GRANT PROPOSALS BY ZLATA GVOZDENOV, PHD

CONTENT

HARVARD MEDICAL SCHOOL
BLAVATNIK INSTITUTE
DEPARTMENT OF BIOLOGICAL CHEMISTRY
AND MOLECULAR PHARMACOLOGY

240 Longwood Avenue
Boston, MA 02115

March 28, 2019

To the Division of Receipt and Referral
National Institutes of Health (NIH)

Application for the Ruth L. Kirschstein National Research Service Award (NRSA) Individual Postdoctoral Fellowship (Parent F32)

To Whom It May Concern:

I am pleased to submit a grant proposal titled "Distinguishing Biological Function from Biological Noise" for consideration under the NIH Research Grant Program (Ruth L. Kirschstein National Research Service Award (NRSA), Individual Postdoctoral Fellowship (Parent F32)) with PA number **PA-19-188**. This research project focuses on fundamental transcriptional mechanisms, which is of immense relevance for scientific and medical fields.

The letters of recommendation will be sent from:

Prof. Dr. Brian C. Freeman
University of Illinois, Urbana-Champaign
Department of Cell and Developmental Biology

Prof. Dr. Johannes Buchner
Technische Universität München
Department Chemie

Prof. Dr. Jie Chen
University of Illinois, Urbana-Champaign
Department of Cell and Developmental Biology

Prof. Dr. Stephen Buratowski
Harvard Medical School
Department of Biological Chemistry and Molecular Pharmacology

Thank you for your consideration.
Sincerely,
Zlata Gvozdenov, PhD

March 28, 2019

Distinguishing biological function from biological noise

Project Summary/Abstract

Are all synthesized transcripts that occur *in vivo* functional? Are some transcripts mechanistic byproducts that lack functional significance? Are all DNA binding events *in vivo* biologically meaningful? The hypothesis is that some of these events are not biologically functional. The term “biological noise”, the flip-side of biological function, is defined as reproducible, non-functional events that take place in living cell due to lack of fidelity. To date, no systematic, genome-scale experiments that measure (and distinguish) biological function from the erroneous events (i.e. noise) have been reported. Here, novel series of experiments to measure biological noise are proposed. Random, functionally irrelevant DNA will be introduced in *Saccharomyces cerevisiae* and standardized assays (ChIP-Seq, RNA-Seq, Net-Seq, TIF-Seq) that measure biological events (protein-DNA binding, transcription, transcript isoform heterogeneity) will be performed. The frequencies and magnitudes of the measured biological events on random DNA and endogenous DNA will be compared. High similarities between the measured events on random and endogenous DNA will be indicative of lack of function and high biological noise. Conversely, low frequency occurrence of the events on random DNA compared to endogenous DNA will indicate low biological noise. The measured noise on the functionally irrelevant DNA will be further used to compute the extent of genome-wide endogenous transcription, DNA binding and transcript isoform heterogeneity that lacks biological significance. For the first time, the selectivity of the biological processes will be investigated on chromosome-sized random DNA as a random selection tool *in vivo* to define associated DNA features. The proposed work will advance our understanding of transcriptional mechanism via experimental measurements of transcriptional precision and fidelity on a genome-wide scale at an unprecedented depth. This work will have broad implications in human biology and health.

Building on the investigator’s background in nuclear proteostasis, expertise in the field of genomics and bioinformatics will be gained. The project will be conducted in a word-renowned transcription/chromatin group, providing a strong foundation for the applicant’s future independent research as a principal investigator in this field.

Specific Aims

While traditional mRNA synthesis commences at promoter elements and ends at terminator regions, cellular transcription also takes place outside of that framework. In *Saccharomyces cerevisiae* and in higher eukaryotes, transcripts have been identified that originate or terminate within open reading frames (ORFs), intergenic sequences and elsewhere outside of promoter regions [1-4]. It is not clear how often these transcripts occur, whether they serve a particular purpose or whether they are just a mechanistic byproduct.

The presence of all transcripts requires – by definition – the presence of DNA-bound transcription factors. Transcription factors bind to DNA elements and modulate the recruitment of RNA Polymerase, thereby influencing the extent and timing of RNA synthesis in their vicinity. While transcription factors typically display a good correlation to their linked transcriptional activity, many transcription factor binding events are not directly connected to any transcriptional activity [5]. It is unclear whether this type of DNA binding activity (to sites devoid of transcription) represents bona fide functional events (whose function is unknown) or non-functional noise.

Besides transcriptional activation, eukaryotic gene expression is regulated on the level of individual RNA molecules. Through 5' and 3' UTR site selections, which are independent of each other, a typical cell gives rise to a multitude of mRNAs isoforms [6-8]. The choices of many 5' UTR start sites for capping and 3' ends for polyadenylation was proposed to regulate stability, localization and translation of the same-gene mRNA transcripts [9-11]. While it is likely that the extensive repertoire of 5' and 3' isoforms within traditional RNA pool is functionally important, it is also conceivable that some of this extensive heterogeneity is an outcome of an imperfect biological process.

The hypothesis of this work is that some proportion of the observed events (generated transcripts, protein DNA-binding, polyadenylation pattern, etc.) can be classified as lacking a biological function. This proposal aims to distinguish biological function from "biological noise" – the concept coined by my mentor as an antonym to biological function [12]. The biological noise referred to here is characterized by reproducible errors in the system, which are not meaningful, and it is different from the stochastic variations within given cells [12, 13]. As of this writing, no systematic, genome-wide approaches to measure and distinguish biologically functional and erroneous events have been described, and little to no experimental data exists to test the biological noise hypothesis on a broad scale. Here, an experimentally innovative approach to quantify biological noise is based on the fact that the noise can be measured with traditional assays by utilizing random (and functionally irrelevant) DNA elements in *S. cerevisiae* as a proxy for non-functional DNA binding, transcription, 5' and 3' site selection, and nucleosomal positioning. The overall goal of this study is to quantifying the proportion of random, non-functional events (such as protein-DNA

binding, non-functional transcripts, their isoform heterogeneity, and nucleosomal pattern) to experimentally define the precision and fidelity of the transcriptional process. To test the stated hypothesis and to achieve the overall goals, I propose the following Aims:

Aim 1. Construct Strains Harboring Random, Non-Functional DNA

Aim 2. Quantify Transcriptional Activity at Random, Non-Functional DNA

Aim 3. Quantify Protein Binding to Random, Non-Functional DNA

A strain containing random, non-functional DNA – constructed in Aim 1 – will be used in the subsequent aims to perform standardized high-throughput assays (RNA-Seq, NET-Seq, TIF-Seq, ChIP-Seq) to evaluate transcription, including the transcript 5' and 3' ends, protein-DNA binding and nucleosomal positioning. Occurrence (and signal strength) of a biological event on the random, functionally irrelevant DNA would mean that an irrelevant DNA template is sufficient for the process *in vivo* and represents a measure of biological noise. Biological noise will be quantified from the measurements of functionally irrelevant events at random DNA and will be compared to the events occurring at functional, endogenous DNA. The noise computed on the random DNA will be subtracted from the endogenous DNA to deliver a quantitative estimate of total functional biological events. The advantage will be taken from the random, non-functional and long DNA *in vivo* to precisely characterize selectivity of the biological processes and associated DNA motifs.

This work addresses unique questions about biological events that lack functional significance which were not approached before. The proposed – and novel – experimental approach strives to compare the events at random, non-functional sequences to those seen on endogenous DNA. The approach is also advantageous because it utilizes chromosome-sized random DNA devoid of biological function in conjunction with the numerous standardized, established assays. The work will improve our understanding about fundamental transcriptional mechanisms, which is of considerable relevance for scientific and medical fields.

Research Strategy

Significance: Decades of studies have shed light on transcriptional process in detail. In *Saccharomyces cerevisiae* and other eukaryotes, mRNA synthesis is initiated by RNA Polymerase II from a promoter region upstream of an ORF and terminated at terminator regions. However, mounting evidence is pointing to numerous RNA Polymerase II-generated transcripts that are not initiated from canonical promoters, do not terminate at the termination regions, nor have correct directionality [1-4]. With ~75-85% of the genome being transcribed, the amount of generated transcriptome in the yeast cells was shown to be greater and more complex than previously appreciated [7, 14-16]. Based on the published works on genome-wide distribution of RNA Polymerase II, my mentor hypothesized that only ~10% of the estimated 12,000 RNA Polymerase II elongating molecules are involved in the generation of conventional mRNAs in *S. cerevisiae* [12]. If 90% of elongating RNA Polymerase II do not produce traditional mRNAs, what function do they perform? One possibility is that transcription is a rather erroneous process that could produce transcripts with no biological roles. While theoretically predicted [12], little to no experimental data exists to test this hypothesis on a genome-wide scale. Furthermore, no appropriate experimental measures were utilized to measure and distinguish biologically functional from the erroneous transcripts. This work proposes a novel experimental approach to address erroneous transcription on a genome-wide scale. It allows for *in vivo* probing of the selectivity of the fortuitous transcriptional processes on a large and random, functionally irrelevant DNA pool.

Activators, DNA-binding transcription factors, are required for RNA Polymerase II transcription *in vivo*. A single DNA-binding transcription factor can bind to a multitude of different sites across the genome and affect the expression of numerous mRNAs [17]. However, not all transcription factor binding events appear to regulate transcription in a functionally relevant manner. For instance, Gal4 induces the transcription of genes involved in galactose metabolism when galactose is used as a carbon source [18]. However, Gal4 also binds to the locations that are not linked to the utilization of galactose and where differential Gal4 binding is not correlated with differential mRNA levels [5]. It is unknown whether the DNA-binding events across the genome that do not appear to affect transcription represent bone fide functional events (whose function is unknown) or biological noise. A DNA-bound transcription factor also has the potential to activate transcription in a region where the synthesized mRNA has no functional significance. Given a ~250 bp regulatory region, there is a ~12% chance to observer a specific 6-mer, a typical transcription factor binding consensus motif [19]. With estimated ~200-300 transcription factors in *S. cerevisiae* [20], it is very likely that a single regulatory region contains transcription factor binding motifs. To the best of my and my mentor's knowledge, no direct experimental quantification of biological noise for DNA-binding proteins has ever been published. The interactions between transcriptional activators and DNA are of great relevance for transcriptional regulation and

comprehensive investigation of non-functional protein-DNA binding events would contribute to our understanding of global transactivation specificity.

Gene expression is also regulated via the choice of transcripts' 5' and 3' UTRs. Based on the genome-wide technique that assesses relative transcript isoforms (TIFs) abundance by simultaneously considering both 5' and 3' ends, a maximum of ~500 TIFs per gene was estimated [16]. 5' UTRs vary in length, but are generally shorter than 3' UTRs, with a median of ~50 bp [7, 15]. Even though the functional importance of 5' UTRs is believed to reside in the regulation of protein translation [21], our understanding of 5' site selection divergence and its functional consequences is limited. Regulation of 3' UTRs is at least as complex. Once mRNA synthesis continues into the 3' UTR, RNA Polymerase II terminates transcription and the mRNA is polyadenylated [22]. Polyadenylation sites are largely heterogeneous with the major polyadenylated site representing only 20% of all mRNA molecules for a gene [8]. In addition, the length of the polyA tails is estimated to vary up to ~ 250 nt [23]. As a consequence, differential 3' UTR sites and polyA tail lengths greatly contribute to mRNAs isoform heterogeneity. Due to different turnover rates and function, the existence of these isoforms were suggested to have important implications for the regulation of the gene expression [8, 10, 11, 24]. However, it is not clear whether all isoforms are functionally important or whether a portion of diversity is an outcome of imperfect termination or polyadenylation. Interestingly, polyadenylation sites outside of the 3' UTR were observed [8, 16] and were suggested to be biased towards 5' of the ORF [8]. Polyadenylation of these shorter transcripts was proposed to be different from the polyadenylation that targets transcripts for cellular degradation [25] and similar to the conventional 3' UTR polyadenylation. Because so many short transcripts are generated from promoters or 5' regions of protein-coding genes as compared to the rest of the genome [3, 4], it is possible that polyadenylation at these transcripts is due to higher abundance of the generated transcripts (due to greater rate of transcription) and it is not biologically meaningful per se. The proposed project here will assess whether transcribed irrelevant DNA can be polyadenylated and whether some isoforms (both 5' and 3') will be more abundant than others. Because the DNA analyzed is functionally irrelevant, this work will assess the selectivity and fidelity of 5' or 3' site utilization and of the polyadenylation process *in vivo*.

Transcription depends on a permissive, transcription factor-accessible environment, typically with nucleosomes positioned outside the start sites. Transcription is readily initiated at the promoter regions which display low nucleosomal density [26]. However, ~75-90% of genomic DNA is wrapped in nucleosomes [27], making the DNA accessibility challenging for transcriptional machinery. This organization limits spurious transcriptional initiation, and failure in maintenance of chromatin structure results in disrupted (open) chromatin and increased cryptic transcription initiation [28]. If the transcripts generated within coding regions have no functional role but are just a mechanistic error, it can be speculated that chromatin status can contribute to increased

biological noise. The experimental validation of the possibility that nucleosomal organization contributes to the events with no biological implications has not been demonstrated. The question raised here is whether the open chromatin, in addition to initiating production of functional mRNAs, could be subject to higher transcriptional noise. By utilizing functionally irrelevant DNA, this work will test whether biological noise ("incorrect" transcription) occurs more frequently at open chromatin locations at random, functionally irrelevant DNA compared to the less open chromatin.

Chromatin landscape depends on the *cis*-acting DNA sequence elements, *trans*-acting factors and transcriptional activity [29]. While ~50% nucleosomal pattern is proposed to be determined by the DNA sequence [27], transcriptional processes are correlated with the establishment of nucleosomal patterns [30]. Precisely positioned nucleosomal patterns within promoters or 5' ORF locations decrease unidirectionally in transcribed regions [31]. Interestingly, transcription on foreign, fortuitous coding regions not functionally relevant to the host cell was associated with typical +1 nucleosome positioning array [30]. This suggests that biological nucleosomal pattern without apparent functional relevance can be formed on non-functional sequences. Our understanding on whether and how much of the nucleosomal pattern is functionally important is limited. Here, I will address whether the establishment of certain chromatin regions is an inherent cellular property, and whether establishment of these chromatin structures is functionally relevant. Utilizing completely random and non-functional DNA sequence, nucleosomal positioning will be examined and defined *in vivo* and compared to endogenous chromatin.

Goal summary: Relative to the wealth of data proposing biological functions of transcriptional process, little is known about non-functional events under regular conditions. Despite the relevance of non-meaningful mechanistic errors to understand fidelity of the transcription, the concept of biological noise was not given much consideration. There is a considerable need to distinguish functional biological events from those that are not meaningful to better define transcriptional fidelity. To achieve this, I propose following aims:

Aim 1. Construct Strains Harboring Random, Non-Functional DNA

Aim 2. Quantify Transcriptional Activity of Random, Non-Functional DNA

Aim 3. Quantify Protein Binding to Random, Non-Functional DNA

Innovation: My current group utilized a functional evolutionary approach to separate the contributions of DNA and species-specific factors to biological processes [8, 30, 32]. Sequences from different foreign yeast species introduced into *S. cerevisiae* were used to study determinants of nucleosome positioning, directionality of transcription from

promoters and 3' isoform utilization. Introduction of foreign DNA proposed here is conceptually similar. Performing various high-throughput assay (ChIP-Seq, RNA-Seq, NET-Seq, TIF-Seq) on randomly generated DNA sequences (the sequences that are not specific to any species and that are functionally irrelevant) is experimentally innovative. While short random sequence oligonucleotides were used more than 30 years ago to determine consensus motifs [33], this study will be the first one to perform measurements of biological nose on random, non-functional DNA on a large scale. We believe this approach is advantageous because it utilizes traditional assays that have been well established for the measurements of biological events. Combining the outputs of various assays, which complement each other, on a novel tool (random, non-functional DNA) will allow us to identify and quantify biological noise with unprecedented depth and precision.

Approach: The approach is based on the fact that long, functionally irrelevant DNA sequence can be introduced into living cells. To distinguish biological function from biological noise, I will measure DNA events (protein-DNA binding, transcription, transcript 5' and 3' ends and nucleosomal positioning) on the irrelevant DNA by performing standardized assays and compare the outcomes to the endogenous *S. cerevisiae* DNA. I will subtract the computed noise from the endogenous signals to define the proportion of functional events. Even though a large number of experiments is proposed here, it should be noted that these are standardized and relatively rapid experiments. The prerequisite for these experiments is a new tool – a strain with a large random, non-functional DNA sequence (described in Aim 1). This construct will be used in subsequent Aims.

Aim 1. Construct Strains Harboring Random, Non-Functional DNA.

The experimental basis for measuring non-functional events will be established by generating strains that have random, functionally irrelevant DNA. Two standard methods, each with their own advantages, will be employed to assemble random DNA.

i) *In vivo assembly of defined random DNA sequence chromosomes via homologous recombination.*

Two random 150-200 kb DNA sequences with varying GC/AT contents were generated computationally: one 150-200 kb DNA construct contains 50% GC content and one 40% GC (60% AT) content. The GC contents were varied with the aim to provide completely random DNA (50% GC) and random DNA which is closer to mimicking the composition of *S. cerevisiae* genomic DNA (40% GC). This will be needed in the later steps (Aims 2-3) when evaluating the importance of DNA sequence for biological noise. The *in situ* generated sequences were split into overlapping DNA fragments, which will be commercially synthesized. The following procedure involves very standardized *in vivo* assembly of the overlapping DNA fragments [34-36], which

has a high success rate owing to the avid homologous recombination in yeast. *In vivo* assembly of genomes from synthetic DNA fragments is used on a regular basis by the Broad Institute DNA construction and assembly team. Generation of a larger DNA construct here is similar, except I will utilize the tactics for the assembly of defined random DNA. From consultations with experts at the Broad Institute, there is no doubt that this method will work.

Briefly, the 150-200 kb constructs will be assembled *in vivo* using transformation associated recombination (TAR) cloning [34-36]. This involves yeast homologous recombination of co-transformed overlapping DNA fragments and a TAR vector to generate a circular chromosome *in vivo* [34-36]. The vector contains both yeast artificial chromosome (YAC) and bacterial artificial chromosome (BAC) elements, which allows for maintenance in both species (chromosome assembly in yeast and generation of high copy number DNA in bacteria). Due to the size of the construct and the starting lengths of the fragments, the assembly will be split in two parts. Firstly, commercially synthesized 1-2 kb DNA fragments will be used for the assembly of the intermediate products of ~20-30 kb with TAR. These constructs will be verified with pairs of diagnostic primers that detect successful homologous recombination at the junctions. Secondly, the intermediate constructs will be further used to assemble a 150-200 kb YAC with TAR. The chromosome will be isolated and sequenced, and the two generated strains (50% and 40% GC content) will be used for the subsequent Aims.

ii) In vitro assembly of random DNA sequence chromosomes via ligation of degenerate oligo mixes. This method was initially developed in our laboratory [37]. The basic unit is an oligo with defined 6-8 nt and random sequence (~200 nt) made by mixed oligonucleotide synthesis – the random oligo synthesis that utilizes nucleotide mixtures instead of defined nucleotides (i.e. 50% GC and AT, or 40% GC and 60% AT). The chromosome-sized random DNA assembly from these single-stranded (ss) oligos will involve: hybridization of commercial oligos at defined priming (complementary) 6-8 nt to create short-duplex (6-8 bp) at 3' end and long (~200 nt) random ss 5' ends; Klenow extension by mutually primed synthesis to create double-stranded (ds) DNA (~400 bp); and ligation of the heterogeneous ds DNA to generate larger random DNA constructs. The advantage of this method is that I will obtain great DNA sequence diversity and long DNA constructs at low cost. The occurrence of repetitive 6-8mers in the generated random DNA due to the initially defined priming (overlapping) sequences will be attenuated by using many different priming sequences. The ligated large DNA constructs (>100 kb) will be cloned into YAC, sequenced, and introduced in yeast, as our group reported earlier [30]. Alternatively, different short random DNA constructs (~10 kb cloned into YAC) can be introduced in different yeast cells, which can be pooled before proceeding to the high-throughput experiments. In this way, for instance, if pooling 50 different strains with different random DNA sequences, the level of sequencing signal for a given random DNA will correspond to $1/50^{th}$ of the actual signal. Given the multitude of options in synthetic DNA technologies and the flexibility in modifying yeast endogenous chromosomal sequences, it is highly unlikely that I will not be able to introduce irrelevant DNA sequence into yeast.

Aim 2. Quantify Transcriptional Activity of Random, Non-Functional DNA.

Transcription events will be quantified at random DNA construct from Aim 1 and compared to endogenous DNA. This would permit a quantitative assessment of the proportion of non-functional transcripts, including transcript isoforms, relative to the total cellular transcriptome. This will allow for a better understanding of the frequency of transcriptional noise at multiple levels. Computed noise information from multiple random DNA regions will be averaged and subtracted away from endogenous *S. cerevisiae* transcripts to yield a map of net functional transcription. From the chromosome-sized random, functionally irrelevant DNA sequence, DNA features associated with selectivity of the transcription will be defined *in vivo*.

Aim 2A. Quantify Steady-State and Nascent Transcripts of Random, Non-Functional DNA.

Steady-state RNA levels and actively elongating RNA Polymerase II will be measured on the randomized, non-functional DNA from Aim 1 and compared to the endogenous DNA of *S. cerevisiae* from the same experiment. RNA-Seq will be implemented to examine the steady-state transcriptome [7], as we have done before [19, 30]. Briefly, total RNA will be isolated from the strains with irrelevant DNA from Aim 1 and will be subject to polyA selected, strand specific RNA-Seq. NET-Seq, which we also implemented earlier [32], will be performed to track actively elongating RNA Polymerase II transcripts, which involves both stable and unstable transcripts [38] (the later are poorly detected with RNA-Seq). For NET-Seq, RNA Polymerase II subunit Rpb3 will be epitope tagged in the strains from Aim 1, and will be used for pull-downs of the RNA Polymerase II together with co-purified RNA. Released RNA will be ligated to a linker, a cDNA library will be prepared and the samples will be subjected to the high-throughput sequencing. Both RNA-Seq and NET-Seq reads will be mapped to the *S. cerevisiae* genome and the introduced random sequences. The usage of both RNA- and NET-Seq in parallel would allow to assess relationship between steady-state RNA levels and active RNA Polymerase II elongation.

I fully expect transcription to occur on random DNA because i) transcription readily occurs on YAC carrying foreign yeast species DNA [30, 32], some of which is likely non-functional; ii) RNA Polymerase II initiation specificity is hypothesized to differ between maximally activated promoters and random sequences by a factor of 10^4 [12]; iii) short random DNA templates were transcribed at surprisingly high level *in vitro*, based on my experience.

DATA ANALYSIS will focus on the frequency and magnitude of the events (total steady-state and nascent reads) per kb generated *in vivo* on functionally irrelevant DNA (transcript coverage) as well as transcript directionality. The measured number of reads per kb windows will be used to generate a mean value as an estimate for transcriptional noise at random DNA. These results will be compared to endogenous DNA reads. The average sequencing reads per kb of random DNA could represent a certain percentage

of the average value for endogenous RNAs [30]. This would provide quantification how much of the irrelevant transcription occurs in the living cells compared to the total endogenous transcription.

Measured irrelevant transcription at random DNA will be further used to predict transcriptional noise at any given genomic locus. Because different random DNA regions could differ in transcription frequencies and magnitudes, a transcriptional noise probability will be computed that takes into account those different values. At any given endogenous locus, a high ratio of transcriptional noise probability of random DNA/endogenous DNA transcription will indicate high transcriptional noise. Conversely, a low ratio of random DNA/endogenous DNA transcription will be highly suggestive of predominantly functional transcription (and low transcriptional noise) at the endogenous locus. The final goal is to subtract transcription computed on random DNA from the endogenous transcription to calculate net functional transcription. The expectation is that, after subtraction, numerous (and previously presumed to be functional) transcripts in the dataset will be re-classified as biological noise.

While subtracting biological noise, not all DNA regions will be treated equally. Biological noise contributions of different DNA sequences, protein-DNA binding, chromatin status and isoform heterogeneity will be quantified in this project at random DNA (Aims 2-3) and integrated into calculation of the overall endogenous transcriptional noise. The information about the transcripts' location (the sequence signature) on the random DNA in conjunction with isoform heterogeneity (Aim 2B), protein-DNA binding status (Aim 3A), and chromatin status (Aim 3B) will be correlated with the transcriptional noise values at random DNA. The combination of features (e.g. DNA sequence and chromatin status with certain noise values) on random DNA will be compared to endogenous DNA with the similar features to predict biologically irrelevant transcriptional events. The endogenous regions with greater similarity to random DNA with high noise values will be treated as having higher biological noise and higher values will be subtracted. The usage of random DNA with defined different GC contents (Aim 1) would thus allow to test for the effect of DNA sequence on the transcriptional noise. From this part, we will learn whether and how much of transcription in biological systems is an erroneous process, escorted by specific or inadvertent DNA or other features.

The possibility of rapid cellular elimination of the transcripts originating from random, functionally irrelevant DNA will be addressed as well. Higher signal from NET-Seq than from RNA-Seq would suggest that transcripts are actively degraded. If NET-Seq/RNA-Seq read ratio is higher for random RNA than for the functional RNA, then it can be concluded that capacity of the cell to eliminate non-functional transcripts has evolved to attenuate transcriptional noise.

Aim 2B. Assess 5' and 3' Site Selection of the Transcripts from Random, Non-Functional DNA

Here, the choices of 5' and 3' (polyadenylation) sites will be tested on random, non-functional DNA. This part will address the question whether distinct transcript isoforms can occur on irrelevant DNA and whether functionally irrelevant DNA can be polyadenylated. Similarities between endogenous and random DNA would suggest that isoform heterogeneity and polyadenylation lack functional significance. The quantified amount of similarity would define how much of the observed endogenous isoforms are likely to be the consequence of biological noise. In addition, global random selection of the DNA consensus motif associated with the 3' and 5' choices will be precisely defined *in vivo*.

RNA-Seq and NET-Seq provide a read coverage for a given location (cumulative signal for many RNA molecules). As such, these methods are not suitable for obtaining comprehensive information about single RNA molecules and corresponding isoforms. TIF-Seq method enables simultaneous detection of 3' and 5' ends of single RNA molecules (capped and polyadenylated RNA molecules) by utilizing 5'-3' circularization step prior to sequencing [16]. This technique will be especially useful for the investigation of the random sequence DNA because i) transcripts on random DNA are expected to vary in both start and termination site selection, and TIF-Seq allows for obtaining quantitative information for both concurrently; ii) both polyadenylated and non-polyadenylated transcripts can be analyzed by TIF-Seq, which will allow for the determination of whether polyadenylation on random DNA sequence occurs at similar levels to that seen for endogenous *S. cerevisiae* transcripts. The technique will provide information on relative abundance of specific isoforms, such as RNA isoforms originating from irrelevant DNA with respect to endogenous DNA. However, as a paired-end-like approach, TIF-Seq is suitable for detection of shorter RNA molecules. I will complement isoform studies with 3'READS (3′ region extraction and deep sequencing) approach, which is independent of the transcript size. 3′READS method identifies genome-wide polyA sites and it is used to quantify the relative abundance of the 3′ mRNA isoforms, as we previously described [39]. Overall, isoforms characteristics (5' and 3' site selection, 3' polyadenylation, isoform abundance, length and degradation (NET-Seq/RNA-Seq ratio)) will be compared to the endogenous isoforms in a similar fashion to that described in Aim 2A. The outcomes of these analyses will shed light onto the functional importance of isoform heterogeneity *in vivo*. Of note, the genome-wide studies on transcript isoforms are currently the major project in our group conducted by the bioinformatics experts who pioneered the methodologies.

Aim 3. Quantify Protein Binding to Random, Non-Functional DNA.

Here, measurements of protein-DNA interactions on randomly generated DNA sequences will be utilized to distinguish and quantify biologically irrelevant protein-DNA binding *in vivo*. DNA binding events on irrelevant DNA compared to the endogenous DNA would allow for the assessment of total relative non-functional protein-DNA binding that constitutes biological noise. Similarly, interaction between octamer and DNA will be investigated on functionally irrelevant DNA to define nucleosomal pattern, which would determine the extent of functional implications of a chromatin status.

Aim 3A. Quantify Binding of Transcription Factors and RNA Polymerase II to Random, Non-Functional DNA.

DNA-binding proteins that will be assayed are transcription factors such as Rap1, Reb1, Gal4 and Hsf1, as well as the RNA Polymerase II largest subunit and the TATA box binding protein TBP. These transcription factors were chosen due to the fact that they are some of the prominent transcription factors – they are well-studied but lack the aspect about non-specific DNA binding events, which would greatly contribute to understanding their operation. RNA Polymerase II and TBP will be used here because, and in combination with Aim 2A, it will be possible to quantify how much of the DNA binding by RNA Polymerase II constitutes productive and non-productive binding (binding that correlates or does not correlate with transcription).

Chromatin immunoprecipitation (ChIP) experiments, as initially optimized in our laboratory [40], targeting transcription factors, RNA Polymerase II and TBP will be utilized to test DNA binding in the strains from Aim 1 in a high-throughput manner (ChIP-Seq). Commercial antibodies are available for these proteins, and some transcription factors are also commonly studied as epitope-tagged fusion proteins from TAP-tag collection. The ChIP-Seq data will be analyzed using standardized programs and compared to the previously published reports for consistency. ChIP-Seq is very standardized technique, and the proteins that will be analyzed here were evaluated with similar approaches by us and others [41-46], so I do not anticipate problems.

Data analysis will be performed similarly as described in Aim 2A. In general, because sequence-specific transcription factors are used, it is expected that the irrelevant binding would occur mainly at the locations containing transcription factor-specific consensus motifs. Nucleosome positioning might be competitive for the DNA binding and the magnitude of the transcription factor binding can depend on the nucleosomal occupancy (Aim 3B). The information about these two biological events (transcription factor binding and nucleosome occupancy) will be factored in for the evaluation of transcription factor-DNA binding that constitutes biological noise (e.g. a less strong protein binding at the locus with strong nucleosomal occupancy will score similarly to the protein with high DNA binding signal and low nucleosomal occupancy). Different proteins tested here could also deliver different levels of biological noise. Whereas TBP sequence specificity is expected to be lower than for the mentioned transcription factors, no sequence specificity will be anticipated for RNA Polymerase II. The results from the Aim 2A will be further combined here to evaluate how much of RNA Polymerase II binding constitutes productive transcription with respect to the endogenous and functionally irrelevant DNA templates. Subtraction of biological noise form the genome-wide DNA binding datasets will allow for a more accurate and detailed identification of consensus binding sites for the various sequence-specific DNA binding proteins. From this part, we will learn of the extent to which protein-DNA binding in biological systems is an erroneous process.

Aim 3B. Assess Nucleosomal Pattern at Random, Non-Functional DNA

Nucleosome ChIP-Seq [47] and potentially DNaseI-Seq [48], which was implemented by myself before [49], will be used to assay chromatin status on random, functionally irrelevant chromatin from Aim 1. The outcomes will be compared to endogenous yeast chromatin. The aim is to evaluate the propensity of the cell to generate certain chromatin landscape with no functional implications. Data analysis will be performed similarly as described in Aim 2A. Regarding result interpretation, I expect to see some of these outcomes: i) Random DNA does not have conserved dinucleotide periodicity preferential for bending of DNA around octamer [27], which will render less well established nucleosomal positioning and occupancy across the cell population. We will learn from this outcome whether inherent, biological DNA sequence is crucial for (dense) nucleosomal positioning. ii) Regions of less well positioned nucleosomes or open chromatin can enhance transcription factor binding through exposing consensus motifs, and might possess higher transcriptional activity (this information will be available from the Aims 2A and 3A). This outcome would indicate that, while nucleosome depleted regions are important for traditional transcription initiation, they are also subject to high biological noise (transcription of non-functional transcripts). iii) While random chromatin is not expected to be organized like a functional chromatin, prevalently transcribed regions might attain the organization resembling +1 nucleosome-like pattern. This involves relatively positioned +1 nucleosome with decaying downstream nucleosomal pattern similar to functional genic region. The expectation is based on the observation that transcription elongation is associated with this nucleosomal pattern [30]. This outcome would underscore the importance of transcription for the nucleosome positioning and vice versa. In sum, identifying similarities and differences in nucleosome organization between random sequences and endogenous DNA region could determine how much of the established chromatin landscape is functionally irrelevant (inherent cellular property regardless of the functional relevance), including how much of the nucleosomal positioning is associated with the transcriptional process. Finally, I will define DNA elements associated with nucleosomal positioning using random DNA as a means of *in vivo* random selection.

Conclusion: The observation that erroneous events with no biological function could happen in living cells is rather interesting but unaddressed phenomenon. The approach I propose here combines a random sequence artificial chromosome with well-established genome-wide assays to directly measure and identify biological noise on a global scale. In addition and as part of my work, I plan on establishing a database of biological noise for the various assays described here. I can envision this biological noise database to be quite useful to researchers, especially as a mean to cross-reference a specific locus (or many loci) with biological noise-subtracted transcription, RNA isoform or protein-binding levels. Finally, using the strains generated in Aim 1, standardized and relatively rapid experiments can be performed in the future to assay the contribution of biological noise to a variety of biological processes such as posttranslational modifications, chromatin remodeling, DNA replication and many other pathways.

Scope of the work: I am aware that this proposal addresses many different biological questions. However, these biological processes (as well as concomitant experimental work) are related. The proposed experiments are very standardized and are all actively done in our laboratory. With the exception of the new strains with random DNA sequence, there are no unusual, complicated or extensively time-consuming experiments, which guarantees completion of the project within the designated time frame.

Bibliography and References Cited

1. Steinmetz EJ, Conrad NK, Brow DA, Corden JL. 2001. RNA-binding protein Nrd1 directs poly(A)-independent 3'-end formation of RNA polymerase II transcripts. *Nature* **413**: 327–331.

2. Kaplan CD, Laprade L, Winston F. 2003. Transcription elongation factors repress transcription initiation from cryptic sites. *Science* **301**: 1096–1099.

3. Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Münster S, Camblong J, Guffanti E, Stutz F, Huber W, Steinmetz LM. 2009. Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**: 1033–1037.

4. Neil H, Malabat C, D'Aubenton-Carafa Y, Xu Z, Steinmetz LM, Jacquier A. 2009. Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature* **457**: 1038–1042.

5. Gvozdenov Z, Kolhe J, Freeman BC. The Hsp90 molecular chaperone regulates the transcription factor network controlling the chromatin landscape, (*submitted to PNAS*).

6. Zhang Z, Dietrich FS. 2005. Mapping of transcription start sites in *Saccharomyces cerevisiae* using 5' SAGE. *Nucleic Acids Res* **33**: 2838–2851.

7. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**: 1344–1349.

8. Moqtaderi Z, Geisberg JV, Jin Y, Fan X, Struhl K. 2013. Species-specific factors mediate extensive heterogeneity of mRNA 3′ ends in yeasts. *Proc Natl Acad Sci U S A* **110**: 11073–11078.

   PMCID: PMC3703967

9. Mignone F, Gissi C, Liuni S, Pesole G. 2002. Untranslated regions of mRNAs. *Genome Biol* **3**: reviews00041–0004.10.

10. Geisberg JV, Moqtaderi Z, Fan X, Ozsolak F, Struhl K. 2014. Global analysis of mRNA isoform half-lives reveals stabilizing and destabilizing elements in yeast. *Cell* **156**: 812–824. PMCID: PMC3939777

11. Wu X, Bartel DP. 2017. Widespread influence of 3'-end structures on mammalian mRNA processing and stability. *Cell* **169**: 905–917.

12. Struhl K. 2007. Transcriptional noise and the fidelity of initiation by RNA Polymerase II. *Nat Struct Mol Biol* **14**: 103–105.

13. Raser JM, O'Shea EK. 2005. Noise in gene expression: origins, consequences and control. *Science* 309: 2010–2013.

14. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, et al. 2005. The transcriptional landscape of the mammalian genome. *Science* **309**: 1559–1563.

15. David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM. 2006. A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci U S A* **103**: 5320–5325.

16. Pelechano V, Wei W, Steinmetz LM. 2013. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* **497**: 127–131.

17. Hahn JS, Hu Z, Thiele DJ, Iyer VR. 2004. Genome-wide analysis of the biology of stress responses through heat shock transcription factor. *Mol Cell Biol* **24**: 5249–5256.

18. Ptashne M, Gann A. 1997. Transcriptional activation by recruitment. *Nature* **386**: 569–577.

19. Tirosh I, Wong KH, Barkai N, Struhl K. 2011. Extensive divergence of yeast stress responses through transitions between induced and constitutive activation. *Proc Natl Acad Sci U S A* **108**: 16693–16698.

PMCID: PMC3189053

20. Hughes TR, de Boer CG. 2013. Mapping yeast transcriptional networks. *Genetics* **195**: 9–36.

21. Dvir S, Velten L, Sharon E, Zeevi D, Carey LB, Weinberger A, Segal E. 2013. Deciphering the rules by which 5′-UTR sequences affect protein expression in yeast. *Proc Natl Acad Sci U S A* **110**: E2792–E2801.

22. Zhao J, Hyman L, Moore C. 1999. Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol Mol Biol Rev* **63**: 405–445.

23. Jalkanen AL, Coleman SJ, Wilusz J. 2014. Determinants and implications of mRNA poly(A) tail size–does this protein make my tail look big? *Semin Cell Dev Biol* **34**: 24–32.

24. Di Giammartino DC, Nishida K, Manley JL. 2011. Mechanisms and consequences of alternative polyadenylation. *Mol Cell* **43**: 853–866.

25. Wyers F, Rougemaille M, Badis G, Rousselle JC, Dufour ME, Boulay J, Régnault B, Devaux F, Namane A, Séraphin B, et al. 2005. Cryptic Pol II transcripts are

degraded by a nuclear quality control pathway involving a new poly(A) polymerase. *Cell* **121**: 725–737.

26. Sekinger EA, Moqtaderi Z, Struhl K. 2005. Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. *Mol Cell* **18**: 735–748.

27. Segal E, Fondufe-Mittendorf Y, Chen L, Thåström A, Field Y, Moore IK, Wang J-PZ, Widom J. 2006. A genomic code for nucleosome positioning. *Nature* **442**: 772–778.

28. Kaplan CD, Laprade L, Winston F. 2003. Transcription elongation factors repress transcription initiation from cryptic sites. *Science* **301**: 1096–1099.

29. Struhl K, Segal E. 2013. Determinants of nucleosome positioning. *Nat Struct Mol Biol* **20**: 267–273.

PMCID: PMC3740156

30. Hughes AL, Jin Y, Rando OJ, Struhl K. 2012. A functional evolutionary approach to identify determinants of nucleosome positioning: A unifying model for establishing the genome-wide pattern. *Mol Cell* **48**: 5–15. PMCID: PMC3472102

31. Vaillant C, Palmeira L, Chevereau G, Audit B, d'Aubenton-Carafa Y, Thermes C, Arneodo A. 2010. A novel strategy of transcription regulation by intragenic nucleosome ordering. *Genome Res* **20**: 59–67.

32. Jin Y, Eser U, Struhl K, Churchman LS. 2017. The ground state and evolution of promoter region directionality. *Cell* **170**: 889–898. PMCID: PMC5576552

33. Oliphant AR, Struhl K. 1987. The use of random-sequence oligonucleotides for determining consensus sequences. *Methods Enzymol* **155**: 568–582.

34. Gibson DG, Benders GA, Andrews-Pfannkoch C, Denisova EA, Baden-Tillson H, Zaveri J, Stockwell TB, Brownley A, Thomas DW, Algire MA, et al. 2008. Complete chemical synthesis, assembly, and cloning of a *Mycoplasma genitalium* genome. *Science* **319**: 1215–1220.

35. Gibson DG, Benders GA, Axelrod KC, Zaveri J, Algire MA, Moodie M, Montague MG, Venter JC, Smith HO, Hutchison III CA. 2008. One-step assembly in yeast of 25 overlapping DNA fragments to form a complete synthetic *Mycoplasma genitalium* genome. *Proc Natl Acad Sci U S A* **105**: 20404–20409.

36. Noskov VN, Karas BJ, Young L, Chuang RY, Gibson DG, Lin YC, Stam J, Yonemoto IT, Suzuki Y, Andrews-Pfannkoch C, et al. 2012. Assembly of large, high G+C bacterial DNA fragments in yeast. *ACS Synth Biol* **1**: 267–273.

37. Oliphant AR, Nussbaum AL, Struhl K. 1986. Cloning of random-sequence oligodeoxynucleotides. *Gene* **44**: 177–183.

38. Churchman LS, Weissman JS. 2011. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469**: 368–373.

39. Jin Y, Geisberg JV, Moqtaderi Z, Ji Z, Hoque M, Tian B, Struhl K. 2015. Mapping 3' mRNA isoforms on a genomic scale. *Curr Protoc Mol Biol* **110**: 4.23.1–4.23.17. PMCID: PMC4397975

40. Kuras L, Struhl K. 1999. Binding of TBP to promoters *in vivo* is stimulated by activators and requires Pol II holoenzyme. *Nature* **399**: 609–613.

41. Koerber RT, Rhee HS, Jiang C, Pugh BF. 2009. Interaction of transcriptional regulators with specific nucleosomes across the *Saccharomyces* genome. *Mol Cell* **35**: 889–902.

42. Rhee HS, Pugh BF. 2011. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147**: 1408–1419.

43. Rhee HS, Pugh BF. 2012. Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature* **483**: 295–301.

44. Park D, Lee Y, Bhupindersingh G, Iyer VR. 2013. Widespread misinterpretable ChIP-Seq bias in yeast. *PLoS One* **8**: e83506.

45. Rossi MJ, Lai WKM, Pugh BF. 2018. Genome-wide determinants of sequence-specific DNA binding of general regulatory factors. *Genome Res* **28**: 497–508.

46. Petrenko N, Jin Y, Dong L, Wong KH, Struhl K. 2019. Requirements for RNA polymerase II preinitiation complex formation *in vivo*. *Elife* **8**: e43654.

47. Jiang C, Pugh BF. 2009. Nucleosome positioning and gene regulation: advances through genomics. *Nat Rev Genet* **10**: 161–172.

48. Hesselberth JR, Chen X, Zhang Z, Sabo PJ, Sandstrom R, Reynolds AP, Thurman RE, Neph S, Kuehn MS, Noble WS, et al. 2009. Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting. *Nat Methods* **6**: 283–289.

49. Echtenkamp FJ, Gvozdenov Z, Adkins NL, Zhang Y, Lynch-Day M, Watanabe S, Peterson CL, Freeman BC. 2016. Hsp90 and p23 molecular chaperones control chromatin architecture by maintaining the functional pool of the RSC chromatin remodeler. *Mol Cell* **64**: 888–899.

50. Struhl K, Stinchcomb DT, Scherer S, Davis RW. 1979. High-frequency transformation of yeast: Autonomous replication of hybrid DNA molecules. *Proc Natl Acad Sci U S A* **76**: 1035–1039.

PMCID: PMC383183

51. Moqtaderi Z, Geisberg JV, Struhl K. 2018. Extensive structural differences of closely related 3’ mRNA isoforms: Links to Pab1 binding and mRNA stability. *Mol Cell* **72**: 849–861. PMCID: PMC6289678

HARVARD MEDICAL SCHOOL
BLAVATNIK INSTITUTE
DEPARTMENT OF BIOLOGICAL CHEMISTRY
AND MOLECULAR PHARMACOLOGY

240 Longwood Avenue
Boston, MA 02115

March 30, 2020

To the Division of Receipt and Referral
National Institutes of Health (NIH)

Application for the Ruth L. Kirschstein National Research Service Award (NRSA) Individual Postdoctoral Fellowship (Parent F32)

To Whom It May Concern:

I am pleased to submit a grant proposal titled "Transcriptome-scale, condition-specific regulation of mRNA isoform stability via the 3'UTR" for consideration under the NIH Research Grant Program Ruth L. Kirschstein National Research Service Award (NRSA), Individual Postdoctoral Fellowship (Parent F32) with PA number **PA-19-188**. This research project focuses on fundamental gene regulation mechanisms, which is of immense relevance for scientific and medical fields.

The letters of recommendation will be sent from:

Prof. Dr. Brian C. Freeman
University of Illinois, Urbana-Champaign
Department of Cell and Developmental Biology

Prof. Dr. Johannes Buchner
Technische Universität München
Department Chemie

Prof. Dr. Jie Chen
University of Illinois, Urbana-Champaign
Department of Cell and Developmental Biology

Prof. Dr. Stephen Buratowski
Harvard Medical School
Department of Biological Chemistry and Molecular Pharmacology

Thank you for your consideration.
Sincerely,
Zlata Gvozdenov, PhD

August 3, 2020

Transcriptome-scale, condition-specific regulation of mRNA isoform stability via the 3'UTR

Project Summary/Abstract

Altering environmental conditions leads to reprograming of eukaryotic gene expression. One important cellular process that helps adapt gene expression levels to environmental triggers is the regulation of mRNA stability. mRNAs are degraded by specialized cellular machineries, which have been studied in great detail. However, little data exists to explain what cellular signals determine mRNA longevity in response to changing conditions. Alternative polyadenylation allows an individual gene to give rise to multiple distinct mRNA 3' isoforms. Distinct 3' isoforms from the same gene can have different half-lives, and the steady-state distribution of mRNA 3' isoforms can vary under different growth conditions. It is plausible that a diverse array of isoforms is needed for cells to respond to environmental conditions. In part, different isoforms could allow for prompt modulation of gene expression via regulation of mRNA stability. Until now, there have been no genome-wide studies addressing the importance of different isoform profiles for mRNA stabilities under different conditions. This work will comprehensively examine condition-specific isoforms, isoform properties, and cellular factors involved in the regulation of isoform half-lives. In particular, a catalog of isoform half-lives for different growth conditions will be obtained using techniques optimized and/or developed in the Struhl laboratory for studying 3' isoforms. The isoforms will be subsequently examined for sequence and structural elements. The relevance of any *cis*-features identified will be directly tested in the context of defective *trans*-factors, such as degradation machinery components or putative RNA-binding proteins. The proposed work will advance our understanding of the molecular mechanisms that underlie regulated mRNA decay by investigating isoform half-lives and sequence/structural elements under various conditions on a genome-wide scale. The work is expected to reveal 3' isoform-dependent regulation of mRNA stability by the major degradation pathways and/or RNA-binding proteins in response to environmental triggers.

Building on the investigator's background in nuclear proteostasis, training and expertise in genomics and bioinformatics will be gained. The project will be conducted in a world-renowned transcription group with deep experience in this field, providing a strong foundation for the applicant's future independent research as a principal investigator in the field of gene expression.

Specific Aims

Eukaryotic gene expression is regulated at the levels of mRNA synthesis, modification and degradation. In response to environmental triggers, cells balance the rates of mRNA synthesis and degradation to maintain transcript levels for proper cellular operations. General mechanisms that control bulk mRNA degradation are mostly dependent on exosome with its associated proteins or Rat1/Xrn1, and these pathways have been described in great detail [1]. Despite the extensive work on the processes that modify and target mRNA for degradation, upstream signals that sense, determine and regulate mRNA lifespan are less well understood.

When elongating RNA Polymerase II enters the 3' untranslated region (3'UTR), the nascent RNA is cleaved and polyadenylated, ultimately leading to transcriptional termination [2]. A typical gene has numerous (~60) cleavage/polyadenylation sites, and hence many different 3'UTR isoforms [3-5]. The Struhl laboratory showed that 3' mRNA isoforms, even those differing by a single nucleotide, can have different half-lives [6]. More importantly, by comparing half-lives of neighboring clusters of isoforms, they identified many hundreds of mRNA stabilizing and mRNA destabilizing elements that, respectively, increase or decrease mRNA half-lives [6]. Interestingly, the poly(A) tail plays a critical role in mRNA stability, via hybridization with polyU-rich elements in the 3'UTR and by interacting with Pab1, the poly(A)-binding protein [6, 7]. Remarkably, there are extensive structural differences among closely related 3' mRNA isoforms, and these are linked to Pab1 binding, mRNA structure, and other (few if any identified) 3'UTR-binding proteins [7]. Sequences responsible for these extensive structural and functional differences are evolutionarily conserved, indicating biological importance [7].

Very little is known about the regulation of mRNA stability under different environmental or developmental conditions. There are a few examples of regulated mRNA stability during development, oncogenesis and other pathophysiological conditions, and these are often associated with condition-specific polyadenylation [8-10]. However, while it is highly likely that regulated mRNA stability will be an important form of gene regulation, this has never been addressed in a comprehensive manner. The approach developed by the Struhl laboratory to study mRNA stability and mRNA structure of 3' isoforms makes it straightforward to address this issue on a transcriptome-wide scale in an experimental system (yeast) that will take advantage of sophisticated molecular genetic techniques as well as detailed physiological knowledge.

This proposal aims to identify condition-specific 3'UTR *cis*- and *trans*-factors that control mRNA isoform stability. The overall goal of the project is to understand the mechanisms that regulate mRNA half-life and contribute to the regulation of gene expression. To test the stated hypothesis and to achieve the overall goals, I propose the following Aims:

Aim 1. Test and create a catalog of condition-specific 3' isoform half-lives

Aim 2. Identify sequence and structural elements contributing to condition-specific 3' isoform stability

Aim 3. Identify cellular factors that are involved in regulation of condition-specific 3' isoform half-lives

In Aim 1, the half-lives of 3' mRNA isoforms under various physiological conditions will be measured by the method developed in the Struhl laboratory. From this data and in Aim 2A, mRNA stabilizing and destabilizing elements under each condition will be identified, and a comparison between the conditions will identify condition-specific stability elements. This catalog of condition-specific stability elements will be the basis for Aim 2B, in which I will use the DREADS method (also developed in the Struhl laboratory) to determine the *in vivo* structures of 3' isoforms on a transcriptome scale. From the half-life and structural data, we will learn the extent to which 3'UTR sequences and structural features contribute to differential mRNA half-lives under various growth conditions. Of particular interest are common features (either sequence or structural) for elements that mediate a specific form of regulation. Genetic experiments involving mRNA derivatives with mutations in the mRNA stability elements will confirm their functional importance. In Aim 3, I will perform analyses similar to those of Aims 1 and 2 in mutant strains that lack specific RNA-binding proteins or proteins that are part of the RNA degradation machineries. In particular, I will pay attention to how the sequence elements identified in Aim 2 might interact with these factors. The goal is to identify proteins that are important for regulated mRNA stability.

Overall, the work in this proposal examines how 3' isoform differences can regulate mRNA degradation rates in response to environmental triggers. On a broader scale, this work aims to improve our understanding of fundamental mechanisms controlling eukaryotic gene expression.

Research Strategy

Significance: mRNA degradation is an important component for gene regulation as up to 50% of changes in gene expression in response to cellular signals can occur at the level of mRNA stability [11]. Messenger RNA degradation pathways and machineries are well-conserved among eukaryotes [1]. Degradation is generally initiated by dissociation of Pab1, the major poly(A)-binding protein and removal of polyadenylated tail from transcripts 3' ends by the Pan2/Pan3 or Ccr4/Not complexes, which is followed by exosome-mediated 3' to 5' mRNA degradation [12-15]. Alternatively, mRNA degradation can be initiated with 5' decapping and proceeded with 5' to 3' exonuclease degradation by Xrn1 [16, 17]. While general mRNA degradation pathways have been studied in great detail, our understanding of how mRNA sequences within the 3'UTR and proteins that interact with these sequences determine the transcripts' longevity is limited. In mammalian cells, RNA stability can be regulated by sequence-specific RNA-binding proteins and microRNAs (miRNAs), and hence can be regulated in response to environmental conditions and developmental status [18, 19]. mRNA isoforms within a given gene sometimes exhibit different half-lives depending on environmental conditions [9], possibly due to RNA-binding proteins. In yeast, stabilities of individual and functionally related mRNAs can be regulated in response to environmental conditions such as rapid shifts in carbon sources and cellular stress [20, 21]. However, mechanistic understanding and the overall biological significance of regulated mRNA decay in yeast is limited.

Over the past 6 years, the Struhl laboratory has developed innovative approaches to study mRNA stability on the transcriptome level in yeast [6, 7]. The basic method for measuring mRNA half-lives involves rapidly depleting Pol II from the nucleus via anchor-away [22, 23]. Coupling this approach to 3'READS, a method that identifies poly(A) sites [24], the Struhl laboratory measured the half-lives of >20,000 mRNA 3' isoforms [6], something that has never been done before. Furthermore, and unlike previous approaches, this method can be applied to essentially any environmental or genetic condition. Of particular interest, isoforms within a "cluster" (defined arbitrarily as occurring over a <30 nt window, with a maximum gap of 10 nt between consecutive members) have similar half-lives, but different clusters within the same gene can have different half-lives [6]. On this basis, hundreds of sequences conferring stabilization or destabilization of mRNAs were identified in the context of wild-type genes [6].

One class of stabilizing elements is a polyU sequence that can confer increased stability upon introduction into ectopic transcripts [6]. The polyU element inhibits association of poly(A) binding protein (Pab1), by hybridizing to poly(A) tails, the substrate of Pab1, thereby revealing an unexpected and general role of the poly(A) tail in mRNA stability [6]. In addition, genetic experiments indicated that double-stranded structures at 3' ends are a major determinant of mRNA stability [6].

In addition to this new approach to study mRNA half-lives, the Struhl laboratory pioneered new methods to obtain transcriptome-scale structural information (DREADS technique based on DMS modification) and protein binding (CLIP-seq based on immunoprecipitation) on individual 3' mRNA isoforms *in vivo* [7]. Strikingly, near-

identical mRNA isoforms can possess dramatically different structures throughout the 3'UTR [7]. Analyses of identical mRNAs in different species or refolded *in vitro* indicate that structural differences *in vivo* are often due to *trans*-acting factors [7]. The level of Pab1 binding to poly(A) containing isoforms is surprisingly variable, and differences in Pab1 binding correlate with the extent of structural variation for closely-spaced isoforms [7]. A pattern encompassing single-strandedness near the 3' terminus, double-strandedness of the poly(A) tail, and low Pab1 binding is associated with mRNA stability [6, 7]. Thus, individual 3' mRNA isoforms can be remarkably different physical entities *in vivo*. Sequences responsible for isoform-specific structures, differential Pab1 binding, and mRNA stability are evolutionarily conserved, indicating biological function [7].

The approaches described above make it possible to address the key subject of the proposal, namely regulation of mRNA stability. In particular, it is now possible to identify mRNA stabilizing and destabilizing elements with the 3'UTR under any physiological condition. A comparison of such elements under multiple conditions will identify condition-specific stability elements. Furthermore, DREADS and CLIP-Seq can be performed under the same conditions, which makes it possible to directly connect mRNA stability, mRNA structure, and protein binding in response to regulatory stimuli. Ultimately, the goal is to elucidate the molecular mechanisms that underlie regulated mRNA decay, an important and very understudied aspect of gene regulation. To achieve this, I propose following 3 aims:

Aim 1. Test and create a catalog of condition-specific 3' isoform half-lives

Aim 2. Identify sequence and structural elements contributing to condition-specific 3' isoform stability

Aim 3. Identify cellular factors that are involved in regulation of condition-specific 3' isoform half-lives

Innovation: The Struhl laboratory has developed powerful and innovative methodology to measure half-lives, protein-binding, and structures of 3' mRNA isoforms on a transcriptome scale. With these methods, they have identified many hundreds of stabilizing and destabilizing elements in mRNAs. The major innovation in this proposal is to extend this work to a variety of environmental conditions to identify mRNA stability elements whose activity is regulated. While there are indications in the literature that stability of 3' isoforms can be differentially regulated by various environmental triggers, this study will be the first one to identify condition-specific mRNA stability elements on a global scale. Furthermore, combining these mRNA half-life experiments with DREADS (determines structures of 3' isoforms) and CLIP-READS (determines protein-binding of 3' isoforms) in wildtype and mutant yeast strains will allow us to unravel crucial aspects about the regulation of mRNA longevity.

Approach: The approach for measuring half-lives of 3' mRNA isoforms on a transcriptome scale involves conditional and rapid depletion of RNA Polymerase II to eliminates new RNA synthesis followed by measuring the levels of 3' mRNA isoforms (3'READS) and isoform structure (DREADS) at various times after the transcriptional

shutoff. Isoform stabilities along with sequence and structural elements that affect stability will be catalogued and analyzed for a reasonable number of conditions (YPD, galactose, YPD supplemented with sorbitol, and yeast minimal media). Then, experiments will be performed in the absence of degradation components, or putative RNA-binding proteins, to identify isoforms and elements affected by particular components under specific conditions. By perturbing mRNA stabilities via application of several external stimuli, this work will improve our understanding about the factors that determine mRNA isoform longevity.

As regulated mRNA stability has been observed in a number of specific cases, it is virtually certain that global analysis of mRNA isoform half-lives will result in identification of condition-specific stability elements. Even though many experiments (and hence samples) are proposed here, these are standardized and relatively rapid experiments that are routinely performed in our laboratory. In fact, most of the assays and computational analyses were modified and/or developed in our group. There are no difficult or risky steps for the experimental or bioinformatical pipelines. From the consultations with the experts who are directly familiar with this work, about a year will be needed for the completion of Aims 1 and 2. This leaves ample time for Aim 3, which does incorporate experiments (and the amount of work) described in Aim 1 and 2. Once Aims 1 and 2 are completed, I will choose a narrower pool of relevant isoform sequence and/or structural elements to examine in Aim 3.

Aim 1. Test and create a catalog of condition-specific 3' isoform half-lives.

The basic hypothesis of this proposal is that a change in growth conditions would differentially affect the stabilities of certain 3' mRNA isoforms through condition-specific stability elements. This will be examined by measuring 3' isoform half-lives under different growth conditions. The goal is to create a genome-wide catalog of 3' isoform half-lives under various conditions in order to identify condition-specific stability elements. These will be used in the subsequent aims to identify RNA sequences, structural elements, RNA-binding proteins, and genetic requirements that are important for condition-specific regulation of 3' isoform half-lives.

Yeast will be grown in YPD (complete yeast medium), galactose (different carbon source), sorbitol (osmotic stress), and minimal media (lack of nutrients and stress). These conditions were chosen due to the fact that they are commonly used yeast growth conditions with well-studied regulatory pathways, yet isoform half-life data is unavailable. In addition, a condition such as a carbon source switch leads to drastic changes in mRNA turnover of functionally related mRNAs and RNA-binding proteins distributions [21, 25], making it an excellent model to study conditional isoform half-lives.

To measure the half-life of transcript isoforms, I will shut off new mRNA synthesis by depleting the essential RNA Polymerase II subunit Rpb1, and measure the levels of the existing transcripts over time for the given growth conditions. To achieve this, I will combine the two methods our group described earlier [6, 24]. Briefly, a previously published RNA Polymerase II shutoff strain [22] will be grown for several generations in

different media (YPD, YPG, YPD with 1 M sorbitol, and yeast minimal media). With rapamycin addition, the modified RNA Polymerase II largest subunit FKBP-Rpb1 will be exported from the nucleus to the cytoplasm of these cells [22, 23]. Total RNA will be isolated prior to rapamycin-dependent Rpb1 depletion, as well as at 20, 40, 60, 90, 120 min intervals post rapamycin addition. Those time intervals were experimentally optimized to obtain half-life information on thousands of mRNA isoforms in yeast [6]. Then, 3'READS approach will be used to identify genome-wide polyadenylation sites and to quantify the relative abundance of the 3′ mRNA isoforms as described [24]. The experiments will be performed in two replicates and at the sequencing depths our laboratory determined previously as optimal for the heterogeneous isoform studies (~20 million reads per sample) [4, 6]. Conditional depletion of RNA Polymerase II in conjunction with 3'READS will allow for the simultaneous assessment of the distribution, abundance and half-lives of 3' isoforms for different growth conditions.

Data analysis: Firstly, generated data will be compared to the available data sets our group published for consistency [4, 6]. This will include the comparison of the steady-state (pre-rapamycin addition/0 time point) genome-wide 3' isoform distributions and abundances for all genes in published conditions. Our laboratory already established analytical pipelines that tabulate frequencies of 3' transcript isoforms which will greatly expedite my analysis. Secondly, I will calculate half-lives for isoforms with sufficiently high expression levels [6] under the specified conditions. I estimate to obtain half-life data for ~10,000-15,000 3' isoforms [6], but this number may vary with different conditions. I will group 3' isoforms in various classes based on the change of their decay rates in different conditions. Changes in 3'UTR isoform stabilities as a function of changed condition with respect to the reference condition (YPD) will deliver condition-specific 3' isoforms. Meaningful isoform stability changes would be the ones where the half-life of an isoform is x-fold higher or lower than the defined median isoform half-life for that condition. Conversely, identical decay rates for the 3' isoforms, regardless of the conditions, will represent a pool of condition-independent 3'UTR isoform cases. The resulting catalog will provide organized information about the genome-wide and condition-specific mRNA 3' isoform stabilities.

Feasibility and possible pitfalls: Technically, the anchor-away method generally works very well, but it is less effective under some conditions (e.g. heat shock). We can test whether this uncommon problem is occurring by inefficient mRNA decay and by measuring Pol II occupancy on a genome-wide level using established methods [26, 27]. The laboratory has performed numerous anchor-away experiments on many factors under different conditions, so this is not a significant problem. Conceptually, the Aim assumes that condition-specific differences in isoform half-lives and condition-specific stability elements will exist, but this is virtually certain to be the case given that there are already specific examples in multiple organisms. While it is hard to predict how many condition-specific elements will be uncovered and under which conditions, there is no doubt that we will identify them. Hence, there is no doubt that Aim 1 will be completed successfully.

Aim 2. Identify sequence and structural elements contributing to condition-specific 3' isoform stability.

The hypothesis is that isoforms whose half-life vary by conditions might be different in structure. Furthermore, isoform pairs whose end points mark stability elements might also vary in stability according to condition. Here, I will examine the dependencies of different condition-specific mRNA isoforms with varying half-lives catalogued in Aim 1 on the sequence and structural elements. The goal is to identify stabilizing and/or destabilizing elements of condition-specific isoforms with distinct decay rates. The results of this part will reveal conditional 3'UTR *cis*-elements associated with mRNA longevity.

Aim 2A. Identify sequence elements for 3' isoforms with differential, condition-specific half-lives.

3'UTR sequences for the isoforms with condition-dependent half-lives will be analyzed. The goal is to identify isoform elements associated with the given condition-specific isoform longevity status. These elements can be located upstream of the shortest and between short and long same-gene isoform pairs with similar and differential condition-specific half-lives. For identical 3' isoforms with half-lives compared across different conditions, sequence elements upstream of the 3' isoform end will be considered. In addition, using the approach pioneered by our laboratory [6] that is based on the behavior of isoform clusters (i.e. closely related isoforms with the same half-lives), I will analyze condition-specific stability elements which are the sequences that lie between clusters with different half-lives.

Identified elements for isoforms with differing half-lives will be evaluated in terms of the nucleotide compositions and lengths. Given a pool of elements flanked between (or located upstream of) the condition-specific 3' isoform pairs with increased (or reduced) half-lives, the nucleotide frequency distributions will be calculated to define the overrepresented stabilizing (or destabilizing) elements' features. Then, motif-finding tools such as MEME [28], or an alternative for the detection of RNA-binding protein motifs [29], will be implemented to look for the commonly occurring motifs among the same type of regulatory elements (i.e. stabilizing, destabilizing, or neutral elements). In conjunction with this, I will rely on numerous annotated RNA-binding protein atlases [30-33] to scan identified isoform elements and their motifs for known RNA-binding protein targets. This motif/RNA-binding protein detection is needed because in the subsequent Aim 3B, I will investigate whether RNA-binding proteins targeting these isoform motifs regulate the isoform stabilities. Finally, the elements common to the same type of regulation will be investigated for conservation using conservation-detection software [34]. As a control to all these analyses, I will use isoform pairs with half-lives unchanged in any of the conditions. Overall, common elements for the isoforms with condition-dependent and independent half-life changes will be identified. These analyses will reveal whether and which condition-dependent or -independent isoform pairs with same or different half-lives are enriched in stabilizing or destabilizing elements, or also neutral elements. As in the previous Aim 1, there are no alternatives suggested for this Aim as detection of some condition-specific elements is virtually certain. Methods used to detect stabilizing/destabilizing sequence elements or any motifs are very straightforward.

Aim 2B. Perform structural analysis of 3' isoforms with differential, condition-specific half-lives.

Structural information about *in vivo* 3'UTR transcript isoforms with distinct half-lives under different conditions will be obtained. The dimethyl sulfate (DMS) region extraction and deep sequencing (DREADS) technique, which was developed in our laboratory [7], will be used to study condition-specific structure of mRNA 3' isoforms (described in Aim 1). The measurement of 3' isoform structures as a function of time and condition will be performed similarly to Aim 1, except that mRNA collected before and at the different time points post transcription shutoff will be subject to DREADS. Briefly, DREADS technique exploits DMS reactivity towards A and C mRNA residues, preventing reverse transcriptase passage past the modified residue during the sequencing library construction. Given limited DMS treatment, a population of different cDNA truncated molecules is generated. Comparing the frequencies of A/C occurrences between DMS treated and untreated control, and subtracting the untreated control background, the information about 3'UTR reactivity profile is obtained, as described [7].

Further data analysis on condition-specific 3' isoform structures will be performed similarly as described in Aims 1 and 2A. Isoforms reactivity profiles will be compared across the 3' isoform pairs with condition-dependent half-lives. Focus will be on already detected 3' isoforms with conditions-specific differential half-lives (Aim 1) and their stabilizing/destabilizing sequence elements (Aim 2A). Unlike in the previous Aims, comparison of structural information for identical isoforms will be performed over common sequences immediately upstream of isoform endpoints. For identical isoforms, the relationship of structure and isoform stability will be examined by correlating the number of reactive residues in a fixed nucleotide window to the relative stability in each isoform across all conditions. Correlation of changes in isoforms' structures with either increased or decreased half-lives would provide strong evidence for the importance of isoform-specific structure in governing isoform turnover.

Substantial changes in DMS reactivity were shown to occur due to the RNA-protein binding changes more so than due to the RNA folding changes [7]. Here, protein binding to a linear RNA sequence (3'READS data) and the structural changes (DREADS data) due to altered RNA-protein binding will be considered during the 3'UTR structural analysis. Similarly to the previous Aim 2A, sequence- and structure-based motif-finding programs [28, 29, 35-37] will be employed on structural elements (both linear sequences and their DMS reactivities) that affect stability and on neutral element controls. This will help to link specific RNA-binding proteins to isoform structure and stability.

At this point, I will synthesize genome-wide data sets on condition-specific 3' isoform half-lives, and stabilizing and destabilizing sequence and structural elements that will have been tabulated mostly using analytical tools our laboratory developed. Isoforms with specific sequence/structure signatures will be sorted from the most changed to the least changed condition-dependent half-life values. Data mining, specifically association rule learning, will be used to correlate isoform features (3' sequence, structure, RNA-binding protein (motif), condition, and half-life), as well as the combination of those

features, to identify overrepresented condition-specific elements and contexts important for mRNA stability. From this part we will learn whether the isoforms with differential (and also comparable) half-lives contain protected (i.e. structural) elements, and what these are. This part is complementary to Aim 2A: for instance, if identical isoforms have different half-lives in different conditions, this part will demonstrate whether the 3' isoforms have different 3'UTR structures, in which case the structure could be the underlying reason for the change in stability. While it is possible that degradation of molecules with similar reactivity profiles is stochastic, structures from more stable isoforms are likely to differ from less stable ones, and will be reported here. The identified isoforms and elements with differential condition-specific half-life values will be used in the subsequent Aim 3 to define regulatory mechanisms of 3' isoform stability. The narrower pool of isoforms to pursue further will be isoforms with prominent, condition-dependent half-live changes, with identified stabilizing/destabilizing sequence and changed structural elements, and ideally known RNA-binding protein motifs within these elements. In this pool, I will also include control isoforms with no half-life changes.

Aim 2C. Perform genetic analysis of condition-specific 3' isoform sequences and structures.

To verify the significance of identified 3'UTR sequence and structural elements from Aims 2A and 2B for isoform half-lives, genetic experiments will be performed. A reasonable number of identified condition-specific 3'UTRs stabilizing/destabilizing sequence and structural elements will be mutated in the original isoforms using CRISP-Cas9 mediated genome editing tactics [38] and verified using conventional PCR/Sanger sequencing. Similarly, identified 3'UTRs stabilizing/destabilizing elements will be inserted at other genomic locations (e.g. 3' ends of genes with no differential, condition-specific stabilities). Half-lives and structures of these constructs will be determined using transcription shutoff in conjunction with 3'READS and DREADS, as described in Aims 1 and 2. Not all samples will be subject to high-throughput sequencing but some (where possible) will be analyzed with quantitative PCR. The results of the genetic experiments will directly confirm whether the identified isoform sequence or secondary structures are determinants of isoform stabilities. Besides addressing whether the elements are sufficient for the regulatory function, the constructs will be needed for Aim 3B to test the importance of given RNA-binding proteins for regulation of isoform stability. Elements that can be functionally transposed to other regions are very likely to be binding sites for proteins [7].

Aim 3. Identify cellular factors that are involved in regulation of condition-specific 3' isoform half-lives.

The hypothesis here is that conditions-specific mRNA stability reflects differential actions of RNA-binding proteins that interact with specific subsets of mRNAs. It is formally possible that regulation could be mediated via deadenylation or through the major degradation pathways, but this likely affects many (perhaps most) mRNAs. Then, 3'UTR RNA-binding proteins might compete with the secondary 3'UTR structure, or might entirely lack the ability to bind to an isoform. In either scenario, the mRNP's ability

to bind will differ and could affect the transcript degradation rate. Here, the factors whose loss affects differential mRNA stability will be identified.

Aim 3A. Test the influence of RNA degradation machineries on condition-specific 3' isoform stability.

Stabilities of 3' isoforms under different growth conditions will be tested in the absence of the factors involved in degradation of mRNA. The factors of interest will be the exosome subunit Rrp6, Xrn1 nuclease, and non-essential subunits from Ccr4-Not deadenylase and Pan2-Pan3 complexes. The reason why these targets were chosen is because they are the most crucial components related to the regulation of mRNA decay. Given a spectrum of different isoform sequences/structures, which have differential half-lives, it is conceivable that these mRNA isoform elements could connect to the prominent degradation pathways. It can be hypothesized that the degradation machineries are selective for a specific 3'UTR sequence/structure and perhaps influence isoform half-lives. 3'UTRs can also have a sequence/structure code which can be interpreted by specific degradation factors to control the degradation timing. In either case, the half-life of the isoforms with certain 3'UTR elements could be differently impacted in the absence of a degradation component, which would be reflected by altered isoform half-life dynamics. While a general effect on mRNA stability might be observed for some factors, the results will reveal whether and which factors connect to differential isoform stabilities.

Non-essential proteins or subunits of the complexes mentioned above will be deleted in the RNA Polymerase II anchor-away compatible strain and the knock-out strains (5 total) will be verified (described in Aim 2C). The growth of the mutant strains will be tested by measuring doubling time with optical density (or counting the number of cells if cell morphology is adversely affected). These knock-outs are ubiquitously used and I do not anticipate problems, but deletions that render lethality (for target growth conditions) will be eliminated. The strains will be grown under differential conditions and the isoform half-lives and structures will be measured with transcription shutoff in combination with 3'READS and DREADS, as described in Aims 1 and 2. Sequence and structural elements will be analyzed, as described in Aim 2, and compared to the respective wild types. Of note, not all conditions implemented in Aim 1 will end up being utilized here but only the ones that exhibit extensive condition-specific variation in isoform turnover, as identified in Aim 1. Similarly, not all 3' isoforms, isoform half-lives, sequence and structural elements will be analyzed *de novo*, but only condition-specific 3' isoforms with identified stabilizing/destabilizing sequence and/or structural elements, as narrowed down in Aims 1 and 2. From this part we will learn whether the half-lives of condition-specific isoforms with specific 3'UTR elements change in the absence of degradation components. There are no alternatives suggested for this Aim because observation of no effect of select degradation components on isoform-half lives under the given condition is a straightforward answer. It is also possible that all targets will be similarly affected, in which case, we will learn that degradation machineries do not specifically target only certain 3' isoforms in a condition-specific manner.

Aim 3B. Test the influence of RNA binding proteins on condition-specific 3' isoforms stability.

Though possible, it is unlikely that RNA secondary structure is regulated with *trans* RNA-RNA interaction because these are rare in yeast [39]. The most plausible explanation as to why highly similar same-gene mRNA isoforms have differential half-lives is due to proteins that interact with specific RNA sequence and/or structure and alter isoform stability in response to the conditions. In corroboration, RNA-binding protein motifs and corresponding RNA-binding proteins have been linked to differential mRNA stabilities, and to functionally depend or associate with mRNA degradation components [40]. The goal of this part is to identify and validate such protein targets.

RNA-binding protein targets are obtained from Aim 2. There are ~500-800 RNA-binding proteins [41, 42], however, only ~20 have known RNA-binding motifs [7, 31, 32, 33, 40, 43]. The focus will be on these RNA-binding proteins (with known RNA-binding consensus sequence) whose motifs co-occur within the identified stabilizing/destabilizing sequence and/or structural elements for condition-specific 3' isoform pairs with differential half-lives. Of particular interest are decay elements that are differentially protected by DMS and are associated with the mentioned RNA-binding protein motifs. Coincidence of such elements with an isoform's condition-dependent stability elements can be an indicative that a potential target RNA-binding protein is involved in the regulation of this mRNA and these proteins will be prioritized. It is anticipated that most affected 3' isoforms will be the ones corresponding to the genes involved in galactose metabolism, osmoregulation and stress.

Similarly to Aim 3A, the contributions of these annotated RNA-binding proteins (identified in Aim 2) to mRNA isoform stability will be examined by deleting these RNA-binding proteins of interest (~5-10 protein targets). These deletion strains, made and verified as in Aim 2C and 3A, will be grown under differential conditions and the isoform half-lives and structures will be measured with transcription shutoff in combination with 3'READS and DREADS, as described in Aims 1 and 2. Sequence and structural elements will be analyzed, as described in Aim 2, and compared to the respective wild types. Only the relevant growth conditions will be utilized and only a narrower pool of isoforms and elements will be in focus. The expectation is that the depletion of target RNA-binding protein will lead to the change in isoform 3'UTR structure, and thereby its half-life, likely due to the loss of the 3' isoform-protein binding.

The differential binding of RNA-binding protein will be validated next. CLIP-READS, a technique that combines crosslinking and immunoprecipitation (CLIP) with 3'READS to provide a genome-wide map of protein-3'-isoform binding [7], will be utilized. Briefly, cells grown under the conditions specified in Aim 1 (and 2) will be irradiated with UV light in order to crosslink mRNAs and proteins. With protein- or fusion tag-specific antibodies (proteins will be epitope tagged as in Aim 2C), protein−mRNA complexes will be isolated, mRNA released, and subjected to 3'READS. Data analysis will focus on differential protein binding at 3' isoforms as a function of changed condition. The expectation is that the pattern of the target RNA-protein binding changes at condition-

specific 3'UTR stabilizing/destabilizing elements, resulting in the change of target isoform half-life.

Further quick experiments will be performed to directly and functionally link the given proteins to the condition-specific elements and regulated mRNA half-life. Genetic constructs engineered in Aim 2C, in which condition-specific stabilizing/destabilizing elements (with identified target RNA-binding protein motifs) are mutated or transposed at various genomic locations, will be used to directly test whether a) RNA-binding protein is lost at mutated stabilizing/destabilizing elements correlating with the loss of condition-specific change in element's structure and isoform half-life; b) RNA-binding protein is gained at other locations where stabilizing/destabilizing elements have been transplanted (unregulated 3'UTRs replaced with regulated), and the structure and half-life of the corresponding modified 3' isoform is increased/decreased in a condition-specific manner; c) knocking out RNA-binding protein reverses the effect of gained differential, condition-specific protein binding and changed half-life and the structure of genetic constructs.

The genetic constructs' structures and half-lives were determined in Aim 2C. Here, CLIP-READS experiments will be complemented to measure condition-specific protein binding to these RNA elements. Then, RNA-binding protein deletion strains with genetic constructs will be grown in reference and other target condition(s), and will be subject to CLIP-READS and transcription shutoff in conjunction with 3'READS and DREADS, and analyzed as described earlier. The expectations of these experiments are that change in condition-specific RNA-protein binding, besides at the original sites, also changes at the sites where target regulatory elements are transplanted, affecting the structure and stability of the engineered isoforms. Deletion of the target RNA-binding protein would reverse the observed changes in 3'UTR structure and isoform half-life, both at the original and the isoform with unregulated 3'UTRs replaced with regulatory elements. Similarly, mutated regulatory elements would be expected to lose RNA-binding protein and change of growth conditions would not affect the isoform structure and stability. Overall, these experiments will elucidate the molecular mechanisms of mRNA decay regulation via RNA-binding proteins targeting 3' isoform elements.

Feasibility and possible pitfalls: If no proteins are successfully validated with the described strategy, I will implement biochemical approaches. I will express and purify RNA binding proteins or use whole cell extracts and examine the binding to *in-vitro* synthesized RNA with and without elements of interest (band-shifts assays). I will directly test the interaction between isoforms-specific RNA elements and RNA-binding proteins. Regulatory 3'UTR elements that undergo structural changes and can be functionally transposed are certainly expected to be regulated by protein(s) [7].

Scope of the work and timeline: I am aware that this proposal is dependent upon a lot of genome-wide transcript isoform information, and that successful completion of the project requires the extensive synthesis of differential transcript isoforms data with half-life and structural data. However, the genome-wide studies on transcript isoforms are currently the major project in our group conducted by the experimental and bioinformatics experts who pioneered the methodologies. The proposed experiments

and computational analysis are all well established and actively done in our laboratory (and much of the proposed experimental work here is rather related to each other and to our group's current work). There are no unusual, complicated or extensively time-consuming experiments, which guarantees the completion of the project within the designated time frame. The experts in our group, who are very familiar with the experimental and analytical part, predict that completion of Aims 1 and 2 will take about a year. This leaves almost 2 years for Aim 3, which is desirable given that experimental and analytical load of Aim 3 is higher than that for Aims 1 and 2.

This work will advance our understanding of the condition-specific regulation of mRNA stabilities via utilization of differential mRNA 3' isoforms. While numerous mRNA target elements will be identified and characterized within the scope of this project, further work will be needed to elucidate the mechanisms by which regulation of isoform stability takes place *in vivo*. By setting specific aims for further work, this project allows for long-term planning and is well-suited for establishing an independent research program.

Introduction to Resubmission

This resubmitted proposal seeks to improve the previously reviewed submission, pending final resolution. Reviewers were very positive about all the experiments and approaches (the proposal was described as “well-written”, “rigorous”, “well-designed”, “justified”, “significant” and with “high quality of the proposed science” and “high likelihood of success”). I thank reviewers for several helpful comments regarding prioritization that have been incorporated into this resubmission. A major concern for some reviewers was a vague training plan (mostly the mentor’s) and the use of institutional resources; these sections were rewritten with more details. Unfortunately, the points (mostly) regarding training plan/professional development, which reviewers noted as missing, were contained in the application (as listed below), but are now organized in a better way.

Reviewers gave useful comments regarding the number/priority of knock-outs and the number of 3’ isoforms with detectable half-life based on previous studies. I integrated these details to make the aims more specific. Regarding “descriptions of how knock-out proteins will be validated or whether they have an impact on survival or growth”, the testing of the correct knock-out strains and testing of the strains’ growth/viability is incorporated. I would like to point out that my studies are not limited from a data-analysis perspective on whether mutants grow slower as mRNA half-lives of a slower growing cell population (either due to growth conditions or to knock-out strain) are normalized against the respective doubling time.

Reviewers had some concerns that the proposal was somewhat ambitious, despite my previous training and the presence of the founding 3’ isoform experts in the Struhl lab. I would add to this that during my past postdoctoral period I completed a project (written for publication and not related to the current proposal) where I successfully utilized a crucial technique for 3’ isoform studies. Not only were all experiments done entirely by myself, but also the computational analysis which involved the generation of novel pipelines. This is to say that I already have extensive expertise for some of the experimental and analytical parts needed to complete this project within the designated timeframe. Reviewer 3 had an impression that I did not perform the PhD bioinformatics analysis independently, even though I stated that for my later two PhD projects the entire computational analysis (and actually most of it for the first PhD project) was performed completely by myself. With the reviewer’s remark, though, I caught myself that I was not explicit enough when describing my additional computational experience gained during my postdoctoral project, and this was amended. Finally, while I will identify some trans-targets (mRNA-binding proteins responsible for differential 3’ isoform half-life) within the scope of this proposal, the project allows for continuation into

an independent research program, and oncoming research is meant to go beyond the funding period.

Some aspects of the review are not correct. Reviewer 1 wrote that I have one first-author manuscript but I have two first-author manuscripts (excluding a first-author review, which Reviewer 1 did count). I hope a mistakenly viewed publication record is not the reason why Reviewer 1 scored the Fellowship Applicant 2 while the others gave 1.

Reviewers 2 and 3 were concerned about my usage of institutional resources and career development opportunities. I thank reviewers for the comments – I rewrote the plan, but I was sad to notice reviewers raising points for which information was in the application. Reviewer 3 meant I did not mention presenting at seminars in the Boston area but I wrote that I will present my work at these seminars. Regarding my plans for attending conferences, the same reviewer stated that I did not say which conferences and when I would be presenting. However, under the second postdoctoral year in Activities Planned, I wrote that I would attend conferences in the field of gene regulation and share my work with other experts. Nevertheless, I added the names and dates of the conferences and used the verb “presenting” more frequently to avoid confusions. Reviewer 3 also asked what additional resources I would use from Harvard’s Postdoc Office for, e.g., transitioning into applying for jobs. In Activities Planned, I wrote that I would attend seminars and workshops related to the academic job search (provided by the Postdoc Office, as noted on the same page) and that the Office organizes seminars on writing and teaching statements (under Other Resources in Facilities). Reviewer 3 meant I did not explain how I would take advantage of other labs, while I wrote that I have the opportunity to talk to and learn from Professor Buratowski (Institutional Environment), that we can use Professor Buratowski lab’s chromatography columns or CHEF in Fred Winston lab (Equipment), and I also stated that we receive thorough feedback from these transcription labs on our joint meetings. Nevertheless, I rewrote in a clearer way about the benefits from our neighbors (in Institutional Environment). Reviewer 2 noted that specific examples of professional development activities were not described (and referred to the ones described only in Goals and Objectives), while I did describe some examples and listed almost all activities offered (in e.g., Other Resources, Activities Planned). Reviewer 1 meant there was no mention of training/practicing general teaching skills while I mentioned attending seminars on teaching and pedagogy and designing class material for MEDscience students. Overall, in this resubmission, I was much more descriptive and specific about professional development and benefiting from the institutional environment. For a thorough new training plan, please see Goals for the Fellowship Training, Activities Planned, Description of Institutional Environment and Commitment to Training, and Facilities and other Resources.

Bibliography and References Cited

1. Houseley J, Tollervey D. 2009. The Many Pathways of RNA Degradation. *Cell* **136**: 763–776.

2. Zhao J, Hyman L, Moore C. 1999. Formation of mRNA 3' Ends in Eukaryotes: Mechanism, Regulation, and Interrelationships with Other Steps in mRNA Synthesis. *Microbiol Mol Biol Rev* **63**: 405–445.

3. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, et al. 2005. The Transcriptional Landscape of the Mammalian Genome. *Science* **309**: 1559–1563.

4. Moqtaderi Z, Geisberg JV, Jin Y, Fan X, Struhl K. 2013. Species-Specific Factors Mediate Extensive Heterogeneity of mRNA 3′ Ends in Yeasts. *Proc Natl Acad Sci U S A* **110**: 11073–11078.

   PMCID: PMC3703967

5. Pelechano V, Wei W, Steinmetz LM. 2013. Extensive Transcriptional Heterogeneity Revealed by Isoform Profiling. *Nature* **497**: 127–131.

6. Geisberg JV, Moqtaderi Z, Fan X, Ozsolak F, Struhl K. 2014. Global Analysis of mRNA Isoform Half-Lives Reveals Stabilizing and Destabilizing Elements in Yeast. *Cell* **156**: 812–824.

   PMCID: PMC3939777

7. Moqtaderi Z, Geisberg JV, Struhl K. 2018. Extensive Structural Differences of Closely Related 3' mRNA Isoforms: Links to Pab1 Binding and mRNA Stability. *Mol Cell* **72**: 849–861. PMCID: PMC6289678

8. Misquitta CM, Iyer VR, Werstiuk ES, Grover AK. 2001. The Role of 3'-Untranslated Region (3'-UTR) Mediated mRNA Stability in Cardiovascular Pathophysiology. *Mol Cell Biochem* **224**: 53–67.

9. Mayr C, Bartel DP. 2009. Widespread Shortening of 3'UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. *Cell* **138**: 673–684.

10. Ulitsky I, Shkumatava A, Jan CH, Subtelny AO, Koppstein D, Bell GW, Sive H, Bartel DP. 2012. Extensive Alternative Polyadenylation during Zebrafish Development. *Genome Res* **22**: 2054–2066.

11. Cheadle C, Fan J, Cho-Chung YS, Werner T, Ray J, Do L, Gorospe M, Becker KG. (2005). Stability Regulation of mRNA and the Control of Gene Expression. *Ann N Y Acad Sci* **1058**: 196–204.

12. Caponigro G, Parker R. 1995. Multiple Functions for the Poly(A)-Binding Protein in mRNA Decapping and Deadenylation in Yeast. *Genes Dev* **19**: 2421–2432.

13. Brown CE, Tarun SZ Jr, Boeck R, Sachs AB. 1996. PAN3 Encodes a Subunit of the Pab1p-Dependent Poly(A) Nuclease in *Saccharomyces cerevisiae*. *Mol Cell Biol* **16**: 5744–5753.

14. Tucker M, Valencia-Sanchez MA, Staples RR, Chen J, Denis CL, Parker R. 2001. The Transcription Factor Associated Ccr4 and Caf1 Proteins Are Components of the Major Cytoplasmic mRNA Deadenylase in *Saccharomyces cerevisiae*. *Cell* **104**: 377–386.

15. Anderson JS, Parker R. 1998. The 3' to 5' Degradation of Yeast mRNAs is a General Mechanism for mRNA Turnover that Requires the SKI2 DEVH Box Protein and 3' to 5' Exonucleases of the Exosome Complex. *EMBO J* **17**: 1497–1506.

16. Hsu CL, Stevens A. 1993. Yeast Cells Lacking 5'-->3' Exoribonuclease 1 Contain mRNA Species That Are Poly(A) Deficient and Partially Lack the 5' Cap Structure. *Mol Cell Biol* **13**: 4826–4835.

17. Beelman CA, Stevens A, Caponigro G, Lagrandeur TE, Hatfield L, Fortner DM, Parker R. 1996. An Essential Component of the Decapping Enzyme Required for Normal Rates of mRNA Turnover. *Nature* **382**: 642–646.

18. Eulalio A, Huntzinger E, Nishihara T, Rehwinkel J, Fauser M, Izaurralde E. 2009. Deadenylation is a Widespread Effect of miRNA Regulation. *RNA* **15**: 21–32.

19. Hitti E, Khabar KS. 2012. Sequence Variations Affecting AU-Rich Element Function and Disease. *Front Biosci* **17**: 1846–1860.

20. Vasudevan S, Peltz SW. 2001. Regulated ARE-Mediated mRNA Decay in *Saccharomyces cerevisiae*. *Mol Cell* **7**: 1191–1200.

21. Munchel SE, Shultzaberger RK, Takizawa N, Weis K. 2011. Dynamic Profiling of mRNA Turnover Reveals Gene-Specific and System-Wide Regulation of mRNA Decay. *Mol Biol Cell* **22**: 2787–2795.

22. Haruki H, Nishikawa J, Laemmli UK. 2008. The Anchor-Away Technique: Rapid, Conditional Establishment of Yeast Mutant Phenotypes. *Mol Cell* **31**: 925–932.

23. Fan X, Geisberg JV, Wong KH, Jin Y. 2011. Conditional Depletion of Nuclear Proteins by the Anchor Away System. *Curr Protoc Mol Biol* **Chapter 13**: Unit13.10B. PMCID: PMC3076635

24. Jin Y, Geisberg JV, Moqtaderi Z, Ji Z, Hoque M, Tian B, Struhl K. 2015. Mapping 3' mRNA Isoforms on a Genomic Scale. *Curr Protoc Mol Biol* **110**: 4.23.1–4.23.17. PMCID: PMC4397975

25. Jamonnak N, Creamer TJ, Darby MM, Schaughency P, Wheelan SJ, Corden JL. 2011. Yeast Nrd1, Nab3, and Sen1 Transcriptome-Wide Binding Maps Suggest Multiple Roles in Post-Transcriptional RNA Processing. *RNA* **17**: 2011–2025.

26. Coller J. 2008. Methods to Determine mRNA Half-Life in *Saccharomyces cerevisiae*. *Methods Enzymol* **448**: 267–84.

27. Kuras L, Struhl K. 1999. Binding of TBP to promoters *in vivo* is stimulated by activators and requires Pol II holoenzyme. *Nature* **399**: 609–613.

28. Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**: 28–36.

29. Achar A, Sætrom P. 2015. RNA Motif Discovery: A Computational Overview. *Biol Direct* **10**: 61.

30. Hogan DJ, Riordan DP, Gerber AP, Herschlag D, Brown PO. 2008. Diverse RNA-binding Proteins Interact with Functionally Related Sets of RNAs, Suggesting an Extensive Regulatory System. *PLoS Biol* **6**: e255.

31. Tuck AC, Tollervey D. 2013. A Transcriptome-Wide Atlas of RNP Composition Reveals Diverse Classes of mRNAs and lncRNAs. *Cell* **154**: 996–1009.

32. Freeberg MA, Han T, Moresco JJ, Kong A, Yang YC, Lu ZJ, Yates JR, Kim JK. 2013. Pervasive and Dynamic Protein Binding Sites of the mRNA Transcriptome in *Saccharomyces cerevisiae*. *Genome Biol* **14**: R13.

33. Riordan DP, Herschlag D, Brown PO. 2011. Identification of RNA Recognition Elements in the *Saccharomyces cerevisiae* Transcriptome. *Nucleic Acids Res* **39**: 1501–1509.

34. Ramani R, Krumholz K, Huang YF, Siepel A. 2019. PhastWeb: A Web Interface for Evolutionary Conservation Scoring of Multiple Sequence Alignments Using PhastCons and PhyloP. *Bioinformatics* **35**: 2320–2322.

35. *https://genie.weizmann.ac.il/pubs/rnamotifs08/rnamotifs08_predict.html*

36. *https://rna.urmc.rochester.edu/RNAstructureWeb/Servers/Predict3/Predict3.html*

37. *http://rna.tbi.univie.ac.at/*

38. Ryan OW, Poddar S, Cate JH. 2016. CRISPR-Cas9 Genome Engineering in *Saccharomyces cerevisiae* Cells. *Cold Spring Harb Protoc* **2016**: pdb.prot086827.

39. Aw JG, Shen Y, Wilm A, Sun M, Lim XN, Boon KL, Tapsin S, Chan YS, Tan CP, Sim AY, et al. 2016. *In Vivo* Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation. *Mol Cell* **62**: 603–617.

40. Cheng J, Maier KC, Avsec Ž, Rus P, Gagneur J. 2017. *Cis*-Regulatory Elements Explain Most of the mRNA Stability Variation across Genes in Yeast. *RNA* **23**: 1648–1659.

41. Glisovic T, Bachorik JL, Yong J, Dreyfuss G. 2008. RNA-Binding Proteins and Post-Transcriptional Gene Regulation. *FEBS Lett* **582**: 1977–1986.

42. Matia-González AM, Laing EE, Gerber AP. 2015. Conserved mRNA-Binding Proteomes in Eukaryotic Organisms. *Nat Struct Mol Biol* **22**: 1027–1033.

43. She R, Chakravarty AK, Layton CJ, Chircus LM, Andreasson JOL, Damaraju N, McMahon PL, Buenrostro JD, Jarosz DF, Greenleaf WJ. 2017. Comprehensive and Quantitative Mapping of RNA-Protein Interactions across a Transcribed Eukaryotic Genome. *Proc Natl Acad Sci U S A* **114**: 3619–3624.

June 3, 2022

Transcriptional and chromatin regulation via distal to proximal enhancer looping

Transcriptional enhancers were discovered more than 40 years ago as genetic elements that activate transcription from long distances away from RNA polymerase II promoters (Banerji et al., 1981). Enhancers are composed of multiple binding sites for transcriptional activator proteins that function by recruiting chromatin modifying activities and Mediator, a complex that directly interacts with RNA polymerase II (Shlyueva et al., 2014). The transcriptional activators bound to enhancers act synergistically, and the combinatorial nature of enhancers is critical for mammalian cells to mediate billions of possible regulatory inputs from a few thousand DNA-binding transcription factors (TFs) (Carey, 1998).

The key mechanistic question is how transcriptional enhancers function at long and variable distances from promoters. The general answer is that this occurs by DNA looping mediated by interactions between proteins bound at enhancers and those bound near promoters (Shlyueva et al., 2014). Such DNA looping via protein-protein interactions was discovered in bacteria ~40 years ago (Dunn et al., 1984) and is a well-established physical mechanism. Using the Hi-C (High-throughput Chromosome Conformation Capture) approach, billions of DNA loops have been detected in human cells (Jin et al., 2013, Rao et al., 2014). In some cases, DNA loops are regulated under specific physiological conditions and have been correlated with binding by transcriptional activator proteins (ChIP-seq) and gene expression (RNA-seq) (Jin et al., 2013, Phanstiel et al., 2017).

Remarkably, and despite numerous studies, very little is known about the proteins and protein domains that mediate the specific loops necessary for transcriptional activation at a distance. This is a critical issue because the specificity of loop formation underlies transcriptional regulation, and the prevailing view in the field that enhancers loop to core promoters via Mediator (or potentially some other general transcription factor) is incorrect. In particular, Mediator is essential for all Pol II transcription (Petrenko et al., 2017), which means it cannot contribute to enhancer specificity. In a seminal experiment, Nolis et al (2009) demonstrated that loops must involve activator proteins bound to enhancers and to promoter-proximal sequences that are distinct from core promoter elements (e.g. TATA and Initiator), and that artificial connections between such proteins suffice for activation. Using a clever approach, Deng et al (2012) confirmed this idea in the true chromosomal context and identified a domain within Ldb1, GATA1-interacting protein that mediated the loop. More recently, Sun et al (2021) explicitly showed that the Pol II machinery and Mediator does not affect looping. Thus, the challenge is to understand the basis of loop formation beyond these very few examples.

In this proposal, we will develop a general approach to identify proteins and protein domains that mediate loop formation and transcriptional activation. The essence of this approach is to recruit enhancer-binding proteins of interest to novel sites via fusion to heterologous DNA-binding domain (e.g. via yeast Gal4 or endonuclease-deficient Cas9 with guide RNAs). The resulting fusion proteins should form new loops (detected by HiC) that depend on the enhancer-binding protein (or domain of this protein) of interest. Furthermore, by regulating the expression of the fusion protein, it will be possible to detect the immediate transcriptional effects, thereby directly linking the protein domain to loop formation and to gene

activation. In principle, the approach will be applicable to many transcriptional activator proteins and will identify the key domains that mediate looping specificity.

Aim 1. Engineering cell lines with time- and location-controllable TFs domains.

The hybrid proteins will consist of the following 3 components. First, it will include either the yeast Gal4 DNA-binding domain or endonuclease-deficient Cas9. The Gal4 DNA-binding domain will direct the protein to accessible Gal4 binding sites, and the Cas9 derivative will direct the protein to desired sites using appropriate guide RNAs. Guide RNAs will be varied in sequence and length to direct the hybrid protein to single or multiple chosen sites. Second, the hybrid protein will contain the putative looping TF of interest (I have chosen 10 to start with, all of which are expressed in K562 cells) but lacking its own DNA binding domain. By lacking its own DNA-binding domain, the TF will not go to its normal target sites, ensuring that all binding is directed by Gal4 or Cas9. Third, the hybrid protein will contain ERT2, a derivative of the human estrogen receptor ligand binding domain that mediates rapid import into the nucleus in response to tamoxifen. The Struhl laboratory has used ERT2 fusion proteins to analyze the kinetics of TBP and SP1 binding to target sites in human cells on a genome-wide scale (Hasegawa and Struhl, 2019; 2021). Key controls for these experiments are proteins that lack the desired TF sequences (i.e. Gal4 and Cas9 domains alone) as well as Gal4 and Cas9 derivatives that contain artificial activation domains that are not expected to form loops.

Constructs expressing the hybrid proteins will be integrated in K562 genome using lentiviral vectors. K562 cells were the primary cell line of the ENCODE project, and it has by far the most whole genome binding data for TFs, chromatin-modifying activities, and histone modifications. This knowledge will be essential in interpreting the data we obtain. Western blotting in the presence or absence of tamoxifen will be performed to address whether the hybrid protein is expressed and behaves as expected. In this regard, the ERT2 feature of the ideal system described above is not essential for the experiments, but rapid induction of the hybrid proteins is best for distinguishing direct from indirect effects, especially with respect to transcription. As an alternative to ERT2 hybrids, we will express the hybrid proteins from a tetracycline-regulated promoter.

Aim 2. Genome-wide binding, looping, and transcriptional activation mediated by the hybrid proteins.

Given cell lines that express the hybrid proteins generated in aim 1, the first step is to determine their genome-wide binding profiles by ChIP-seq. The control experiments involving the Gal4 and Cas9 domains (with guide RNAs) alone will define the recruitment sites of these heterologous domains. For Gal4, it is likely that there will be a few thousand such sites. This has the advantage of analyzing multiple genomic regions for loop formation, but it has the disadvantage of sorting out the loops. The Cas9-based approach will have fewer target sites (depends on the guide RNAs used) and has the reciprocal advantages and disadvantages.

ChIP-seq involves formaldehyde crosslinking, so it detects protein-DNA and protein-protein interactions associated with specific genomic regions. Thus, comparative ChIP-seq analysis of hybrid and control proteins (antibodies against Gal4, Cas9, or epitope tag) will identify putative regions to which the hybrid protein loops (i.e. ChIP-seq peaks of the hybrid

proteins not seen with the control protein). ChIP-seq mapping is sufficient to determine the location of the sites to which the hybrid protein loops to ~100 bp, meaning it can be localized to specific distal and proximal enhancers. Motifs and ENCODE protein binding data will provide excellent clues to the proteins at these looped sites that may be involved in the looping.

To provide more direct evidence for looping, we will perform chromatin conformation capture experiments between specific sites (3C), between one site and multiple looped sites (4C) or all possible loops (HiC) (Han et al., 2018). These different versions map loops to various levels of resolution, and the choice will depend on the experiment. Of interest are 4C experiments that can determine looping for an individual Gal4- or Cas9- based site of the hybrid protein at good resolution in the background of multiple binding sites. We will also perform ChIA-PET, a technique that combines ChIP and Hi-C, to directly link binding of a TF to a specific loop (Han et al., 2018).

Lastly, we will link protein-binding and loop formation to transcriptional output by performing RNA-Seq experiments under the same conditions. Most experiments that assess transcriptional effects dependent on a given protein are complicated by the possibility of indirect effects. To circumvent this problem, we will measure transcription shortly after induction of the hybrid protein (tamoxifen for ERT2-containing proteins; doxycycline for hybrid proteins expressed from an inducible promoter). Taken together, our approach will identify DNA loops mediated by proteins of interest as well as determine whether these loops activate transcription. As such, they will provide new and critical information about enhancer function on a much larger scale than the very few examples currently known.

Aim 3. Future experiments.

The one-year timeframe of the fellowship is suitable for setting up the approach and carrying out the ChIP-seq, 4C or Hi-C, and RNA-seq experiments on a limited number of TFs. The longer-term goal, which I hope to pursue as an independent PI, will be extended to many more TFs, as each TF mediates a different aspect of biology. Mechanistically, the information will be refined by localizing the domains within the TF that mediate the looping interaction, biochemically characterizing the key protein-protein interactions, and using genetic approaches (e.g. siRNA-mediated depletion of proteins, modification of DNA sequence elements within enhancers) to confirm the functional significance of factors, protein domains, and DNA sequences involved in looping and transcriptional activation. In summary, this proposal should address a fundamental aspect of enhancer function that is very poorly understood at present.

Bibliography and References Cited

Banerji J, Rusconi S, Schaffner W. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell. 1981 Dec;27(2 Pt 1):299-308.

Carey M. The enhanceosome and transcriptional synergy. Cell. 1998 Jan 9;92(1):5-8.

Deng W, Lee J, Wang H, Miller J, Reik A, Gregory PD, Dean A, Blobel GA. Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. Cell. 2012 Jun 8;149(6):1233-44.

Dunn TM, Hahn S, Ogden S, Schleif RF. An operator at -280 base pairs that is required for repression of araBAD operon promoter: addition of DNA helical turns between the operator and promoter cyclically hinders repression. Proc Natl Acad Sci U S A. 1984 Aug;81(16):5017-20.

Han J, Zhang Z, Wang K. 3C and 3C-based techniques: the powerful tools for spatial genome organization deciphering. Mol Cytogenet. 2018 Mar 9;11:21.

Hasegawa Y, Struhl K. Promoter-specific dynamics of TATA-binding protein association with the human genome. Genome Res. 2019 Dec;29(12):1939-1950.

Hasegawa Y, Struhl K. Different SP1 binding dynamics at individual genomic loci in human cells. Proc Natl Acad Sci U S A. 2021 Nov 16;118(46):e2113579118.

Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen CA, Schmitt AD, Espinoza CA, Ren B. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature. 2013 Nov 14;503(7475):290-4.

Nolis IK, McKay DJ, Mantouvalou E, Lomvardas S, Merika M, Thanos D. Transcription factors mediate long-range enhancer-promoter interactions. Proc Natl Acad Sci U S A. 2009 Dec 1;106(48):20222-7.

Petrenko N, Jin Y, Wong KH, Struhl K. Evidence that Mediator is essential for Pol II transcription, but is not a required component of the preinitiation complex in vivo. Elife. 2017 Jul 12;6:e28447.

Phanstiel DH, Van Bortle K, Spacek D, Hess GT, Shamim MS, Machol I, Love MI, Aiden EL, Bassik MC, Snyder MP. Static and Dynamic DNA Loops form AP-1-Bound Activation Hubs during Macrophage Development. Mol Cell. 2017 Sep 21;67(6):1037-1048.e6.

Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014 Dec 18;159(7):1665-80.

Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet. 2014 Apr;15(4):272-86.

HARVARD MEDICAL SCHOOL
DEPARTMENT OF BIOLOGICAL CHEMISTRY
AND MOLECULAR PHARMACOLOGY

KEVIN STRUHL
DAVID WESLEY GAISER PROFESSOR

240 LONGWOOD AVENUE
BOSTON, MASS. 02115
617-432-2104
FAX 617-432-2529
EMAIL kevin@hms.harvard.edu

June 1, 2022

Letter of recommendation for **Zlata Gvozdenov** for the Merck fellowship

I enthusiastically support Zlata Gvozdenov, a postdoctoral fellow in my laboratory for 3.5 years, for the Merck fellowship. In addition to her scientific talent and productivity, Zlata's funding history makes her a superb candidate for the fellowship.

Zlata's first project was very creative. It involved a novel approach to distinguishing biological function from biological noise by performing functional experiments (chromatin, transcription, mRNA characterization) in yeast on a large segment of random-sequence DNA. The essential idea is that anything occurring on random-sequence DNA would be, by definition, biological noise. Without going into details, Zlata discovered that transcriptional noise occurs at high levels in yeast, at least as high and generally higher than transcriptional levels of typical yeast genes. In addition, her results provided new insights into how chromatin (nucleosome occupancy and positioning) and transcription patterns (5' and 3' end formation) arise from the evolved yeast genome. Zlata gave a superb and well-received talk on this in data club. A near final draft of this work has been written, and I anticipate submitted it to *Mol. Cell* in the next month or so.

Although I think this project was creative and interesting (as did reviewers of my MIRA grant who gave me a 10), Zlata had trouble getting funded by a very conservative study section. She eventually received an NIH postdoctoral grant, but for ridiculous reasons they cut it to 17 months instead of the usual 3 years. Based on this experience and legitimate worries about how future experiments in this area would be received, Zlata decided (with my advice) a few months ago to start a new and creative project that addresses a major scientific question but is somewhat more conventional.

Zlata's project for the Merck fellowship addresses a fundamental question about enhancers, that is remarkably understudied. Simply put, while DNA looping is well established for how enhancers work at a distance, almost nothing is known about which protein mediate the loops and activate transcription. Hard to believe, but true. Zlata has designed a clever new and general approach that directly addresses this issue. I'm sure she will successfully cary it out and make a major contribution to the enhancer field, which of course means a major contribution to gene regulation and biology. Success will also set her up very well to go on the job market for an academic position, something she desires.

Zlata is a very hard and dedicated worker who gets a wide variety of experiments to work. I would expect that she will be ready to hit the job market as an independent investigator in 1-2 years, a timeframe that is in excellent accord with the timing of the Merck fellowship and the likely status of her project. While the decision to award a Merck fellowship should of course be based on the individual and the project, I note that from a financial perspective, the Merck fellowship would substitute for the 19 months of funding Zlata lost due to the vagaries of NIH policies. But most importantly, Zlata is a superb candidate for a Merck fellowship, and I hope she is successful in receiving it.

Sincerely yours,

Kevin Struhl

Technische Universität München | Lichtenbergstr. 4 | 85747 Garching

March 3, 2019

Letter of recommendation for Dr. Zlata Gvozdenov

This letter is in support of Dr. Zlata Gvozdenov's application for t**he NRSA Individual Postdoctoral Fellowship Award (F32).** Dr. Gvozdenov is an exceptionally talented postdoctoral research fellow working in Dr. Kevin Struhl laboratory at Harvard Medical School. Zlata was officially a doctoral student under my supervision at the Chair of Biotechnology, Technical University of Munich (TUM), Germany. Her practical work was performed under the supervision of Prof. Dr. Brian Freeman at the University of Illinois at Urbana-Champaign.

Predoctoral achievements

I have known Dr. Gvozdenov since her Master studies – I interviewed her for the Biochemistry Master program at the TUM. She completed her Bachelor degree at the University of Salzburg, in Austria. She had top academic evaluations, great enthusiasm and ambition, as well as astonishing research experience for an undergraduate student. She has been exceptionally eloquent in both German and English – none of which are her native languages. During her Master studies at the TUM, Dr. Gvozdenov continued to excel in theoretical and practical aspects. She was one of our best students who was holding the most prestigious German scholarships (the Scholarship for Scientific Education and Training by DAAD). She was trained by some of the most successful scientists on various topic of biochemistry, molecular biology, bioorganic (and computational) chemistry and biophysics. This strong experimental basis was an excellent foundation for her further work, a view justified by her extremely productive graduate research.

Doctoral achievements

For her PhD, Dr. Gvozdenov investigated the roles of molecular chaperone in the nucleus. Molecular chaperones are typically associated with the cytosolic polypeptide chain folding and are by large understudied in the nuclear realms. Dr. Gvozdenov identified novel roles of molecular chaperones in the nucleus. She completed 3 big projects during her PhD research and showed that 1. Molecular chaperones Hsp90 and p23 regulated chromatin dynamics, 2. Molecular chaperone Hsp90 control chromatin landscape by regulating transcription factor activities globally, 3. CCT/TRiC molecular chaperonin controls transcription. The results of

the first project are published with Dr. Gvozdenov as the co-first author in the high impact journal *Molecular Cell*, with the article being on the cover. The results of the second project are published last year with Dr. Gvozdenov as the first author in *Journal of Molecular Biology*. The third project is in the final phase of preparation for submission. All of these projects are highly innovative. Her last, major project is exceptional breakthrough which shows that "cytosolic" chaperonin (eukaryotic GroEL/ES homolog) is a key transcriptional regulator. This project has an important clinical application – it links chaperone regulation in the nucleus to neurological disorders such as Huntington 's disease. From my interactions with Zlata as her official supervisor, I am aware that she was enormously devoted while completing all steps of the projects from conceptualization, experimental work, computational analyses and data interpretation, data organization to the manuscripts writing. In all these steps, Dr. Gvozdenov has been very independent, efficient and innovative. Her independence and hard-work are especially pronounced for the last two projects: while consulting with Dr. Freeman, Dr. Zlata Gvozdenov completed all steps from experimental work to creative organization of complex data and manuscript writing. In addition to the manuscripts reporting on the results of her research, Dr. Gvozdenov also wrote an excellent book chapter about molecular chaperones in the nucleus together with her supervisor, identifying historical studies that were linking molecular chaperones to the nucleus. To the best of my knowledge, there are no similar reviews available. In summary, Dr. Gvozdenov's graduate research work has resulted in important contributions to the field of molecular chaperones and the regulation of gene transcription. Her research has led to unexpected discoveries, opening new areas. With three high impact first author publications and one book chapter, Dr. Gvozdenov has outstanding graduate schools credentials that rank her as our top graduate student.

Mentoring and communication

As a senior graduate researcher Dr. Gvozdenov has successfully mentored an undergraduate student and trained new graduate students joining Dr. Freeman's group. Zlata also excels in leadership skills as reflected by her parallel career in martial arts for over 20 years, where she has trained hundreds of students. Dr. Gvozdenov is a passionate and articulate speaker – her talks are very well structured and, regardless of the material complexity, captivating and easy to follow. She presented her graduate school work several times at national and international conferences, as well as on numerous seminars, and she was very well remembered by giving impressive talks. The universities Dr. Gvozdenov attended or laboratories she did research in were spread in numerous countries with different cultural background, and each time she easily adjusted as reflected by the excellent evaluations (grades and/or letters).

Concluding statement

From discussion with Zlata, I know that becoming a principal investigator in academia is her genuine career choice. According to her scientific maturity and talent, Dr. Gvozdenov is an outstanding applicant whom I see as a successful PI. Her in-depth familiarity with the literature and the way she gets almost any experiment to work in her hands are all remarkable. She is extremely hard-working, productive, talented, and very innovative (amongst the others, developed new experimental tactics to study molecular chaperones in the nucleus, as described in her thesis). Dr. Gvozdenov is a passionate scientist with good human relations who is rather respected and admired by her peers. I am also very well acquainted with her perseverance, a trait that is very important to succeed. **NIH** postdoctoral fellowship would complete her exceptional achievements and further support her in pursuing an independent investigator career.

Johannes Buchner

Technische Universität München
Fakultät für Chemie
Lehrstuhl für Biotechnologie

Univ.-Prof. Dr. Johannes Buchner
Lichtenbergstr. 4
85747 Garching

Tel. +49.89.289.13341
Fax +49.89.289.13345

johannes.buchner@tum.de
www.**biotech**.ch.**tum**.de
www.tum.de