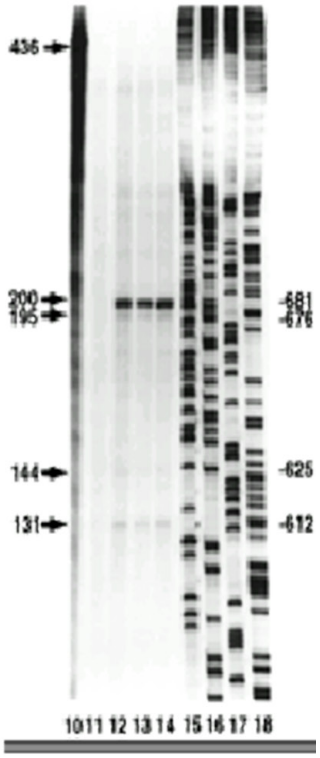
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Swiss Institute of
Bioinformatics

EPD



Eukaryotic Promoter Database
30th Anniversary Symposium

Promoter Research: Past, Present and Future

29-30 September 2016
Conference Centre Starling Hotel-EPFL, Lausanne

Short Talks Abstracts

1. **Tugce Bilgin Sonay**, UZH: Tandem repeats in promoters confer gene expression divergence
2. **Paul Gagniuc**, UPB - Romania: A pattern approach for promoter research
3. **Mahmoud Ibrahim**, MDC-Berlin - Germany: Promoter Dynamics during Directed Differentiation of Motor Neurons
4. **Dominik Hartl**, FMI: Dissecting sequence determinants of CpG island promoter activity
5. **Sarah Natalia Mapelli**, IOR: Promoter-proximal transcripts: a genome-wide prediction and functional validation study
6. **Maroun Bou Sleiman**, EPFL: The alternative splicing landscape of the Drosophila gut upon enteric infection

Tandem repeats in promoters confer gene expression divergence

Tugce Bilgin Sonay, Andreas Wagner
Institute of Evolutionary Biology and Environmental Scienes, University of Zurich, Zurich, Switzerland
The Swiss Institute of Bioinformatics, Lausanne, Switzerland

INTRODUCTION

Tandem Repeats (TRs) are DNA tracts in which a short base-pair motif is repeated several times in tandem. They are among the most variable loci, experiencing mutations in the number of repeat units that are 100 to 100,000 times more frequent than point mutations [1], which makes them an important source of genetic variation. Genome-wide surveys of TR variation have been scarce due to the technical difficulties derived from short-read technology. Recent advances in sequencing technology, together with the development of computational tools, have permitted analysis of several thousand loci from multiple individuals in a cost-effective manner [2].

In eukaryotes, TRs located in promoters, tend to occur in genes associated with transcriptional regulation, DNA binding, protein–protein binding, and developmental processes [3], suggesting a regulatory role for TRs. In fact, TRs are abundant in human promoters [4] and emerge as good candidates for a type of genomic variation that can directly alter gene expression [5]. Accumulating evidence from exhaustive genetic studies has already shown that TR variation has dramatic phenotypic effects. An especially remarkable example in mammals, regards features of a dog's snout: the degree of dorsoventral nose bend and midface length correlate with the length of two tandem repeats in a gene that regulates bone formation [6].

Because gene expression changes might contribute to the fundamental differences between humans and other species [7], it is imperative to study mechanisms that may permit rapid expression changes on short evolutionary time scales. We therefore explored the impact of TRs on gene expression evolution in three species, human, chimpanzee and rhesus macaque by using in total 30,275 TRs (repeat unit length 2-50 base pairs).

RESULTS AND DISCUSSION

We showed for the first time impact of TRs on gene expression divergence between human and other primates. More specifically, we observed an association between repeats in gene promoters and increased expression divergence, an observation that was robust to changes in the method used to identify tandem repeats and to assess gene expression divergence.

To identify all genes with TRs in the promoters, we used a set of 13,035 one-to-one

orthologous genes in the reference genome assemblies of human, chimpanzee, and macaque. We found that on average 29-31% of these genes harbored TRs in their 5 kb upstream of transcription start sites. To evaluate the functionality of our promoter definition, we first checked the overlap between the TRs and DNase hypersensitive sites [8], which are accessible regions of DNA, associated with gene regulatory elements. We found a significant enrichment of TRs in DNase hypersensitive sites (P-value=10^{-350}), suggesting that a substantial part of repeat sequences could potentially be involved in gene regulation.

Based on this premise, we used publicly available RNA-seq gene expression data [9] to assess whether genes that contain TRs in their promoters have higher expression divergence than those that do not. In order to avoid noise and bias for organ-specific gene expression variation differences, we took a phylogenetic approach and performed a bootstrap-like resampling analysis, where gene expression values were sampled from different individuals of a species (see Methods). We computed two different expression distance matrices of (1000 replicates) × (3 species pairs) for each organ and employed these matrices to construct neighbor-joining gene expression trees. We found that the total tree length of genes with repeats was significantly greater in all organs ($P<10^{-200}$ except for liver, where P=0.02) (Figure 1).

When changing the distance of the upstream regions considered, we found that repeat-containing genes diverged more rapidly in their expression, and this difference was most pronounced for repeats within 1 kbp upstream of the transcription start site (95% CI: 0.0145, 0.0146). The difference got progressively smaller as we included repeats that are further away from the transcription start site (95% CI for windows of length 10 kbp: 0.003, 0.006; 15 kbp: 0.0021, 0.0024; 20 kbp: 0.0004, 0.0007). This observation might be explained through core promoter modules occurring preferentially close to this site and exerting a strong influence over transcriptional regulation [10].

We then wondered whether the observed association between TRs and expression divergence is simply due to relaxed selection. We therefore performed multiple analyses to compare the level of selection in genes and their promoters with and without TRs. Between these two sets of genes, we found no significant difference in the sequence divergence of the coding sequences, or of the promoter sequences. Moreover, number of recombination hotspots in repeat containing promoters were statistically indistinguishable from that of other promoters suggesting that genes with TRs do not experience more recombination events than genes without TRs. We also found that promoters with TRs are not in close proximity of particular chromosome locations, such as centromeres or telomeres. Next, as CpG islands have gene regulatory roles through epigenetic mechanisms [11], we repeated our analyses while removing 216 TRs that overlap a CpG island, which did not make a difference in the observed association between TRs and expression divergence. Altogether, all these different approaches suggest no co-factor or evidence for a more relaxed selection in genes with TRs in their promoters compared to those without.

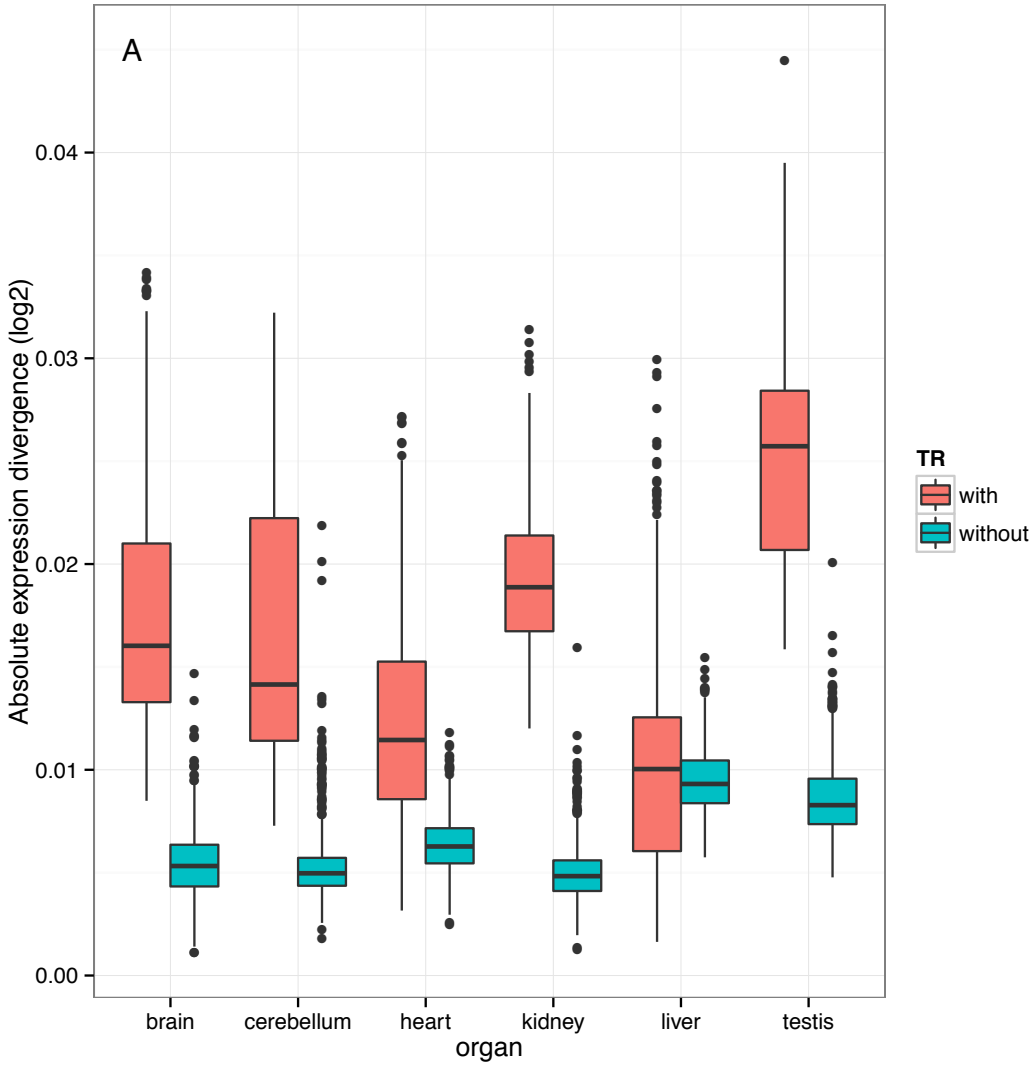


Figure 1. Boxplot of total tree lengths of genes with repeats (thick lines) and genes without repeats (thin lines). Horizontal lines in the middle of each box mark the median, edges of boxes correspond to the 25th and 75th percentiles, and whiskers cover 99.3% of the data points.

Our results provide further motivation for future studies to clarify the exact role of these genes in primate evolution, and the extent to which repeats may have been involved in their regulation. In a seminal paper, King and Wilson [7] observed about humans and chimpanzees that "their macromolecules are so alike that regulatory mutations may account for their biological differences." Since then, we have learned that such mutations, and in particular mutations that cause gene expression change, are indeed important in the evolution of primates and other organisms. Our work showed for the first time in eukaryotes that TRs, a type of sequence with unusually high mutability, are a relevant class of regulatory mutations that might contribute to such species differences. After the publication of this work [12], two new studies on human TR variation showed concordant observations on gene expression regulation [13, 14] in human populations.

METHODS

Tandem Repeat Identification

We identified tandem repeats in the GRCh38, panTro2.1.4 and mmul_1 reference genomes, corresponding respectively to human, chimpanzee and rhesus macaque using Tandem Repeat Finder (TRF) (Benson 1999) v2.30, with parameters "2 5 5 80 10 80 5".

Gene expression

The gene expression data we used were based on RNA sequencing of ~3.2 billion 76-base pair Illumina Genome Analyser IIx reads [9]: one-to-one gene expression measurements from multiple primates, where each gene's expression had been measured in six different organs (brain, cerebellum, heart, kidney, liver, testis) for between 1 and 6 individuals per species. From this data set, we used RNA-seq based expression values of all 13,035 one-to-one gene orthologs from humans, chimpanzees and macaques. We obtained DNA sequences of the genes in our expression data set through the BioMart tool of Ensembl [15], using human annotation version GRCh37.p10, chimpanzee annotation CHIMP2.1.4, and macaque annotation MMUL_1.

We computed statistical differences in gene expression divergence with a bootstrap-like resampling procedure, where we sampled gene expression values from different individuals of a species to create 1000 replicate data sets (n=13,035) for each organ, and species. We partitioned gene pairs in each such data set into two groups: gene pairs where genes of a given species contained tandem repeats in promoter, and gene pairs without such repeats. We then computed, separately for genes in the two groups, a pairwise matrix of Euclidean gene expression distance [16] between all genes in a pair of species.

Overall, we created 12 separate expression distance matrices of size (1000×3), for two gene subsets based on repeat presence and for six organs. We used these matrices to construct gene expression trees using the neighbor-joining approach (implemented in the 'ape' package [17] in R. We used the branch lengths of the trees we constructed as a measure of gene expression divergence. To test the null-hypothesis that the expression divergences (branch lengths) of the 1000 sampled trees were significantly different between the two gene subsets for each organ, we used paired *t*-tests (N=1000, df=n-1 unless otherwise mentioned). All P values are reported after Bonferroni correction for multiple testing and were found to be robust to number of bootstrap replicates.

REFERENCES

1. Legendre M, Pochet N, Pak T, Verstrepen KJ: **Sequence-based estimation of minisatellite and microsatellite repeat variability**. *Genome Res* 2007, **17**:1787–1796.

2. Willems TF, Gymrek M, Highnam G, Mittelman D, Erlich Y: **The landscape of human STR variation**. *Genome Res* 2014:gr.177774.114–.

3. Gemayel R, Vinces MD, Legendre M, Verstrepen KJ: **Variable Tandem Repeats Accelerate Evolution of Coding and Regulatory Sequences.** *Annu Rev Genet* 2010.

4. Sawaya S, Bagshaw A, Buschiazzo E, Kumar P, Chowdhury S, Black MA, Gemmell N: **Microsatellite Tandem Repeats Are Abundant in Human Promoters and Are Associated with Regulatory Elements**. *PLoS One* 2013, **8**.

5. Vinces MD, Legendre M, Caldara M, Hagihara M, Verstrepen KJ: **Unstable tandem repeats in promoters confer transcriptional evolvability.** *Science* 2009, **324**:1213–6.

6. Fondon JW, Garner HR: **Molecular origins of rapid and continuous morphological evolution.** *Proc Natl Acad Sci U S A* 2004, **101**:18058–18063.

7. King MC, Wilson AC: **Evolution at two levels in humans and chimpanzees**. *Science (80-)* 1975, **188**:107–116.

8. Material SO, Web S, Press H, York N, Nw A: **The ENCODE (ENCyclopedia Of DNA Elements) Project.** *Science* 2004, **306**:636–40.

9. Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M, Albert FW, Zeller U, Khaitovich P, Grützner F, Bergmann S, Nielsen R, Pääbo S, Kaessmann H: **The evolution of gene expression levels in mammalian organs**. *Nature* 2011, **478**:343–348.

10. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman M V, Romano LA: **The evolution of transcriptional regulation in eukaryotes.** *Mol Biol Evol* 2003, **20**:1377–1419.

11. Deaton AM, Bird A: **CpG islands and the regulation of transcription.** *Genes Dev* 2011, **25**:1010–1022.

12. Bilgin Sonay T, Carvalho T, Robinson M, Greminger M, Krutzen M, Comas D, Highnam G, Mittelman DA, Sharp AJ, Marques-Bonet T, Wagner A: **Tandem repeat variation in human and great ape populations and its impact on gene expression divergence.** *Genome Res* 2015, **25**:1591–9.

13. Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, Daly MJ, Price AL, Pritchard JK, Sharp AJ, Erlich Y: **Abundant contribution of short tandem repeats to gene expression variation in humans**. *Nat Genet* 2015, **48**:22–29.

14. Quilez J, Guilmatre A, Garg P, Highnam G, Gymrek M, Erlich Y, Joshi RS, Mittelman D, Sharp AJ: **Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans**. *Nucleic Acids Res* 2016:gkw219.

15. Kinsella RJ, Kähäri A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, Kersey P, Flicek P: **Ensembl BioMarts: a hub for data retrieval across taxonomic space**. *Database J Biol databases curation* 2011, **2011**:9.

16. Tirosh I, Weinberger A, Carmi M, Barkai N: **A genetic signature of interspecies variations in gene expression.** *Nat Genet* 2006, **38**:830–834.

17. Paradis E, Claude J, Strimmer K: **APE: Analyses of Phylogenetics and Evolution in R language.** *Bioinformatics* 2004, **20**:289–290.

A pattern approach for promoter research

Paul A. Gagniuc[1,2], Constantin Ionescu-Tirgoviste[1], Cristian Guja[1]

[1] National Institute of Diabetes, Nutrition and Metabolic Diseases ”N.C. Paulescu”, Romania

[2] Institute of Genetics, University of Bucharest, Bucharest, 060101, Romania

Introduction

Further investigations into biological complexity require diverse analysis methods. In the past, the stochastic approaches and the sequence alignment algorithm have been and still are the most common methods in the field of DNA research. However, many known methods currently adopted for promoter research rarely lead to eloquent conclusions or far reaching ties with other biological phenomena. With the development of Eukaryotic Promoter Database (EPD) many other mathematical ideas were put into practice to slowly decrypt one of the most challenging regions of the genome, namely the promoter regions. Here, our goal is to introduce a new method that has yielded interesting results in the past.

Materials and Methods

Our method is based on two-dimensional patterns of gene promoters. A two-dimensional pattern consists of a set of points whose coordinates are plotted according to the information drawn from the DNA sequence of a promoter region [1]. For each point the x-coordinate is represented by the percentage of cytosine and guanine, while the y-coordinate is represented by a new type of value, namely Kappa Index of Coincidence. The relationship between gene promoters is evaluated by a comparative analysis of these patterns and by considering how they cluster within a global distribution. Thus, in the resulting distribution the overlapping promoters (or close positions) indicate a common usage of molecules involved in the transcription process [1].

Results

Over several years, the DNA pattern method was extensively tested based on EPD promoters. It seems that DNA patterns reflect the interaction between molecules associated with specific promoter regions [1]. For instance, our first large-scale study on EPD promoters (in 2012) indicated a total of 10 classes of gene promoters in eukaryotes [1]. Another interesting link has been made between *Homo Sapiens* gene promoters and known studies related to the structure of the cell nucleus [2]. A more direct approche of this method revealed several observations with reference to the connection between common phenotypes encountered in medical practice, such as: T1D (Type 1 Diabetes), T2D (Type 2 Diabetes), autoimmunity or obesity [3-7]. Promoters of genes associated with these phenotypes have been studied and the findings may be presented in short as follows: 1) promoters of genes associated with the two

main phenotypes of diabetes seem to be associated to a third intermediary phenotype, named intermediary diabetes mellitus (IDM) [3,4]. 2) Promoters of genes associated with T1D and the promoters of genes associated with other autoimmune diseases, share common transcription factors [5]. 3) IDM and obesity genes are equipped with promoters prone to common triggers [6,7]. 4) IDM and obesity associated genes (with promoters similar in structure) are functionally receptive to transcription factors specific to both T1D and T2D phenotypes [3,6,7]. It is becoming increasingly evident that our scientific community requires *real next-generation analysis methods* and not only next-generation technological methods. By next-generation we should understand new approaches and not old methods slightly perfected. Our analysis method is largely unexplored outside the study of gene promoters. Here, some highlights of different fundamental and specific studies related to current medical problems have been mentioned from a DNA pattern perspective. Further developments in the same direction include promoters of genes associated with different types of cancers. There is no doubt in our vision that gene promoters are behind many future results meant to cover the huge gaps that exist between various biological processes.

References

1. Gagniuc and Ionescu-Tirgoviste: *Eukaryotic genomes may exhibit up to 10 generic classes of gene promoters*. BMC Genomics 2012, 13:512.
2. Gagniuc and Ionescu-Tîrgovişte. *Gene promoters show chromosome specificity and reveal chromosome territories in humans*, BMC Genomics 2013, 14:278.
3. Ionescu-Tîrgovişte C*, Gagniuc PA*, Guja C (2015) *Structural Properties of Gene Promoters Highlight More than Two Phenotypes of Diabetes*. PLoS ONE 10(9): e0137950.
4. Constantin Ionescu-Tirgoviste, Paul Gagniuc, Cristian Guja. *Intermediary diabetes mellitus (IDM): a new pathology between boundaries*. European Association for the Study of Diabetes (EASD), Diabetologia 2014, 2014.
5. Cristian Guja, Paul Gagniuc, Constantin Ionescu-Tirgoviste. *The promoters of genes may be the closest link between type 1 diabetes and other autoimmune diseases*. European Association for the Study of Diabetes (EASD), Diabetologia 2014, 2014.
6. Constantin Ionescu-Tirgoviste, Paul Gagniuc, Cristian Guja. *The promoters of genes associated with type 1 and type 2 diabetes seem to have some specific features*. Diabetologia 2012, Volume 55, Issue 1 Supplement, pp 1-538. DOI 10.1007/s00125-012-2688-9.
7. Paul A. Gagniuc, Constantin Ionescu-Tirgoviste, Elvira Gubceac, Cristian Guja. *A correlation between Intermediary Diabetes Mellitus and Obesity*. 4th International Symposium on ADIPOBIOLOGY and ADIPOPHARMACOLOGY (ISAA). Bucharest, Romania, October 28 - 31, 2015.

Promoter Dynamics during Directed Differentiation of Motor Neurons

Mahmoud M Ibrahim[1], Silvia Velasco[2], Esteban O Mazzoni[2] and Uwe Ohler[1]

[1] *The Berlin Institute for Medical Systems Biology at the Max-Delbrueck-Center for Molecular Medicine, Berlin, Germany*

[2] *Department of Biology, New York University, New York City, NY, USA.*

Introduction

Embryonic stem cells are known to feature bivalent chromatin states, where the nucleosomes adjacent to a promoter feature both active (H3K4me3) and repressive (H3K27me3) histone modifications. This arrangement has been hypothesized to keep the promoter in a “poised” state allowing for timely response to promoter activation signals. However, it is not clear how bivalent chromatin is resolved and whether it is required for subsequent gene activation during differentiation.

Studying the resolution and activation / repression of promoter chromatin states over time is hindered by the fact that in vitro differentiation relies on studying heterogeneous cell populations going through an inefficient differentiation process (typically only a small percentage of the cells successfully differentiate). This limits the power of the data produced by ChIP-Seq, open chromatin assays and RNA-Seq.

Methods

We take advantage of a highly efficient *in vitro* differentiation system, where ESCs are programmed into spinal motor neurons with >90% efficiency, by inducing the expression of three transcription factors (Ngn2, Isl1 and Lhx3) **[1]** going through a homogeneous differentiation process. Using this system we profile the histone modifications H3K4me3, H3K4me2, H3K27ac and H3K27me3 over the time-course of the differentiation process, using ChIP-Seq **[2]**.

In order to co-cluster multiple histone modifications over the time-course of differentiation, we developed a Bayesian Network model that can take as input multiple time-course histone modification datasets and provides time-course chromatin state clusters, with each cluster representing a the time-course profiles of all analyzed histone modifications together **[2]**.

Results

Using our data and model, we clustered promoter regions into 11 time-course clusters **[2]**. The clusters revealed that promoters are activated and repressed with multiple waves of chromatin dynamics, which is reflected in gene expression data. Bivalent promoters are resolved through an antagonistic switch between H3K27me3 and H3K27ac, while genes repressed during differentiation feature an opposite switch ending in a bivalent chromatin state at the end of the differentiation. However, not all activated genes start in a bivalent state and not all repressed genes feature a switch to bivalency.

On average, ESC bivalent genes that were activated during differentiation did show accelerated gene activated during differentiation compared to activated genes that did not start in a bivalent state. However, this was, at least partially, explained by the chromatin dynamics of their nearby enhancers.

Our results indicate that bivalency might not be a required prerequisite for subsequent activation of all promoters during differentiation and cell fate programming. The resolution and establishment of

bivalent and repressed promoters during differentiation and programming might act to fine-tune the response kinetics of certain promoters relative to others depending on the dynamics of promoter-enhancer interactions occurring during differentiation.

References

[1] Esteban O Mazzoni, Shaun Mahony, Michael Closser, Carolyn A Morrison, Stephane Nedelec, Damian J Williams, Disi An, David K Gifford and Hynek Wichterle. **2013**. Synergistic binding of transcription factors to cell-specific enhancers programs motor neuron identity. *Nature Neuroscience*. 16, 1219–1227.

[2] Silvia Velasco, Mahmoud M Ibrahim, Akshay Kakumanu, Görkem Garipler, Begum Aydin, Mohamed Ahmed Al-Sayegh, Antje Hirsekorn, Farah Abdul-Rahman, Rahul Satija, Uwe Ohler, Shaun Mahony and Esteban O Mazzoni. **2016.** A multi-step transcriptional and chromatin state cascade underlies motor neuron programming. *in revision*.

Dissecting sequence determinants of CpG island promoter activity

Dominik Hartl, Dirk Schübeler
Friedrich Miescher Institute for Biomedical Research, Basel

Cellular diversity is created by specific gene expression patterns. Cell type specificity is encoded within regulatory elements such as promoters and enhancers and acted upon by transcription factors (TFs). The majority of promoters in mammalian genomes display high densities of CpG dinucleotides, a sequence feature termed CpG island (CGI). These CGI promoters regulate 90% of all transcripts generated by RNA Polymerase II, and while the majority of the genes driven by CGI promoters are ubiquitously expressed (house-keeping genes, HKGs), 30% are only active in specific tissues (e.g. the hox genes). If and how CpG richness in CGIs contributes to their activity of gene expression remains elusive and is a key question of this project.
More specifically, we systematically determine DNA sequence components controlling the transcriptional activity of CGI promoters *in vivo* using murine stem cells (ESCs).
For this purpose we established a high throughput reporter assay to quantify the transcriptional activity of hundreds of synthesized promoter sequence variants in parallel at a defined chromosomal locus. Testing of a library of variants revealed that high CpG density at CGI promoters is necessary but not sufficient for transcriptional activity and that not only the density itself but also context of CpGs does play a role for transcriptional activity. Moreover through comprehensive mutation of putative TF motifs we can show that low complexity motifs, such as Sp1, have the highest impact on transcription. We will discuss these findings in light of current models of CpG island regulation.

Promoter-proximal transcripts: a genome-wide prediction and functional validation study

Sarah Natalia Mapelli, Sara Napoli, Giuseppina Pisignano, Giuseppina M. Carbone and Carlo V. Catapano
Tumor Biology and Experimental Therapeutics Program, Institute of Oncology Research (IOR), and Oncology Institute of Southern Switzerland (IOSI), and USI

Long noncoding RNAs (lncRNAs) are emerging as important players in the epigenetic machinery with key roles in development and diseases. Recent RNA-sequencing studies have revealed frequent low-abundance transcripts in the promoter regions that might act as cis-regulatory elements of the adjacent genes. However, the frequency and function of these promoter-proximal noncoding transcripts remain largely unknown. In this study we applied transcripts prediction strategies to a collection of global nuclear run-on sequencing (GRO-seq) datasets from human cell lines to perform a survey of promoter-proximal transcripts in the human genome. We discovered several promoters with promoter-proximal transcripts (≤2 kb from the gene transcription start site) and highly dynamic patterns of transcription both in the sense and antisense orientation relative to the neighboring genes in the cell lines examined. The diversity of the relationships between promoter-proximal transcripts and expression of the neighboring genes suggested their possible involvement in both transcriptional activation and silencing depending on the promoter and cell context.
For selected genes the presence and orientation of the predicted promoter-proximal transcripts were confirmed by qRT-PCR and strand-specific RT-PCR in human prostate cell lines. Small interfering RNAs (siRNAs) were used next to target the promoter-proximal transcripts and assess their functional impact on the neighboring gene expression. Modulation of the adjacent genes was observed in a promoter transcript-dependent and cell-specific manner leading to either transcriptional activation or repression. Collectively, these findings demonstrate the widespread presence of sense and antisense promoter-proximal transcripts and provide evidence of their involvement in transcriptional gene regulation in a cell context dependent fashion. Understanding the contribution of promoter-proximal transcripts to epigenetic regulatory networks may uncover novel mechanisms of disease and provide the basis for novel epigenetic drug discovery.

The alternative splicing landscape of the *Drosophila* gut upon enteric infection

Bou Sleiman M[1], Andreani T[4], Frochaux M[1], Osman D[3], Lemaitre B[2], Guigo R[5], Deplancke B[1]

[1] Institute of Bioengineering, EPFL, [2] Global Health Institute, EPFL, [3] Azm Center for Research in Biotechnology and its Applications, EDST, Lebanese University, Tripoli, Lebanon, [4] Institute of Molecular Biology, Mainz, Germany, [5] Centre for Genomic Regulation, Barcelona, Spain.

Abstract

RNA Splicing is a key mechanism that not only generates protein diversity, but contributes to the fine tuning of the transcriptome. This ability to diversify and control the transcriptional output of the genome may facilitate how the organism adapts to a changing environment. We employ a systems approach in the study of isoform ratios in the infected and uninfected guts of females from 38 inbred lines of Drosophila melanogaster. We find that infection leads to extensive and consistent differences in isoform ratios, which result in a more diverse transcriptome, that is skewed toward longer transcripts, due to longer 5'UTRs. Additionally, we establish a role for genetic variation in mediating inter-individual differences, with splicing Quantitative Trait Loci being more numerous in the infected state and preferentially located in the 5' end of transcripts and directly upstream of the splice donor sites. Moreover, we find a general increase in intron retention events concentrated in 5' ends of transcripts. The length, CG content, and RNA Polymerase II occupancy of the retained introns suggest that they have exon-like characteristics and are possibly being translated. Finally, we show that the sequences of retained introns are enriched with the Lark/RBM4 RNA-binding motif, pointing to a role of Lark in mediating the gut defense response. For the first time, we describe a link between splicing and the gut's response to enteric infection, which could have general implications on gene regulation and protein translation.

1.1 Results

1.1.1 Enteric infection leads to extensive changes in transcript isoform ratios

We have previously measured the resistance of 140 Drosophila Genetic Reference Panel (DGRP) lines to enteric infection with *Pseudomonas entomophila* (*P.e.*)[2]. In this study, we selected 38 DGRP lines, 20 of which are susceptible and 18 resistant to *P.e.* enteric infection (**Fig. 1a**). Whole guts of adult female flies were sequenced in the infected and uninfected condition (total of 76 samples). To gain insight into the changes in the isoform composition of each gene after infection, we used a multivatiate distance-based approach described in Gonzalez-Porta et al. (2012) [3]. Of the 1877 genes that passed filtering, 40% were significantly changed after infection (**Fig. 1b**, p-value of

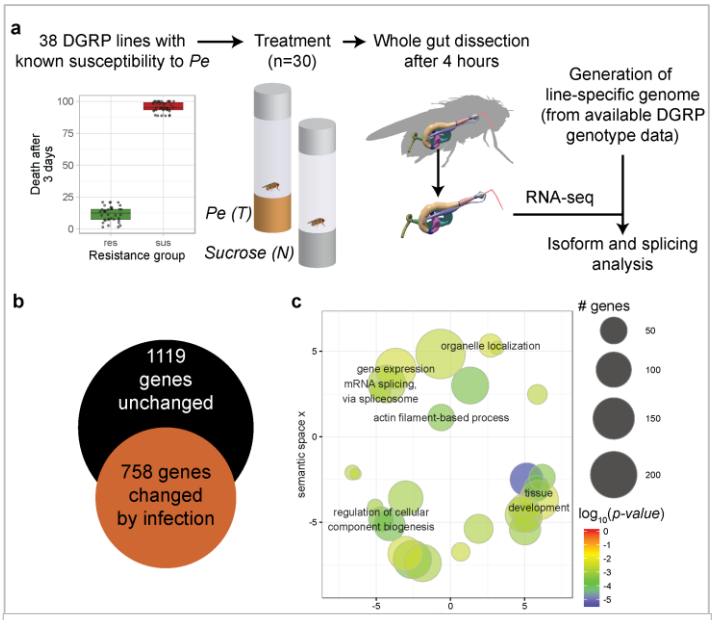


Figure 1. (a) The general experimental design of the study. **(b)** Venn diagram of the number of genes whose isoform ratios are significantly altered after infection. **(c)** Graphical representation of enriched biological process gene ontology terms based on the list of genes whose isoform ratios are differentially expressed after infection.

homogeneity > 0.05, BH-corrected p-value < 0.05, effect size > 0.2). Interestingly, only 25% of the significant genes based on splicing ratios are known to be differentially expressed after infection, suggesting that gene-level differential expression could overlook important aspects of the gut transcriptional response to enteric infection. A gene ontology analysis shows that genes associated with RNA-metabolism, organelle organization and biogenesis, and epithelial tissue development are enriched within this set (**Fig. 1c**). Interestingly, the set of genes we obtained is not enriched with immunity gene ontology terms. This could possibly be due to different regulatory restraints imposed on genes involved in the immediate immune response (i.e. in the resistance mechanisms [4]), many of which are typically switched on and massively produced after infection, versus genes involved in homeostasis (i.e. the tolerance mechanisms [4]), which are required to function in both conditions, albeit with different dynamics.

1.1.2 Post-infection transcripts tend to be longer, mainly due to the production of longer 5' UTR

We next sought to characterize the effect of the splicing differences on the length of the produced transcripts. In order to do that, we estimated an effective length measure for each gene. Briefly, for each gene in each sample, we estimated the effective transcript length weighted mean of its individual transcripts (taking into account the effect of insertions and deletions) by their expression ratios. Similarly, we extended this method to individual features within the transcript, namely the 5'UTR, 3' UTR, and the coding sequence. Then we compared the effective lengths before and after infection to obtain the number of genes who have an increased, decreased, and an unchanged effective length. Interestingly, while the effect of natural variation, namely insertions and deletions, on the coefficient of variation in feature length was most prominent in 3' UTRs, the effect of infection on the effective length of genes was strongest in 5' UTRs. To show which feature contributes to the effective length change the most, we performed a similar analysis, this time calculating the transcript length effective change differences after the removal of a certain feature. Indeed, the removal of 5'UTR length and not the predicted polypeptide or 3' UTR abolished this skew in the proportions (**Fig. 2b**). Together, these results suggest that infection-induced differences in transcript ratios affect 5' UTRs the most and favor the production of the isoforms with longer 5'UTRs.

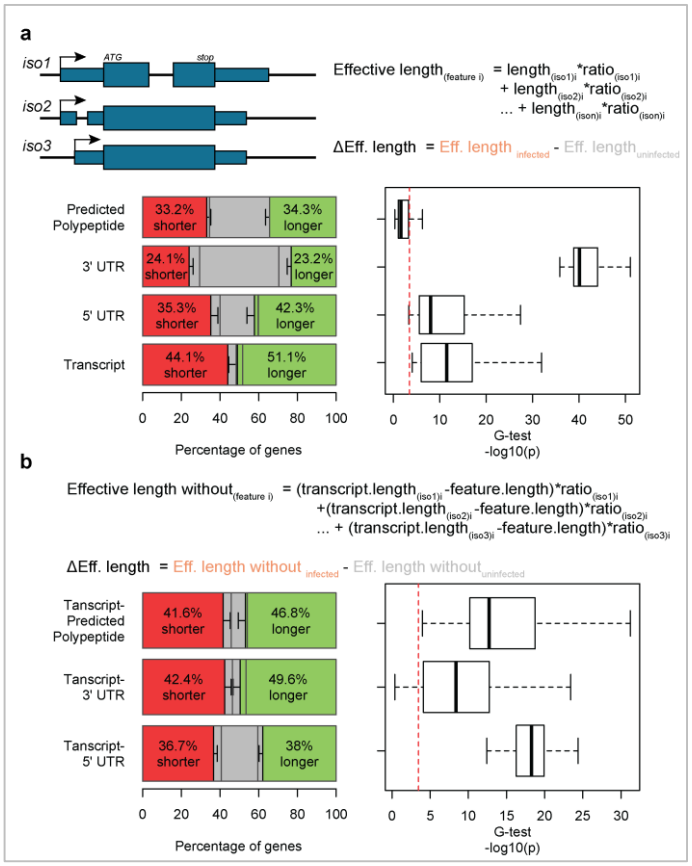


Figure 2 (a) The percentage of genes that are shorted, longer, or unchanged in effective length after infection. Error bars are SD. A null distribution was generated by performing 100 permutations of sample shuffling. The grey bars are average obtained by permutations. Repeated G-tests were used to compare length change in each line to the null distribution. Boxplots show the $-\log_{10}$(p-values) of the tests (dotted red line is the Bonferroni-corrected *p-value* threshold) **(b)** The effective length of each gene without either the predicted polypeptides, 3'UTRs, or 5'UTRs.

1.1.3 The effect of natural variation on splicing is increased after infection.

We next sought to establish a link between genetic variation and these transcript levels. To achieve this, we identified splicing quantitative trait loci (sQTLs) in the two infection states. Specifically, for each gene, we looked variations within a 10kb window, that correlate with the shift in its isoform ratios. For that, we used SQTLseekeR[1], which employs a similar statistical methodology as the one we used to detect significant differences in transcript ratios. We identified 499 and 839 naïve- and treated-specific sQTLs, and 395 sQTLs that are common to both conditions (**Fig. 3a**). Interestingly, there were around 50% more sQTLs in the treated state. Additionally, the number of genes affected by sQTLs in the treated condition is almost double that of the naïve condition (108 *vs.* 65 genes). However, there is a similar number of genes with significantly different post-infection splicing ratios that are in the naïve (13), treated (16), and shared group (20), indicating that infection response genes are not more likely to be affected by sQTLs upon infection. Together, this suggests that the effect of natural variation on splicing is more pronounced after infection, and that line-specific differences can be more readily detected in the infected state. To obtain insights into which biological processes are affected by variation in splicing ratios, we performed separate gene ontology enrichment of the three sets of genes. **Figure 3b** shows a single graphical representation of the three GO enrichment results. In the naïve state, GO terms related to transcription and splicing as well as development and nitrogen compound metabolic processes are enriched. In the treated state, other categories emerge, namely the detection of stimulus, cell adhesion, and carbohydrate metabolic processes. Both conditions share categories related to cellular homeostasis (specifically ion homeostasis) and energy derivation by oxidation of organic compounds.

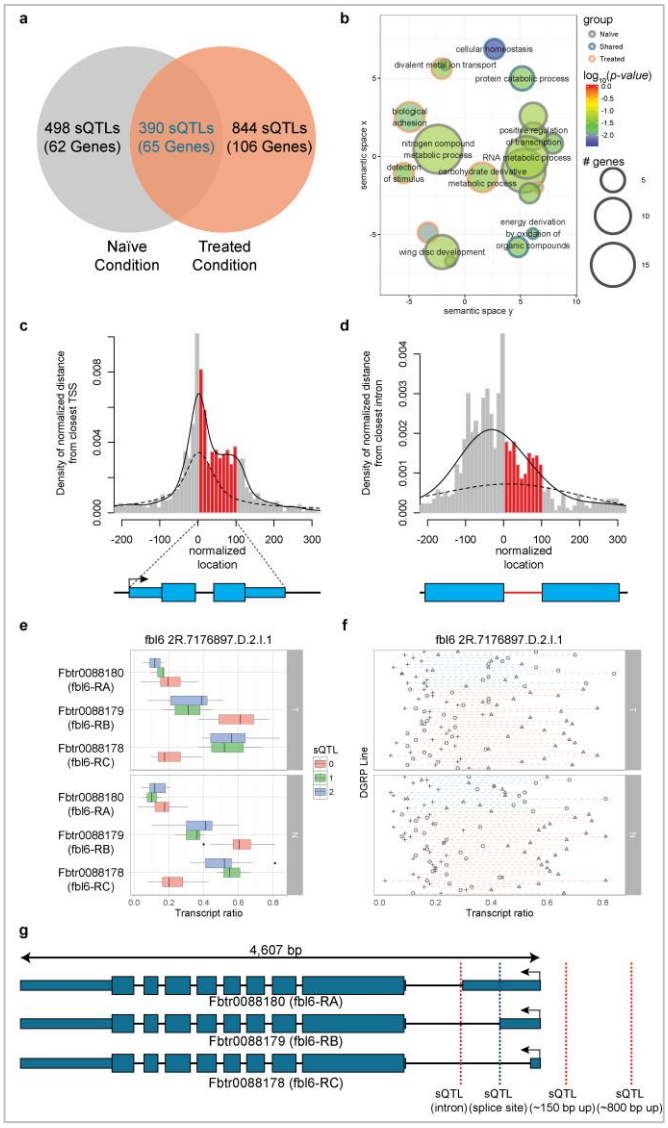


Figure 3. (**a**) Venn diagram showing the result of the *cis*-sQTL analysis (and number of associated genes) using sQTLseekeR[1] (BH adjusted p-value < 0.01, maximum difference in ratio > 0.1). **(b)** GO enrichment of the genes in the *cis*-sQTL results. **(c)** Metaplot of the pooled *cis*-sQTL results with respect to normalized gene length, and **(d)** intron length. Solid lines represent the density of *cis*-sQTLs, while dashed lines represent a random sample of 500,000 variants that are within 10kb of a gene. **(e)** Isoform ratios of a gene (*fbl6*) that has a *cis*-sQTL on one of its splice sites. The expression levels are grouped by allele of the sQTL, with 0,1,2 being reference, heterozygous, and alternate alleles. N and T are naïve and treated conditions. **(f)** The isoform ratios by DGRP line in the two conditions. The shape of the point indicates the isoform and the colour of the dashed line indicates genotype. **(g)** Gene diagram of *fbl6* showing its multiple linked *cis*-sQTLs (blue dashed line corresponds to the *cis*-sQTL in the e-f).

Next, we examined the locations of the sQTLs in relation to the gene they are associated with. We used two approaches to obtain metaplots: a gene-centric and intron-centric approach. **Fig. 3c** shows

that it is more likely to find sQTLs at 5' ends of genes, as well as within the gene bodies. In the intron-centric approach, sQTLs exhibit a sharp peak at the 5' end, with the highest peak immediately upstream of the intron (**Fig. 3d**). There are more sQTLs upstream than downstream, and the number of sQTLs drops sharply right after the intron. This data suggests that natural variation affecting splicing could be doing so by causing differences in the signals required for splicing, predominantly around the 5' splice site. One such example of sQTL is in the gene *fbl6*, which has multiple sQTLs, one of which is exactly at the 5' splice site (**Fig. 3e-g**). However, not all sQTLs could be assigned such a direct mechanism of action as this example, and some might have subtler effects by affecting exonic and intronic splicing enhancers (ESEs and ISEs). Interestingly, 70% of the sQTLs overlapped a predicted enhancer, which is 10% higher than the maximum predicted through permutations. Taken together, our sQTL data shows that we can detect effects of natural variation on splicing, even more in the infected state, and that these effects could be due to direct changes in splice sites, as well as other mechanisms predominantly at or around the splice donor site.

1.1.4 Intron retention is increased following infection across a natural population and Retained introns have exon-like characteristics

We next looked at the effect of infection on intron retention. **Fig. 4a** shows intron retention events that are significant in more than 4 lines. There is a high degree of overlap among the DGRP lines, as well as between the DGPR and w^{1118} data. suggesting that this phenomenon is not random across the genome, but affects a specific set of introns. Interestingly, a metaplot of the location of retained and spliced introns shows that the density of retained introns is very high at the 5' end of transcripts, which could at least partly explain why longer UTRs are being produced after infection (**Fig. 4c**).

In terms of length, retained introns tended to be shorter than their spliced counterparts (**Fig. 4d**). In addition to that, their GC contents were higher than those of the spliced introns, and consequently the difference in GC content between the introns and their flanking exons is lower (**Fig. 4e**). Interestingly, the retained introns also had a higher RNA polymerase II occupancy before and after infection (**Fig. 4f**). Interestingly, we found enrichment of many RNA-binding motifs (RBMs) in the spliced introns, but very few RBMs in the retained ones (**Fig. 4g**).

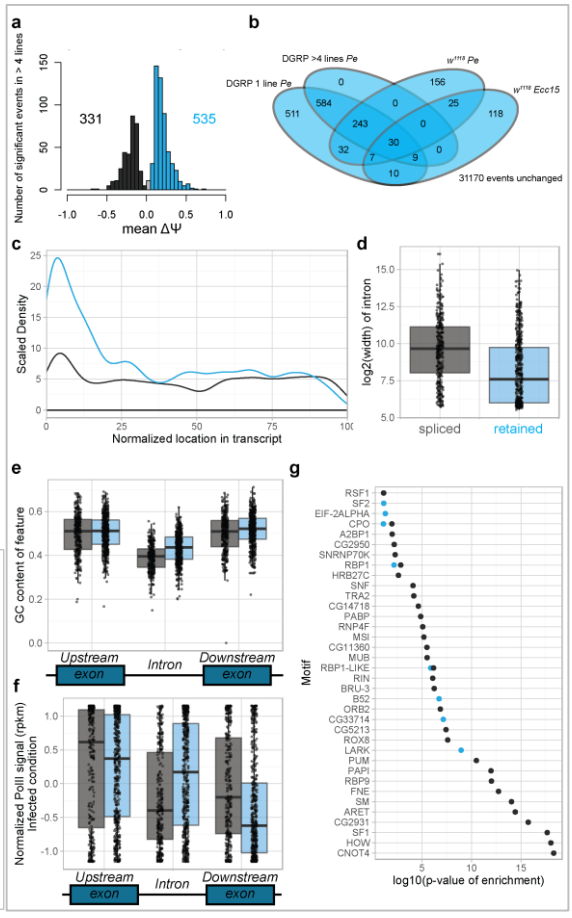


Figure 4 Blue and grey are retained and spliced out introns, respectively. **(a)** Histogram of delta PSI values of intron retention events whose PSI values are significantly different after infection in at least 4 DGRP lines. **(b)** Venn diagram of the overlap between the sets of events that are significant in 1 DGRP line, at least 4 DGRP lines, w^{1118} strain infected with *Pe*, and w^{1118} strain infected with *Ecc15.* **(c)** The density of the intron retention events along the normalized length of the gene. **(d)** Length of introns (in log2) in significant intron retention events. **(e)** GC content of those introns and their flanking exons. **(f)** Normalized PolII ChIP-seq signal of introns and their flanking exons in the *P.e.* infected state. **(g)** The enrichment of *Drosophila* RBMs.

1.2 Discussion

While there are several examples of interactions between splicing and cell stress, there have been very few genome-wide studies addressing the issue [5]. In this study, we show that infection leads to widespread and consistent splicing changes in 38 *Drosophila* strains. Many of the major differences we observe are at the level of 5' UTRs, which means that infection-induced splicing changes could have consequences on regulation, rather than strictly generating protein diversity. In times of stress, the gut might be producing transcripts coding for the same protein species, albeit with different spatial and temporal dynamics. One important aspect of the gut response to pathogenic bacteria is the general inhibition of translation, which has been previously shown to be dependent on the activation of GCN2 kinase. Activated GCN2 kinase phosphorylates the alpha subunit of the eukaryotic initiation factor (eIF2α), which leads to inhibition of translation initiation. Paradoxically, and specifically after cellular stress, some proteins like ATF4 and ATF5 rely on upstream open reading frames (uORFs) to circumvent translational inhibition [6, 7, 8]. The presence of uORFs generally inhibits the main ORF, unless they are found in specific configurations and in certain cellular conditions, like in the cases of ATF4 and ATF5 after stress-induced eiF2α phosphorylation. It is possible that the production of longer 5' UTRs, through intron retention or alternative TSS choice, could introduce upstream open reading frames (uORFs), as well as other elements, further contributing to this inhibition of translation [9, 10, 11, 12]. This also opens the possibility for the production of isoforms that are resistant to inhibition of translation or even isoforms whose translation efficiency is enhanced in stress conditions. For instance, it has been shown that the presence of uORFs in 5'UTRs could affect the recruitment of an isoform to polyribosomes, thus contributing the the translation efficiency [13]. Therefore, the poor correlations observed between transcript levels and protein abundances in other systems, could be due to the fact that splicing has been consistently ignored.

Bibliography :1.Monlong J, Calvo M, Ferreira PG, Guigó R. Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nat Commun* **5**, (2014).2.Bou Sleiman MS, Osman D, Massouras A, Hoffmann AA, Lemaitre B, Deplancke B. Genetic, molecular and physiological basis of variation in Drosophila gut immunocompetence. *Nat Commun* **6**, (2015).3.Gonzàlez-Porta M, Calvo M, Sammeth M, Guigó R. Estimation of alternative splicing variability in human populations. *Genome Research* **22**, 528-538 (2012).4.Schneider DS, Ayres JS. Two ways to survive infection: what resistance and tolerance can teach us about treating infectious diseases. *Nat Rev Immunol* **8**, 889-895 (2008).5.Shalgi R, Hurt Jessica A, Lindquist S, Burge Christopher B. Widespread Inhibition of Posttranscriptional Splicing Shapes the Cellular Transcriptome following Heat Shock. *Cell Reports* **7**, 1362-1370 (2014).6. Watatani Y*, et al.* Stress-induced Translation of ATF5 mRNA Is Regulated by the 5'-Untranslated Region. *Journal of Biological Chemistry* **283**, 2543-2553 (2008).7.Hatano M*, et al.* The 5'-untranslated region regulates ATF5 mRNA stability via nonsense-mediated mRNA decay in response to environmental stress. *FEBS Journal* **280**, 4693-4707 (2013).8.Vattem KM, Wek RC. Reinitiation involving upstream ORFs regulates ATF4 mRNA translation in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 11269-11274 (2004).9. Johnstone TG, Bazzini AA, Giraldez AJ. Upstream ORFs are prevalent translational repressors in vertebrates. *The EMBO journal* **35**, 706-723 (2016).10.Calvo SE, Pagliarini DJ, Mootha VK. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 7507-7512 (2009).11.Wethmar K. The regulatory potential of upstream open reading frames in eukaryotic gene expression. *Wiley Interdisciplinary Reviews: RNA* **5**, 765-768 (2014).12.Waern K, Snyder M. Extensive Transcript Diversity and Novel Upstream Open Reading Frame Regulation in Yeast. *G3: Genes|Genomes|Genetics* **3**, 343-352 (2013).13.Sterne-Weiler T*, et al.* Frac-seq reveals isoform-specific recruitment to polyribosomes. *Genome Research* **23**, 1615-1623 (2013).