to Allosteric Transitions. *Science*, 154(3748): 475–483, October 1966. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.154.3748.475.

[31] François Jacob and Jacques Monod. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3(3):318–356, June 1961. ISSN 0022-2836. doi: 10.1016/S0022-2836(61)80072-7.

[32] Andrew A. Travers and Richard R. Burgess. Cyclic Re-use of the RNA Polymerase Sigma Factor. *Nature*, 222(5193):537, May 1969. ISSN 1476-4687. doi: 10.1038/222537a0.

[33] Mark SB Paget and John D Helmann. The Σ70 family of sigma factors. *Genome Biology*, 4(1):203, 2003. ISSN 1465-6906.

[34] S Kustu, E Santero, J Keener, D Popham, and D Weiss. Expression of sigma 54 (ntrA)-dependent genes is probably united by a common mechanism. *Microbiological Reviews*, 53 (3):367–376, September 1989. ISSN 0146-0749.

[35] M. J. Merrick. In a class of its own — the RNA polymerase sigma factor σ;54 (σN). *Molecular Microbiology*, 10(5):903–909, 1993. ISSN 1365-2958. doi: 10.1111/j.1365-2958.1993.tb00961.x.

[36] Victoria Shingler. Signal sensing by Σ54-dependent regulators: Derepression as a control mechanism. *Molecular Microbiology*, 19(3):409–416, 1996. ISSN 1365-2958. doi: 10.1046/j.1365-2958.1996.388920.x.

[37] Rachel Anne Mooney, Seth A. Darst, and Robert Landick. Sigma and RNA Polymerase: An On-Again, Off-Again Relationship? *Molecular Cell*, 20(3):335–345, November 2005. ISSN 1097-2765. doi: 10.1016/j.molcel.2005.10.015.

[38] D. Pribnow. Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proceedings of the National Academy of Sciences*, 72(3):784–788, March 1975. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.72.3.784.

[39] Peter H. Seeburg, Christiane Nusslein, and Heinz Schaller. Interaction of RNA Polymerase with Promoters from Bacteriophage fd. *European Journal of Biochemistry*, 74(1):107–113, March 1977. ISSN 0014-2956, 1432-1033. doi: 10.1111/j.1432-1033.1977.tb11372.x.

[40] Gerald Z. Hertz and Gary D. Stormo. [2] Escherichia coli promoter sequences: Analysis and prediction. In *Methods in Enzymology*, volume 273 of *RNA Polymerase and Associated Factors Part A*, pages 30–42. Academic Press, January 1996. doi: 10.1016/S0076-6879(96) 73004-5.

[41] W. Ross, K. K. Gosink, J. Salomon, K. Igarashi, C. Zou, A. Ishihama, K. Severinov, and R. L. Gourse. A third recognition element in bacterial promoters: DNA binding by the alpha subunit of RNA polymerase. *Science*, 262(5138):1407–1413, November 1993. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.8248780.

[42] H Krämer, M Niemöller, M Amouyal, B Revet, B von Wilcken-Bergmann, and B Müller-Hill. Lac repressor forms loops with linear DNA carrying two suitably spaced lac operators. *The EMBO Journal*, 6(5):1481–1491, May 1987. ISSN 0261-4189.

[43] Nicole A. Becker, Justin P. Peters, Troy A. Lionberger, and L. James Maher. Mechanism of promoter repression by Lac repressor–DNA loops. *Nucleic Acids Research*, 41(1):156–166, January 2013. ISSN 0305-1048. doi: 10.1093/nar/gks1011.

[44] Mitchell Lewis. Allostery and the lac Operon. *Journal of Molecular Biology*, 425(13): 2309–2316, July 2013. ISSN 0022-2836. doi: 10.1016/j.jmb.2013.03.003.

[45] A. Kolb, S. Busby, H. Buc, S. Garges, and S. Adhya. TRANSCRIPTIONAL REGULATION BY cAMP AND ITS RECEPTOR PROTEIN. *Annual Review of Biochemistry*, 62(1):749–797, 1993. doi: 10.1146/annurev.bi.62.070193.003533.

[46] I Pastan and S Adhya. Cyclic adenosine 5'-monophosphate in Escherichia coli. *Bacteriological Reviews*, 40(3):527–551, September 1976. ISSN 0005-3678.

[47] Steve Busby and Richard H Ebright. Transcription activation by catabolite activator protein (CAP). *Journal of Molecular Biology*, 293(2):199–213, October 1999. ISSN 0022-2836. doi: 10.1006/jmbi.1999.3161.

[48] Socorro Gama-Castro, Heladia Salgado, Alberto Santos-Zavaleta, Daniela Ledezma-Tejeida, Luis Muñiz-Rascado, Jair Santiago García-Sotelo, Kevin Alquicira-Hernández, Irma Martínez-Flores, Lucia Pannier, Jaime Abraham Castro-Mondragón, Alejandra Medina-Rivera, Hilda Solano-Lira, César Bonavides-Martínez, Ernesto Pérez-Rueda,

Shirley Alquicira-Hernández, Liliana Porrón-Sotelo, Alejandra López-Fuentes, Anastasia Hernández-Koutoucheva, Víctor Del Moral-Chávez, Fabio Rinaldi, and Julio Collado-Vides. RegulonDB version 9.0: High-level integration of gene regulation, coexpression, motif clustering and beyond. *Nucleic Acids Research*, 44(D1):D133–D143, April 2016. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkv1156.

[49] Agustino Martıńez-Antonio and Julio Collado-Vides. Identifying global regulators in transcriptional regulatory networks in bacteria. *Current Opinion in Microbiology*, 6(5):482–489, October 2003. ISSN 1369-5274. doi: 10.1016/j.mib.2003.09.002.

[50] J. V. Platko, D. A. Willins, and J. M. Calvo. The ilvIH operon of Escherichia coli is positively regulated. *Journal of Bacteriology*, 172(8):4563–4570, August 1990. ISSN 0021-9193, 1098-5530. doi: 10.1128/jb.172.8.4563-4570.1990.

[51] M. Felice De, C. T. Lago, C. H. Squires, and J. M. Calvo. Acetohydroxy acid synthase isoenzymes of Escherichia coli K12 and Salmonella typhimurium. *Annales de microbiologie*, 133(2):251–256, 1982. ISSN 0300-5410.

[52] J. C. Andrews and S. A. Short. Opp-lac Operon fusions and transcriptional regulation of the Escherichia coli trp-linked oligopeptide permease. *Journal of Bacteriology*, 165(2):434–442, February 1986. ISSN 0021-9193, 1098-5530. doi: 10.1128/jb.165.2.434-442.1986.

[53] D. A. Willins, C. W. Ryan, J. V. Platko, and J. M. Calvo. Characterization of Lrp, and Escherichia coli regulatory protein that mediates a global response to leucine. *Journal of Biological Chemistry*, 266(17):10768–10774, June 1991. ISSN 0021-9258, 1083-351X.

[54] S. A. Haney, J. V. Platko, D. L. Oxender, and J. M. Calvo. Lrp, a leucine-responsive protein, regulates branched-chain amino acid transport genes in Escherichia coli. *Journal of Bacteriology*, 174(1):108–115, January 1992. ISSN 0021-9193. doi: 10.1128/jb.174.1.108-115.1992.

[55] Travis H. Tani, Arkady Khodursky, Robert M. Blumenthal, Patrick O. Brown, and Rowena G. Matthews. Adaptation to famine: A family of stationary-phase genes revealed by microarray analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 99(21):13471–13476, October 2002. ISSN 0027-8424. doi: 10.1073/pnas.212510999.

[56] Shaolin Chen and Joseph M Calvo. Leucine-induced Dissociation of Escherichia coli Lrp Hexadecamers to Octamers. *Journal of Molecular Biology*, 318(4):1031–1042, May 2002. ISSN 0022-2836. doi: 10.1016/S0022-2836(02)00187-0.

[57] Benjamin R. Hart and Robert M. Blumenthal. Unexpected Coregulator Range for the Global Regulator Lrp of Escherichia coli and Proteus mirabilis. *Journal of Bacteriology*, 193(5): 1054–1064, March 2011. ISSN 0021-9193, 1098-5530. doi: 10.1128/JB.01183-10.

[58] Byung-Kwan Cho, Christian L. Barrett, Eric M. Knight, Young Seoub Park, and Bernhard Ø. Palsson. Genome-scale reconstruction of the Lrp regulatory network in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 105 (49):19462–19467, December 2008. ISSN 0027-8424. doi: 10.1073/pnas.0807227105.

[59] Stephanie de los Rios and John J. Perona. Structure of the Escherichia coli Leucine-responsive Regulatory Protein Lrp Reveals a Novel Octameric Assembly. *Journal of Molecular Biology*, 366(5):1589–1602, March 2007. ISSN 0022-2836. doi: 10.1016/j.jmb.2006.12.032.

[60] Brian T. Wilhelm and Josette-Renée Landry. RNA-Seq—quantitative measurement of expression through massively parallel RNA-sequencing. *Methods*, 48(3):249–257, July 2009. ISSN 1046-2023. doi: 10.1016/j.ymeth.2009.03.016.

[61] David S. Johnson, Ali Mortazavi, Richard M. Myers, and Barbara Wold. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science*, 316(5830):1497–1502, June 2007. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1141319.

[62] Joel Rozowsky, Ghia Euskirchen, Raymond K. Auerbach, Zhengdong D. Zhang, Theodore Gibson, Robert Bjornson, Nicholas Carriero, Michael Snyder, and Mark B. Gerstein. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature Biotechnology*, 27(1):66–75, January 2009. ISSN 1546-1696. doi: 10.1038/nbt.1518.

[63] Valerie Hower, Steven N. Evans, and Lior Pachter. Shape-based peak identification for ChIP-Seq. *BMC Bioinformatics*, 12(1):15, December 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-15.

[64] Jianxing Feng, Tao Liu, Bo Qin, Yong Zhang, and Xiaole Shirley Liu. Identifying ChIP-seq enrichment using MACS. *Nature Protocols*, 7(9):1728–1740, September 2012. ISSN 1750-2799. doi: 10.1038/nprot.2012.101.

[65] Yanxiao Zhang, Yu-Hsuan Lin, Timothy D. Johnson, Laura S. Rozek, and Maureen A. Sartor. PePr: A peak-calling prioritization pipeline to identify consistent or differential peaks from replicated ChIP-Seq data. *Bioinformatics*, 30(18):2568–2575, September 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu372.

[66] Byung-Kwan Cho, Donghyuk Kim, Eric M. Knight, Karsten Zengler, and Bernhard O. Palsson. Genome-scale reconstruction of the sigma factor network in Escherichia coli: Topology and functional states. *BMC Biology*, 12:4, January 2014. ISSN 1741-7007. doi: 10.1186/1741-7007-12-4.

[67] Dongling Zheng, Chrystala Constantinidou, Jon L. Hobman, and Stephen D. Minchin. Identification of the CRP regulon using in vitro and in vivo transcriptional profiling. *Nucleic Acids Research*, 32(19):5874–5893, October 2004. ISSN 0305-1048. doi: 10.1093/nar/gkh908.

[68] Donghyuk Kim, Sang Woo Seo, Ye Gao, Hojung Nam, Gabriela I. Guzman, Byung-Kwan Cho, and Bernhard O. Palsson. Systems assessment of transcriptional regulation on central carbon metabolism by Cra and CRP. *Nucleic Acids Research*, 2018. doi: 10.1093/nar/gky069.

[69] Constance Holden. Alliance Launched to Model E. coli. *Science*, 297(5586):1459–1460, August 2002. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.297.5586.1459a.

[70] Xin Fang, Anand Sastry, Nathan Mih, Donghyuk Kim, Justin Tan, James T. Yurkovich, Colton J. Lloyd, Ye Gao, Laurence Yang, and Bernhard O. Palsson. Global transcriptional regulatory network for Escherichia coli robustly connects gene expression to transcription factor activities. *Proceedings of the National Academy of Sciences*, 114(38):10286–10291, September 2017. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1702581114.

[71] Talukder Ali Azam and Akira Ishihama. Twelve Species of the Nucleoid-associated Protein from Escherichia coli SEQUENCE RECOGNITION SPECIFICITY AND DNA BINDING AFFINITY. *Journal of Biological Chemistry*, 274(46):33105–33113, December 1999. ISSN 0021-9258, 1083-351X. doi: 10.1074/jbc.274.46.33105.

[72] Shane C. Dillon and Charles J. Dorman. Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. *Nature Reviews Microbiology*, 8(3):185–195, March 2010. ISSN 1740-1534. doi: 10.1038/nrmicro2261.

[73] Beth A. Shen and Robert Landick. Transcription of Bacterial Chromatin. *Journal of Molecular Biology*, May 2019. ISSN 0022-2836. doi: 10.1016/j.jmb.2019.05.041.

[74] T. L. Megraw and C. B. Chae. Functional complementarity between the HMG1-like yeast mitochondrial histone HM and the bacterial histone-like protein HU. *Journal of Biological Chemistry*, 268(17):12758–12763, June 1993. ISSN 0021-9258, 1083-351X.

[75] Jean-François Briat, Sylvie Letoffe, Régis Mache, and Josette Rouviere-Yaniv. Similarity between the bacterial histone-like protein HU and a protein from spinach chloroplasts. *FEBS Letters*, 172(1):75–79, June 1984. ISSN 0014-5793. doi: 10.1016/0014-5793(84)80877-7.

[76] Talukder Ali Azam, Akira Iwata, Akiko Nishimura, Susumu Ueda, and Akira Ishihama. Growth Phase-Dependent Variation in Protein Composition of the Escherichia coli Nucleoid. *Journal of Bacteriology*, 181(20):6361–6370, October 1999. ISSN 0021-9193, 1098-5530.

[77] Talukder Ali Azam, Sota Hiraga, and Akira Ishihama. Two types of localization of the DNA-binding proteins within the Escherichia coli nucleoid. *Genes to Cells*, 5(8):613–626, 2000. ISSN 1365-2443. doi: 10.1046/j.1365-2443.2000.00350.x.

[78] Tiffany Vora, Alison K. Hottes, and Saeed Tavazoie. Protein Occupancy Landscape of a Bacterial Genome. *Molecular Cell*, 35(2):247–253, July 2009. ISSN 1097-2765. doi: 10.1016/j.molcel.2009.06.035.

[79] William Wiley Navarre, Steffen Porwollik, Yipeng Wang, Michael McClelland, Henry Rosen, Stephen J. Libby, and Ferric C. Fang. Selective Silencing of Foreign DNA with Low GC Content by the H-NS Protein in Salmonella. *Science*, 313(5784):236–238, July 2006. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1128794.

[80] Sacha Lucchini, Gary Rowley, Martin D. Goldberg, Douglas Hurd, Marcus Harrison, and Jay C. D. Hinton. H-NS Mediates the Silencing of Laterally Acquired Genes in Bacteria. *PLOS Pathogens*, 2(8):e81, August 2006. ISSN 1553-7374. doi: 10.1371/journal.ppat.0020081.

[81] Benjamin Lang, Nicolas Blot, Emeline Bouffartigues, Malcolm Buckle, Marcel Geertz, Claudio O. Gualerzi, Ramesh Mavathur, Georgi Muskhelishvili, Cynthia L. Pon, Sylvie Rimsky, Stefano Stella, M. Madan Babu, and Andrew Travers. High-affinity DNA binding sites for H-NS provide a molecular basis for selective silencing within proteobacterial genomes. *Nucleic Acids Research*, 35(18):6330–6337, September 2007. ISSN 0305-1048. doi: 10.1093/nar/gkm712.

[82] Robert Landick, Joseph T Wade, and David C Grainger. H-NS and RNA polymerase: A love–hate relationship? *Current Opinion in Microbiology*, 24:53–59, April 2015. ISSN 1369-5274. doi: 10.1016/j.mib.2015.01.009.

[83] Maarten C. Noom, William W. Navarre, Taku Oshima, Gijs J. L. Wuite, and Remus T. Dame. H-NS promotes looped domain formation in the bacterial chromosome. *Current Biology*, 17(21):R913–R914, November 2007. ISSN 0960-9822. doi: 10.1016/j.cub.2007.09.005.

[84] Christina Kahramanoglou, Aswin S. N. Seshasayee, Ana I. Prieto, David Ibberson, Sabine Schmidt, Jurgen Zimmermann, Vladimir Benes, Gillian M. Fraser, and Nicholas M. Luscombe. Direct and indirect effects of H-NS and Fis on global gene expression control in Escherichia coli. *Nucleic Acids Research*, 39(6):2073–2091, March 2011. ISSN 0305-1048. doi: 10.1093/nar/gkq934.

[85] Remus T. Dame, Maarten C. Noom, and Gijs J. L. Wuite. Bacterial chromatin organization by H-NS protein unravelled using dual DNA manipulation. *Nature*, 444(7117):387–390, November 2006. ISSN 1476-4687. doi: 10.1038/nature05283.

[86] Stefan T. Arold, Paul G. Leonard, Gary N. Parkinson, and John E. Ladbury. H-NS forms a superhelical protein scaffold for DNA condensation. *Proceedings of the National Academy of Sciences*, 107(36):15728–15732, September 2010. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1006966107.

[87] Beth A. Boudreau, Daniel R. Hron, Liang Qin, Ramon A. van der Valk, Matthew V. Kotlajich, Remus T. Dame, and Robert Landick. StpA and Hha stimulate pausing by RNA polymerase by promoting DNA-DNA bridging of H-NS filaments. *Nucleic Acids Research*, 46(11):5525–5546, June 2018. ISSN 1362-4962. doi: 10.1093/nar/gky265.

[88] W Ross, J F Thompson, J T Newlands, and R L Gourse. E.coli Fis protein activates ribosomal RNA transcription in vitro and in vivo. *The EMBO Journal*, 9(11):3733–3742, November 1990. ISSN 0261-4189.

[89] Anton J. Bokal IV, Wilma Ross, and Richard L. Gourse. The Transcriptional Activator Protein FIS: DNA Interactions and Cooperative Interactions with RNA Polymerase at theEscherichia coli rrnBP1 Promoter. *Journal of Molecular Biology*, 245(3):197–207, January 1995. ISSN 0022-2836. doi: 10.1006/jmbi.1994.0016.

[90] Michael L. Opel, Kimberly A. Aeling, Walter M. Holmes, Reid C. Johnson, Craig J. Benham, and G. Wesley Hatfield. Activation of transcription initiation from a stable RNA promoter by a Fis protein-mediated DNA structural transmission mechanism. *Molecular Microbiology*, 53(2):665–674, 2004. ISSN 1365-2958. doi: 10.1111/j.1365-2958.2004.04147.x.

[91] Sarah M McLeod, Sarah E Aiyar, Richard L Gourse, and Reid C Johnson. The C-terminal domains of the RNA polymerase α subunits: Contact site with fis and localization during co-activation with CRP at the Escherichia coli proP P2 promoter1 1Edited by M. Gottesman. *Journal of Molecular Biology*, 316(3):517–529, February 2002. ISSN 0022-2836. doi: 10.1006/jmbi.2001.5391.

[92] Carolina Sousa, Victor de Lorenzo, and Angel Cebolla. Modulation of gene expression through chromosomal positioning in Escherichia coli. *Microbiology*, 143(6):2071–2078, 1997. doi: 10.1099/00221287-143-6-2071.

[93] Jack A. Bryant, Laura E. Sellars, Stephen J. W. Busby, and David J. Lee. Chromosome position effects on gene expression in Escherichia coli K-12. *Nucleic Acids Research*, 42 (18):11383–11392, October 2014. ISSN 0305-1048. doi: 10.1093/nar/gku828.

[94] O. L. Miller, Barbara A. Hamkalo, and C. A. Thomas. Visualization of Bacterial Genes in Action. *Science*, 169(3943):392–395, July 1970. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.169.3943.392.

[95] Giulia Oliva, Tobias Sahr, and Carmen Buchrieser. Small RNAs, 5′ UTR elements and RNA-binding proteins in intracellular bacteria: Impact on metabolism and virulence. *FEMS Microbiology Reviews*, 39(3):331–349, May 2015. ISSN 1574-6976. doi: 10.1093/femsre/fuv022.

[96] Elke Van Assche, Sandra Van Puyvelde, Jos Vanderleyden, and Hans P. Steenackers. RNA-binding proteins involved in post-transcriptional regulation in bacteria. *Frontiers in Microbiology*, 6, March 2015. ISSN 1664-302X. doi: 10.3389/fmicb.2015.00141.

[97] Melissa J. Moore. From Birth to Death: The Complex Lives of Eukaryotic mRNAs. *Science*, 309(5740):1514–1518, September 2005. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1111443.

[98] Anand Ramanathan, G. Brett Robb, and Siu-Hong Chan. mRNA capping: Biological functions and applications. *Nucleic Acids Research*, 44(16):7511–7526, September 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw551.

[99] Thomas Gonatopoulos-Pournatzis and Victoria H. Cowling. Cap-binding complex (CBC). *Biochemical Journal*, 457(2):231–242, January 2014. ISSN 0264-6021, 1470-8728. doi: 10.1042/BJ20131214.

[100] Yeon Lee and Donald C. Rio. Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annual Review of Biochemistry*, 84(1):291–323, 2015. doi: 10.1146/annurev-biochem-060614-034316.

[101] Thomas R. Cech, Arthur J. Zaug, and Paula J. Grabowski. In vitro splicing of the ribosomal RNA precursor of tetrahymena: Involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell*, 27(3, Part 2):487–496, December 1981. ISSN 0092-8674. doi: 10.1016/0092-8674(81)90390-1.

[102] Evan C. Merkhofer, Peter Hu, and Tracy L. Johnson. Introduction to Cotranscriptional RNA Splicing. *Methods in molecular biology (Clifton, N.J.)*, 1126:83–96, 2014. ISSN 1064-3745. doi: 10.1007/978-1-62703-980-2_6.

[103] Nick J. Proudfoot. Ending the message: Poly(A) signals then and now. *Genes & Development*, 25(17):1770–1782, January 2011. ISSN 0890-9369, 1549-5477. doi: 10.1101/gad.17268411.

[104] Ran Elkon, Alejandro P. Ugalde, and Reuven Agami. Alternative cleavage and polyadenylation: Extent, regulation and function. *Nature Reviews Genetics*, 14(7):496–506, July 2013. ISSN 1471-0064. doi: 10.1038/nrg3482.

[105] Jay R. Greenberg. High Stability of Messenger RNA in Growing Cultured Cells. *Nature*, 240(5376):102–104, November 1972. ISSN 1476-4687. doi: 10.1038/240102a0.

[106] Hyeshik Chang, Jaechul Lim, Minju Ha, and V. Narry Kim. TAIL-seq: Genome-wide Determination of Poly(A) Tail Length and 3′ End Modifications. *Molecular Cell*, 53(6): 1044–1052, March 2014. ISSN 1097-2765. doi: 10.1016/j.molcel.2014.02.007.

[107] Richard J. Jackson, Christopher U. T. Hellen, and Tatyana V. Pestova. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nature Reviews Molecular Cell Biology*, 11(2):113–127, February 2010. ISSN 1471-0080. doi: 10.1038/nrm2838.

[108] Christine Mayr. Regulation by 3′-Untranslated Regions. *Annual Review of Genetics*, 51(1): 171–194, 2017. doi: 10.1146/annurev-genet-120116-024704.

[109] Lynne E. Maquat, Woan-Yuh Tarn, and Olaf Isken. The Pioneer Round of Translation: Features and Functions. *Cell*, 142(3):368–374, August 2010. ISSN 0092-8674. doi: 10.1016/j.cell.2010.07.022.

[110] Anna Łabno, Rafał Tomecki, and Andrzej Dziembowski. Cytoplasmic RNA decay pathways - Enzymes and mechanisms. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1863(12):3125–3147, December 2016. ISSN 0167-4889. doi: 10.1016/j.bbamcr.2016.09.023.

[111] Elmar Wahle and G. Sebastiaan Winkler. RNA decay machines: Deadenylation by the Ccr4–Not and Pan2–Pan3 complexes. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1829(6):561–570, June 2013. ISSN 1874-9399. doi: 10.1016/j.bbagrm.2013.01.003.

[112] Martine A. Collart. The Ccr4-Not complex is a key regulator of eukaryotic gene expression. *Wiley Interdisciplinary Reviews: RNA*, 7(4):438–454, 2016. ISSN 1757-7012. doi: 10.1002/wrna.1332.

[113] Akio Yamashita, Tsung-Cheng Chang, Yukiko Yamashita, Wenmiao Zhu, Zhenping Zhong, Chyi-Ying A. Chen, and Ann-Bin Shyu. Concerted action of poly(A) nucleases and decapping enzyme in mammalian mRNA turnover. *Nature Structural & Molecular Biology*, 12 (12):1054, December 2005. ISSN 1545-9985. doi: 10.1038/nsmb1016.

[114] Marcos Arribas-Layton, Donghui Wu, Jens Lykke-Andersen, and Haiwei Song. Structural and functional control of the eukaryotic mRNA decapping machinery. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1829(6):580–589, June 2013. ISSN 1874-9399. doi: 10.1016/j.bbagrm.2012.12.006.

[115] Roy Parker. RNA Degradation in *Saccharomyces cerevisae*. *Genetics*, 191(3):671–702, July 2012. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.111.137265.

[116] Aleksander Chlebowski, Michał Lubas, Torben Heick Jensen, and Andrzej Dziembowski. RNA decay machines: The exosome. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1829(6):552–560, June 2013. ISSN 1874-9399. doi: 10.1016/j.bbagrm.2013.01.006.

[117] Christopher Iain Jones, Maria Vasilyevna Zabolotskaya, and Sarah Faith Newbury. The $5' \rightarrow 3'$ exoribonuclease XRN1/Pacman and its functions in cellular processes and development. *Wiley Interdisciplinary Reviews: RNA*, 3(4):455–468, 2012. ISSN 1757-7012. doi: 10.1002/wrna.1109.

[118] Rafal Tomecki and Andrzej Dziembowski. Novel endoribonucleases as central players in various pathways of eukaryotic RNA metabolism. *RNA*, 16(9):1692–1724, January 2010. ISSN 1355-8382, 1469-9001. doi: 10.1261/rna.2237610.

[119] Olaf Isken, Yoon Ki Kim, Nao Hosoda, Greg L. Mayeur, John W. B. Hershey, and Lynne E. Maquat. Upf1 Phosphorylation Triggers Translational Repression during Nonsense-Mediated mRNA Decay. *Cell*, 133(2):314–327, April 2008. ISSN 0092-8674. doi: 10.1016/j.cell.2008.02.030.

[120] Lasse Peters and Gunter Meister. Argonaute Proteins: Mediators of RNA Silencing. *Molecular Cell*, 26(5):611–623, June 2007. ISSN 1097-2765. doi: 10.1016/j.molcel.2007.05.001.

[121] Julius Brennecke, Alexander Stark, Robert B. Russell, and Stephen M. Cohen. Principles of MicroRNA–Target Recognition. *PLOS Biology*, 3(3):e85, February 2005. ISSN 1545-7885. doi: 10.1371/journal.pbio.0030085.

[122] Tim A. Rand, Sean Petersen, Fenghe Du, and Xiaodong Wang. Argonaute2 Cleaves the Anti-Guide Strand of siRNA during RISC Activation. *Cell*, 123(4):621–629, November 2005. ISSN 0092-8674. doi: 10.1016/j.cell.2005.10.020.

[123] Stefanie Jonas and Elisa Izaurralde. Towards a molecular understanding of microRNA-mediated gene silencing. *Nature Reviews Genetics*, 16(7):421–433, July 2015. ISSN 1471-0056. doi: 10.1038/nrg3965.

[124] Vikram Agarwal, George W Bell, Jin-Wu Nam, and David P Bartel. Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, 4:e05005, August 2015. ISSN 2050-084X. doi: 10.7554/eLife.05005.

[125] Robin C. Friedman, Kyle Kai-How Farh, Christopher B. Burge, and David P. Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, 19(1):92–105, January 2009. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.082701.108.

[126] Stefanie Gerstberger, Markus Hafner, and Thomas Tuschl. A census of human RNA-binding proteins. *Nature Reviews Genetics*, 15(12):829–845, December 2014. ISSN 1471-0056. doi: 10.1038/nrg3813.

[127] Robin P. Wharton and Gary Struhl. RNA regulatory elements mediate control of Drosophila body pattern by the posterior morphogen nanos. *Cell*, 67(5):955–967, November 1991. ISSN 0092-8674. doi: 10.1016/0092-8674(91)90368-9.

[128] P D Zamore, J R Williamson, and R Lehmann. The Pumilio protein binds RNA through a conserved domain that defines a new class of RNA-binding proteins. *RNA*, 3(12):1421–1433, December 1997. ISSN 1355-8382.

[129] Robin P Wharton, Junichiro Sonoda, Tammy Lee, Michelle Patterson, and Yoshihiko Murata. The Pumilio RNA-Binding Domain Is Also a Translational Regulator. *Molecular Cell*, 1(6):863–872, May 1998. ISSN 1097-2765. doi: 10.1016/S1097-2765(00)80085-4.

[130] C Nusslein-Volhard, H. Frohnhofer, and R Lehmann. Determination of anteroposterior polarity in Drosophila. *Science*, 238(4834):1675–1681, December 1987. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.3686007.

[131] Jamie Van Etten, Trista L. Schagat, Joel Hrit, Chase Weidmann, Justin Brumbaugh, Joshua J. Coon, and Aaron C. Goldstrohm. Human Pumilio proteins recruit multiple deadenylases to efficiently repress messenger RNAs. *Journal of Biological Chemistry*, page jbc.M112.373522, September 2012. ISSN 0021-9258, 1083-351X. doi: 10.1074/jbc.M112.373522.

[132] Xiaoqiang Wang, Juanita McLachlan, Phillip D. Zamore, and Traci M. Tanaka Hall. Modular Recognition of RNA by a Human Pumilio-Homology Domain. *Cell*, 110(4):501–512, August 2002. ISSN 0092-8674. doi: 10.1016/S0092-8674(02)00873-5.

[133] Aaron C. Goldstrohm, Traci M. Tanaka Hall, and Katherine M. McKenney. Post-transcriptional Regulatory Functions of Mammalian Pumilio Proteins. *Trends in Genetics*, 34(12):972–990, December 2018. ISSN 0168-9525. doi: 10.1016/j.tig.2018.09.006.

[134] Jennifer A. Bohn, Jamie L. Van Etten, Trista L. Schagat, Brittany M. Bowman, Richard C. McEachin, Peter L. Freddolino, and Aaron C. Goldstrohm. Identification of diverse target RNAs that are functionally regulated by human Pumilio proteins. *Nucleic Acids Research*, 46(1):362–386, January 2018. ISSN 0305-1048. doi: 10.1093/nar/gkx1120.

[135] Alessia Galgano, Michael Forrer, Lukasz Jaskiewicz, Alexander Kanitz, Mihaela Zavolan, and André P. Gerber. Comparative Analysis of mRNA Targets for Human PUF-Family Proteins Suggests Extensive Interaction with the miRNA Regulatory System. *PLOS ONE*, 3(9):e3164, September 2008. ISSN 1932-6203. doi: 10.1371/journal.pone.0003164.

[136] Peng Jiang, Mona Singh, and Hilary A. Coller. Computational assessment of the cooperativity between RNA binding proteins and MicroRNAs in Transcript Decay. *PLoS computational biology*, 9(5):e1003075, 2013. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003075.

[137] Wayne O. Miles, Katrin Tschöp, Anabel Herr, Jun-Yuan Ji, and Nicholas J. Dyson. Pumilio facilitates miRNA regulation of the E2F3 oncogene. *Genes & Development*, 26(4):356–368, February 2012. ISSN 0890-9369. doi: 10.1101/gad.182568.111.

[138] Grace M. Kroner, Michael B. Wolfe, and Peter L. Freddolino. Escherichia coli Lrp Regulates One-Third of the Genome via Direct, Cooperative, and Indirect Routes. *Journal of Bacteriology*, 201(3):e00411–18, February 2019. ISSN 0021-9193, 1098-5530. doi: 10.1128/JB.00411-18.

[139] Chang-Ho Baek, Shifeng Wang, Kenneth L. Roland, and Roy Curtiss. Leucine-Responsive Regulatory Protein (Lrp) Acts as a Virulence Repressor in Salmonella enterica Serovar Typhimurium. *Journal of Bacteriology*, 191(4):1278–1292, February 2009. ISSN 0021-9193, 1098-5530. doi: 10.1128/JB.01142-08.

[140] Michael D. Engstrom and Harry L. T. Mobley. Regulation of Expression of Uropathogenic Escherichia coli Nonfimbrial Adhesin TosA by PapB Homolog TosR in Conjunction with H-NS and Lrp. *Infection and Immunity*, 84(3):811–821, February 2016. ISSN 0019-9567. doi: 10.1128/IAI.01302-15.

[141] Angelina Cordone, Sacha Lucchini, Maurilio De Felice, and Ezio Ricca. Direct and indirect control of Lrp on LEE pathogenicity genes of Citrobacter rodentium. *FEMS Microbiology Letters*, 325(1):64–70, December 2011. ISSN 0378-1097. doi: 10.1111/j.1574-6968.2011.02411.x.

[142] R. P. S. Parti, Rahul Shrivastava, S. Srivastava, A. R. Subramanian, Raja Roy, Brahm S. Srivastava, and Ranjana Srivastava. A transposon insertion mutant of Mycobacterium fortuitum attenuated in virulence and persistence in a murine infection model that is complemented by Rv3291c of Mycobacterium tuberculosis. *Microbial Pathogenesis*, 45(5):370–376, November 2008. ISSN 0882-4010. doi: 10.1016/j.micpath.2008.08.008.

[143] Elizabeth A. Hussa, Ángel M. Casanova-Torres, and Heidi Goodrich-Blair. The Global Transcription Factor Lrp Controls Virulence Modulation in Xenorhabdus nematophila. *Journal of Bacteriology*, 197(18):3015–3025, September 2015. ISSN 0021-9193, 1098-5530. doi: 10.1128/JB.00272-15.

[144] Wei Lin, Gabriela Kovacikova, and Karen Skorupski. The quorum sensing regulator HapR downregulates the expression of the virulence gene transcription factor AphA in Vibrio cholerae by antagonizing Lrp- and VpsR-mediated activation. *Molecular Microbiology*, 64(4):953–967, 2007. ISSN 1365-2958. doi: 10.1111/j.1365-2958.2007.05693.x.

[145] Shaolin Chen, Michele H. Rosner, and Joseph M. Calvo. Leucine-regulated self-association of leucine-responsive regulatory protein (Lrp) from Escherichia coli. *Journal of Molecular Biology*, 312(4):625–635, September 2001. ISSN 0022-2836. doi: 10.1006/jmbi.2001.4955.

[146] Shaolin Chen, Zhiqi Hao, Eva Bieniek, and Joseph M. Calvo. Modulation of Lrp action in Escherichia coli by leucine: Effects on non-specific binding of Lrp to DNA. *Journal of*

Molecular Biology, 314(5):1067–1075, December 2001. ISSN 0022-2836. doi: 10.1006/jmbi.2000.5209.

[147] Dawn M. Graunke, Russell O. Pieper, and Albert J. Fornace. Presetting of chromatin structure and transcription factor binding poise the human GADD45 gene for rapid transcriptional up-regulation. *Nucleic Acids Research*, 27(19):3881–3890, October 1999. ISSN 0305-1048. doi: 10.1093/nar/27.19.3881.

[148] Gu Xiao, David White, and Jill Bargonetti. P53 binds to a constitutively nucleosome free region of the mdm2 gene. *Oncogene*, 16(9):1171, March 1998. ISSN 1476-5594. doi: 10.1038/sj.onc.1201631.

[149] Arie B. Brinkman, Thijs J. G. Ettema, Willem M. De Vos, and John Van Der Oost. The Lrp family of transcriptional regulators. *Molecular Microbiology*, 48(2):287–294, 2003. ISSN 1365-2958. doi: 10.1046/j.1365-2958.2003.03442.x.

[150] D. A. Willins and J. M. Calvo. In vitro transcription from the Escherichia coli ilvIH promoter. *Journal of Bacteriology*, 174(23):7648–7655, December 1992. ISSN 0021-9193, 1098-5530. doi: 10.1128/jb.174.23.7648-7655.1992.

[151] Peter J. Park. ChIP–seq: Advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680, October 2009. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg2641.

[152] Tomohiro Shimada, Natsumi Saito, Michihisa Maeda, Kan Tanaka, and Akira Ishihama. Expanded roles of leucine-responsive regulatory protein in transcription regulation of the Escherichia coli genome: Genomic SELEX screening of the regulation targets. *Microbial Genomics*, 1(1), 2015. doi: 10.1099/mgen.0.000001.

[153] Meina Neumann and Silke Leimkühler. Heavy metal ions inhibit molybdoenzyme activity by binding to the dithiolene moiety of molybdopterin in Escherichia coli. *The FEBS Journal*, 275(22):5678–5689, 2008. ISSN 1742-4658. doi: 10.1111/j.1742-4658.2008.06694.x.

[154] Frederick C. Neidhardt, Philip L. Bloch, and David F. Smith. Culture Medium for Enterobacteria. *Journal of Bacteriology*, 119(3):736–747, September 1974. ISSN 0021-9193, 1098-5530.

[155] Hani Goodarzi, Olivier Elemento, and Saeed Tavazoie. Revealing Global Regulatory Perturbations across Human Cancers. *Molecular Cell*, 36(5):900–911, December 2009. ISSN 1097-2765. doi: 10.1016/j.molcel.2009.11.016.

[156] Erik R. Zinser and Roberto Kolter. Prolonged Stationary-Phase Incubation Selects forlrp Mutations in Escherichia coliK-12. *Journal of Bacteriology*, 182(15):4361–4365, August 2000. ISSN 0021-9193, 1098-5530. doi: 10.1128/JB.182.15.4361-4365.2000.

[157] Y. Cui, Q. Wang, G. D. Stormo, and J. M. Calvo. A consensus sequence for binding of Lrp to DNA. *Journal of Bacteriology*, 177(17):4872–4880, September 1995. ISSN 0021-9193, 1098-5530. doi: 10.1128/jb.177.17.4872-4880.1995.

[158] Shaolin Chen, Maria Iannolo, and Joseph M. Calvo. Cooperative Binding of the Leucine-Responsive Regulatory Protein (Lrp) to DNA. *Journal of Molecular Biology*, 345(2):251–264, January 2005. ISSN 0022-2836. doi: 10.1016/j.jmb.2004.10.047.

[159] J. H. Rex, B. D. Aronson, and R. L. Somerville. The tdh and serA operons of Escherichia coli: Mutational analysis of the regulatory elements of leucine-responsive genes. *Journal of Bacteriology*, 173(19):5944–5953, October 1991. ISSN 0021-9193, 1098-5530. doi: 10.1128/jb.173.19.5944-5953.1991.

[160] Mikhail Pachkov, Ionas Erb, Nacho Molina, and Erik van Nimwegen. SwissRegulon: A database of genome-wide annotations of regulatory sites. *Nucleic Acids Research*, 35 (suppl_1):D127–D131, January 2007. ISSN 0305-1048. doi: 10.1093/nar/gkl857.

[161] Charles E. Grant, Timothy L. Bailey, and William Stafford Noble. FIMO: Scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, April 2011. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr064.

[162] Larry Reitzer and Barbara L. Schneider. Metabolic Context and Possible Physiological Themes of Σ54-Dependent Genes in Escherichia coli. *Microbiology and Molecular Biology Reviews*, 65(3):422–444, September 2001. ISSN 1092-2172, 1098-5557. doi: 10.1128/MMBR.65.3.422-444.2001.

[163] Marina Caldara, Daniel Charlier, and Raymond Cunin. The arginine regulon of Escherichia coli: Whole-system transcriptome analysis discovers new genes and provides an integrated view of arginine regulation. *Microbiology*, 152(11):3343–3354, 2006. doi: 10.1099/mic.0.29088-0.

[164] Alexandros K. Kiupakis and Larry Reitzer. ArgR-Independent Induction and ArgR-Dependent Superinduction of the astCADBE Operon in Escherichia coli. *Journal of Bacteriology*, 184(11):2940–2950, June 2002. ISSN 0021-9193, 1098-5530. doi: 10.1128/JB.184.11.2940-2950.2002.

[165] E. B. Newman and Rongtuan Lin. LEUCINE-RESPONSIVE REGULATORY PROTEIN: A Global Regulator of Gene Expression in E. Coli. *Annual Review of Microbiology*, 49(1):747–775, 1995. doi: 10.1146/annurev.mi.49.100195.003531.

[166] Q. Wang and J. M. Calvo. Lrp, a major regulatory protein in Escherichia coli, bends DNA and can organize the assembly of a higher-order nucleoprotein structure. *The EMBO Journal*, 12(6):2495–2501, 1993. ISSN 1460-2075. doi: 10.1002/j.1460-2075.1993.tb05904.x.

[167] Akira Ishihama, Tomohiro Shimada, and Yukiko Yamazaki. Transcription profile of Escherichia coli: Genomic SELEX search for regulatory targets of transcription factors. *Nucleic Acids Research*, 44(5):2058–2074, March 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw051.

[168] Ümit Pul, Reinhild Wurm, and Rolf Wagner. The Role of LRP and H-NS in Transcription Regulation: Involvement of Synergism, Allostery and Macromolecular Crowding. *Journal of Molecular Biology*, 366(3):900–915, February 2007. ISSN 0022-2836. doi: 10.1016/j.jmb.2006.11.067.

[169] Stacey N. Peterson and Norbert O. Reich. Competitive Lrp and Dam Assembly at the pap Regulatory Region: Implications for Mechanisms of Epigenetic Regulation. *Journal of Molecular Biology*, 383(1):92–105, October 2008. ISSN 0022-2836. doi: 10.1016/j.jmb.2008.07.086.

[170] Matthew F. Traxler, Vineetha M. Zacharia, Stafford Marquardt, Sean M. Summers, Huyen-Tran Nguyen, S. Elizabeth Stark, and Tyrrell Conway. Discretely calibrated regulatory loops controlled by ppGpp partition gene induction across the 'feast to famine' gradient in Escherichia coli. *Molecular Microbiology*, 79(4):830–845, 2011. ISSN 1365-2958. doi: 10.1111/j.1365-2958.2010.07498.x.

[171] Jean Bouvier, Sylvie Gordia, Gabriele Kampmann, Roland Lange, Regine Hengge-Aronis, and Claude Gutierrez. Interplay between global regulators of Escherichia coli : Effect of RpoS, Lrp and H-NS on transcription of the gene osmC. *Molecular Microbiology*, 28(5): 971–980, 1998. ISSN 1365-2958. doi: 10.1046/j.1365-2958.1998.00855.x.

[172] Frédéric Colland, Mechthild Barth, Regine Hengge-Aronis, and Annie Kolb. σ factor selectivity of Escherichia coli RNA polymerase: Role for CRP, IHF and Lrp transcription factors. *The EMBO Journal*, 19(12):3028–3037, June 2000. ISSN 0261-4189, 1460-2075. doi: 10.1093/emboj/19.12.3028.

[173] Clement M. Potel, Miao-Hsia Lin, Albert J. R. Heck, and Simone Lemeer. Widespread bacterial protein histidine phosphorylation revealed by mass spectrometry-based proteomics. *Nature Methods*, 15(3):187–190, March 2018. ISSN 1548-7105. doi: 10.1038/nmeth.4580.

[174] Josue Baeza, James A. Dowell, Michael J. Smallegan, Jing Fan, Daniel Amador-Noguez, Zia Khan, and John M. Denu. Stoichiometry of Site-specific Lysine Acetylation in an Entire Proteome. *Journal of Biological Chemistry*, 289(31):21326–21338, January 2014. ISSN 0021-9258, 1083-351X. doi: 10.1074/jbc.M114.581843.

[175] Kirill A. Datsenko and Barry L. Wanner. One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proceedings of the National Academy of Sciences*, 97(12):6640–6645, June 2000. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.120163297.

[176] Sayed K. Goda and Nigel P. Minton. A simple procedure for gel electrophoresis and Northern blotting of RNA. *Nucleic Acids Research*, 23(16):3357–3358, August 1995. ISSN 0305-1048. doi: 10.1093/nar/23.16.3357.

[177] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):pp. 10–12, May 2011. ISSN 2226-6089. doi: 10.14806/ej.17.1.200.

[178] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, page btu170, April 2014. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btu170.

[179] Simon Andrews. FastQC: A quality control tool for high throughput sequence data, 2010.

[180] Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Käller. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19): 3047–3048, October 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw354.

[181] Peter L. Freddolino, Sasan Amini, and Saeed Tavazoie. Newly Identified Genetic Variations in Common Escherichia coli MG1655 Stock Cultures. *Journal of Bacteriology*, 194(2): 303–306, January 2012. ISSN 0021-9193. doi: 10.1128/JB.06087-11.

[182] Ben Langmead and Steven L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4):357–359, April 2012. ISSN 1548-7091. doi: 10.1038/nmeth.1923.

[183] 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*, 25(16):2078–2079, August 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp352.

[184] Frank R. Hampel. The Influence Curve and its Role in Robust Estimation. *Journal of the American Statistical Association*, 69(346):383–393, June 1974. ISSN 0162-1459. doi: 10.1080/01621459.1974.10482962.

[185] Peter J. Rousseeuw and Christophe Croux. Alternatives to the Median Absolute Deviation. *Journal of the American Statistical Association*, 88(424):1273–1283, December 1993. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.1993.10476408.

[186] Wei Simko. R package ”corrplot”: Visualization of a Correlation Matrix, 2017.

[187] Eric Jones, Travis Oliphant, Pearu Peterson, and others. SciPy: Open Source Scientific Tools for Python, 2001.

[188] Yosef Hochberg and Yoav Benjamini. More powerful procedures for multiple significance testing. *Statistics in Medicine*, 9(7):811–818, 1990. ISSN 1097-0258. doi: 10.1002/sim.4780090710.

[189] Qunhua Li, James B. Brown, Haiyan Huang, and Peter J. Bickel. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics*, 5(3):1752–1779, September 2011. ISSN 1932-6157. doi: 10.1214/11-AOAS466.

[190] Nicolas L. Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525–527, May 2016. ISSN 1546-1696. doi: 10.1038/nbt.3519.

[191] Harold Pimentel, Nicolas L. Bray, Suzette Puente, Páll Melsted, and Lior Pachter. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nature Methods*, 14(7): 687–690, July 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4324.

[192] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome biol*, 11(10):R106, 2010.

[193] Maureen K. Thomason, Thorsten Bischler, Sara K. Eisenbart, Konrad U. Förstner, Aixia Zhang, Alexander Herbig, Kay Nieselt, Cynthia M. Sharma, and Gisela Storz. Global Transcriptional Start Site Mapping Using Differential RNA Sequencing Reveals Novel Antisense RNAs in Escherichia coli. *Journal of Bacteriology*, 197(1):18–28, January 2015. ISSN 0021-9193, 1098-5530. doi: 10.1128/JB.02096-14.

[194] Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[195] Hadley Wickham. *Ggplot2: Elegant Graphics for Data Analysis*. September 2015.

[196] J. D. Hunter. Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering*, 9(3):90–95, May 2007. ISSN 1521-9615. doi: 10.1109/MCSE.2007.55.

[197] Scott A. Scholz, Rucheng Diao, Michael B. Wolfe, Elayne M. Fivenson, Xiaoxia Nina Lin, and Peter L. Freddolino. High-Resolution Mapping of the Escherichia coli Chromosome Reveals Positions of High and Low Transcription. *Cell Systems*, 8(3):212–225.e9, March 2019. ISSN 2405-4712. doi: 10.1016/j.cels.2019.02.004.

[198] Virginia S. Lioy, Axel Cournac, Martial Marbouty, Stéphane Duigou, Julien Mozziconacci, Olivier Espéli, Frédéric Boccard, and Romain Koszul. Multiscale Structuring of the E. coli Chromosome by Nucleoid-Associated and Condensin Proteins. *Cell*, 172(4):771–783.e18, February 2018. ISSN 0092-8674. doi: 10.1016/j.cell.2017.12.027.

[199] Somenath Bakshi, Heejun Choi, and James C. Weisshaar. The spatial biology of transcription and translation in rapidly growing Escherichia coli. *Frontiers in Microbiology*, 6, 2015. ISSN 1664-302X. doi: 10.3389/fmicb.2015.00636.

[200] Qian Chai, Bhupender Singh, Kristin Peisker, Nicole Metzendorf, Xueliang Ge, Santanu Dasgupta, and Suparna Sanyal. Organization of Ribosomes and Nucleoids in Escherichia coli Cells during Growth and in Quiescence. *Journal of Biological Chemistry*, 289(16): 11342–11352, April 2014. ISSN 0021-9258, 1083-351X. doi: 10.1074/jbc.M114.557348.

[201] Ding Jun Jin and Julio E. Cabrera. Coupling the distribution of RNA polymerase to global gene regulation and the dynamic structure of the bacterial nucleoid in Escherichia coli. *Journal of Structural Biology*, 156(2):284–291, November 2006. ISSN 1047-8477. doi: 10.1016/j.jsb.2006.07.005.

[202] Tung B. K. Le, Maxim V. Imakaev, Leonid A. Mirny, and Michael T. Laub. High-Resolution Mapping of the Spatial Organization of a Bacterial Chromosome. *Science*, 342(6159):731–734, November 2013. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1242059.

[203] Martial Marbouty, Antoine Le Gall, Diego I. Cattoni, Axel Cournac, Alan Koh, Jean-Bernard Fiche, Julien Mozziconacci, Heath Murray, Romain Koszul, and Marcelo Nollmann. Condensin- and Replication-Mediated Bacterial Chromosome Folding and Origin Condensation Revealed by Hi-C and Super-resolution Imaging. *Molecular Cell*, 59(4):588–602, August 2015. ISSN 1097-2765. doi: 10.1016/j.molcel.2015.07.020.

[204] Xindan Wang, Tung B. K. Le, Bryan R. Lajoie, Job Dekker, Michael T. Laub, and David Z. Rudner. Condensin promotes the juxtaposition of DNA flanking its loading site in Bacillus subtilis. *Genes & Development*, 29(15):1661–1675, January 2015. ISSN 0890-9369, 1549-5477. doi: 10.1101/gad.265876.115.

[205] Julio E. Cabrera and Ding J. Jin. Active Transcription of rRNA Operons Is a Driving Force for the Distribution of RNA Polymerase in Bacteria: Effect of Extrachromosomal Copies of rrnB on the In Vivo Localization of RNA Polymerase. *Journal of Bacteriology*, 188(11): 4007–4014, June 2006. ISSN 0021-9193, 1098-5530. doi: 10.1128/JB.01893-05.

[206] Tamas Gaal, Benjamin P. Bratton, Patricia Sanchez-Vazquez, Alexander Sliwicki, Kristine Sliwicki, Andrew Vegel, Rachel Pannu, and Richard L. Gourse. Colocalization of distant chromosomal loci in space in E. coli: A bacterial nucleolus. *Genes & Development*, 30(20): 2272–2285, October 2016. ISSN 0890-9369, 1549-5477. doi: 10.1101/gad.290312.116.

[207] Jonathan R. Beckwith, Ethan R. Signer, and Wolfgang Epstein. Transposition of the Lac Region of E. coli. *Cold Spring Harbor Symposia on Quantitative Biology*, 31:393–401, January 1966. ISSN 0091-7451, 1943-4456. doi: 10.1101/SQB.1966.031.01.051.

[208] M. B. Schmid and J. R. Roth. Gene location affects expression level in Salmonella typhimurium. *Journal of Bacteriology*, 169(6):2872–2875, June 1987. ISSN 0021-9193, 1098-5530. doi: 10.1128/jb.169.6.2872-2875.1987.

[209] Stephen Cooper and Charles E. Helmstetter. Chromosome replication and the division cycle of Escherichia coli Br. *Journal of Molecular Biology*, 31(3):519–540, February 1968. ISSN 0022-2836. doi: 10.1016/0022-2836(68)90425-7.

[210] Millicent Masters. The frequency of P1 transduction of the genes of Escherichia coli as a function of chromosomal position: Preferential transduction of the origin of replication. *Molecular and General Genetics MGG*, 155(2):197–202, January 1977. ISSN 1432-1874. doi: 10.1007/BF00393160.

[211] Dena H. S. Block, Razika Hussein, Lusha W. Liang, and Han N. Lim. Regulatory consequences of gene translocation in bacteria. *Nucleic Acids Research*, 40(18):8979–8992, October 2012. ISSN 0305-1048. doi: 10.1093/nar/gks694.

[212] Enoch Yeung, Aaron J. Dy, Kyle B. Martin, Andrew H. Ng, Domitilla Del Vecchio, James L. Beck, James J. Collins, and Richard M. Murray. Biophysical Constraints Arising from Compositional Context in Synthetic Gene Networks. *Cell Systems*, 5(1):11–24.e12, July 2017. ISSN 2405-4712. doi: 10.1016/j.cels.2017.06.001.

[213] Charles J. Dorman. Dna Supercoiling and Bacterial Gene Expression. *Science Progress*, 89 (3-4):151–166, August 2006. ISSN 0036-8504. doi: 10.3184/003685006783238317.

[214] Elisa Brambilla and Bianca Sclavi. Gene Regulation by H-NS as a Function of Growth Conditions Depends on Chromosomal Position in Escherichia coli. *G3: Genes, Genomes, Genetics*, 5(4):605–614, April 2015. ISSN 2160-1836. doi: 10.1534/g3.114.016139.

[215] Agustino Martínez-Antonio, Alejandra Medina-Rivera, and Julio Collado-Vides. Structural and functional map of a bacterial nucleoid. *Genome Biology*, 10(12):247, December 2009. ISSN 1474-760X. doi: 10.1186/gb-2009-10-12-247.

[216] Sam Meyer, Sylvie Reverchon, William Nasser, and Georgi Muskhelishvili. Chromosomal organization of transcription: In a nutshell. *Current Genetics*, 64(3):555–565, June 2018. ISSN 1432-0983. doi: 10.1007/s00294-017-0785-5.

[217] Ying-Ja Chen, Peng Liu, Alec A. K. Nielsen, Jennifer A. N. Brophy, Kevin Clancy, Todd Peterson, and Christopher A. Voigt. Characterization of 582 natural and synthetic terminators and quantification of their design constraints. *Nature Methods*, 10(7):659–664, July 2013. ISSN 1548-7105. doi: 10.1038/nmeth.2515.

[218] D. Clavel, G. Gotthard, D. von Stetten, D. De Sanctis, H. Pasquier, G. G. Lambert, N. C. Shaner, and A. Royant. Structural analysis of the bright monomeric yellow-green fluorescent protein mNeonGreen obtained by directed evolution. *Acta Crystallographica Section D: Structural Biology*, 72(12):1298–1307, December 2016. ISSN 2059-7983. doi: 10.1107/S2059798316018623.

[219] Sriram Kosuri, Daniel B. Goodman, Guillaume Cambray, Vivek K. Mutalik, Yuan Gao, Adam P. Arkin, Drew Endy, and George M. Church. Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proceedings of the National Academy of Sciences*, 110(34):14024–14029, August 2013. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1301301110.

[220] Douglas W. Selinger, Rini Mukherjee Saxena, Kevin J. Cheung, George M. Church, and Carsten Rosenow. Global RNA Half-Life Analysis in Escherichia coli Reveals Positional Patterns of Transcript Degradation. *Genome Research*, 13(2):216–223, January 2003. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.912603.

[221] Huiyi Chen, Katsuyuki Shiroguchi, Hao Ge, and Xiaoliang Sunney Xie. Genome-wide study of mRNA degradation and transcript elongation in Escherichia coli. *Molecular Systems Biology*, 11(5):808, May 2015. ISSN 1744-4292, 1744-4292. doi: 10.15252/msb.20159000.

[222] Jonathan A. Bernstein, Arkady B. Khodursky, Pei-Hsun Lin, Sue Lin-Chao, and Stanley N. Cohen. Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. *Proceedings of the National Academy of Sciences*, 99(15):9697–9702, July 2002. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.112318199.

[223] N. R. Lomb. Least-squares frequency analysis of unequally spaced data. *Astrophysics and Space Science*, 39(2):447–462, February 1976. ISSN 1572-946X. doi: 10.1007/BF00648343.

[224] J. D. Scargle. Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysical Journal*, 263:835–853, December 1982. ISSN 0004-637X. doi: 10.1086/160554.

[225] Lisa Postow, Christine D. Hardy, Javier Arsuaga, and Nicholas R. Cozzarelli. Topological domain structure of the Escherichia coli chromosome. *Genes & Development*, 18(14):1766–1779, July 2004. ISSN 0890-9369, 1549-5477. doi: 10.1101/gad.1207504.

[226] William S. Cleveland and Susan J. Devlin. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, 83(403):596–610, September 1988. ISSN 0162-1459. doi: 10.1080/01621459.1988.10478639.

[227] R R Sinden and D E Pettijohn. Chromosomes in living Escherichia coli cells are segregated into domains of supercoiling. *Proceedings of the National Academy of Sciences of the United States of America*, 78(1):224–228, January 1981. ISSN 0027-8424.

[228] Avantika Lal, Amlanjyoti Dhar, Andrei Trostel, Fedor Kouzine, Aswin S. N. Seshasayee, and Sankar Adhya. Genome scale patterns of supercoiling in a bacterial chromosome. *Nature Communications*, 7:11055, March 2016. ISSN 2041-1723. doi: 10.1038/ncomms11055.

[229] Jeffrey R Moffitt, Shristi Pandey, Alistair N Boettiger, Siyuan Wang, and Xiaowei Zhuang. Spatial organization shapes the turnover of a bacterial transcriptome. *eLife*, 5, May 2016. ISSN 2050-084X. doi: 10.7554/eLife.13065.

[230] Ana I. Prieto, Christina Kahramanoglou, Ruhi M. Ali, Gillian M. Fraser, Aswin S. N. Seshasayee, and Nicholas M. Luscombe. Genomic analysis of DNA binding and gene regulation by homologous nucleoid-associated proteins IHF and HU in Escherichia coli K12. *Nucleic Acids Research*, 40(8):3524–3537, April 2012. ISSN 0305-1048. doi: 10.1093/nar/gkr1236.

[231] Mohan C. Joshi, David Magnan, Timothy P. Montminy, Mark Lies, Nicholas Stepankiw, and David Bates. Regulation of Sister Chromosome Cohesion by the Replication Fork Tracking Protein SeqA. *PLOS Genetics*, 9(8):e1003673, August 2013. ISSN 1553-7404. doi: 10.1371/journal.pgen.1003673.

[232] Frederick R. Blattner, Guy Plunkett, Craig A. Bloch, Nicole T. Perna, Valerie Burland, Monica Riley, Julio Collado-Vides, Jeremy D. Glasner, Christopher K. Rode, George F. Mayhew, Jason Gregor, Nelson Wayne Davis, Heather A. Kirkpatrick, Michael A. Goeden, Debra J. Rose, Bob Mau, and Ying Shao. The Complete Genome Sequence of Escherichia coli K-12. *Science*, 277(5331):1453–1462, September 1997. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.277.5331.1453.

[233] Rachel A. Mooney, Sarah E. Davis, Jason M. Peters, Jennifer L. Rowland, Aseem Z. Ansari, and Robert Landick. Regulator Trafficking on Bacterial Transcription Units In Vivo. *Molecular Cell*, 33(1):97–108, January 2009. ISSN 1097-2765. doi: 10.1016/j.molcel.2008.12.021.

[234] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis,

John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25, May 2000. ISSN 1546-1718. doi: 10.1038/75556.

[235] Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research*, 45 (D1):D331–D338, January 2017. ISSN 0305-1048. doi: 10.1093/nar/gkw1108.

[236] Sean S. J. Heng, Oliver Y. W. Chan, Bryan M. H. Keng, and Maurice H. T. Ling. Glucan Biosynthesis Protein G Is a Suitable Reference Gene in Escherichia coli K-12. https://www.hindawi.com/journals/isrn/2011/469053/, 2011.

[237] Waseem Akhtar, Johann de Jong, Alexey V. Pindyurin, Ludo Pagie, Wouter Meuleman, Jeroen de Ridder, Anton Berns, Lodewyk F. A. Wessels, Maarten van Lohuizen, and Bas van Steensel. Chromatin Position Effects Assayed by Thousands of Reporters Integrated in Parallel. *Cell*, 154(4):914–927, August 2013. ISSN 0092-8674. doi: 10.1016/j.cell.2013.07.018.

[238] Da-Eun Jeong, Younju So, Soo-Young Park, Seung-Hwan Park, and Soo-Keun Choi. Random knock-in expression system for high yield production of heterologous protein in Bacillus subtilis. *Journal of Biotechnology*, 266:50–58, January 2018. ISSN 0168-1656. doi: 10.1016/j.jbiotec.2017.12.007.

[239] S. L. French and O. L. Miller. Transcription mapping of the Escherichia coli chromosome by electron microscopy. *Journal of Bacteriology*, 171(8):4207–4216, August 1989. ISSN 0021-9193, 1098-5530. doi: 10.1128/jb.171.8.4207-4216.1989.

[240] Brian J. Paul, Wilma Ross, Tamas Gaal, and Richard L. Gourse. rRNA Transcription in Escherichia coli. *Annual Review of Genetics*, 38(1):749–770, November 2004. ISSN 0066-4197. doi: 10.1146/annurev.genet.38.072902.091347.

[241] Byung-Kwan Cho, Eric M. Knight, Christian L. Barrett, and Bernhard Ø Palsson. Genome-wide analysis of Fis binding in Escherichia coli indicates a causative role for A-/AT-tracts. *Genome Research*, 18(6):900–910, January 2008. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.070276.107.

[242] Christine A. Hirvonen, Wilma Ross, Christopher E. Wozniak, Erin Marasco, Jennifer R. Anthony, Sarah E. Aiyar, Vanessa H. Newburn, and Richard L. Gourse. Contributions of UP Elements and the Transcription Factor FIS to Expression from the Seven rrn P1 Promoters inEscherichia coli. *Journal of Bacteriology*, 183(21):6305–6314, November 2001. ISSN 0021-9193, 1098-5530. doi: 10.1128/JB.183.21.6305-6314.2001.

[243] Yunfeng Gao, Yong Hwee Foo, Ricksen S. Winardhi, Qingnan Tang, Jie Yan, and Linda J. Kenney. Charged residues in the H-NS linker drive DNA binding and gene silencing in single cells. *Proceedings of the National Academy of Sciences*, 114(47):12560–12565, November 2017. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1716721114.

[244] Aathmaja Anandhi Rangarajan and Karin Schnetz. Interference of transcription across H-NS binding sites and repression by H-NS. *Molecular Microbiology*, 108(3):226–239, 2018. ISSN 1365-2958. doi: 10.1111/mmi.13926.

[245] Jason D. Buenrostro, Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang, and William J. Greenleaf. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, 10(12):1213–1218, December 2013. ISSN 1548-7105. doi: 10.1038/nmeth.2688.

[246] Charles J. Dorman. H-NS-like nucleoid-associated proteins, mobile genetic elements and horizontal gene transfer in bacteria. *Plasmid*, 75:1–11, September 2014. ISSN 0147-619X. doi: 10.1016/j.plasmid.2014.06.004.

[247] Ferric C Fang and Sylvie Rimsky. New insights into transcriptional regulation by H-NS. *Current Opinion in Microbiology*, 11(2):113–120, April 2008. ISSN 1369-5274. doi: 10.1016/j.mib.2008.02.011.

[248] Koichi Higashi, Toru Tobe, Akinori Kanai, Ebru Uyar, Shu Ishikawa, Yutaka Suzuki, Naotake Ogasawara, Ken Kurokawa, and Taku Oshima. H-NS Facilitates Sequence Diversification of Horizontally Transferred DNAs during Their Integration in Host Chromosomes. *PLOS Genetics*, 12(1):e1005796, January 2016. ISSN 1553-7404. doi: 10.1371/journal.pgen.1005796.

[249] Sabrina S. Ali, Jeremy Soo, Chitong Rao, Andrea S. Leung, David Hon-Man Ngai, Alexander W. Ensminger, and William Wiley Navarre. Silencing by H-NS Potentiated the Evolution of Salmonella. *PLOS Pathogens*, 10(11):e1004500, November 2014. ISSN 1553-7374. doi: 10.1371/journal.ppat.1004500.

[250] Ebru Uyar, Ken Kurokawa, Mika Yoshimura, Shu Ishikawa, Naotake Ogasawara, and Taku Oshima. Differential Binding Profiles of StpA in Wild-Type and hns Mutant Cells: A Comparative Analysis of Cooperative Partners by Chromatin Immunoprecipitation-Microarray Analysis. *Journal of Bacteriology*, 191(7):2388–2391, April 2009. ISSN 0021-9193, 1098-5530. doi: 10.1128/JB.01594-08.

[251] Kevin S. Lang, Ashley N. Hall, Christopher N. Merrikh, Mark Ragheb, Hannah Tabakh, Alex J. Pollock, Joshua J. Woodward, Julia E. Dreifus, and Houra Merrikh. Replication-Transcription Conflicts Generate R-Loops that Orchestrate Bacterial Stress Survival and Pathogenesis. *Cell*, 170(4):787–799.e18, August 2017. ISSN 00928674. doi: 10.1016/j.cell.2017.07.044.

[252] M Demerec and P E Hartman. Complex Loci in Microorganisms. *Annual Review of Microbiology*, 13(1):377–406, October 1959. ISSN 0066-4227. doi: 10.1146/annurev.mi.13.100159.002113.

[253] Jeffrey Lawrence. Selfish operons: The evolutionary impact of gene clustering in prokaryotes and eukaryotes. *Current Opinion in Genetics & Development*, 9(6):642–648, December 1999. ISSN 0959-437X. doi: 10.1016/S0959-437X(99)00025-8.

[254] Howard Ochman, Jeffrey G. Lawrence, and Eduardo A. Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299, May 2000. ISSN 1476-4687. doi: 10.1038/35012500.

[255] AliAzam Talukder and Akira Ishihama. Growth phase dependent changes in the structure and protein composition of nucleoid in Escherichia coli. *Science China Life Sciences*, 58 (9):902–911, September 2015. ISSN 1869-1889. doi: 10.1007/s11427-015-4898-0.

[256] Tony Kouzarides. Chromatin Modifications and Their Function. *Cell*, 128(4):693–705, February 2007. ISSN 0092-8674. doi: 10.1016/j.cell.2007.02.005.

[257] Christian Lanctôt, Thierry Cheutin, Marion Cremer, Giacomo Cavalli, and Thomas Cremer. Dynamic genome architecture in the nuclear space: Regulation of gene expression in three dimensions. *Nature Reviews Genetics*, 8(2):104–115, February 2007. ISSN 1471-0064. doi: 10.1038/nrg2041.

[258] Peter P. Cherepanov and Wilfried Wackernagel. Gene disruption in Escherichia coli: TcR and KmR cassettes with the option of Flp-catalyzed excision of the antibiotic-resistance determinant. *Gene*, 158(1):9–14, January 1995. ISSN 0378-1119. doi: 10.1016/0378-1119(95)00193-A.

[259] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10: R25, 2009. ISSN 1474-760X. doi: 10.1186/gb-2009-10-3-r25.

[260] Nathan C. Shaner, Gerard G. Lambert, Andrew Chammas, Yuhui Ni, Paula J. Cranfill, Michelle A. Baird, Brittney R. Sell, John R. Allen, Richard N. Day, Maria Israelsson, Michael W. Davidson, and Jiwu Wang. A bright monomeric green fluorescent protein derived from *Branchiostoma lanceolatum*. *Nature Methods*, 10(5):407–409, May 2013. ISSN 1548-7105. doi: 10.1038/nmeth.2413.

[261] Amin Espah Borujeni, Anirudh S. Channarasappa, and Howard M. Salis. Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic Acids Research*, 42(4):2646–2659, February 2014. ISSN 0305-1048. doi: 10.1093/nar/gkt1139.

[262] Tomoya Baba, Takeshi Ara, Miki Hasegawa, Yuki Takai, Yoshiko Okumura, Miki Baba, Kirill A. Datsenko, Masaru Tomita, Barry L. Wanner, and Hirotada Mori. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: The Keio collection. *Molecular Systems Biology*, 2(1):2006.0008, January 2006. ISSN 1744-4292, 1744-4292. doi: 10.1038/msb4100050.

[263] Douglas Hanahan, Joel Jessee, and Fredric R. Bloom. [4] Plasmid transformation of Escherichia coli and other bacteria. In *Methods in Enzymology*, volume 204 of *Bacterial Genetic Systems*, pages 63–113. Academic Press, January 1991. doi: 10.1016/0076-6879(91) 04006-A.

[264] H.S. Girgis, Y. Liu, W.S. Ryu, and S. Tavazoie. A comprehensive genetic characterization of bacterial motility. *PLoS Genetics*, 3(9):1644–1660, 2007. doi: 10.1371/journal.pgen.0030154.

[265] Confidence intervals for sample autocorrelation - MATLAB & Simulink. The MathWorks, Inc., 2018.

[266] Thomas P. Robitaille, Erik J. Tollerud, Perry Greenfield, Michael Droettboom, Erik Bray, Tom Aldcroft, Matt Davis, Adam Ginsburg, Adrian M. Price-Whelan, Wolfgang E. Kerzendorf, Alexander Conley, Neil Crighton, Kyle Barbary, Demitri Muna, Henry Ferguson, Frédéric Grollier, Madhura M. Parikh, Prasanth H. Nair, Hans M. Günther, Christoph Deil, Julien Woillez, Simon Conseil, Roban Kramer, James E. H. Turner, Leo Singer, Ryan Fox, Benjamin A. Weaver, Victor Zabalza, Zachary I. Edwards, K. Azalee Bostroem, D. J. Burke, Andrew R. Casey, Steven M. Crawford, Nadia Dencheva, Justin Ely, Tim Jenness, Kathleen Labrie, Pey Lian Lim, Francesco Pierfederici, Andrew Pontzen, Andy Ptak, Brian Refsdal, Mathieu Servillat, and Ole Streicher. Astropy: A community Python package for astronomy. *Astronomy & Astrophysics*, 558:A33, October 2013. ISSN 0004-6361, 1432-0746. doi: 10.1051/0004-6361/201322068.

[267] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software*, 33(1):1–22, 2010. ISSN 1548-7660.

[268] Michael B. Wolfe, Aaron C. Goldstrohm, and Peter L. Freddolino. Global analysis of RNA metabolism using bio-orthogonal labeling coupled with next-generation RNA sequencing. *Methods*, December 2018. ISSN 1046-2023. doi: 10.1016/j.ymeth.2018.12.001.

[269] Nicole L. Garneau, Jeffrey Wilusz, and Carol J. Wilusz. The highways and byways of mRNA decay. *Nature Reviews Molecular Cell Biology*, 8(2):113–126, February 2007. ISSN 1471-0080. doi: 10.1038/nrm2104.

[270] Jonathan Houseley and David Tollervey. The Many Pathways of RNA Degradation. *Cell*, 136(4):763–776, February 2009. ISSN 0092-8674. doi: 10.1016/j.cell.2009.01.019.

[271] David P. Bartel. Metazoan MicroRNAs. *Cell*, 173(1):20–51, March 2018. ISSN 0092-8674. doi: 10.1016/j.cell.2018.03.006.

[272] T C Santiago, I J Purvis, A J Bettany, and A J Brown. The relationship between mRNA stability and length in Saccharomyces cerevisiae. *Nucleic Acids Research*, 14(21):8347–8360, November 1986. ISSN 0305-1048.

[273] M. Nonet, C. Scafe, J. Sexton, and R. Young. Eucaryotic RNA polymerase conditional mutant that rapidly ceases mRNA synthesis. *Molecular and Cellular Biology*, 7(5):1602–1611, May 1987. ISSN 0270-7306, 1098-5549. doi: 10.1128/MCB.7.5.1602.

[274] D. Herrick, R. Parker, and A. Jacobson. Identification and comparison of stable and unstable mRNAs in Saccharomyces cerevisiae. *Molecular and Cellular Biology*, 10(5):2269–2284, January 1990. ISSN 0270-7306, 1098-5549. doi: 10.1128/MCB.10.5.2269.

[275] J. Ross. mRNA stability in mammalian cells. *Microbiological Reviews*, 59(3):423–450, January 1995. ISSN 1092-2172, 1098-5557.

[276] Alistair J. P. Brown and Francis A. Sagliocco. mRNA Abundance and Half-Life Measurements. In *Yeast Protocols*, Methods in Molecular Biology™, pages 277–295. Humana Press, 1996. ISBN 978-0-89603-319-1. doi: 10.1385/0-89603-319-8:277.

[277] Michael D. Cleary, Christopher D. Meiering, Eric Jan, Rebecca Guymon, and John C. Boothroyd. Biosynthetic labeling of RNA with uracil phosphoribosyltransferase allows cell-specific microarray analysis of mRNA synthesis and decay. *Nature Biotechnology*, 23(2): 232–237, February 2005. ISSN 1546-1696. doi: 10.1038/nbt1061.

[278] Sarah E. Munchel, Ryan K. Shultzaberger, Naoki Takizawa, and Karsten Weis. Dynamic profiling of mRNA turnover reveals gene-specific and system-wide regulation of mRNA decay. *Molecular Biology of the Cell*, 22(15):2787–2795, August 2011. ISSN 1059-1524, 1939-4586. doi: 10.1091/mbc.e11-01-0028.

[279] R. Knüppel, C. Kuttenberger, and S. Ferreira-Cerca. Toward Time-Resolved Analysis of RNA Metabolism in Archaea Using 4-Thiouracil. *Frontiers in microbiology*, 8:286–286, 2017. ISSN 1664-302X. doi: 10.3389/fmicb.2017.00286.

[280] L. Dolken, Z. Ruzsics, B. Radle, C. C. Friedel, R. Zimmer, J. Mages, R. Hoffmann, P. Dickinson, T. Forster, P. Ghazal, and U. H. Koszinowski. High-resolution gene expression profiling for simultaneous kinetic parameter analysis of RNA synthesis and decay. *RNA*, 14(9): 1959–1972, July 2008. ISSN 1355-8382. doi: 10.1261/rna.1136108.

[281] Michal Rabani, Joshua Z Levin, Lin Fan, Xian Adiconis, Raktima Raychowdhury, Manuel Garber, Andreas Gnirke, Chad Nusbaum, Nir Hacohen, Nir Friedman, Ido Amit, and Aviv Regev. Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nature Biotechnology*, 29(5):436–442, May 2011. ISSN 1087-0156, 1546-1696. doi: 10.1038/nbt.1861.

[282] Björn Schwanhäusser, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, and Matthias Selbach. Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342, May 2011. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature10098.

[283] William T. Melvin, Helen B. Milne, Alison A. Slater, Hamish J. Allen, and Hamish M. Keir. Incorporation of 6-Thioguanosine and 4-Thiouridine into RNA. *European Journal of Biochemistry*, 92(2):373–379, December 1978. ISSN 1432-1033. doi: 10.1111/j.1432-1033.1978.tb12756.x.

[284] Jessica Spitzer, Markus Hafner, Markus Landthaler, Manuel Ascano, Thalia Farazi, Greg Wardle, Jeff Nusbaum, Mohsen Khorshid, Lukas Burger, Mihaela Zavolan, and Thomas Tuschl. Chapter Eight - PAR-CLIP (Photoactivatable Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation): A Step-By-Step Protocol to the Transcriptome-Wide Identification of Binding Sites of RNA-Binding Proteins. In Jon Lorsch, editor, *Methods in Enzymology*, volume 539 of *Laboratory Methods in Enzymology: Protein Part B*, pages 113–161. Academic Press, January 2014. doi: 10.1016/B978-0-12-420120-0.00008-6.

[285] H. Tani, R. Mizutani, K. A. Salam, K. Tano, K. Ijiri, A. Wakamatsu, T. Isogai, Y. Suzuki, and N. Akimitsu. Genome-wide determination of RNA stability reveals hundreds of short-lived noncoding transcripts in mammals. *Genome Research*, 22(5):947–956, May 2012. ISSN 1088-9051. doi: 10.1101/gr.130559.111.

[286] J. L. Aspden and R. J. Jackson. Differential effects of nucleotide analogs on scanning-dependent initiation and elongation of mammalian mRNA translation in vitro. *RNA*, 16(6): 1130–1137, June 2010. ISSN 1355-8382. doi: 10.1261/rna.1978610.

[287] E. R. Pfefferkorn and L. C. Pfefferkorn. Specific labeling of intracellular Toxoplasma gondii with uracil. *The Journal of Protozoology*, 24(3):449–453, August 1977. ISSN 0022-3921.

[288] Leslie Gay, Kate V. Karfilis, Michael R. Miller, Chris Q. Doe, and Kryn Stankunas. Applying thiouracil tagging to mouse transcriptome analysis. *Nature Protocols*, 9(2):410–420, February 2014. ISSN 1750-2799. doi: 10.1038/nprot.2014.023.

[289] Christina Chatzi, Yingyu Zhang, Rongkun Shen, Gary L. Westbrook, and Richard H. Goodman. Transcriptional Profiling of Newly Generated Dentate Granule Cells Using TU Tagging Reveals Pattern Shifts in Gene Expression during Circuit Integration,. *eNeuro*, 3(1), March 2016. ISSN 2373-2822. doi: 10.1523/ENEURO.0024-16.2016.

[290] Michael D. Cleary. Chapter 19 Cell Type–Specific Analysis of mRNA Synthesis and Decay In Vivo with Uracil Phosphoribosyltransferase and 4-thiouracil. In *Methods in Enzymology*, volume 448, pages 379–406. Elsevier, 2008. ISBN 978-0-12-374378-7. doi: 10.1016/S0076-6879(08)02619-0.

[291] Erik J. Sontheimer. Site-specific RNA crosslinking with 4-thiouridine. *Molecular Biology Reports*, 20(1):35–44, July 1994. ISSN 0301-4851, 1573-4978. doi: 10.1007/BF00999853.

[292] Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano, Anna-Carina Jungkamp, Mathias Munschauer, Alexander Ulrich, Greg S. Wardle, Scott Dewell, Mihaela Zavolan, and Thomas Tuschl. Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell*, 141(1):129–141, April 2010. ISSN 00928674. doi: 10.1016/j.cell.2010.03.009.

[293] Stephen Zamenhof and Gertrude Griboff. Incorporation of Halogenated Pyrimidines into the Deoxyribonucleic Acids of Bacterium Coli and its Bacteriophages. *Nature*, 174(4424): 306–307, August 1954. ISSN 1476-4687. doi: 10.1038/174306a0.

[294] M. L. Eidinoff, L. Cheong, and M. A. Rich. Incorporation of Unnatural Pyrimidine Bases into Deoxyribonucleic Acid of Mammalian Cells. *Science*, 129(3362):1550–1551, June 1959. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.129.3362.1550.

[295] H. G. Gratzner. Monoclonal antibody to 5-bromo- and 5-iododeoxyuridine: A new reagent for detection of DNA replication. *Science*, 218(4571):474–475, October 1982. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.7123245.

[296] S.Raza Haider, Gloria Juan, Frank Traganos, and Zbigniew Darzynkiewicz. Immunoseparation and Immunodetection of Nucleic Acids Labeled with Halogenated Nucleotides. *Experimental Cell Research*, 234(2):498–506, August 1997. ISSN 00144827. doi: 10.1006/excr.1997.3644.

[297] M. Ohtsu, M. Kawate, M. Fukuoka, W. Gunji, F. Hanaoka, T. Utsugi, F. Onoda, and Y. Murakami. Novel DNA Microarray System for Analysis of Nascent mRNAs. *DNA Research*, 15(4):241–251, May 2008. ISSN 1340-2838, 1756-1663. doi: 10.1093/dnares/dsn015.

[298] Michael D. Best. Click Chemistry and Bioorthogonal Reactions: Unprecedented Selectivity in the Labeling of Biological Molecules. *Biochemistry*, 48(28):6571–6584, July 2009. ISSN 0006-2960, 1520-4995. doi: 10.1021/bi9007726.

[299] C. Y. Jao and A. Salic. Exploring RNA transcription and turnover in vivo by using click chemistry. *Proceedings of the National Academy of Sciences*, 105(41):15779–15784, October 2008. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0808480105.

[300] Kaito Abe, Tomoaki Ishigami, Ann-Bin Shyu, Shigeo Ohno, Satoshi Umemura, and Akio Yamashita. Analysis of interferon-beta mRNA stability control after poly(I:C) stimulation using RNA metabolic labeling by ethynyluridine. *Biochemical and Biophysical Research Communications*, 428(1):44–49, November 2012. ISSN 0006291X. doi: 10.1016/j.bbrc.2012.09.144.

[301] Takashi Ideue, Shungo Adachi, Takao Naganuma, Akie Tanigawa, Tohru Natsume, and Tetsuro Hirose. U7 small nuclear ribonucleoprotein represses histone gene transcription in cell cycle-arrested cells. *Proceedings of the National Academy of Sciences*, 109(15):5693–5698, April 2012. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1200523109.

[302] Dezhong Qu, Li Zhou, Wei Wang, Zhe Wang, Guoxin Wang, Weilin Chi, and Biliang Zhang. 5-Ethynylcytidine as a new agent for detecting RNA synthesis in live cells by "click" chemistry. *Analytical Biochemistry*, 434(1):128–135, March 2013. ISSN 00032697. doi: 10.1016/j.ab.2012.11.023.

[303] Naoki Hida, Mohamed Y. Aboukilila, Dana A. Burow, Rakesh Paul, Marc M. Greenberg, Michael Fazio, Samantha Beasley, Robert C. Spitale, and Michael D. Cleary. EC-tagging allows cell type-specific RNA analysis. *Nucleic Acids Research*, 45(15):e138–e138, September 2017. ISSN 0305-1048. doi: 10.1093/nar/gkx551.

[304] Xichen Bao, Xiangpeng Guo, Menghui Yin, Muqddas Tariq, Yiwei Lai, Shahzina Kanwal, Jiajian Zhou, Na Li, Yuan Lv, Carlos Pulido-Quetglas, Xiwei Wang, Lu Ji, Muhammad J. Khan, Xihua Zhu, Zhiwei Luo, Changwei Shao, Do-Hwan Lim, Xiao Liu, Nan Li, Wei Wang, Minghui He, Yu-Lin Liu, Carl Ward, Tong Wang, Gong Zhang, Dongye Wang, Jianhua Yang, Yiwen Chen, Chaolin Zhang, Ralf Jauch, Yun-Gui Yang, Yangming Wang, Baoming Qin, Minna-Liisa Anko, Andrew P. Hutchins, Hao Sun, Huating Wang, Xiang-Dong Fu, Biliang Zhang, and Miguel A. Esteban. Capturing the interactome of newly transcribed RNA. *Nature Methods*, 15(3):213–220, March 2018. ISSN 1548-7105. doi: 10.1038/nmeth.4595.

[305] Ian A. Roundtree, Molly E. Evans, Tao Pan, and Chuan He. Dynamic RNA Modifications in Gene Expression Regulation. *Cell*, 169(7):1187–1200, June 2017. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2017.05.045.

[306] Mai Sun, Björn Schwalb, Daniel Schulz, Nicole Pirkl, Stefanie Etzold, Laurent Larivière, Kerstin C. Maier, Martin Seizl, Achim Tresch, and Patrick Cramer. Comparative dynamic transcriptome analysis (cDTA) reveals mutual feedback between mRNA synthesis and degradation. *Genome Research*, 22(7):1350–1359, January 2012. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.130161.111.

[307] Andrew Lugowski, Beth Nicholson, and Olivia Selfridge Rissland. DRUID: A pipeline for transcriptome-wide measurements of mRNA stability. *RNA*, page rna.062877.117, February 2018. ISSN 1355-8382, 1469-9001. doi: 10.1261/rna.062877.117.

[308] Michelle T. Paulsen, Artur Veloso, Jayendra Prasad, Karan Bedi, Emily A. Ljungman, Brian Magnuson, Thomas E. Wilson, and Mats Ljungman. Use of Bru-Seq and BruChase-Seq for genome-wide assessment of the synthesis and stability of RNA. *Methods*, 67(1):45–54, May 2014. ISSN 1046-2023. doi: 10.1016/j.ymeth.2013.08.015.

[309] Benjamin Neymotin, Rodoniki Athanasiadou, and David Gresham. Determination of in vivo RNA kinetics using RATE-seq. *RNA*, 20(10):1645–1652, January 2014. ISSN 1355-8382, 1469-9001. doi: 10.1261/rna.045104.114.

[310] Erin E. Duffy, Michael Rutenberg-Schoenberg, Catherine D. Stark, Robert R. Kitchen, Mark B. Gerstein, and Matthew D. Simon. Tracking Distinct RNA Populations Using Efficient and Reversible Covalent Chemistry. *Molecular Cell*, 59(5):858–866, September 2015. ISSN 10972765. doi: 10.1016/j.molcel.2015.07.023.

[311] Louis Chaiet and Wolf Frank J. The properties of streptavidin, a biotin-binding protein produced by Streptomycetes. *Archives of Biochemistry and Biophysics*, 106:1–5, January 1964. ISSN 0003-9861. doi: 10.1016/0003-9861(64)90150-X.

[312] Erin E. Duffy and Matthew D. Simon. Enriching s4U-RNA Using Methane Thiosulfonate (MTS) Chemistry. *Current Protocols in Chemical Biology*, 8(4):234–250, December 2016. ISSN 2160-4762. doi: 10.1002/cpch.12.

[313] Emanuel Wyler, Jennifer Menegatti, Vedran Franke, Christine Kocks, Anastasiya Boltengagen, Thomas Hennig, Kathrin Theil, Andrzej Rutkowski, Carmelo Ferrai, Laura Baer, Lisa Kermas, Caroline Friedel, Nikolaus Rajewsky, Altuna Akalin, Lars Dölken, Friedrich Grässer, and Markus Landthaler. Widespread activation of antisense transcription of the host genome during herpes simplex virus 1 infection. *Genome Biology*, 18(1), December 2017. ISSN 1474-760X. doi: 10.1186/s13059-017-1329-5.

[314] Linda Warfield, Srinivas Ramachandran, Tiago Baptista, Didier Devys, Laszlo Tora, and Steven Hahn. Transcription of Nearly All Yeast RNA Polymerase II-Transcribed Genes Is Dependent on Transcription Factor TFIID. *Molecular Cell*, 68(1):118–129.e5, October 2017. ISSN 10972765. doi: 10.1016/j.molcel.2017.08.014.

[315] Leighton J. Core, Joshua J. Waterfall, and John T. Lis. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science (New York, N.Y.)*, 322(5909):1845–1848, December 2008. ISSN 0036-8075. doi: 10.1126/science.1162228.

[316] Naoto Imamachi, Hidenori Tani, Rena Mizutani, Katsutoshi Imamura, Takuma Irie, Yutaka Suzuki, and Nobuyoshi Akimitsu. BRIC-seq: A genome-wide approach for determining RNA stability in mammalian cells. *Methods*, 67(1):55–63, May 2014. ISSN 10462023. doi: 10.1016/j.ymeth.2013.07.014.

[317] Victor N. Ierusalimsky and Pavel M. Balaban. Long-living RNA in the CNS of terrestrial snail. *RNA Biology*, 15(2):207–213, February 2018. ISSN 1547-6286. doi: 10.1080/15476286.2017.1411460.

[318] Katherine C. Palozola, Greg Donahue, Hong Liu, Gregory R. Grant, Justin S. Becker, Allison Cote, Hongtao Yu, Arjun Raj, and Kenneth S. Zaret. Mitotic transcription and waves of gene reactivation during mitotic exit. *Science*, 358(6359):119–122, October 2017. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aal4671.

[319] Anders Holmberg, Anna Blomstergren, Olof Nord, Morten Lukacs, Joakim Lundeberg, and Mathias Uhlén. The biotin-streptavidin interaction can be reversibly broken using water at elevated temperatures. *ELECTROPHORESIS*, 26(3):501–510, February 2005. ISSN 1522-2683. doi: 10.1002/elps.200410070.

[320] Veronika A Herzog, Brian Reichholf, Tobias Neumann, Philipp Rescheneder, Pooja Bhat, Thomas R Burkard, Wiebke Wlotzka, Arndt von Haeseler, Johannes Zuber, and Stefan L Ameres. Thiol-linked alkylation of RNA to assess expression dynamics. *Nature Methods*, 14 (12):1198–1204, September 2017. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.4435.

[321] Edward Yang, Erik van Nimwegen, Mihaela Zavolan, Nikolaus Rajewsky, Mark Schroeder, Marcelo Magnasco, and James E Darnell Jr. Decay Rates of Human mRNAs: Correlation With Functional Characteristics and Sequence Attributes. *Genome Research*, page 11, 2003.

[322] L. V. Sharova, A. A. Sharov, T. Nedorezov, Y. Piao, N. Shaik, and M. S.H. Ko. Database for mRNA Half-Life of 19 977 Genes Obtained by DNA Microarray Analysis of Pluripotent and Differentiating Mouse Embryonic Stem Cells. *DNA Research*, 16(1):45–58, January 2009. ISSN 1340-2838, 1756-1663. doi: 10.1093/dnares/dsn030.

[323] M. T. Paulsen, A. Veloso, J. Prasad, K. Bedi, E. A. Ljungman, Y.-C. Tsan, C.-W. Chang, B. Tarrier, J. G. Washburn, R. Lyons, D. R. Robinson, C. Kumar-Sinha, T. E. Wilson, and M. Ljungman. Coordinated regulation of synthesis and stability of RNA during the acute TNF-induced proinflammatory response. *Proceedings of the National Academy of Sciences*, 110(6):2240–2245, February 2013. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1219192110.

[324] E N Nikolov and M D Dabeva. Re-utilization of pyrimidine nucleotides during rat liver regeneration. *Biochemical Journal*, 228(1):27–33, May 1985. ISSN 0264-6021.

[325] Rob Phillips, Jane Kondev, Julie Theriot, Hernan G. Garcia, and Nigel Orme. *Physical Biology of the Cell*. London ; New York, NY : Garland Science, [2013], 2013. ISBN 978-0-8153-4450-6.

[326] Hailey B. Lefkofsky, Artur Veloso, and Mats Ljungman. Transcriptional and post-transcriptional regulation of nucleotide excision repair genes in human cells. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 776:9–15, June 2015. ISSN 0027-5107. doi: 10.1016/j.mrfmmm.2014.11.008.

[327] E. M. Tank, C. Figueroa-Romero, L. M. Hinder, K. Bedi, H. C. Archbold, X. Li, K. Weskamp, N. Safren, X. Paez-Colasante, C. Pacut, S. Thumma, M. T. Paulsen, K. Guo, J. Hur, M. Ljungman, E. L. Feldman, and S. J. Barmada. Abnormal RNA stability in amyotrophic lateral sclerosis. *Nature Communications*, 9(1):2845, July 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-05049-z.

[328] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J. Gaffney, Laura L. Elo, Xuegong Zhang, and Ali Mortazavi. A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17, 2016. ISSN 1474-7596. doi: 10.1186/s13059-016-0881-8.

[329] Radmila Hrdlickova, Masoud Toloue, and Bin Tian. RNA-Seq methods for transcriptome analysis. *Wiley interdisciplinary reviews. RNA*, 8(1), January 2017. ISSN 1757-7004. doi: 10.1002/wrna.1364.

[330] 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. ISSN 1474-760X. doi: 10.1186/gb-2013-14-4-r36.

[331] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, January 2013. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bts635.

[332] Giacomo Baruzzo, Katharina E Hayer, Eun Ji Kim, Barbara Di Camillo, Garret A FitzGerald, and Gregory R Grant. Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nature Methods*, December 2016. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.4106.

[333] Fritz J. Sedlazeck, Philipp Rescheneder, and Arndt von Haeseler. NextGenMap: Fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics*, 29(21):2790–2791, November 2013. ISSN 1367-4803. doi: 10.1093/bioinformatics/btt468.

[334] Daehwan Kim, Ben Langmead, and Steven L. Salzberg. HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*, 12(4):357–360, April 2015. ISSN 1548-7105. doi: 10.1038/nmeth.3317.

[335] Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. Salmon: Fast and bias-aware quantification of transcript expression using dual-phase inference. *Nature methods*, 14(4):417–419, April 2017. ISSN 1548-7091. doi: 10.1038/nmeth.4197.

[336] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R. Kelley, Harold Pimentel, Steven L. Salzberg, John L. Rinn, and Lior Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, 7(3):562–578, March 2012. ISSN 1754-2189. doi: 10.1038/nprot.2012.016.

[337] Mihaela Pertea, Geo M Pertea, Corina M Antonescu, Tsung-Cheng Chang, Joshua T Mendell, and Steven L Salzberg. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology*, 33(3):290–295, March 2015. ISSN 1087-0156. doi: 10.1038/nbt.3122.

[338] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169, January 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu638.

[339] Bo Li and Colin N. Dewey. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12(1):323, December 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-323.

[340] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15:550, December 2014. ISSN 1474-760X. doi: 10.1186/s13059-014-0550-8.

[341] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, January 2010. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp616.

[342] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, April 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv007.

[343] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

[344] Björn Schwalb, Daniel Schulz, Mai Sun, Benedikt Zacher, Sebastian Dümcke, Dietmar E. Martin, Patrick Cramer, and Achim Tresch. Measurement of genome-wide RNA synthesis and decay rates with Dynamic Transcriptome Analysis (DTA). *Bioinformatics*, 28(6):884–885, March 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts052.

[345] Andrew Lugowski, Beth Nicholson, and Olivia S. Rissland. Determining mRNA half-lives on a transcriptome-wide scale. *Methods*, 137:90–98, March 2018. ISSN 10462023. doi: 10.1016/j.ymeth.2017.12.006.

[346] Joseph Russo, Adam M. Heck, Jeffrey Wilusz, and Carol J. Wilusz. Metabolic labeling and recovery of nascent RNA to accurately quantify mRNA stability. *Methods*, 120:39–48, May 2017. ISSN 10462023. doi: 10.1016/j.ymeth.2017.02.003.

[347] Maureen Cronin, Krishna Ghosh, Frank Sistare, John Quackenbush, Vincent Vilker, and Catherine O'Connell. Universal RNA Reference Materials for Gene Expression. *Clinical Chemistry*, 50(8):1464–1471, August 2004. ISSN 0009-9147, 1530-8561. doi: 10.1373/clinchem.2004.035675.

[348] External RNA Controls Consortium. Proposed methods for testing and selecting the ERCC external RNA controls. *BMC Genomics*, 6(1):150, November 2005. ISSN 1471-2164. doi: 10.1186/1471-2164-6-150.

[349] The External RNA Controls Consortium, Shawn C. Baker, Steven R. Bauer, Richard P. Beyer, James D. Brenton, Bud Bromley, John Burrill, Helen Causton, Michael P. Conley, Rosalie Elespuru, Michael Fero, Carole Foy, James Fuscoe, Xiaolian Gao, David Lee Gerhold, Patrick Gilles, Federico Goodsaid, Xu Guo, Joe Hackett, Richard D. Hockett, Pranvera Ikonomi, Rafael A. Irizarry, Ernest S. Kawasaki, Tamma Kaysser-Kranich, Kathleen Kerr, Gretchen Kiser, Walter H. Koch, Kathy Y. Lee, Chunmei Liu, Z. Lewis Liu, Anne Lucas, Chitra F. Manohar, Garry Miyada, Zora Modrusan, Helen Parkes, Raj K. Puri, Laura Reid, Thomas B. Ryder, Marc Salit, Raymond R. Samaha, Uwe Scherf, Timothy J. Sendera, Robert A. Setterquist, Leming Shi, Richard Shippy, Jesus V. Soriano, Elizabeth A. Wagar, Janet A. Warrington, Mickey Williams, Frederike Wilmer, Mike Wilson, Paul K. Wolber, Xiaoning Wu, and Renata Zadro. The External RNA Controls Consortium: A progress report. *Nature Methods*, 2:731–734, October 2005. ISSN 1548-7105. doi: 10.1038/nmeth1005-731.

[350] Simon A. Hardwick, Ira W. Deveson, and Tim R. Mercer. Reference standards for next-generation sequencing. *Nature Reviews Genetics*, 18(8):473–484, August 2017. ISSN 1471-0064. doi: 10.1038/nrg.2017.44.

[351] Seqc/Maqc-Iii Consortium, Zhenqiang Su, Paweł P. Łabaj, Sheng Li, Jean Thierry-Mieg, Danielle Thierry-Mieg, Wei Shi, Charles Wang, Gary P. Schroth, Robert A. Setterquist, John F. Thompson, Wendell D. Jones, Wenzhong Xiao, Weihong Xu, Roderick V. Jensen, Reagan Kelly, Joshua Xu, Ana Conesa, Cesare Furlanello, Hanlin Gao, Huixiao Hong, Nadereh Jafari, Stan Letovsky, Yang Liao, Fei Lu, Edward J. Oakeley, Zhiyu Peng, Craig A. Praul, Javier Santoyo-Lopez, Andreas Scherer, Tieliu Shi, Gordon K. Smyth, Frank Staedtler, Peter Sykacek, Xin-Xing Tan, E. Aubrey Thompson, Jo Vandesompele, May D. Wang, Jian Wang, Russell D. Wolfinger, Jiri Zavadil, Scott S. Auerbach, Wenjun Bao, Hans Binder, Thomas Blomquist, Murray H. Brilliant, Pierre R. Bushel, Weimin Cai, Jennifer G. Catalano, Ching-Wei Chang, Tao Chen, Geng Chen, Rong Chen, Marco Chierici, Tzu-Ming Chu, Djork-Arné Clevert, Youping Deng, Adnan Derti, Viswanath Devanarayan, Zirui Dong, Joaquin Dopazo, Tingting Du, Hong Fang, Yongxiang Fang, Mario Fasold, Anita Fernandez, Matthias Fischer, Pedro Furió-Tari, James C. Fuscoe, Florian Caimet, Stan Gaj, Jorge Gandara, Huan Gao, Weigong Ge, Yoichi Gondo, Binsheng Gong, Meihua Gong, Zhuolin Gong, Bridgett Green, Chao Guo, Lei Guo, Li-Wu Guo, James Hadfield,

Jan Hellemans, Sepp Hochreiter, Meiwen Jia, Min Jian, Charles D. Johnson, Suzanne Kay, Jos Kleinjans, Samir Lababidi, Shawn Levy, Quan-Zhen Li, Li Li, Li Li, Peng Li, Yan Li, Haiqing Li, Jianying Li, Shiyong Li, Simon M. Lin, Francisco J. López, Xin Lu, Heng Luo, Xiwen Ma, Joseph Meehan, Dalila B. Megherbi, Nan Mei, Bing Mu, Baitang Ning, Akhilesh Pandey, Javier Pérez-Florido, Roger G. Perkins, Ryan Peters, John H. Phan, Mehdi Pirooznia, Feng Qian, Tao Qing, Lucille Rainbow, Philippe Rocca-Serra, Laure Sambourg, Susanna-Assunta Sansone, Scott Schwartz, Ruchir Shah, Jie Shen, Todd M. Smith, Oliver Stegle, Nancy Stralis-Pavese, Elia Stupka, Yutaka Suzuki, Lee T. Szkotnicki, Matthew Tinning, Bimeng Tu, Joost van Delft, Alicia Vela-Boza, Elisa Venturini, Stephen J. Walker, Liqing Wan, Wei Wang, Jinhui Wang, Jun Wang, Eric D. Wieben, James C. Willey, Po-Yen Wu, Jiekun Xuan, Yong Yang, Zhan Ye, Ye Yin, Ying Yu, Yate-Ching Yuan, John Zhang, Ke K. Zhang, Wenqian Zhang, Wenwei Zhang, Yanyan Zhang, Chen Zhao, Yuanting Zheng, Yiming Zhou, Paul Zumbo, Weida Tong, David P. Kreil, Christopher E. Mason, and Leming Shi. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature Biotechnology*, 32(9):903–914, September 2014. ISSN 1546-1696. doi: 10.1038/nbt.2957.

[352] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012. ISSN 1476-4687. doi: 10.1038/nature11247.

[353] Stephen G. Landt, Georgi K. Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley E. Bernstein, Peter Bickel, James B. Brown, Philip Cayting, Yiwen Chen, Gilberto DeSalvo, Charles Epstein, Katherine I. Fisher-Aylor, Ghia Euskirchen, Mark Gerstein, Jason Gertz, Alexander J. Hartemink, Michael M. Hoffman, Vishwanath R. Iyer, Youngsook L. Jung, Subhradip Karmakar, Manolis Kellis, Peter V. Kharchenko, Qunhua Li, Tao Liu, X. Shirley Liu, Lijia Ma, Aleksandar Milosavljevic, Richard M. Myers, Peter J. Park, Michael J. Pazin, Marc D. Perry, Debasish Raha, Timothy E. Reddy, Joel Rozowsky, Noam Shoresh, Arend Sidow, Matthew Slattery, John A. Stamatoyannopoulos, Michael Y. Tolstorukov, Kevin P. White, Simon Xi, Peggy J. Farnham, Jason D. Lieb, Barbara J. Wold, and Michael Snyder. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*, 22(9):1813–1831, January 2012. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.136184.111.

[354] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Statist. Soc. B*, 57(1):289–300, 1995.

[355] Hidenori Tani and Nobuyoshi Akimitsu. Genome-wide technology for determining RNA stability in mammalian cells: Historical perspective and recent advantages based on modified nucleotide labeling. *RNA Biology*, 9(10):1233–1238, October 2012. ISSN 1547-6286, 1555-8584. doi: 10.4161/rna.22036.

[356] Takeo Wada and Attila Becskei. Impact of Methods on the Measurement of mRNA Turnover. *International Journal of Molecular Sciences*, 18(12):2723, December 2017. doi: 10.3390/ijms18122723.

[357] Marvin Wickens, David S. Bernstein, Judith Kimble, and Roy Parker. A PUF family portrait: 3′UTR regulation as a way of life. *Trends in Genetics*, 18(3):150–157, March 2002. ISSN 0168-9525. doi: 10.1016/S0168-9525(01)02616-6.

[358] Ruth Lehmann and Christiane Nüsslein-Volhard. Involvement of the pumilio gene in the transport of an abdominal signal in the Drosophila embryo. *Nature*, 329(6135):167, September 1987. ISSN 1476-4687. doi: 10.1038/329167a0.

[359] Danislav S. Spassov and Roland Jurecic. Cloning and comparative sequence analysis of PUM1 and PUM2 genes, human members of the Pumilio family of RNA-binding proteins. *Gene*, 299(1):195–204, October 2002. ISSN 0378-1119. doi: 10.1016/S0378-1119(02) 01060-0.

[360] Mark Fox, Jun Urano, and Renee A. Reijo Pera. Identification and characterization of RNA sequences to which human PUMILIO-2 (PUM2) and deleted in Azoospermia-like (DAZL) bind. *Genomics*, 85(1):92–105, January 2005. ISSN 0888-7543. doi: 10.1016/j.ygeno.2004.10.003.

[361] Dong Chen, Wei Zheng, Aiping Lin, Katherine Uyhazi, Hongyu Zhao, and Haifan Lin. Pumilio 1 Suppresses Multiple Activators of p53 to Safeguard Spermatogenesis. *Current Biology*, 22(5):420–425, March 2012. ISSN 0960-9822. doi: 10.1016/j.cub.2012.01.039.

[362] John P. Vessey, Lucia Schoderboeck, Ewald Gingl, Ettore Luzi, Julia Riefler, Francesca Di Leva, Daniela Karra, Sabine Thomas, Michael A. Kiebler, and Paolo Macchi. Mammalian Pumilio 2 regulates dendrite morphogenesis and synaptic function. *Proceedings of the National Academy of Sciences*, 107(7):3222–3227, February 2010. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0907128107.

[363] Henrike Siemen, Damien Colas, H. Craig Heller, Oliver Brüstle, and Renee A. Reijo Pera. Pumilio-2 Function in the Mouse Nervous System. *PLOS ONE*, 6(10):e25932, October 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0025932.

[364] Vincenzo A. Gennarino, Ravi K. Singh, Joshua J. White, Antonia De Maio, Kihoon Han, Ji-Yoen Kim, Paymaan Jafar-Nejad, Alberto di Ronza, Hyojin Kang, Layal S. Sayegh, Thomas A. Cooper, Harry T. Orr, Roy V. Sillitoe, and Huda Y. Zoghbi. Pumilio1 Haploinsufficiency Leads to SCA1-like Neurodegeneration by Increasing Wild-Type Ataxin1 Levels. *Cell*, 160(6):1087–1098, March 2015. ISSN 0092-8674. doi: 10.1016/j.cell.2015.02.012.

[365] Meng Zhang, Dong Chen, Jing Xia, Wenqi Han, Xiekui Cui, Nils Neuenkirchen, Gretchen Hermes, Nenad Sestan, and Haifan Lin. Post-transcriptional regulation of mouse neurogenesis by Pumilio proteins. *Genes & Development*, 31(13):1354–1369, July 2017. ISSN 0890-9369, 1549-5477. doi: 10.1101/gad.298752.117.

[366] Siraj K. Zahr, Guang Yang, Hilal Kazan, Michael J. Borrett, Scott A. Yuzwa, Anastassia Voronova, David R. Kaplan, and Freda D. Miller. A Translational Repression Complex in Developing Mammalian Neural Stem Cells that Regulates Neuronal Specification. *Neuron*, 97(3):520–537.e6, February 2018. ISSN 08966273. doi: 10.1016/j.neuron.2017.12.045.

[367] Hongxin Dong, Mengyi Zhu, Liping Meng, Yan Ding, Ding Yang, Shanshan Zhang, Wenan Qiang, Daniel W. Fisher, Eugene Yujun Xu, Hongxin Dong, Mengyi Zhu, Liping Meng, Yan Ding, Ding Yang, Shanshan Zhang, Wenan Qiang, Daniel W. Fisher, and Eugene Yujun Xu. Pumilio2 regulates synaptic plasticity via translational repression of synaptic receptors in mice. *Oncotarget*, 5(0), January 2018. ISSN 1949-2553. doi: 10.18632/oncotarget.24345.

[368] Ryo Narita, Kiyohiro Takahasi, Etsu Murakami, Emi Hirano, Seiji P. Yamamoto, Mitsutoshi Yoneyama, Hiroki Kato, and Takashi Fujita. A Novel Function of Human Pumilio Proteins in Cytoplasmic Sensing of Viral Infection. *PLOS Pathogens*, 10(10):e1004417, October 2014. ISSN 1553-7374. doi: 10.1371/journal.ppat.1004417.

[369] Michèle Brocard, Sarika Khasnis, C. David Wood, Claire Shannon-Lowe, and Michelle J. West. Pumilio directs deadenylation-associated translational repression of the cyclin-dependent kinase 1 activator RGC-32. *Nucleic Acids Research*, 46(7):3707–3725, April 2018. ISSN 0305-1048. doi: 10.1093/nar/gky038.

[370] Martijn Kedde, Marieke van Kouwenhove, Wilbert Zwart, Joachim A. F. Oude Vrielink, Ran Elkon, and Reuven Agami. A Pumilio-induced RNA structure switch in p27-3′ UTR controls miR-221 and miR-222 accessibility. *Nature Cell Biology*, 12(10):1014–1020, October 2010. ISSN 1476-4679. doi: 10.1038/ncb2105.

[371] Sungyul Lee, Florian Kopp, Tsung-Cheng Chang, Anupama Sataluri, Beibei Chen, Sushama Sivakumar, Hongtao Yu, Yang Xie, and Joshua T. Mendell. Noncoding RNA NORAD Regulates Genomic Stability by Sequestering PUMILIO Proteins. *Cell*, 164(1–2): 69–80, January 2016. ISSN 0092-8674. doi: 10.1016/j.cell.2015.12.017.

[372] Cécile Naudin, Aurore Hattabi, Fabio Michelet, Ayda Miri-Nezhad, Aissa Benyoucef, Françoise Pflumio, François Guillonneau, Serge Fichelson, Isabelle Vigon, Isabelle Dusanter-Fourt, and Evelyne Lauret. PUMILIO/FOXP1 signaling drives expansion of hematopoietic stem/progenitor and leukemia cells. *Blood*, 129(18):2493–2506, May 2017. ISSN 0006-4971, 1528-0020. doi: 10.1182/blood-2016-10-747436.

[373] Ailone Tichon, Rotem Ben-Tov Perry, Lovorka Stojic, and Igor Ulitsky. SAM68 is required for regulation of Pumilio by the NORAD long noncoding RNA. *Genes & Development*, 32 (1):70–78, January 2018. ISSN 0890-9369, 1549-5477. doi: 10.1101/gad.309138.117.

[374] Vincenzo A. Gennarino, Elizabeth E. Palmer, Laura M. McDonell, Li Wang, Carolyn J. Adamski, Amanda Koire, Lauren See, Chun-An Chen, Christian P. Schaaf, Jill A. Rosenfeld, Jessica A. Panzer, Ute Moog, Shuang Hao, Ann Bye, Edwin P. Kirk, Pawel Stankiewicz, Amy M. Breman, Arran McBride, Tejaswi Kandula, Holly A. Dubbs, Rebecca Macintosh, Michael Cardamone, Ying Zhu, Kevin Ying, Kerith-Rae Dias, Megan T. Cho, Lindsay B. Henderson, Berivan Baskin, Paula Morris, Jiang Tao, Mark J. Cowley, Marcel E. Dinger, Tony Roscioli, Oana Caluseriu, Oksana Suchowersky, Rani K. Sachdev, Olivier Lichtarge, Jianrong Tang, Kym M. Boycott, J. Lloyd Holder, and Huda Y. Zoghbi. A Mild PUM1 Mutation Is Associated with Adult-Onset Ataxia, whereas Haploinsufficiency Causes Developmental Delay and Seizures. *Cell*, 172(5):924–936.e11, February 2018. ISSN 00928674. doi: 10.1016/j.cell.2018.02.006.

[375] Xiaoqiang Wang, Phillip D. Zamore, and Traci M. Tanaka Hall. Crystal Structure of a Pumilio Homology Domain. *Molecular Cell*, 7(4):855–865, April 2001. ISSN 1097-2765. doi: 10.1016/S1097-2765(01)00229-5.

[376] Shuyun Dong, Yang Wang, Caleb Cassidy-Amstutz, Gang Lu, Rebecca Bigler, Mark R. Jezyk, Chunhua Li, Traci M. Tanaka Hall, and Zefeng Wang. Specific and Modular Binding Code for Cytosine Recognition in Pumilio/FBF (PUF) RNA-binding Domains. *Journal of Biological Chemistry*, 286(30):26732–26742, July 2011. ISSN 0021-9258, 1083-351X. doi: 10.1074/jbc.M111.244889.

[377] Zachary T. Campbell, Cary T. Valley, and Marvin Wickens. A protein-RNA specificity code enables targeted activation of an endogenous human transcript. *Nature Structural & Molecular Biology*, 21(8):732–738, August 2014. ISSN 1545-9985. doi: 10.1038/nsmb.2847.

[378] Adam R. Morris, Neelanjan Mukherjee, and Jack D. Keene. Ribonomic Analysis of Human Pum1 Reveals cis-trans Conservation across Species despite Evolution of Diverse mRNA Target Sets. *Molecular and Cellular Biology*, 28(12):4093–4103, June 2008. ISSN 0270-7306, 1098-5549. doi: 10.1128/MCB.00155-08.

[379] Eric L. Van Nostrand, Gabriel A. Pratt, Alexander A. Shishkin, Chelsea Gelboin-Burkhart, Mark Y. Fang, Balaji Sundararaman, Steven M. Blue, Thai B. Nguyen, Christine Surka, Keri Elkins, Rebecca Stanton, Frank Rigo, Mitchell Guttman, and Gene W. Yeo. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature Methods*, 13(6):508–514, June 2016. ISSN 1548-7091. doi: 10.1038/nmeth.3810.

[380] Chase A. Weidmann, Nathan A. Raynard, Nathan H. Blewett, Jamie Van Etten, and Aaron C. Goldstrohm. The RNA binding domain of Pumilio antagonizes poly-adenosine binding protein and accelerates deadenylation. *RNA*, 20(8):1298–1319, January 2014. ISSN 1355-8382, 1469-9001. doi: 10.1261/rna.046029.114.

[381] Tzu-Fang Lou, Chase A. Weidmann, Jordan Killingsworth, Traci M. Tanaka Hall, Aaron C. Goldstrohm, and Zachary T. Campbell. Integrated analysis of RNA-binding protein complexes using in vitro selection and high-throughput sequencing and sequence specificity landscapes (SEQRS). *Methods*, 118-119:171–181, April 2017. ISSN 10462023. doi: 10.1016/j.ymeth.2016.10.001.

[382] C. Tuerk and L. Gold. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, 249(4968):505–510, August 1990. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.2200121.

[383] Timothy L. Bailey, Nadya Williams, Chris Misleh, and Wilfred W. Li. MEME: Discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research*, 34(suppl 2): W369–W373, January 2006. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkl198.

[384] Arttu Jolma, Teemu Kivioja, Jarkko Toivonen, Lu Cheng, Gonghong Wei, Martin Enge, Mikko Taipale, Juan M. Vaquerizas, Jian Yan, Mikko J. Sillanpää, Martin Bonke, Kimmo

Palin, Shaheynoor Talukder, Timothy R. Hughes, Nicholas M. Luscombe, Esko Ukkonen, and Jussi Taipale. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Research*, 20(6):861–873, June 2010. ISSN 1088-9051. doi: 10.1101/gr.100552.109.

[385] Arttu Jolma, Jian Yan, Thomas Whitington, Jarkko Toivonen, Kazuhiro R. Nitta, Pasi Rastas, Ekaterina Morgunova, Martin Enge, Mikko Taipale, Gonghong Wei, Kimmo Palin, Juan M. Vaquerizas, Renaud Vincentelli, Nicholas M. Luscombe, Timothy R. Hughes, Patrick Lemaire, Esko Ukkonen, Teemu Kivioja, and Jussi Taipale. DNA-Binding Specificities of Human Transcription Factors. *Cell*, 152(1-2):327–339, January 2013. ISSN 00928674. doi: 10.1016/j.cell.2012.12.009.

[386] Daniel Dominguez, Peter Freese, Maria S. Alexis, Amanda Su, Myles Hochman, Tsultrim Palden, Cassandra Bazile, Nicole J. Lambert, Eric L. Van Nostrand, Gabriel A. Pratt, Gene W. Yeo, Brenton R. Graveley, and Christopher B. Burge. Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Molecular Cell*, 70(5):854–867.e9, June 2018. ISSN 1097-2765. doi: 10.1016/j.molcel.2018.05.001.

[387] J. Matthew Taliaferro, Nicole J. Lambert, Peter H. Sudmant, Daniel Dominguez, Jason J. Merkin, Maria S. Alexis, Cassandra A. Bazile, and Christopher B. Burge. RNA Sequence Context Effects Measured In Vitro Predict In Vivo Protein Binding and Regulation. *Molecular Cell*, 64(2):294–306, October 2016. ISSN 1097-2765. doi: 10.1016/j.molcel.2016.08.035.

[388] Olivier Elemento, Noam Slonim, and Saeed Tavazoie. A Universal Framework for Regulatory Element Discovery across All Genomes and Data Types. *Molecular Cell*, 28(2): 337–350, October 2007. ISSN 1097-2765. doi: 10.1016/j.molcel.2007.09.027.

[389] Uttam Roymondal, Shibsankar Das, and Satyabrata Sahoo. Predicting Gene Expression Level from Relative Codon Usage Bias: An Application to Escherichia coli Genome. *DNA Research*, 16(1):13–30, February 2009. ISSN 1340-2838. doi: 10.1093/dnares/dsn029.

[390] Ulrike Brandt-Bohne, Douglas R. Keene, Fletcher A. White, and Manuel Koch. MEGF9: A novel transmembrane protein with a strong and developmentally regulated expression in the nervous system. *Biochemical Journal*, 401(2):447–457, January 2007. ISSN 0264-6021, 1470-8728. doi: 10.1042/BJ20060691.

[391] Richard S Jope and Gail V. W Johnson. The glamour and gloom of glycogen synthase kinase-3. *Trends in Biochemical Sciences*, 29(2):95–102, February 2004. ISSN 0968-0004. doi: 10.1016/j.tibs.2003.12.004.

[392] Olga C. Jorge-Torres, Karolina Szczesna, Laura Roa, Carme Casal, Louisa Gonzalez-Somermeyer, Marta Soler, Cecilia D. Velasco, Pablo Martínez-San Segundo, Paolo Petazzi, Mauricio A. Sáez, Raúl Delgado-Morales, Stephane Fourcade, Aurora Pujol, Dori Huertas, Artur Llobet, Sonia Guil, and Manel Esteller. Inhibition of Gsk3b Reduces Nfkb1 Signaling and Rescues Synaptic Activity to Improve the Rett Syndrome Phenotype in Mecp2 -Knockout Mice. *Cell Reports*, 23(6):1665–1677, May 2018. ISSN 22111247. doi: 10.1016/j.celrep.2018.04.010.

[393] Lindsey N. Kent and Gustavo Leone. The broken cycle: E2F dysfunction in cancer. *Nature Reviews Cancer*, page 1, May 2019. ISSN 1474-1768. doi: 10.1038/s41568-019-0143-7.

[394] Bo Wang, Shu-hao Hsu, Xinmei Wang, Huban Kutay, Hemant Kumar Bid, Jianhua Yu, Ramesh K. Ganju, Samson T. Jacob, Mariia Yuneva, and Kalpana Ghoshal. Reciprocal regulation of microRNA-122 and c-Myc in hepatocellular cancer: Role of E2F1 and transcription factor dimerization partner 2: Wang et al. *Hepatology*, 59(2):555–566, February 2014. ISSN 02709139. doi: 10.1002/hep.26712.

[395] Svetlana Lebedeva, Marvin Jens, Kathrin Theil, Björn Schwanhäusser, Matthias Selbach, Markus Landthaler, and Nikolaus Rajewsky. Transcriptome-wide Analysis of Regulatory Interactions of the RNA-Binding Protein HuR. *Molecular Cell*, 43(3):340–352, August 2011. ISSN 1097-2765. doi: 10.1016/j.molcel.2011.06.008.

[396] Jun Wang, Yan Guo, Huili Chu, Yaping Guan, Jingwang Bi, and Baocheng Wang. Multiple Functions of the RNA-Binding Protein HuR in Cancer Progression, Treatment Responses and Prognosis. *International Journal of Molecular Sciences*, 14(5):10015–10041, May 2013. doi: 10.3390/ijms140510015.

[397] Kelly D. Sullivan, Thomas E. Mullen, William F. Marzluff, and Eric J. Wagner. Knockdown of SLBP results in nuclear retention of histone mRNA. *RNA*, 15(3):459–472, March 2009. ISSN 1355-8382. doi: 10.1261/rna.1205409.

[398] Kent L. Rossman, Channing J. Der, and John Sondek. GEF means go: Turning on RHO GTPases with guanine nucleotide-exchange factors. *Nature Reviews Molecular Cell Biology*, 6 (2):167, February 2005. ISSN 1471-0080. doi: 10.1038/nrm1587.

[399] Liyana Ahmad, Shen-Ying Zhang, Jean-Laurent Casanova, and Vanessa Sancho-Shimizu. Human TBK1: A Gatekeeper of Neuroinflammation. *Trends in molecular medicine*, 22(6): 511–527, June 2016. ISSN 1471-4914. doi: 10.1016/j.molmed.2016.04.006.

[400] Dragana Jankovic, Paolo Gorello, Ting Liu, Sabire Ehret, Roberta La Starza, Cecile Desjobert, Florent Baty, Martin Brutsche, Padma-Sheila Jayaraman, Alessandra Santoro, Christina Mecucci, and Juerg Schwaller. Leukemogenic mechanisms and targets of a NUP98/HHEX fusion in acute myeloid leukemia. *Blood*, 111(12):5672–5682, June 2008. ISSN 0006-4971, 1528-0020. doi: 10.1182/blood-2007-09-108175.

[401] Jian Xu, Vijay G. Sankaran, Min Ni, Tobias F. Menne, Rishi V. Puram, Woojin Kim, and Stuart H. Orkin. Transcriptional silencing of γ-globin by BCL11A involves long-range interactions and cooperation with SOX6. *Genes & Development*, 24(8):783–798, April 2010. ISSN 0890-9369, 1549-5477. doi: 10.1101/gad.1897310.

[402] Xavier Caubit, Paolo Gubellini, Joris Andrieux, Pierre L. Roubertoux, Mehdi Metwaly, Bernard Jacq, Ahmed Fatmi, Laurence Had-Aissouni, Kenneth Y. Kwan, Pascal Salin, Michèle Carlier, Agne Liedén, Eva Rudd, Marwan Shinawi, Catherine Vincent-Delorme, Jean-Marie Cuisset, Marie-Pierre Lemaitre, Fatimetou Abderrehamane, Bénédicte Duban, Jean-François Lemaitre, Adrian S. Woolf, Detlef Bockenhauer, Dany Severac, Emeric

Dubois, Ying Zhu, Nenad Sestan, Alistair N. Garratt, Lydia Kerkerian-Le Goff, and Laurent Fasano. *TSHZ3* deletion causes an autism syndrome and defects in cortical projection neurons. *Nature Genetics*, 48(11):1359–1369, November 2016. ISSN 1546-1718. doi: 10.1038/ng.3681.

[403] Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, 15 (3):651–674, September 2006. ISSN 1061-8600. doi: 10.1198/106186006X133933.

[404] Silke Janitza, Carolin Strobl, and Anne-Laure Boulesteix. An AUC-based permutation variable importance measure for random forests. *BMC Bioinformatics*, 14(1):119, December 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-119.

[405] Erin L. Sternburg, Jason A. Estep, Daniel K. Nguyen, Yahui Li, and Fedor V. Karginov. Antagonistic and cooperative AGO2-PUM interactions in regulating mRNAs. *Scientific Reports*, 8(1):15316, October 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-33596-4.

[406] Daniel Dominguez, Peter Freese, Maria S. Alexis, Amanda Su, Myles Hochman, Tsultrim Palden, Cassandra Bazile, Nicole J. Lambert, Eric L. Van Nostrand, Gabriel A. Pratt, Gene W. Yeo, Brenton Graveley, and Christopher B. Burge. Sequence, Structure and Context Preferences of Human RNA Binding Proteins. *bioRxiv*, page 201996, October 2017. doi: 10.1101/201996.

[407] Inga Jarmoskaite, Sarah K. Denny, Pavanapuresan P. Vaidyanathan, Winston R. Becker, Johan O. L. Andreasson, Curtis J. Layton, Kalli Kappel, Varun Shivashankar, Raashi Sreenivasan, Rhiju Das, William J. Greenleaf, and Daniel Herschlag. A quantitative and predictive model for RNA binding by human Pumilio proteins. *bioRxiv*, page 403006, August 2018. doi: 10.1101/403006.

[408] Inga Jarmoskaite, Sarah K. Denny, Pavanapuresan P. Vaidyanathan, Winston R. Becker, Johan O. L. Andreasson, Curtis J. Layton, Kalli Kappel, Varun Shivashankar, Raashi Sreenivasan, Rhiju Das, William J. Greenleaf, and Daniel Herschlag. A Quantitative and Predictive Model for RNA Binding by Human Pumilio Proteins. *Molecular Cell*, May 2019. ISSN 1097-2765. doi: 10.1016/j.molcel.2019.04.012.

[409] Oliver Hobert. Gene Regulation by Transcription Factors and MicroRNAs. *Science*, 319 (5871):1785–1786, March 2008. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1151651.

[410] Lulu I T. Korsak, Molly E. Mitchell, Katherine A. Shepard, and Michael R. Akins. Regulation of neuronal gene expression by local axonal translation. *Current genetic medicine reports*, 4(1):16–25, March 2016. ISSN 2167-4876. doi: 10.1007/s40142-016-0085-2.

[411] Gang Lu and Traci M. Tanaka Hall. Alternate Modes of Cognate RNA Recognition by Human PUMILIO Proteins. *Structure*, 19(3):361–367, March 2011. ISSN 0969-2126. doi: 10.1016/j.str.2010.12.019.

[412] Anthony M. Mustoe, Steven Busan, Greggory M. Rice, Christine E. Hajdin, Brant K. Peterson, Vera M. Ruda, Neil Kubica, Razvan Nutiu, Jeremy L. Baryza, and Kevin M. Weeks. Pervasive Regulatory Functions of mRNA Structure Revealed by High-Resolution SHAPE Probing. *Cell*, 173(1):181–195.e18, March 2018. ISSN 0092-8674, 1097-4172. doi: 10.1016/j.cell.2018.02.034.

[413] Megan E. Forrest, Ashrut Narula, Thomas J Sweet, Daniel Arango, Gavin Hanson, James Ellis, Shalini Oberdoerffer, Jeff Coller, and Olivia S Rissland. Codon usage and amino acid identity are major determinants of mRNA stability in humans. *bioRxiv*, December 2018. doi: 10.1101/488676.

[414] Gavin Hanson and Jeff Coller. Codon optimality, bias and usage in translation and mRNA decay. *Nature Reviews Molecular Cell Biology*, 19(1):20–30, January 2018. ISSN 1471-0080. doi: 10.1038/nrm.2017.91.

[415] Bastian Linder, Anya V. Grozhik, Anthony O. Olarerin-George, Cem Meydan, Christopher E. Mason, and Samie R. Jaffrey. Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nature Methods*, 12(8):767–772, August 2015. ISSN 1548-7105. doi: 10.1038/nmeth.3453.

[416] René M. Arvola, Chase A. Weidmann, Traci M. Tanaka Hall, and Aaron C. Goldstrohm. Combinatorial control of messenger RNAs by Pumilio, Nanos and Brain Tumor Proteins. *RNA Biology*, 14(11):1445–1456, November 2017. ISSN 1547-6286. doi: 10.1080/15476286.2017.1306168.

[417] Chase A. Weidmann, Chen Qiu, René M. Arvola, Tzu-Fang Lou, Jordan Killingsworth, Zachary T. Campbell, Traci M. Tanaka Hall, and Aaron C. Goldstrohm. Drosophila Nanos acts as a molecular clamp that modulates the RNA-binding and repression activities of Pumilio. *eLife*, 5:e17096, August 2016. ISSN 2050-084X. doi: 10.7554/eLife.17096.

[418] Ji-Young Youn, Wade H. Dunham, Seo Jung Hong, James D.R. Knight, Mikhail Bashkurov, Ginny I. Chen, Halil Bagci, Bhavisha Rathod, Graham MacLeod, Simon W.M. Eng, Stéphane Angers, Quaid Morris, Marc Fabian, Jean-François Côté, and Anne-Claude Gingras. High-Density Proximity Mapping Reveals the Subcellular Organization of mRNA-Associated Granules and Bodies. *Molecular Cell*, 69(3):517–532.e11, February 2018. ISSN 10972765. doi: 10.1016/j.molcel.2017.12.020.

[419] Timothy L. Bailey, Mikael Boden, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, and William S. Noble. MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Research*, 37(suppl 2):W202–W208, January 2009. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkp335.

[420] Torsten Hothorn, Peter Bühlmann, Sandrine Dudoit, Annette Molinaro, and Mark J. Van Der Laan. Survival ensembles. *Biostatistics*, 7(3):355–373, July 2006. ISSN 1465-4644. doi: 10.1093/biostatistics/kxj011.

[421] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25, December 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-25.

[422] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1): 307, December 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-307.

[423] Jesse Davis and Mark Goadrich. The relationship between Precision-Recall and ROC curves. pages 233–240. ACM Press, 2006. ISBN 978-1-59593-383-6. doi: 10.1145/1143844.1143874.

[424] Debashish Ray, Hilal Kazan, Kate B. Cook, Matthew T. Weirauch, Hamed S. Najafabadi, Xiao Li, Serge Gueroussov, Mihai Albu, Hong Zheng, Ally Yang, Hong Na, Manuel Irimia, Leah H. Matzat, Ryan K. Dale, Sarah A. Smith, Christopher A. Yarosh, Seth M. Kelly, Behnam Nabet, Desirea Mecenas, Weimin Li, Rakesh S. Laishram, Mei Qiao, Howard D. Lipshitz, Fabio Piano, Anita H. Corbett, Russ P. Carstens, Brendan J. Frey, Richard A. Anderson, Kristen W. Lynch, Luiz O. F. Penalva, Elissa P. Lei, Andrew G. Fraser, Benjamin J. Blencowe, Quaid D. Morris, and Timothy R. Hughes. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–177, July 2013. ISSN 0028-0836. doi: 10.1038/nature12311.

[425] Ronny Lorenz, Stephan H. Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6:26, 2011. ISSN 1748-7188. doi: 10.1186/1748-7188-6-26.

[426] Argyris Papantonis and Peter R. Cook. Transcription Factories: Genome Organization and Gene Regulation. *Chemical Reviews*, 113(11):8683–8705, November 2013. ISSN 0009-2665. doi: 10.1021/cr300513p.

[427] Ralph Stadhouders, Enrique Vidal, François Serra, Bruno Di Stefano, François Le Dily, Javier Quilez, Antonio Gomez, Samuel Collombet, Clara Berenguer, Yasmina Cuartero, Jochen Hecht, Guillaume J. Filion, Miguel Beato, Marc A. Marti-Renom, and Thomas Graf. Transcription factors orchestrate dynamic interplay between genome topology and gene regulation during cell reprogramming. *Nature Genetics*, 50(2):238–249, February 2018. ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-017-0030-7.

[428] Melissa J. Fullwood, Mei Hui Liu, You Fu Pan, Jun Liu, Han Xu, Yusoff Bin Mohamed, Yuriy L. Orlov, Stoyan Velkov, Andrea Ho, Poh Huay Mei, Elaine G. Y. Chew, Phillips Yao Hui Huang, Willem-Jan Welboren, Yuyuan Han, Hong Sain Ooi, Pramila N. Ariyaratne, Vinsensius B. Vega, Yanquan Luo, Peck Yean Tan, Pei Ye Choy, K. D. Senali Abayratna Wansa, Bing Zhao, Kar Sian Lim, Shi Chi Leow, Jit Sin Yow, Roy Joseph, Haixia Li, Kartiki V. Desai, Jane S. Thomsen, Yew Kok Lee, R. Krishna Murthy Karuturi, Thoreau Herve, Guillaume Bourque, Hendrik G. Stunnenberg, Xiaoan Ruan, Valere Cacheux-Rataboul, Wing-Kin Sung, Edison T. Liu, Chia-Lin Wei, Edwin Cheung, and Yijun Ruan. An oestrogen-receptor-α-bound human chromatin interactome. *Nature*, 462(7269):58–64, November 2009. ISSN 1476-4687. doi: 10.1038/nature08497.

[429] Mark A. Umbarger, Esteban Toro, Matthew A. Wright, Gregory J. Porreca, Davide Baù, Sun-Hae Hong, Michael J. Fero, Lihua J. Zhu, Marc A. Marti-Renom, Harley H. McAdams, Lucy Shapiro, Job Dekker, and George M. Church. The Three-Dimensional Architecture of a Bacterial Genome and Its Alteration by Genetic Perturbation. *Molecular Cell*, 44(2): 252–264, October 2011. ISSN 1097-2765. doi: 10.1016/j.molcel.2011.09.010.

[430] Cedric Cagliero, Ralph S. Grand, M. Beatrix Jones, Ding J. Jin, and Justin M. O'Sullivan. Genome conformation capture reveals that the Escherichia coli chromosome is organized by replication and transcription. *Nucleic Acids Research*, 41(12):6058–6071, July 2013. ISSN 0305-1048. doi: 10.1093/nar/gkt325.

[431] Bilal El Houdaigui, Raphaël Forquet, Thomas Hindré, Dominique Schneider, William Nasser, Sylvie Reverchon, and Sam Meyer. Bacterial genome architecture shapes global transcriptional regulation by DNA supercoiling. *Nucleic Acids Research*. doi: 10.1093/nar/gkz300.

[432] Josette Rouvière-Yaniv, Moshe Yaniv, and Jacques-Edouard Germond. E. coli DNA binding protein HU forms nucleosome-like structure with circular double-stranded DNA. *Cell*, 17 (2):265–274, June 1979. ISSN 0092-8674. doi: 10.1016/0092-8674(79)90152-1.

[433] Debayan Dey, Valakunja Nagaraja, and Suryanarayanarao Ramakumar. Structural and evolutionary analyses reveal determinants of DNA binding specificities of nucleoid-associated proteins HU and IHF. *Molecular Phylogenetics and Evolution*, 107:356–366, February 2017. ISSN 1055-7903. doi: 10.1016/j.ympev.2016.11.014.

[434] Jacques Oberto, Sabrina Nabti, Valérie Jooste, Hervé Mignot, and Josette Rouviere-Yaniv. The HU Regulon Is Composed of Genes Responding to Anaerobiosis, Acid Stress, High Osmolarity and SOS Induction. *PLOS ONE*, 4(2):e4367, February 2009. ISSN 1932-6203. doi: 10.1371/journal.pone.0004367.

[435] Chase A. Weidmann and Aaron C. Goldstrohm. Drosophila Pumilio Protein Contains Multiple Autonomous Repression Domains That Regulate mRNAs Independently of Nanos and Brain Tumor. *Molecular and Cellular Biology*, 32(2):527–540, January 2012. ISSN 0270-7306, 1098-5549. doi: 10.1128/MCB.06052-11.

[436] Gary D. Stormo and George W. Hartzell. Identifying protein-binding sites from unaligned DNA fragments. *Proceedings of the National Academy of Sciences*, 86(4):1183–1187, 1989.

[437] Matthew T. Weirauch, Atina Cote, Raquel Norel, Matti Annala, Yue Zhao, Todd R. Riley, Julio Saez-Rodriguez, Thomas Cokelaer, Anastasia Vedenko, Shaheynoor Talukder, Dream5 Consortium, Phaedra Agius, Aaron Arvey, Philipp Bucher, Curtis G. Callan Jr, Cheng Wei Chang, Chien-Yu Chen, Yong-Syuan Chen, Yu-Wei Chu, Jan Grau, Ivo Grosse, Vidhya Jagannathan, Jens Keilwagen, Szymon M. Kiełbasa, Justin B. Kinney, Holger Klein, Miron B. Kursa, Harri Lähdesmäki, Kirsti Laurila, Chengwei Lei, Christina Leslie, Chaim Linhart, Anand Murugan, Alena Myšičková, William Stafford Noble, Matti Nykter, Yaron Orenstein, Stefan Posch, Jianhua Ruan, Witold R. Rudnicki, Christoph D. Schmid, Ron Shamir, Wing-Kin Sung, Martin Vingron, Zhizhuo Zhang, Harmen J. Bussemaker, Quaid D.

Morris, Martha L. Bulyk, Gustavo Stolovitzky, and Timothy R. Hughes. Evaluation of methods for modeling transcription factor sequence specificity. *Nature Biotechnology*, 31 (2):126–134, February 2013. ISSN 1546-1696. doi: 10.1038/nbt.2486.

[438] Remo Rohs, Sean M. West, Alona Sosinsky, Peng Liu, Richard S. Mann, and Barry Honig. The role of DNA shape in protein–DNA recognition. *Nature*, 461(7268):1248–1253, October 2009. ISSN 1476-4687. doi: 10.1038/nature08473.

[439] Tsu-Pei Chiu, Federico Comoglio, Tianyin Zhou, Lin Yang, Renato Paro, and Remo Rohs. DNAshapeR: An R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics*, 32(8):1211–1213, April 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv735.

[440] Anthony Mathelier, Beibei Xin, Tsu-Pei Chiu, Lin Yang, Remo Rohs, and Wyeth W. Wasserman. DNA Shape Features Improve Transcription Factor Binding Site Predictions In Vivo. *Cell Systems*, 3(3):278–286.e4, September 2016. ISSN 2405-4712. doi: 10.1016/j.cels.2016.07.001.

[441] Hani Goodarzi, Hamed S. Najafabadi, Panos Oikonomou, Todd M. Greco, Lisa Fish, Reza Salavati, Ileana M. Cristea, and Saeed Tavazoie. Systematic discovery of structural elements governing stability of mammalian messenger RNAs. *Nature*, 485(7397):264–268, May 2012. ISSN 0028-0836. doi: 10.1038/nature11013.

[442] Jessica C. Bowman, Nicholas V. Hud, and Loren Dean Williams. The Ribosome Challenge to the RNA World. *Journal of Molecular Evolution*, 80(3-4):143–161, April 2015. ISSN 0022-2844, 1432-1432. doi: 10.1007/s00239-015-9669-9.

[443] Li Li, Christopher Francklyn, and Charles W. Carter. Aminoacylating Urzymes Challenge the RNA World Hypothesis. *Journal of Biological Chemistry*, 288(37):26856–26863, September 2013. ISSN 0021-9258, 1083-351X. doi: 10.1074/jbc.M113.496125.

[444] Charles W. Carter and Richard Wolfenden. tRNA acceptor stem and anticodon bases form independent codes related to protein folding. *Proceedings of the National Academy of Sciences*, 112(24):7489–7494, June 2015. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1507569112.