Hide and Seq:

Novel approaches to RNA-sequencing data

with a focus on neurodegenerative disease

by

Marko Melnick

Sc.B., University of Colorado at Denver, 2011

Masters in Integrative Physiology, University of Colorado at Boulder, 2018

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirement for the degree of

Doctor of Science

Department of Integrative Physiology

2021

Committee Members

Dr. Marissa Ehringer, Ph.D., Committee Chair

Dr. Christopher Link, Ph.D.

Dr. Robin Dowell, Ph.D.

Dr. Tanya Alderete, Ph.D.

Dr. Jerry A. Stitzel, Ph.D.

Hide and Seq:

Novel approaches to RNA-sequencing data

with a focus on neurodegenerative disease

Thesis directed by Dr. Christopher Link, Ph.D. and Dr. Robin Dowell, Ph.D.

Humanity is still at the early stages of understanding the highly complex molecular mechanisms of neurodegeneration. Along with the mechanisms, how neurodegeneration starts remains largely unknown for many of these diseases. My research focuses on using RNA-sequencing data to understand these mechanisms as well as gain insights into how these diseases might form.

We first found that with heat shock, *Caenorhabditis elegans* worms have nuclear double stranded RNA (dsRNA) foci that appear similar to foci in worms with a knockout of *tdp-1,* a worm homolog of the Amyotrophic Lateral Sclerosis (ALS)-related protein TDP-43. Next, we performed RNA-sequencing of immunoprecipitated dsRNA and noticed that transcription of dsRNA is enriched downstream of some genes. We then created a novel algorithm called Dogcatcher which can capture and quantify downstream of gene regions (DoG) genome-wide. We then biologically validate a DoG identified by Dogcatcher using fluorescence imaging.

We next developed an algorithm called Mystery Miner that uses RNA-seq data to look for the presence of pathogens that might contribute to neurodegeneration. Specifically, we take often discarded non-host reads from RNA-sequencing data and identify microbes present in samples as well as quantify microbes between groups. We then apply Mystery Miner to our novel ALS dataset that consists of over 120 patients from four patient classes (three ALS related, one control). Although we find no convincing evidence for the presence of microbes or a set of microbes that might contribute to neurodegeneration, we do find and biologically confirm the presence of a novel RNA-dependent RNA polymerase-like sequence present in our dataset. Additionally, we apply Mystery Miner to other ALS related datasets in the field and perform a meta-analysis looking for any microbes that might contribute to ALS. Despite multiple types of analysis, we find no convincing evidence for the presence of a microbe or set of microbes in ALS related patient classes for any of the datasets analyzed.

Finally, I create an algorithm called MaDDoG that segments DoGs by partitioning loci at points of change (using the mean count) with subsequent quantification of segments between groups of samples. I first generate synthetic data as a proof of concept for MaDDoG, then apply it genome-wide on a real dataset. I then highlight the many applications of MaDDoG such as distinguishing between genes that have truncated or absent control DoGs, identifying regions at the end of DoGs that are likely transcriptional noise, its flexibility with regards to data with high or low variance of mean count, and look for intron retention or alternative exon usage in genic regions.

I hope any biological insights and algorithms created in my research will help patients suffering from neurodegeneration. Additionally, my algorithms can be applied outside of neurodegeneration and I hope that researchers from highly diverse fields will use my algorithms to gain further insights into many disease processes.

Acknowledgements

I feel truly lucky to be in this position. I have personally seen many graduate students suffer from sexism, mental illness, and inept professors (and this was all pre-pandemic!!!). I am eternally grateful to Dr. Chris Link (Captain Patience) for keeping his “dope slapper” close by during our many conversations and ensuring that some of my “spaghetti on the wall” ideas came to fruition. I would like to thank my bioinformatic mentor Patrick Gonzales, who has been a tremendous teacher with exceptional patience. I would also like to thank Dr. Robin Dowell-Deen, your kindness and mentorship has been the ultimate sanity check. I feel blessed to have made so many friends in the Link and Dowell lab, I will truly miss the daily exciting ideas and belly laughs so please don’t hesitate to contact me no matter the time passed. I am also sorry (not sorry) for asking as many stupid questions as possible.

Aside from academia, my family has been an unshakeable foundation of support. I would like to thank my late Grandmother for her boundless kindness and easy-going attitude, my mother for her almost gag-inducing positivity, my stepmom for her support and love, and my aunt Dr. Deborah for her “go-getter” attitude that inspires me regularly. Everyone else in my family has been equally awesome and I know that I could not have done this without my great support network.

…And we can’t forget my pup, the best buddy and worst headache in the world Kosmo!

Dedication

I would like to dedicate this thesis to my late father Rob Melnick, who always told me to "observe my world" and "get your ass in the chair and do the work!". He was my greatest supporter and never blinked when I told him my crazy plans. You are missed greatly.

"When the power of love overcomes the love of power. The world will know peace."

– Jimmy Hendrix

Chapters

List of Figures

List of Tables

I. Introduction

Summary

Sequencing technologies have drastically changed how we view biology with regards to development and disease. We are just at the start of the personalized medicine revolution and sequencing technologies will be a key factor in tailoring treatments for individuals. These technologies are becoming cheaper every day which brings up certain challenges such as how to process, store, and utilize the information in the most effective manner. Along with personalized medicine, these tools have become vital in uncovering mechanisms of disease, development, and the complex underpinnings of genetic regulation. As we develop new algorithms to process the data, we also uncover new information about the true biological processes that are occurring. The Christopher D. Link lab is primarily focused on neurodegenerative diseases [Alzheimer's disease (AD) and Amyotrophic lateral sclerosis (ALS)], and it is through this lens that we ask questions and then develop tools to answer these questions. With that in mind, this thesis focuses on two areas of research (downstream of gene transcription and bioprospecting), the application of three algorithms to these areas, and additional research of biological merit that is too small to create a stand-alone publication. As background, I will first (Chapter II) cover sequencing technology and common sequencing applications, provide a brief overview of neurodegeneration with a more in-depth look at ALS and Alzheimer's disease, and then provide background information on downstream of gene transcription and bioprospecting. Each subsequent chapter (Chapter III-V) will focus on a novel bioinformatic algorithm and the biological insights gained from its application. We will then end (Chapter VII) with how this work compares to current research and potential applications/improvements that can be pursued for future research along with additional research conducted (Chapter VII).

II. Background

Common sequencing technology and bioinformatic applications

Introduction

Every organism must use RNA or DNA in some capacity. For example, some viruses insert their DNA into the host genome and hijack a host organisms' cellular machinery for replication. For those of us unlucky enough to reproduce by other means, maintaining proper cellular dynamics remains essential for life. In general, DNA is transcribed into RNA, and RNA is translated into protein. Every cellular process is performed by proteins and certain RNAs. In general, capturing and sequencing RNAs in the cell provides expression levels that are an indirect (and often poor) measurement for protein levels. Often, the goal is to find differential gene expression (DGE) by examining how the levels of coding RNAs change after perturbations (such as drug exposure, differentiation, or altered cellular environmental conditions) or between disease conditions.

Methods for detecting DGE have changed significantly over the past five decades. Developed in the 70's, one of the earliest methods for DGE detection is the northern blot. This method consists of running RNA on a gel to reduce secondary structure, hybridizing labelled probes to the RNA of interest, and quantifying readouts from the probes[1]. In the 80's, techniques such as quantitative polymerase chain reaction (qPCR) allowed fast and accurate readouts of RNA levels without the need for setting up gels. Briefly, qPCR for quantifying RNA levels consists reverse transcribing RNA to DNA, designing primers to the region of interest, and measuring the amounts of DNA produced after a set number of amplifications (or more accurately by measuring the incorporation of fluorophores in real time)[2]. In the mid 90's, hybridization-based microarrays enabled a more complete view of the transcriptome by obtaining readouts for thousands of genes at a time. Microarrays suffered from several problems, including printing each array, reliance upon existing knowledge of genomic sequences, limited dynamic range of detection due to background or saturation of signals, and complex normalization methods when comparing DGE between experiments[3]. Despite these limitations, microarrays are heavily utilized in clinical settings because of their long history and known properties. Developed in the mid 2000's[4], RNA sequencing (RNA-seq) has become an invaluable tool for biologists studying genomic function, organismal development, and disease processes. RNA-seq has risen to become the most popular expression assay and has largely overcome the challenges of microarrays and offers additional benefits including detection of lowly-expressed genes[5], alternative splice variants[6], single nucleotide variants[7], as well as illuminating the intricacies of gene-expression regulation via non-coding RNAs[8–10] (Figure II-1)[11].

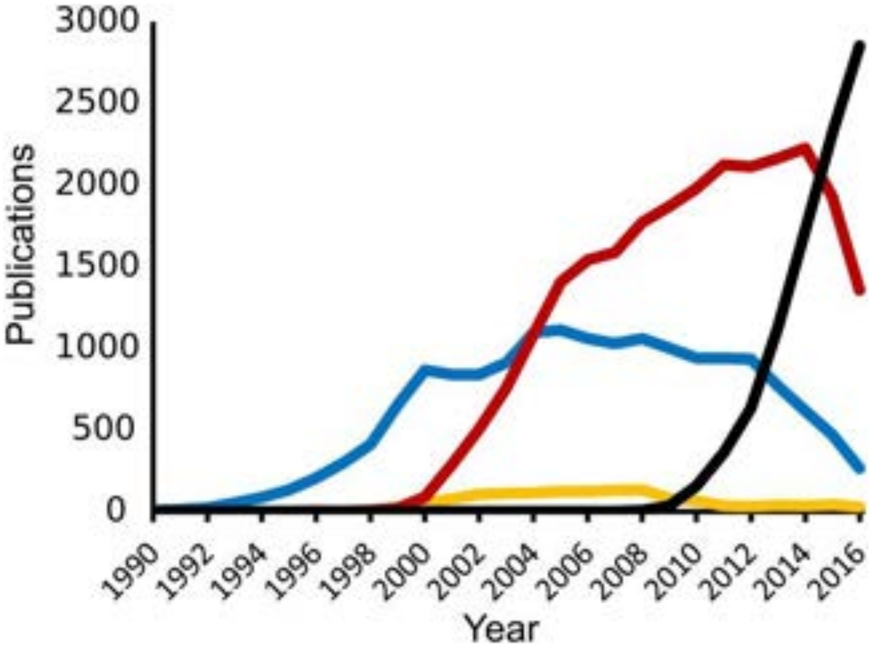

Figure II-1: Transcriptomic technology over time

Line plot of published studies over time (1990-2016) for various key words from Pubmed. RNA-seq technology (black), RNA microarray (red), expressed sequence tag (blue), cap analysis (yellow). Figure from Lowe et al., 2017.

For a typical DGE analysis, this consists of RNA extraction from cells or tissues, library enrichment or depletion (common examples include polyA selection, size filtering, or depletion of ribosomal RNAs), conversion of RNA to complementary DNA (cDNA) by reverse transcription, Sequencing adapter ligation to the ends of the cDNA fragments, amplification by Polymerase chain reaction (PCR), creating “reads” by sequencing, and subsequent computational analysis (Figure II-2 shows overview of a typical RNA-seq experiment).

When moving to computational analysis, these collective molecular decisions converge to produce an output file of sequencing reads of a particular length, numbering in the thousands to billions. Each read has a set of per position quality scores that can be inferred as the confidence in the base call. The choice of RNA isolation method, size selection, and other steps can have tremendous impact on the raw data that is produced. Care must be taken to perform these steps successfully to reduce bias and noise that affect downstream analysis, which may reveal true biological insights. After obtaining files containing reads, a swath of algorithms can be used to transform raw reads into differential expression levels of genes. Briefly, this consists of quality checking and trimming reads, mapping reads to a genome, and quantifying reads over specific genomic regions (genes, intergenic regions, etc.).

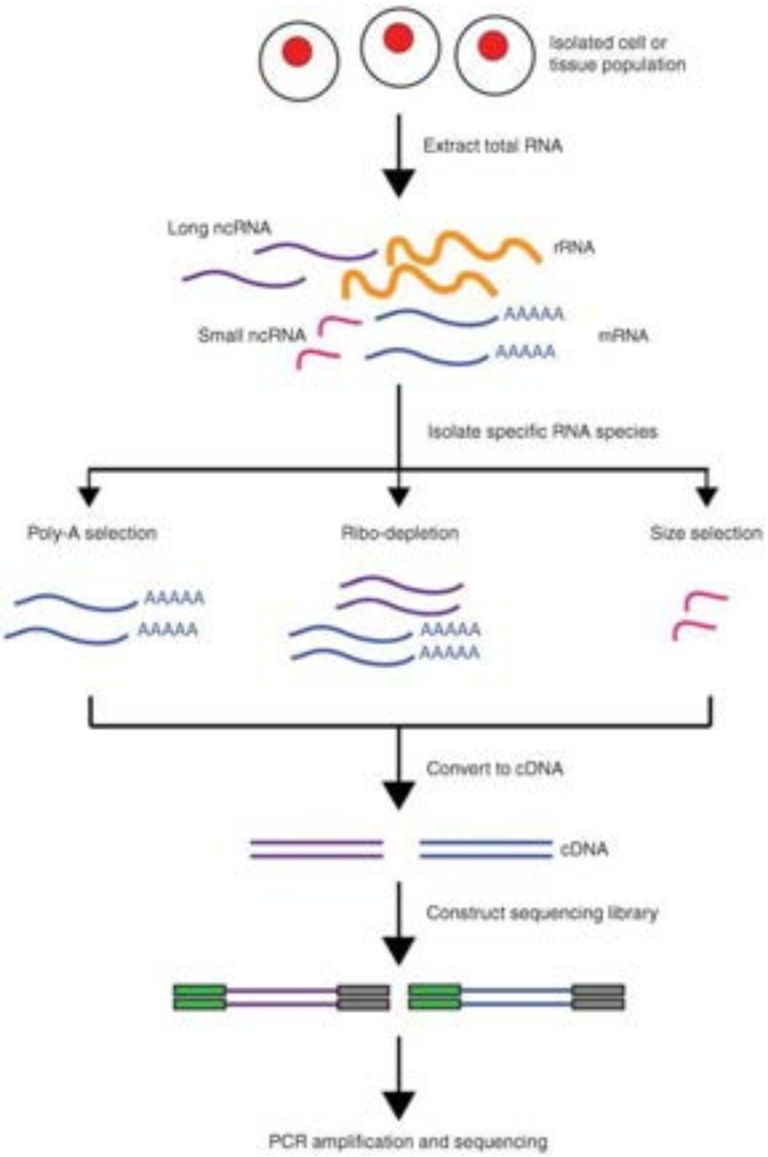

Figure II-2: Overview of a typical RNA-seq experiment

First, RNA is extracted from cells or tissues. Next, RNA is enriched by polyA selection, depleted of ribosomal RNAs, and/or size selected. RNA is then converted to complementary DNA (cDNA) by reverse transcription. Sequencing adapters are then ligated to the ends of the cDNA fragments, amplified by PCR and sequenced. (Figure from Kukurba 2015)[12].

There are well over 100 RNA-seq protocols that have been developed[13], with benefits and drawbacks to each approach. Highlights include using nascent sequencing technologies to capture RNAs being actively transcribed but lacking the capture of steady state RNAs[14], sequencing the RNA-RNA interactome using PARIS or SPLASH but receiving no information on gene expression[15], and exploring gene expression levels at the single-cell level using single-cell sequencing but suffering from high technical variation between samples[16]. Aside from these examples, there are a multitude of methods that can be employed to maximize the use of RNA seq data. Covering every method would be impossible, and instead we will focus on methods used in this thesis including measuring differential gene expression, measuring alternative isoforms and quantifying differential splicing, mapping reads to repetitive regions for differential expression of repetitive elements, identifying and quantifying regions of RNA that have been edited, assembling and quantifying reads that do not match to the host genome to look for the identity of pathogens, and identifying and quantifying regions downstream of genes due to aberrant transcription termination.

Differential gene expression

Once reads have been mapped to regions of interest (usually genes), samples are normalized by library size, and a statistical test is usually applied to test if a given gene has a significant difference in read counts for observed counts vs expected counts (due to random variation). A common technique to test DE genes is to make the assumption that read counts follow a multinomial distribution and are independently sampled from a population with a given fixed fraction of genes[17]. This multinomial was initially approximated using the Poisson distribution, which is a unique one parameter distribution where the variance equals the mean. The main issue with using the Poisson distribution is that it predicts smaller variations than what is seen in the data and does not control type-I error (probability of false discoveries) well[18]. In simpler terms, the observed variance in the count data is usually greater than the mean ("overdispersion") and the Poisson is not capable of modeling this. To address the overdispersion problem, the negative binomial (NB) distribution (equivalent to the gamma-Poisson distribution) is used and because it has a parameter for mean as well as dispersion, it is able to correctly account for overdispersion[19] (Fig. II-3). For large sample sizes it is easy to accurately estimate the dispersion parameter for each gene, in RNA-seq it is common to have sample sizes as low as two or three replicates leading to noisy estimates of the dispersion parameter. One solution to overcoming noisy estimates from low sample size is to assume genes of similar average expression strength have similar dispersion (e.g., borrowing information across genes) and shrink the dispersion to reduce the noise. For example, in DESeq2[1], this is done by first estimating gene-wise dispersion (using maximum likelihood) and fitting a smooth curve which represents the expected dispersion values. Next, gene-wise dispersion estimates are shrunken toward the values on the curve using an empirical Bayes approach which determines the amount of shrinkage based on an estimate of how close true dispersion values tend to be to the fit and the degrees of freedom (shrinkage decreases with increased sample amount)[20] (Figure II-4)[20]. In addition to shrinking dispersion, it is often useful to shrink Log Fold Changes (LFC) in count data because it has inherent heteroskedasticity (variance of LFCs depend on the mean count). To use DESeq2 as another example, it again applies an empirical Bayes procedure by first obtaining maximum-likelihood estimates (MLEs) of the LFCs and then fitting a zero-centered normal distribution to the observed distribution of MLEs over all genes. Next, it uses the distribution as a prior on the LFCs and computes Maximum a posteriori (MAP) estimates as the final estimate of the LFC. Finally, once a model is fit for each gene it is possible to run a statistical test for a significant change in expression. DESeq[1] and edgeR[2] use a variation of the Fisher exact test adopted for the NB distribution (Wald test) that can calculate an exact P-value by conditioning on the total sum to determine if the probability of observing the counts is extreme or more extreme than what is obtained[21]. The P-values are then adjusted (P-adj) for multiple hypothesis testing to control the false discovery rate usually using the procedure from Benjamini and Hochberg[22] . Lastly, a chosen threshold is used to filter genes for significance based on the P-adj value (usually 0.01 or 0.05). Despite the popularity of DESeq2 and edgeR, the choice of which algorithm to use is still open-ended as comparisons show that no method performs optimally in all circumstances[23,24]. In addition, methods are being developed that improve the negative binomial[25] or do not assume a negative binomial distribution (instead use a normal distribution[26] or non-parametric distributions[27]).

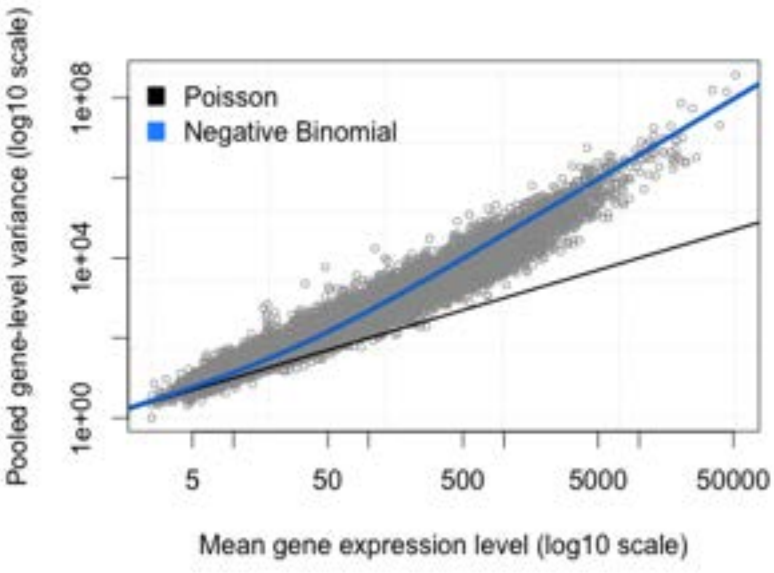

Figure II-3: Poisson vs Negative Binomial

Plot of mean expression of genes compared to pooled gene-level variance. It is clear the Negative Binomial distribution (blue line) fits the observed values better than the Poisson distribution (black line). This gives intuition to why an extra dispersion parameter in the NB is helpful to model the often-observed fact that gene-level variance is higher than the mean gene expression for larger genes ("overdispersion"). Figure from Bottomly et al., 2012[28].

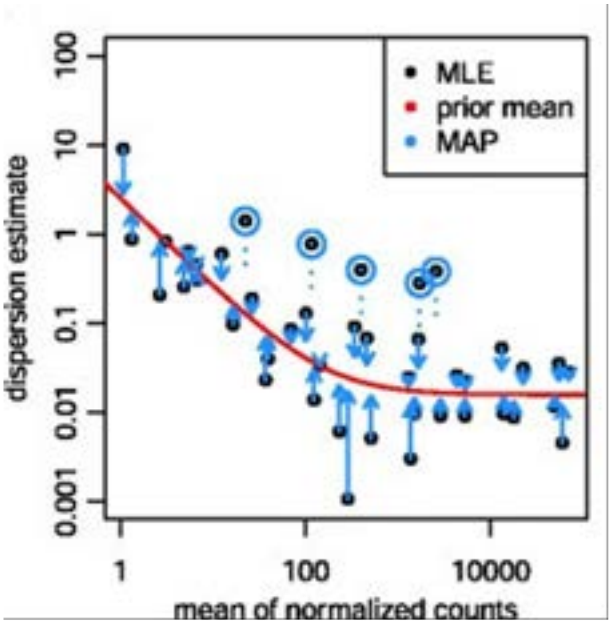

Figure II-4: Dispersion shrinking in DESeq2

Plot of dispersion estimate [i.e. parameter in the negative binomial to estimate gene variability (Y-axis)] vs mean of normalized counts (X-axis). Maximum likelihood estimates (MLE) are obtained using each gene's data (black dots). A curve is then fit to this data to capture the overall trend of the dispersion-mean dependence (red curve). This fit is then used as a prior mean for a second estimation round which gives the Maximum a posteriori (MAP) estimates of dispersion (arrowheads). Black points circled in blue are dispersion outliers and not shrunk toward the prior (red curve). Figure from Love et al., 2014[20].

Alternative Splicing

When sufficient sequencing depth is obtained (40-50 million reads per sample)[29], a less often employed analysis strategy can look for patterns of isoforms usage for a gene and alternative splicing between samples. The majority of protein coding genes in eukaryotes are transcribed into precursor mRNA (pre-mRNA) that consists of protein coding regions (exons) with non-coding regions that are often removed (introns). Introns contain three important sites for splicing, the 5' splice site (5'SS), branch point (BP), and 3' splice site (3'SS). The mechanism of splicing is highly complex but can be broken down into a two-step phosphoryl transfer mechanism (branching and exon ligation) that is catalyzed by the splicesome. The splicesome is also highly complex and consists of 5 small nuclear RNAs (snRNAs) and approximately 100 proteins that assemble denovo on the pre-mRNA whenever splicing occurs[30]. Alternative splicing (AS) is the result of different splice sites being used that results in the production of alternative mature mRNAs[31]. AS greatly expands the diversity of the transcriptome and proteome and is so ubiquitous it is estimated ~95% of multi-exonic genes are alternatively spliced in humans[32]. The five most common types of alternative splicing are alternative 5'SS selection, alternative 3'SS, intron retention, and mutually exclusive exon (Fig II-5)[33].

AS has many roles in development and impacts on disease. It has been shown to contribute to cell differentiation and lineage determination, tissue maintenance, and organ development[34]. AS patterns change in cells during neuronal lineage development and impact neurogenesis, neuronal migration, and synaptogenesis[35]. In fact, even a simple depletion of the splicing factor/RNA-binding protein polypyrimidine tract binding protein 1 (PTBP1) in fibroblasts is sufficient to induce trans-differentiation into neurons[36]. Aside from development, various mechanisms can cause splicing errors in humans leading to disease and include (but not limited to) a single-nucleotide mistake that results in a frameshift and nonsense mediated decay (NMD) of the transcript, the majority of exons (~80%) are small <200bp and can be masked by the larger intronic pool, and the fact that splicing occurs co-transcriptionally and is modulated by the rate of transcriptional elongation by RNA polymerase II (Pol II) which depends on multiple regulatory machineries acting correctly in concert[37]. While correct AS is important for neuronal development, aberrant splicing can cause neuropathological disorders such as frontotemporal dementia with parkinsonism linked to chromosome 17 (FTDP-17) which occurs when mutations affect splicing in the 10th exon of microtubule-associated binding protein tau (MAPT)[38]. Aside from neurodegeneration, mutations in the splicesome can lead to retinal degenerative disorders and hereditary blindness[39], splice site mutations may cause a loss of dystrophin function leading to Duchenne muscular dystrophy, and there is a large body of literature concerning the interplay of AS with immunotherapy and cancer[33]. After decades of research identifying splicing abnormalities that cause an array of diseases, researchers are beginning to see the fruits of their labor as clinical trials are underway that validate therapeutic approaches using small molecules and antisense oligonucleotides[40,41].

As mentioned above, understanding AS in the context of development and disease processes is vital for modern biology and numerous methods exist to categorize and quantify AS. RT-qPCR is the standard quantitative technique used to detect differences in known splice isoforms but suffers from low throughput, large amounts of RNA needed, and false positives due to remaining genomic DNA leftover in the RNA extraction[42]. Other methods include RETF[43], Ligation-based PCR[44], Raman multiplexing[45], and P-RCA[46], but these methods all fall short compared to RNA-seq due to low throughput or inability to detect splice changes genome wide. Algorithms for quantifying AS fall into a few different categories including transcript reconstruction methods[47,48], light-weight pseudoalignment heuristics that quantify transcript abundances[49,50], analysis of differential usage of sub-genic features (exons)[26,51], and algorithms that leverage junction information to infer annotated and novel splicing[52–54]. Choosing which AS algorithm to use remains an open question as a recent cross-comparison using vertebrate data found strengths and weaknesses to each approach[29].

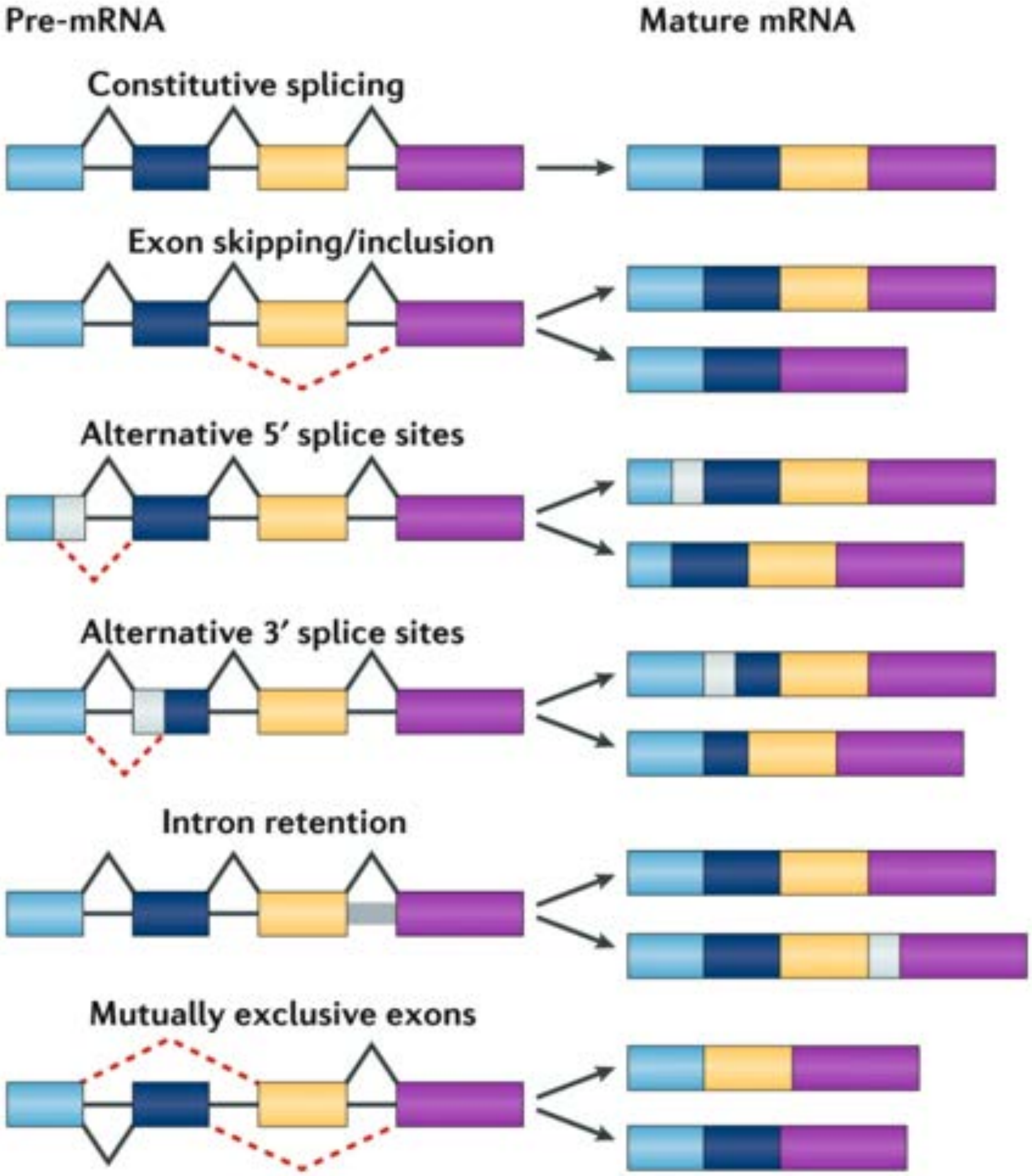

Figure II-5: Common splicing patterns

This figure depicts the five most common splicing patterns. Exons are colored boxes and straight horizontal lines are introns. On the left is pre-mRNA and triangular lines show regions where alternative splicing is possible. On the right is the mature (spliced) mRNA. Figure from Frankiw et al., 2019[33].

Repetitive elements

The next often overlooked analysis consists of using reads that map to multiple locations in the genome, i.e. repetitive regions. These "multi-mapped" reads make up a large proportion (5 to 40%) of total mapped reads and are often ignored upon completion of DGE analysis[55]. These reads were once thought of as "junk", but with development of better algorithms, these reads are now (although still infrequently) being used to answer questions relevant to development and disease. Upon the initial sequencing of the human genome one of the startling discoveries was that ~55% of the genome was made up of repetitive DNA sequences, this figure has since been updated to around two-thirds of the genome[56,57].

These repetitive DNA sequences called repetitive elements (RE) can be characterized into five categories (Fig. II-6). The four minor categories account for ~10% of the genome and include simple sequence repeats, segmental duplications, tandem repeats and satellite DNA sequences, and processed pseudogenes. The last major category are transposable elements (TEs) and account for ~45% of genomic DNA[58]. TEs can be divided based on methods of replication into Class I TEs called retrotransposable elements (RTEs) that use a "copy and paste" mechanism with an RNA intermediate[59] and Class II TEs called DNA transposons that use a "cut and paste" mechanism and use a DNA intermediate[60]. RTEs can be further subdivided into Long-Terminal Repeats (LTR) and non-LTRs. LTRs are also known as endogenous retroviruses (ERVs), comprise 8% of the human genome[56], and are thought to be exogenous viruses that integrated into the host germline in the distant past[61]. Non-LTRs are the only elements believed capable of retrotransposition and consist of Long Interspersed Nuclear Elements (LINEs) or Short Interspersed Nuclear Elements (SINEs)[62].

Although the four minor categories have implications in disease and development[63–65], a large portion of research into REs is focused on TEs due to their increased genomic amount and implications in a variety of diseases[66]. Retrotransposon activity is largely driven by a single family of protein-coding (autonomous) LINE elements known as long interspersed element 1 (L1). Out of the estimate 500,000 copies of L1 in the human genome, the majority are immobile (unable to transpose) leaving a small amount of L1s that are mobile. Of these mobile L1s, it has been shown in a cell culture assay that 84% of all retrotransposition activity is driven by 6 “hot” L1s[67]. L1s have been implicated in autosomal dominant, autosomal recessive, and X-linked genetic disorders, cancer, and autoimmunity[66]. In addition to mobilizing its own RNA, the L1 retrotransposase can mobilize a variety of other RNAs including a SINE called Alu which is the most abundant (by copy number) RTE in the entire genome[56]. Alus have been shown to contribute to neurodegeneration[68], obesity[69], and mental retardation[70]. Finally, human endogenous retroviruses (HERVs) have been implicated in systemic lupus erythematosus[71] and multiple sclerosis (MS)[72].

The most reliable and informative method to investigate REs genome-wide is with high throughput sequencing due in part to how similar a RE analysis is to a DGE analysis. The main issue to overcome is ambiguity in read assignment due to reads that map to more than one location (multi-mapping reads). The first attempts to use multi-mapping reads calculated read coverage across the genome by assigning reads proportionally to all matching regions[73] or assigning them probabilistically to locations based on the local genomic context[74]. The most popular recent algorithms for quantifying REs include salmonTE[75] which essentially uses Salmon[50] and is built for speed, TEtranscripts[76] which assigns reads to genes and REs based on RE hierarchy, and Repenrich[58] which builds indexes of closely related REs and assigns reads to subfamilies. Of these algorithms, salmonTE claims to perform better than the others, although it did no comparison with the updated Repenrich2 algorithm[7].

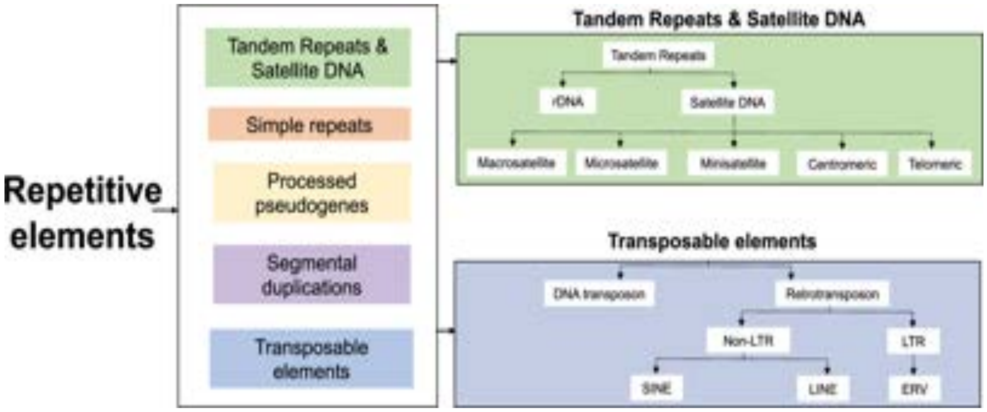

Figure II-6: Repetitive elements in humans

Diagram of repetitive elements in humans. Transposable elements can be divided into DNA transposons or retrotransposons according to the mechanism of transposition (DNA or RNA intermediate). Retrotransposons are the most abundant class in the human genome and consist of Long terminal repeats (LTR) and non-LTRs. LTRs include endogenous retroviruses (ERVs) and non-LTRs can be divided into SINE (Alu elements) and LINEs (LINE1 elements). Figure from Billingsley et al., 2019[77].

RNA Editing

Besides controlling the sequence of mRNA by alternative splicing, another method of regulation occurs by a post-transcriptional modification, which edits the RNA transcript sequence (RNA editing). RNA editing analysis remains uncommon, although discoveries with implications in the fields of immunology as well as using RNA editing in directed therapies have sparked recent interest[78]. In addition, running an RNA expression analysis can be complicated due to constantly changing databases of editing sites (if applicable) and underdeveloped methods for quantifying RNA editing in specific regions and genome wide.

RNA editing was initially discovered in the late 80's from work in the trypanosome[79] and was shortly found to also occur in mammals in a tissue specific manner[80]. This study showed C-to-U editing of apolipoprotein B mRNA with the cytidine deaminase APOBEC1 which transformed the glutamine codon (CAA) to a stop codon (UAA) producing a truncated protein. Aside from C-to-U editing, by far the highest amount of editing is A-to-I conversions[81]. One classic example of A-to-I (Adenosine to Inosine) editing occurs in neurons at position 602 of the glutamate receptor 2 (GluR2) mRNA. Inosine (I) is recognized as guanosine (G) by the ribosome and transforms the CAG codon for glutamine (Q) to CIG (or CGG) for arginine (R). This change makes the GluR2 receptor impermeable to calcium and neutralizes the diffusion of divalent cations[82] (Fig. II-7)[83].

The adenosine deaminase acting on RNA (ADAR) gene family catalyzes A-to-I editing. Of the three members in the mammalian genome, ADAR1 and ADAR2 are expressed at high levels but only ADAR1 contributes to the majority of editing activity[84]. ADARs preferentially edit long double-stranded RNA (dsRNA) duplexes which primarily form from pairs of inverted copies of genomic retro-elements (mainly Alus) in introns or untranslated regions (UTRs) of a transcript.

In fact, it is estimated that >99% of the millions of editing sites in the human genome are located inside Alu repeats[85]. Endogenous ('self') dsRNAs resemble structures found in viruses and can trigger the innate immune system by activating melanoma differentiation-associated protein 5 (MDA5)[86,87]. When activated, MDA5 interacts with the mitochondrial antiviral signaling protein (MAVS) which triggers the interferon response leading to cell damage in the host[88]. ADAR1 editing of these endogenous dsRNAs can disrupt the base pairing and prevent inappropriate activation of MDA5[88]. Additionally, it has been shown in human cells that a knock-out of ADAR1 causes hyperactivation of the dsRNA sensor protein kinase RNA-activated (PKR or EIF2AK2) resulting in translational shut down and cell death[89]. Editing is closely tied to the immune system and ADAR1 mutations have been linked to autoimmune diseases like Aicardi-Goutieres syndrome and systemic lupus erythematosus[90,91]. In addition to immune disorders, RNA editing has been implicated in neurodegenerative diseases[92–94], psychological disorders[95–97], and a multitude of cancers[98].

Profiling RNA editing is still a very active area of research and a variety of bioinformatic programs have been released[99–102]. Initial efforts would map RNA-seq data to a reference genome and/or transcriptome to first identify single nucleotide variants (SNVs) to filter out single nucleotide polymorphisms (SNPs) which required a complete or partial SNP database[103–105]. One algorithm called SNP-free RNA editing Identification Toolkit (SPRINT)[102], utilizes SNV duplets (two consecutive SNV with the same type of variation. i.e., A-to-G and another A-to-G) to distinguish RNA editing sites (RES) from SNPs. SPRINT is also able to distinguish hyper-editing sites (extensive A-to-I RES in a genomic region) and has improved performance in identification of RES to other algorithms[103,106,107]. One recent review of SNP-database dependent algorithms suggested that most algorithms (GIREMI[103], JACUSA[108], RES-Scanner[109], REDItools[101]) perform similarly and that the most reliable way to obtain high-quality RES is to fine tune input parameters[100]. All of the algorithms mentioned above are only for calling edits, the methods of statistically calling differential editing remains a work in progress. A few issues make this problem difficult such as low coverage of reads in repeat elements where the majority of editing occurs, editing happening in some samples with no editing in others, and the aggressive multiple hypothesis correction when considering a large amount of editing locations. In general, it has been suggested to avoid testing differential editing with the normal distribution (avoid Student's t-Test), and many studies have applied the non-parametric Mann-Whitney U test (non-parametric version of Student's t-Test)[110,111]. Briefly, the Mann-Whitney U test is a rank test that compares two populations and tests the null hypothesis that the probability is 50% that a randomly drawn member of the first population will exceed a member of the second population.

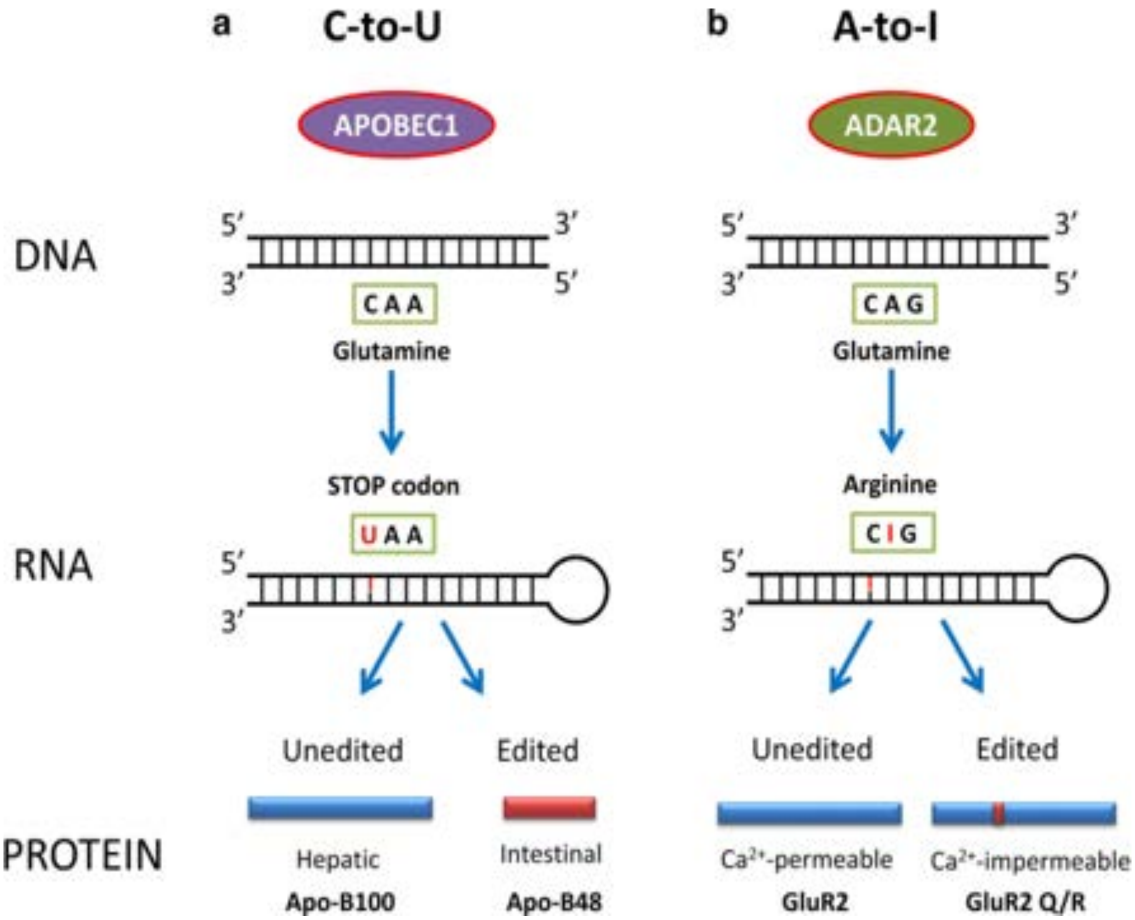

Figure II-7: C-to-U and A-to-I editing

Figure depicting two common cytidine and adenosine deaminases. A. APOBEC1 editing occurs in the gut and is involved in production of apolipoprotein B. C-to-U editing transforms the glutamate to a stop codon and produces a truncated protein. B. A-to-I editing in neurons of the glutamate receptor 2 (GluR2) mRNA. Inosine (I) is recognized as guanosine (G) by the ribosome and transforms the CAG codon for glutamine (Q) to CIG (or CGG) for arginine (R). This change makes the GluR2 receptor impermeable to calcium and neutralizes the diffusion of divalent cations. Figure from Christofi et al., 2019[83].

Conclusion

Given how costly and difficult it is to design and execute an RNA seq study, it behooves researchers to make the most of their data and run these often-overlooked algorithms that may give true biological insights. Additionally, with the ever-increasing number of studies and torrential amount of public RNA seq data being produced each year, these alternative analyses provide opportunities to any entrepreneurial researcher who are conducting meta-analysis studies. Importantly, even when these less commonly utilized RNA-seq analysis techniques are deployed (splicing, repeats, editing), gaps remain in our understanding of RNA-seq data. Specifically, analysis of RNA-seq data largely depends on existing annotation and therefore discovery of novel signals outside of annotated regions (for example, downstream of genes) is rarely considered. Likewise, the reads that fail to map to the genome are typically discarded, yet they may contain evidence of infectious agents. My work focuses on these two, largely ignored, aspects of RNA-seq analysis – specifically in the context of neurodegenerative disease. As such, next I will provide a brief overview of neurodegeneration along with the impact of sequencing on the study of the two most relevant neurodegenerative diseases of my research.

Neurodegeneration

Introduction

Neurodegenerative diseases (ND) cause progressive loss of cognitive and/or motor function and are an ever-increasing burden on patients, families, and communities due to increases in life expectancy worldwide[112]. Cognitive defects occur in AD, frontotemporal dementia (FTD), dementia with Lewy bodies (LBD). Motor system defects occur in ALS, Huntington's disease (HD), and Parkinson's disease (PD). These diseases show diverse clinical manifestations and loss of specific neurons and synapses in distinct brain regions (Fig. II-8)[113]. Despite these differences, many neurodegenerative diseases share common mechanisms or features such as aggregation of proteins[114] or RNA[115], neuroinflammation[116], alternative splicing[117], somatic mutations[118] and variants/loci in genome wide association studies (GWAS)[119–126].

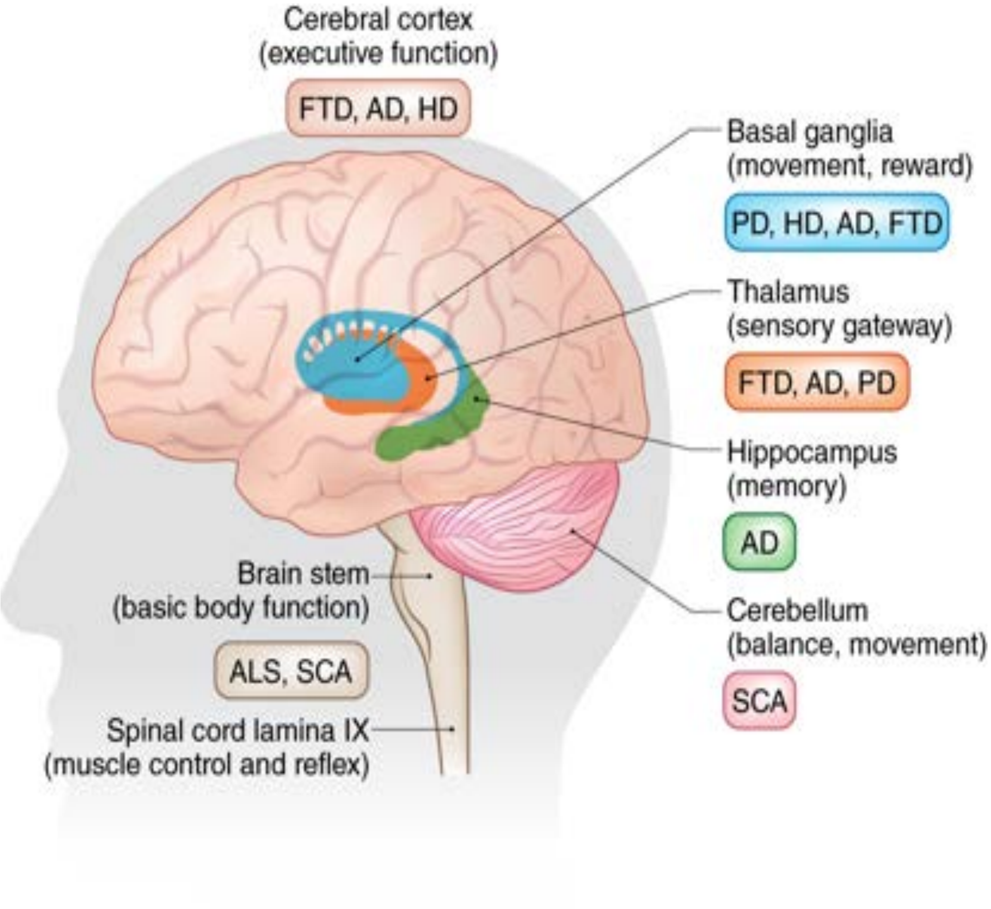

Figure II-8: Brain regions affected in neurodegenerative diseases

Diagram of brain regions where loss of specific neurons and synapses contribute to pathology. There are both distinct and overlapping regions affected from multiple diseases. Cognitive defects occur in Alzheimer’s disease (AD), frontotemporal dementia (FTD), dementia with Lewy bodies (LBD). Motor system defects occur in Amyotrophic lateral sclerosis (ALS), Huntington’s disease (HD), and Parkinson’s disease (PD). Figure from Gan et al., 2018[113].

Connecting genotype to phenotype to clinical diagnosis in neurodegenerative conditions can be a highly complex/convoluted process. Initial efforts in the 80's and 90's traced the genetic cause of HD to a CAG trinucleotide repeat in the gene HTT[127]. It was soon found that mutations in multiple genes can lead to similar clinical entities. For example, mutations in amyloid precursor protein (APP), presenilin 1 (PS1), and PS2[128] in early onset AD, mutations in molecular weight 43 kDA (TDP-43), Superoxide dismutase one (SOD-1), and fused-in sarcoma (FUS) in ALS, and mutations in microtubule-associated protein tau (MAPT) and GRN in frontotemporal dementia (FTD)[129]. The inverse is also true as different mutations in a single gene can lead to multiple diseases. A prime example is the most common genetic cause of ALS and FTD coming from hexanucleotide repeat expansion mutations in intron 1 of C9orf72[130,131] and sometimes it is useful to view these diseases on a spectrum (Fig. III-9)[132].

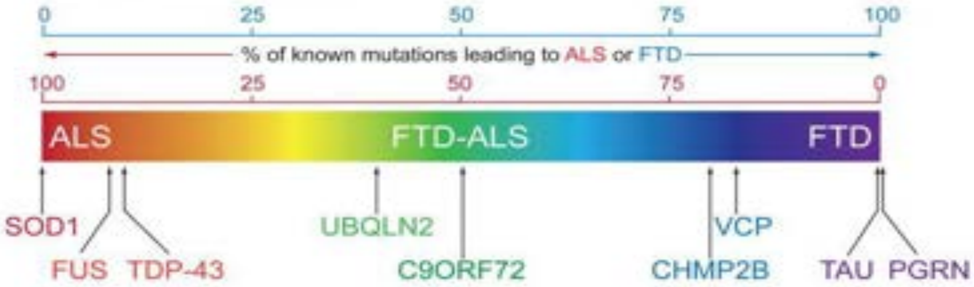

Figure II-9: FTD-ALS spectrum

ALS and FTD are two sides of a broad neurodegenerative disorder with overlapping clinical symptoms. Percentages of known mutations that give rise to ALS (red) or FTD (purple) are plotted. Figure from Ling et al., 2013[132].

Alzheimer's disease

Alzheimer's disease was first characterized by Alois Alzheimer in 1907 when he used the then-new silver staining histopathological technique to examine the brain of one of his patients and found neuritic plaques, neurofibrillary tangles, and amyloid angiopathy (Fig. II-10)[133]. The next decades were spent refining psychological studies of AD and was considered a major public health issue when epidemiological data found it was the 4th fourth leading cause of death in the elderly[134]. During the 1990's and 2000's major genetic risk factors were discovered for familial AD including APP, PS1, and PS2 (as mentioned above) although familial AD only accounts for 1 to 2% of all cases[133]. Later, In sporadic late-onset AD the type ε4 allele of the gene for apolipoprotein E (APOE4) was identified as a common risk factor[135]. APOE4 was found in 50 to 60% of patients with AD and confers a three-fold risk for one copy and eight-fold risk for two copies of the allele[136]. Currently, a definitive diagnosis of AD requires post-mortem evaluation of brain tissue, but strides are being made in diagnosing living patients using cerebrospinal fluid (CSF) and positron emission tomography (PET) biomarkers along with clinical criteria[137]. The causative factors of AD are still not understood and the search for a cure to this disease is frustrating as there are dramatically high clinical failure rates and with no drugs moving past phase 3 in 2020[138].

One of the hallmark mechanisms of putative AD pathology comes from AD-associated amyloid plaques which are mainly composed of the β-amyloid (Aβ) protein. Aβ is produced from protease cleavage of APP. Interestingly, the specific physiological function of APP remains a mystery[139]. APP cleavage by γ-secretase can generate varying chain lengths of Aβ including Aβ40 and Aβ42 which are the main Aβ peptides of the brain[140]. Although Aβ40 is more abundant in the brain, the aggregation prone Aβ42 is the main component of amyloid plaques and has been shown to be neurotoxic[141]. During pathogenesis, Aβ monomers make Aβ aggregates of various unstable oligomers that then form into insoluble fibrillar assemblies of β-strand repeats[142]. Mounting evidence suggests that it is the oligomers that cause toxicity and potential mechanisms of action include membrane damage via pore formation, acting as a pathogenic ligand to various receptors, and oligomers activating the free radical process leading to oxidative stress[143].

In addition to Aβ, human tau (encoded by the MAPT gene) oligomers and fibrils are the main components of neurofibrillary tangles (NFT) and have been shown to induce neurotoxicity[144]. Tau can act as a microtubule binding component that promotes polymerization and stability of microtubules, localizes to axons to promote axonal transport, and is highly expressed in neurons in mammals[145,146]. Tau has six different isoforms in the human brain and is subjected to a multitude of post-translational modifications (PTMs) including phosphorylation, lysine-based PTM, and glycosylation which have been implicated in AD[147–150]. To date it is unknown if NFTs cause neurotoxicity, although there is evidence of neurotoxicity for tau with aberrant PTMs, soluble tau oligomers, and tau fibrils[144]. Interestingly tau is being pursued as a biomarker for AD and can be detected using exosome isolation in CSF and blood in AD patients[151].

How tau and Aβ might cause synergistic pathogenesis is heavily debated. No known mutations in MAPT have been associated with AD suggesting that Aβ aggregation occurs upstream[152,153]. A positive feedback loop has also been suggested as tau can directly bind Aβ to promote Aβ aggregation and Aβ may trigger the transition of tau from a normal to toxic state[154–156]. Some suggest the key may involve immune activation, as it has been shown that Aβ is able to activate several innate immune pathways, incite inflammatory responses, and release cytokines such as interleukin-1β[157]. One example of this from a mouse model of Alzheimer's describes increased tau pathology from upregulation of the interleukin-1β pathway[158]. Although tau is mainly found in neurons, tau deposits have been found in astrocytes of AD brains[159]. Accumulation of tau in astrocytes can alter astrocyte function which induces neuronal degeneration through increasing blood brain barrier (BBB) breakdown and expression of low-molecular weight heat shock proteins[160,161]. A working model suggests that tau pathogenesis is triggered by Aβ in AD, pathogenic tau and Aβ contribute to inflammation, and reactive glial cells (astrocytes) further incite the inflammatory response with subsequent neurodegeneration[162].

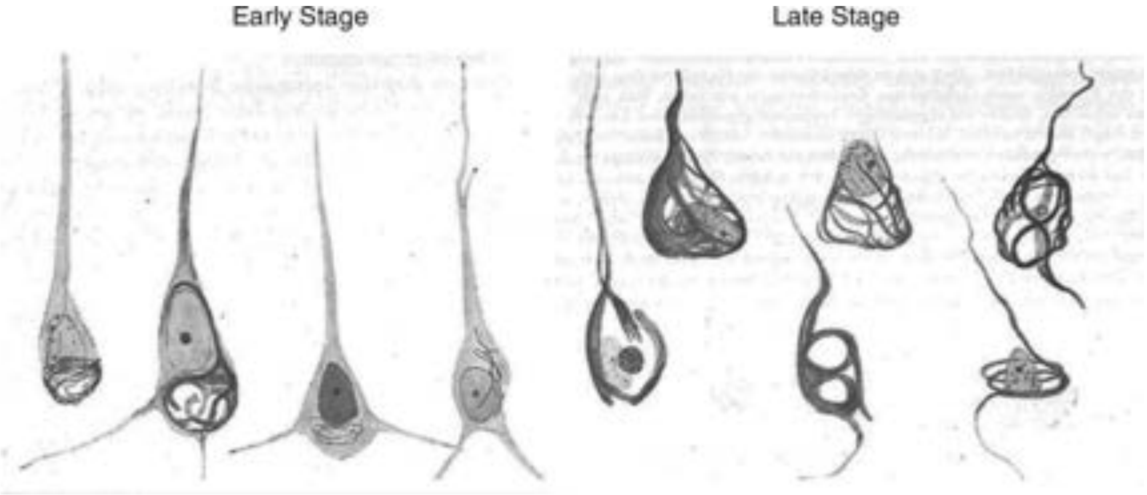

Figure II-10: Alois Alzheimer sketches

Sketches from Alois Alzheimer of histopathological preparations from early and late - stage neurofibrillary tangle pathology from his 1911 paper. Figure from Bondi et al., 2017[133].

Amyotrophic Lateral Sclerosis

Amyotrophic Lateral sclerosis is a fatal motor neuron disease (MND) that was first described in the 1860's by French neurologist Jean-Martin Charcot and is characterized by progressive loss of upper and lower motor neurons at the spinal and bulbar level[163,164]. Roughly 10% of ALS patients have a family history that suggests an autosomal dominant inheritance, and this is classified as familial ALS (fALS), with the remaining 90% of patients classified with sporadic ALS (sALS) because they have no affected family members[165]. Studies on ALS primarily come from European populations and within these populations four genes (TDP-43,FUS,SOD1,C9orf72) account for 70% of fALS[166]. Of these four genes, C9orf72 accounts for up to 30-50% of cases in fALS and 7% of sALS (in all populations)[165]. The cause of ALS is still unknown and has been attributed to genetic risk factors (as mentioned above) as well as risk factors related to lifestyle and environment such as smoking, type 2 diabetes mellitus, exposure to heavy metals, and athletic status[167].

The pathophysiology of ALS is not well understood but hallmarks of the disease include aggregation of ubiquitylated proteins in motor neurons. These aggregations are primarily made up of TDP-43 and ~97% of cases show TDP-43 proteinopathy which is characterized by depletion of TDP-43 in the nucleus and formation of cytoplasmic aggregates leading to both a “loss of function” and “gain of function” model[166]. Mounting evidence suggests pathogenesis from inclusions comes from the TDP-43 C-terminal domain (CTD) being highly disordered and prion-like, carrying most of the ALS-associated TARDP mutations, and CTD fragments being highly cytotoxic and found in ALS-affected brains[168,169]. TDP-43 has also been implicated in AD where aggregates have been found co-localized with NFTs, and a study using mice found TDP-43 and Aβ oligomers were able to cross-seed each other into toxic species[170,171]. TDP-43 is vital for RNA processing and has roles in transcription, translation, mRNA transport, mRNA stabilization, stress granule formation, and alternative splicing[172] (Fig. II-11)[172]. In fact, using genome-wide RNA immunoprecipitation (CLIP-seq) it was found that TDP-43 associates with up to 30% of the entire transcriptome[173]. With deletion of tdp-1, the C. elegans ortholog of TDP-43, it was found to increase the accumulation of dsRNA (in the transcriptome and detected in nuclear foci) as well as enhance the frequency of A-to-I editing in worms. Additionally, TDP-43 was found to limit the formation of dsRNA in human cells[174]. The mechanisms of how exactly TDP-43 limits dsRNA accumulation is still being worked out, but evidence in worms show that deletion of tdp-1 dramatically alters the relocalization of heterochromatin-like protein 2 (HPL-2), the C. elegans ortholog of heterochromatin protein 1 (HP1). Indeed, TDP-1 and HPL-2 were both co-immunoprecipitated in worms and it is thought that TDP-1 recruits HPL-2 co-transcriptionally to repress repetitive element transcription (a major source of dsRNA)[175]. Due to TDP-43's complexity of functions, current challenges remain identifying disease-relevant RNA interactions, pathways impacted from aberrant TDP-43, and identifying the genetic or environmental mechanisms that lead to TDP-43 pathology.

One other source of pathology strongly implicated in ALS is the hexanucleotide repeat expansion (HRE) of C9orf72 (C9ALS). This GGGGCC (G_4C_2) repeat expansion forms RNA with highly stable parallel G-quadruplex structures (G4 RNA) that can aggregate in nuclear foci[176]. How neurodegeneration occurs from HRE of C9orf72 is not well understood but putative mechanisms include loss of C9ORF72 from aborted transcription, bi-directionally transcribed

RNAs from the HREs, repeat-associated non-ATG (RAN) translation of dipeptide repeat proteins (DPRs) from repeat RNAs of the HRE, and loss of function of RNA-binding proteins via sequestration in G4 RNA-containing nuclear foci[176,177]. Interestingly, TDP-43 has been shown to bind G4 RNA structure, transport these RNAs to neurites for local translation, and become sequestered by G4 RNA which may lead to TDP-43 loss of function toxicity[178]. Additionally, it has been shown that in the frontal cortex of patients with C9ALS there is a significant increase in REs (majority are LTRs, LINEs, DNA elements) compared to controls or sALS. It is suggested that these RE differences are the result of dipeptide-repeats inducing chromatin changes, but more research needs to be conducted[179]. Nucleocytoplasmic transport (NCT) defects have also been identified as critical contributors to C9ALS. Current studies suggests that G4 RNAs can cause defects by binding to a key regulator of NCT RanGAP1[180]. How DPRs cause toxicity is more controversial, putative mechanisms include DPRs directly binding nucleoporins inhibiting nuclear pores and DPRs inducing stress granules which sequester NCT transport factors[177]. Nevertheless, it is clear more research needs to be conducted to elucidate the mechanisms of toxicity in C9ALS.

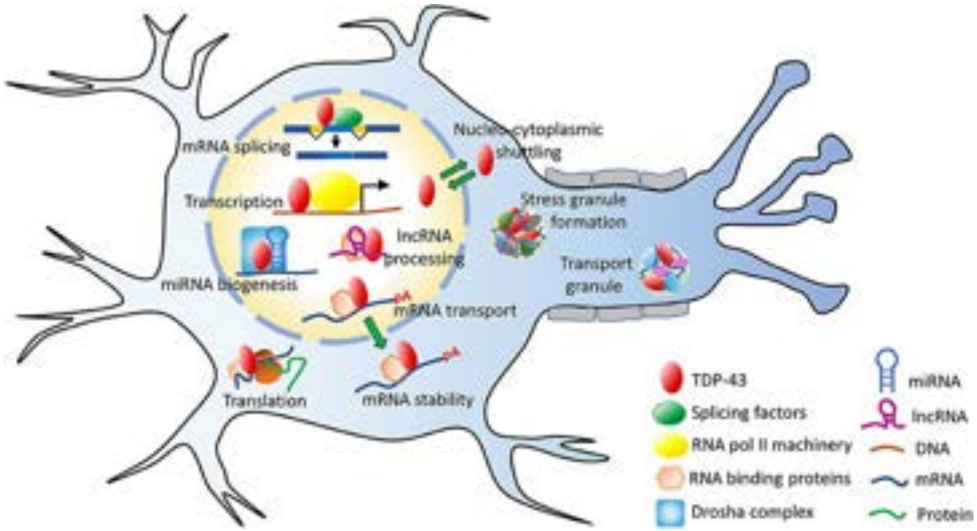

Figure II-11: Roles of TDP-43 in RNA processing

TDP-43 performs several roles in RNA processing. It is located primarily in the nucleus and helps with transcription, splicing, mRNA stability, and miRNA and long non-coding RNA processing. It also acts as a nucleo-cytoplasmic shuttle and in the cytoplasm will assist with stress granule formation and translation. Figure from Prasad et al., 2019[172].

Next generation sequencing in the clinic

As medicine becomes more personalized the dizzying amounts of data generated by next generation sequencing can be overwhelming for researchers and clinicians who are incorporating NGS into patient care. It is often difficult to interpret NGS data and a recent survey of 204 neurologists indicated that 59% of them thought results should have demonstrated clinical utility for diagnosis, prognosis, or treatment. Furthermore, 69% of these neurologists thought results should be limited to genes relevant to a patient’s specific medical condition to limit incidental findings (unrelated information)[181]. Nevertheless, the personalized medicine revolution is upon us and is being utilized for neurodegenerative diseases[182–185]. In a clinical setting, tests for neurodegeneration using NGS are focused on variant calling and can be broken down into three approaches including whole genome sequencing [(WGS), i.e., DNA-seq, RNA-seq], whole-exome sequencing [(WES), all exons], and targeted-panel sequencing [(TPS), select regions and genes][186]. The choice of which approach to use often depends on the disease, costs, and amount of information needed for a diagnosis or treatment. Currently, TPS is the best choice for many neurodegenerative diseases (or WES for rare diseases), but is limited compared to WGS because it must rely on known panel locations and cannot (or rarely) recover information about non-coding RNAs, alternative splicing, or repetitive elements that are becoming increasingly relevant to neurodegeneration (as mentioned in previous sections)[187].

Conclusion

Neurodegeneration will be a looming issue as we face an aging population. The economic impact on nations and burden on health care providers will only increase in the coming years. The mechanisms of how people develop neurodegeneration are largely unknown and what we do know involves highly complex dynamic systems with overlapping pathologies in diseases that are likely the result of compounding failures of the body. Nevertheless, discovering how these diseases arise and finding treatments will be essential to fixing these issues as we move forward. Now that the reader has a better idea of neurodegeneration, we will return to our focus on overlooked analyses of RNA seq data with two largely unknown analyses that have potential implications for neurodegeneration. These uncommon analyses consist of utilizing non-genomic reads to look for aberrant transcription downstream of genes and utilizing non-host reads to search for pathogenic infection.

Downstream of gene transcription

Introduction

This section entails identifying and quantifying non-genic reads in downstream of gene (DoG) regions due to aberrant transcription which is directly relevant to chapter III of this thesis. This area of research is relatively recent (~2015) and thus the underlying mechanisms of how DoGs are made, why certain genes have DoGs, what function they might have, and their implications for disease are largely unknown. This section will start with a brief overview of the process of transcription, discuss the known stressors that cause DoGs, the putative functions of DoG formation, and the putative mechanisms of DoG function.

Transcription termination background

Transcription termination is vital to genomic regulation. Most genes in eukaryotes are transcribed by RNA polymerase II (Pol II) and it is the carboxyl-terminal domain (CTD) of Pol II that interacts with cleavage and polyadenylation factors (CPSF) to generate the polyA tail. Currently, two models are proposed for how pre-mRNA polyA sites (PAS) are involved in transcription termination. The allosteric model describes Pol II sensing the PAS during elongation, which causes a conformational change in the Pol II active site that leads to Pol II release. The other model is called the torpedo model and proposes that the exonuclease Xrn2 is recruited to the PAS and triggers Pol II release when it degrades the downstream transcript and catches up to elongating Pol II[188]. Recent evidence suggests a unified model where dephosphorylation of the elongation factor SPT5 in PAS cleavage slows and commits Pol II to the template strand allowing easy termination by XRN2 (Fig. II-12)[189].

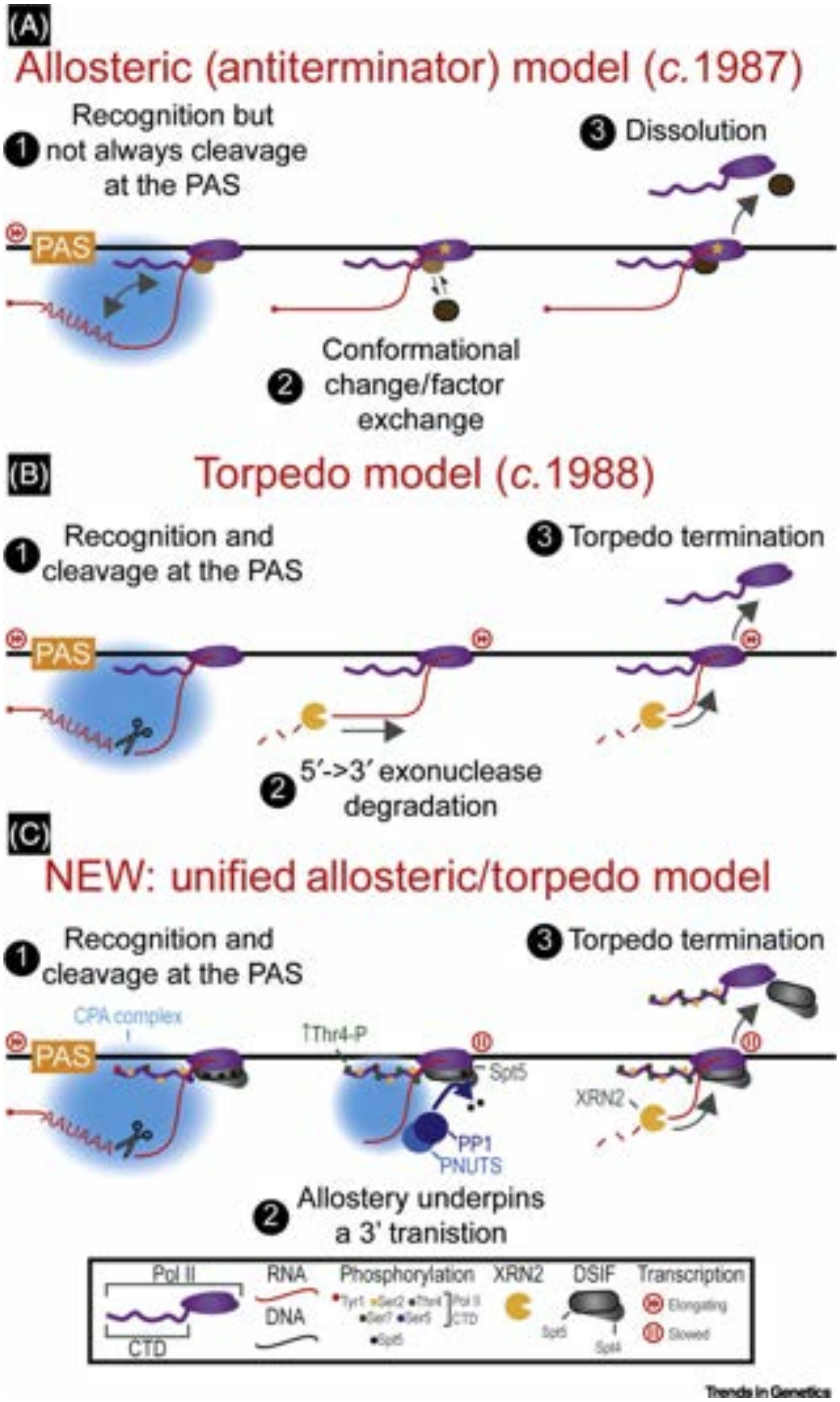
(A)
Allosteric (antiterminator) model (c.1987)
1 Recognition but not always cleavage at the PAS
3 Dissolution
PAS
AAUAAA
2 Conformational change/factor exchange
(B)
Torpedo model (c.1988)
1 Recognition and cleavage at the PAS
3 Torpedo termination
PAS
AAUAAA
2 5′->3′ exonuclease degradation
(C)
NEW: unified allosteric/torpedo model
1 Recognition and cleavage at the PAS
3 Torpedo termination
CPA complex
PAS
AAUAAA
Thr4-P
Spt5
PP1
PNUTS
XRN2
2 Allostery underpins a 3' tranistion
Pol II
CTD
RNA
DNA
Phosphorylation
Tyr1 Ser2 Thr4 Ser7 Ser5 Pol II CTD
Spt5
XRN2
DSIF
Spt5
Spt4
Transcription
Elongating
Slowed
Trends in Genetics

Figure II-12: Models of transcription termination

Two current models for transcription termination. (A) Allosteric model proposes that poly adenylation sites (PAS) can induce a conformational change and induce cleavage factors which help Pol II disassociate from the DNA and induce termination. (B) In the Torpedo model, PAS cleavage induces degradation of Pol II associated RNA that is degraded by the 5’->3’ exonuclease XRN2. (C) Unified model of the Allosteric and Torpedo model. Pol II is slowed by dephosphorylation of SPT5 (via PNUTS/PP1) which acts as an allosteric switch. The switch ensures Pol II stays on the same strand allowing XRN2 to catch up when degrading the polymerase-associated PAS cleavage transcript. Figure from Eaton et al., 2020[189].

Stressors induce downstream of gene transcription

Failure of proper transcription termination is induced by a variety of stressors including hyperosmotic stress, heat shock, oxidative stress, viral infections, and cancer[190–194]. Aberrant termination can lead to read through transcription past the annotated termination sites of genes. These downstream-of-gene (DoG) containing transcripts are long noncoding RNAs that are made minutes after cellular stress, are continuous with upstream mRNAs, and found at ~10% of protein coding genes (Fig. II-13)[190]. Importantly, DoG counts and DoG lengths show low correlation of counts with the upstream gene indicating that gene expression alone cannot explain readthrough induction[191]. DoGs are believed to remain chromatin bound in the nucleus and remain at the site of transcription[195]. The extent to which DoGs are polyadenylated is unknown but a small amount of evidence suggests that they can be both poly and non-polyadenylated[190]. Currently, it is unknown if neurodegenerative diseases induce DoG formation and a meta-analysis to look for DoGs across diseases is much needed in the field.

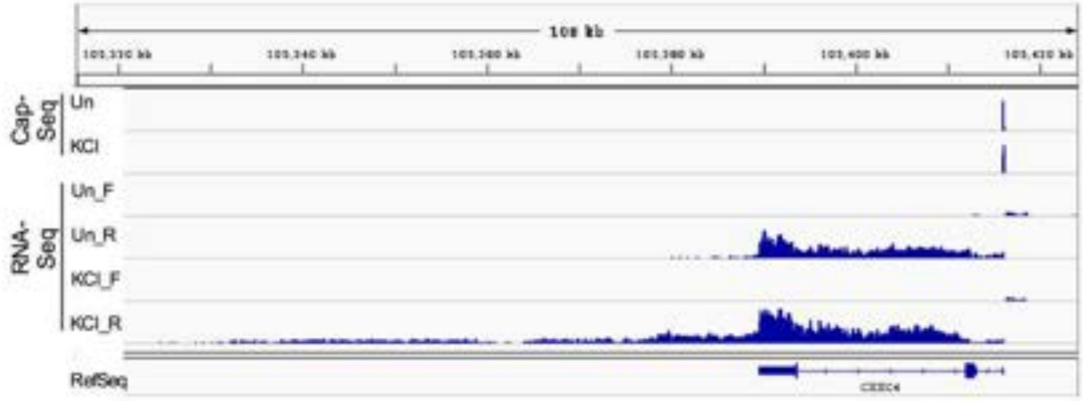

Figure II-13: Downstream of gene induction with osmotic stress

Integrated Genome Viewer (IGV) screenshot of histogram of reads over the CXXC4 gene (5' to 3', right to left). The first two rows show CapSeq which identifies transcription start (TSS) regions and shows no alternative TSS sites are induced with KCL treatment. The bottom four rows show the forward and reverse strand of RNA-seq data and show downstream of gene transcription with KCL treatment. Figure from Vilborg et al., 2015[190].

Mechanisms of downstream of gene induction

The mechanisms of DoG induction upon cellular stress are currently unknown and remain an active area of research. Initial experiments have shown that DoG regions are depleted of PAS suggesting some genes might be "primed" to induce DoGs[190]. How DoGs differ between various stressors is also unknown. It has been found that calcium signaling through IP3R is partially responsible for DoG induction in osmotic stress. With heat shock, it was found that genes with increased heat shock factor 1 (HSF1) binding at promoters showed greater heat shock-induced readthrough and that upon HSF1 depletion via siRNA there is reduced DoG transcription. A recent study of Herpes simplex 1 (HSV-1), found that termination defects are caused by an HSV-1 induced immediate early protein ICP27 inducing a null mRNA 3' processing complex via its interactions with the CPSF complex which blocks/delays cleavage. Interestingly, IPC27 does this by binding to GC-rich sequences upstream of the PAS and this is what delineates correct transcription of viral genes compared to aberrant termination found in host genes[196]. Indeed, there is significant overlap of DoGs from various stressors indicating that there might be a shared mechanism of induction for various stressors[191,197]. Additionally, it has been shown that knocking down CPSF73 which is a subunit of the CPA complex leads to a partial induction of DoGs[190]. Recently, it has been shown that depletion of a catalytic subunit of the integrator complex was able to induce readthrough at hundreds of loci and that these DoGs partially overlap with DoGs induced from osmotic stress[198].

Putative downstream of gene functions

Along with how DoGs are made, what they might do is also unknown. DoGs are reported to stay in the nucleus and one hypothesis suggests they act as nuclear scaffolding to maintain nuclear integrity, but this hypothesis is difficult to confirm due to the thousands of DoGs that might act together as scaffolding[190]. Another potential mechanism involves regulation of gene expression by read through on the opposite strand inducing antisense RNAs. Natural antisense RNAs are pervasive in humans and up to 40% of genes show natural antisense transcription[199]. Antisense transcription can regulate genes by a variety of mechanisms, read through on the opposite strand could form dsRNA leading to editing and degradation via ADAR, collisions of Pol II leads to transcriptional interference and reduced transcription of convergent genes, antisense RNAs can act as masks that protect sense transcripts from being degraded, and antisense transcripts can create dsRNA that is then degraded by the RNA interference pathway (Fig. II-14)[200]**.** One study suggests that read through induced antisense RNAs from convergent genes leads to transcriptional repression and may act as a major mechanism for cells undergoing senescence[194]. Importantly, dsRNA can activate the innate immune system (via Protein Kinase R) and it is possible that DoGs form dsRNA which activate the immune system. Indeed, it has been shown knockout of TDP-43 induces dsRNA accumulation, HSV-1 infection limits the accumulation of dsRNA (mainly viral dsRNA)[201], but if these dsRNAs are formed from DoGs, what effects they have on the immune system, and their unique similarities or differences across stressors remains a mystery.

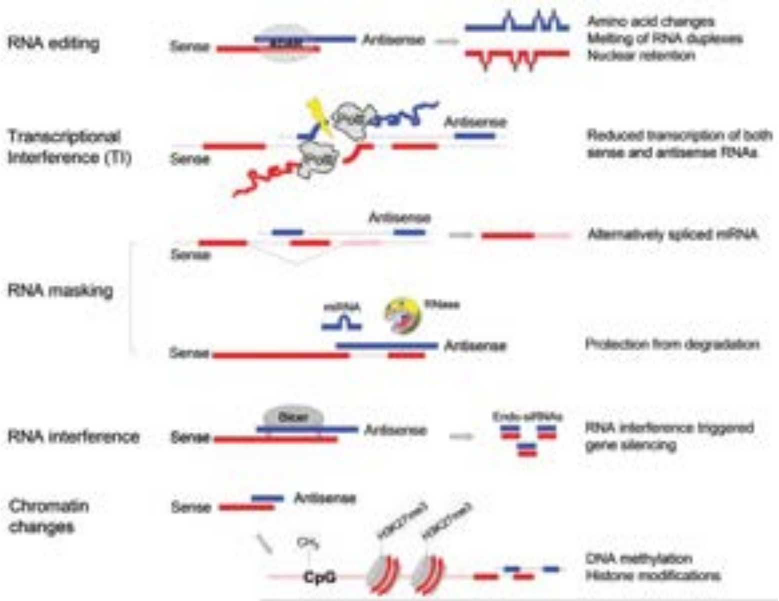

Figure II-14: Cellular mechanisms of natural antisense transcripts

Natural antisense transcripts occur from converging genes on the opposite strands. Antisense transcription can regulate genes by a variety of mechanisms. Transcripts on opposite strands can form dsRNA leading to editing and degradation via Adenosine deaminase acting on RNA (ADAR). Collisions of Pol II leads to transcriptional interference and reduced transcription of convergent genes. Antisense RNAs can act as masks that protect sense transcripts from being degraded, and antisense transcripts can create dsRNA that is then degraded by the RNA interference pathway. Chromatin changes induced by natural antisense transcripts can silenced transcription through repressive chromatin marks such as (H3K9 and H3K27). Figure from Wight et al., 2015[200].

Conclusion

The field of downstream of gene transcription analysis is relatively new and there are many discoveries to be made. Currently, only a few stressors such as osmotic stress, heat shock, viral infection, and a small number of diseases and/or cell types have been tested for DoGs, it will be prudent for researchers of the future to find out if there are additional conditions that induce DoGs. It will also be necessary for researchers to discover the unique or shared mechanisms of DoG induction across stressors, and what function (if any) DoGs have for maintaining proper cellular dynamics. From an algorithmic standpoint, methods of utilizing non-host reads to distinguish regions of true read-through transcription (compared to sampling noise) and accurately quantifying regions of differential DoG transcription between conditions remain areas that are ripe for innovation.

Biomes and bioprospecting

Introduction

Studies of the microbiome have exploded in the last two decades. Currently, what constitutes a healthy microbiome is still unknown and the implications on health and disease are still being discovered. With the ever-decreasing cost of sequencing and improvements in algorithms and techniques for identifying and quantifying biomes, this area has many discoveries to be made. This section will review the general history of microbial detection and quantification, how whole genome sequencing is used, methods of utilizing non-host reads, and pathogens implicated in neurodegenerative diseases.

Brief history of microbiomes

The first study of human-associated microbiota dates back to the 1670s when Antoine van Leeuwenhoek described five different types of bacteria and distinguished differences between body locations and diseases[202]. Since then, initial back of the envelope estimates calculated a 100:1 ratio of bacteria cells to human cells which has since been revised to 1:1 at around 3.0 X 10^{13} cells[203]. Despite the similar amounts of bacterial and human cells, the sheer diversity of the human microbiome is staggering as there is ~45 million non-redundant bacterial genes (compared to ~20,000 in humans) and new strains being discovered each year (150,000 in 2019 alone)[204]. Researching human-associated microbes and the microbiome is still a relatively new field but is vital to understanding health and disease. Although what constitutes a healthy biome is still unknown, disease associated bacteria and aberrant biomes have been implicated in immune dysfunction, asthma, behavioral disorders, cancer, and neurodegenerative disorders[205,206]. Milestones in microbiota research include culturing of anaerobes[207] and development of germ-free mice models[208] in the 1940s, fecal transplants to treat Clostridium difficile infection in the 1950s[209], and development of the Human Microbiome Project in the late 2000s[210].

The identification and classification of microorganisms is a developing field with a variety of methods. Typical laboratory methods for bacterial detection take a long time to process samples, require specialized equipment and employees, and are not available in many countries[211]. Culturing bacteria is one of the primary techniques needed to obtain enough sample for detection but relies on having the correct temperature, inoculation of the specimen, incubation, and culture medium which can be specific for each species[212]. Direct observation of microorganisms through microscopy (with or without staining) is the easiest method and is frequently done in a clinical setting to distinguish gram-positive from gram-negative bacteria[213]. A popular biochemical method for detection is an enzyme-linked immunosorbent assay (ELISA) that detects surface epitopes of bacteria. In addition, Electron microscopy has been used in many first-identifications of novel bacteria or viruses elucidating cell architecture and proteins at the molecular level but is low-throughput[214]. However, the methods mentioned above rely on culturing but not all bacteria can currently be cultured in a laboratory setting. These methods are still in use but are being replaced or supplanted by methods that focus on nucleic acids such as PCR and sequencing technologies.

Sequencing technologies for microbiome analysis

Early efforts using nucleic acids to categorize bacteria were focused on rRNA because it was initially hypothesized and confirmed that rRNAs evolve 100-fold more slowly than protein coding regions in bacteria[215]. It was found that slow and fast evolving regions exist in rRNA, and this allowed researchers to track differences in highly conserved regions to identify phylogenetic relationships that span long evolutionary time (slow) and fast evolving regions suitable for distinguishing bacteria within microbiota (fast). This comparison was initially done with oligonucleotide catalogues but was largely replaced with PCR targeting fast and slow evolving regions. As mentioned in the section on DGE, the development of quantitative real time PCR (qPCR)[216] was a tremendous boon to researchers studying gene expression and this was quickly utilized in microbial detection. The main advantages qPCR provides are fast and high-throughput detection of DNA sequences, low susceptibility to cross-contamination after initial amplification, and a wide dynamic range[217]. One of the main disadvantages of qPCR is that it cannot distinguish between live and dead cells and for this reason it is not widely used in pathogen detection in food because it will amplify DNA left over from non-viable pathogens[218]. Although qPCR is still in use, sequencing technologies greatly improved the throughput and detection capabilities for distinguishing variety and amounts of organisms in a biome.

By far the most widely used system for bacterial community detection is 16S rRNA gene sequencing (Figure II-15)[219]. Typically, nine hypervariable regions (V1-V9) are targeted in the 16s rRNA gene and V1-V3 or V3-5 are sequenced and clustered into bins called Operational Taxonomy Units (OTUs) based on a sequence similarity threshold (usually around 97% to delineate species)[220]. From an OTU cluster a single sequence is selected as representative and all other sequences in the OTU are annotated identically. OTUs are then classified using homology-based approaches using a reference database or prediction based approaches[221]. Some issues arise when using OTUs, in that the 97% species threshold is a rough estimate and is sometimes false. For instance, two different species can share 99% sequence identity (falsely classified as the same species), and a single strain could have multiple copies of the 16S rRNA gene that differ by 5% in certain regions (falsely classified as multiple species)[222,223]. Recently, studies analyzing amplicon sequence variants (ASVs) have shown single-nucleotide differences over sequence regions and improved sensitivity and specificity compared to OTUs[224]. Overall, 16S rRNA sequencing is commonly used because it is cost effective and has established databases and pipelines. Despite these benefits, 16S rRNA shows reduced detection capability of diversity, inability to distinguish between species (usually), and inability to identify all of the kingdoms of life compared to whole genome sequencing (WGS) approaches[225].

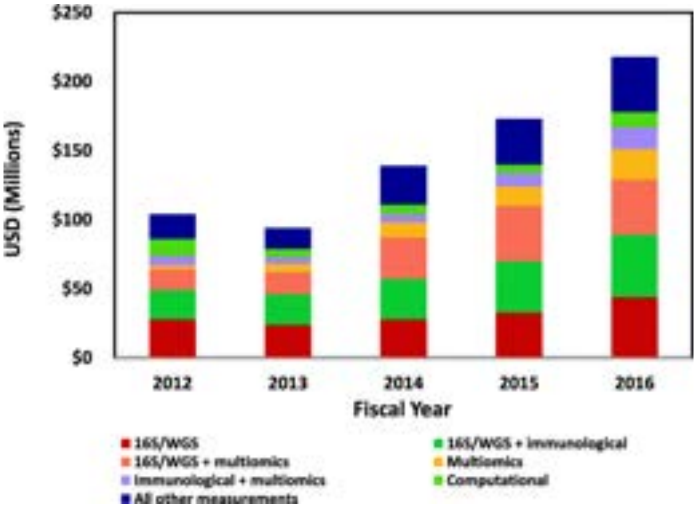

Figure II-15: Types of microbiome studies

The amount in millions spent at the National Institute of Health Human Microbiome Project from 2012-2016. The data consists of six main types and include 16S rRNA analysis, 16S combined with immunological analyses, 16S with multiomic (transcriptomic, proteomic, metabolomic), and multiomic analysis alone. Figure from NIH Human Microbiome Analysis team 2019[219]

With constantly decreasing costs, the field is moving to WGS for metagenomics . Along with the ability to detect of all kingdoms of life, WGS allows researchers to obtain information about the function of genes, genomic structure, and identify novel genes within a community. WGS has been used in diverse fields such as greenhouse gas emission studies to differentiate rumen microbial communities in cows to identify high and low methane emitting cattle phenotypes and one striking example showed that certain global ocean microbial communities share >73% identity with the human gut microbiome[226,227]. An WGS study is typically done by assembling sequenced data into contiguous sequences (contigs and scaffolds) and then identifying and/or quantifying organisms and genes (along with putative proteins) based on the contiguous sequences. Assembly approaches rely on the incorrect assumption that highly similar sequences originate from the same position in the genome and that similar sequences can be "stitched" together. In reality, assembling a genome depends on the length of reads and lengths of repeats being assembled and difficulties can be categorized as trivial (repeats are shorter than read length), computationally intractable (the correct answer requires an exponential number of arrangements of reads), and impossible (insufficient information in a read to identify the correct sequence reconstruction)[228]. Assembly using a "de novo" strategy (no reference genome) is the standard for biome studies and frequently employ the de Bruijn graph-based approach that constructs a graph by reading consecutive kmers (sequences of k bases long) in a read. Benchmarking of RNA-seq assembly algorithms for memory usage, usability, assembly across various organisms, and assembly with viral contamination showed similar top levels of performance from Trinity[229], Trans-ABySS[230], and SPAdes (rna setting)[231].

One of the most untapped avenues in microbial detection is called "bioprospecting" which uses non-biome related RNA-seq studies to search for microbes. In recent studies using the human genome, it was found that 9-20% of reads do not map to the host/reference (human) genome and these "junk" (host-unmapped) reads are utilized in bioprospecting[232,233]. Bioprospecting algorithms are a fairly recent development but can be divided into algorithms that assemble non-host reads into contigs and rely on databases of microbial genomes for quantification[234–237] and those that use non-host contigs and align reads back to the contigs for quantification[238]. With the large amounts of RNA-seq data being produced each year, it is likely that bioprospecting will grow in popularity and might become a standard quality control check for confounding variables in a study or illuminate potential biomarkers of disease.

Pathogens in neurodegenerative diseases

There has been a long history with variable success in the search for pathogens that contribute to neurodegenerative diseases such as Alzheimer's disease (AD)[239–241], Parkinson's disease (PD)[242–244], multiple sclerosis (MS)[245] and ALS[246–250]. In addition, growing interest in the microbiota has shown an increasing role for gut microbiota influencing brain function and vice versa, in a bidirectional communication pathway termed the "microbiota-gut-brain" axis[251]. How the microbiota-gut-brain axis functions is complex as it relates to immune, neural, endocrine, and metabolic pathways, but will likely contribute tremendously to our understanding and prevention of neurodegeneration[252].

One of the diseases with the most amount of evidence for pathogenicity from infection is MS. The two main viruses implicated in MS are Epstein-Barr virus (EBV) and Human Herpes Virus-6 (HHV-6)[253]. Active replication of HHV-6 in MS correlates with a polymorphism in MHC2TA that codes for the Major histocompatibility (MHC) class II transactivator and it is thought that this lowers MHC class II molecules allowing the virus to escape immune detection[254]. In fact, a recent study showed a positive association of MS and higher levels of antibodies to the viral subtype HHV-6A and that this was detected up to 10 years before MS symptoms occurred[255]. For EBV in MS, EBV infected B-cells are the primary putative source of pathogenesis and show increased antigen presentation and have been found in actively demyelinating lesions[256]. In addition to viruses, multiple studies have looked at bacteria or a synergism with bacteria and viruses in MS but all results have been inconclusive[257].

Recently, research on the role of the microbiome in Alzheimer's disease has seen a deluge of studies in the areas of nutrition, sedentary lifestyle, sleep deprivation, and the underlying mechanisms of potential pathogenesis[258]. Intestinal bacteria can excrete functional amyloid peptides and Gram-negative endotoxin/lipopolysaccharide (LPS). One bacterial amyloid peptide called *curli,* is secreted by multiple bacteria in the gut and contains subunits similar to Aβ and is recognized by toll-like receptor (TLR) which can activate cytokines that cross the blood-brain barrier and contribute to neuroinflammation and neurodegeneration[259,260]. The jury is still out on if bacterial amyloids themselves are virulent or not[261], and care must be taken as treatment with inhibitors of bacterial amyloids was shown to aggravate aggregation diseases[262]. LPS has also been implicated in AD and has been found to co-localize with amyloid plaques in AD brains, activate the immune response via TLR2 receptors, and contribute to amyloid plaque formation, myelin injury, and tau phosphorylation[263].

Diverse pathogens have been reported in the blood, cerebrospinal fluid (CSF) and central nervous system (CNS) from ALS patients. For example, bacteria that have been detected include *Cutibacterium acnes, Corynebacterium sp, Fusobacterium nucleatum, Lawsonella clevelandesis,* and *Streptococcus thermophilus* in CSF[264]*,* and mycoplasma in blood[265]. Fungi, including *Candida famata*, *Candida albicans*, *Candida parapsilosis*, *Candida glabrata*, and *Penicillium notatum,* have been detected in CSF, while *Malassezia globosa*, *Cryptococcus neoformans*[249], and *Candida albicans* have been found in various regions of the CNS[249,266,267]. The search for viruses that contribute to ALS pathology is much more extensive and includes studies on herpes virus[247,268], enterovirus[247,269–272], human immunodeficiency virus (HIV)[273,274], and human endogenous retrovirus (HERV-K)[275–277]. Importantly, multiple studies using immunohistochemistry have shown an increased load of various pathogens in ALS samples compared to controls in multiple tissues suggesting these pathogens are present and cannot be simply attributed to contamination[247,249,264,266,267]. Ultimately, the presence of ALS dysbiosis is unresolved and remains an active area of investigation with evidence for[278–282] and against[283] it.

Conclusion

Microbiome research will continue to have a pivotal role in the study of health and disease. As sequencing technologies become cheaper and algorithms improve, we will continue to understand more of the underlying mechanisms that these organisms play in our lives (for good or bad). Additionally, large amounts of unused data (i.e., non-host reads) may be utilized to identify and quantify pathogenic or beneficial organisms relevant to disease. It is clear that further understanding of biomes will be crucial for delineating the causative role of pathogens in disease (if they have one) or as potential biomarkers of disease.

III. Dogcatcher: Heat shock in C. elegans induces downstream of gene transcription and accumulation of double-stranded RNA

Contributions from fellow researchers

I would like to thank Patrick Gonzales for performing the heat shock and J2-immunoprecipitation for RNA sequencing libraries, and consultation for the Sequencing Read Archive (SRA) upload. I would like to thank Joseph Cabral for the imaging and analysis of formation of dsRNA and Heat Shock Factor-1 foci over one hour (Figure 1 from the paper and Fig II-1 in this thesis). I would like to thank Mary A. Allen for consultation on Fluorescent In-Situ Hybridization experiment.

Citation

All or some of this chapter was published as Melnick M, Gonzales P, Cabral J, Allen MA, Dowell RD, Link CD. Heat shock in C. elegans induces downstream of gene transcription and accumulation of double-stranded RNA. PLoS ONE. 2019;14(4). doi:10.1371/journal.pone.0206715

Introduction

Cytoplasmic proteotoxic stress induced by temperatures outside of the optimal range for cells or organisms triggers the heat shock response (HSR)[284]. The response to heat shock is multi-faceted and regulation of both transcription and translation occurs. Transcriptional responses include formation of stress granules, alternative splicing, and aberrant transcriptional termination[190,285–287]. The HSR is a highly conserved transcriptional response and is driven largely by the heat shock transcription factor HSF1[288]. Under basal level conditions, HSF1 is a monomer in the cytoplasm and nucleus. Upon stress, HSF1 undergoes homotrimerization and binds to DNA heat shock elements (HSE) and initiates the transcription of heat shock protein genes[289,290]. In addition, translation of non-heat shock mRNAs is reduced through pausing of translation elongation as well as inhibition of translation initiation[291–293]. Regulation and clearance of misfolded proteins by heat shock proteins has been implicated in neurodegenerative diseases such as Huntington's disease (HD), Parkinson's disease (PD), Alzheimer's disease, and amyotrophic lateral sclerosis[294].

Aside from the canonical binding of HSF1 to HSE loci, heat shock can cause HSE-independent transcriptional changes[285]. In mammalian cells, HSF1 granules colocalize with markers of active transcription where HSF1 binds at satellite II and III repeat regions[295]. In the worm Caenorhabditis elegans, HSF-1 (worm ortholog of HSF1) granules also show markers of active transcription but the putative sites of HSF-1 stress granule binding are unknown[296].

In addition to formation of HSF1 stress granules, heat shock can cause reduced efficiency of transcription termination and the accumulation of normally un-transcribed sequences, designated in the literature as downstream of gene containing transcripts (DoGs)[190]. Recent studies have shown increased antisense transcription when read through transcription goes past the PAS into neighboring genes on opposite strands[191,193,194,297]. Antisense transcription has the potential to modulate gene expression by creation of double-stranded RNA (dsRNA) with subsequent degradation through RNA interference (RNAi)[200].

Previous studies in our lab found deletion of *tdp-1*, the worm ortholog of ALS associated protein TDP-43, results in the accumulation of dsRNA foci[174]. In addition to deletion of *tdp-1*, we discovered that heat shock robustly induced nuclear dsRNA foci in worms. To assay this unexpected formation of dsRNA, we performed strand-specific RNA-seq and strand-specific RNA immunoprecipitation sequencing (RIP-seq) with the J2 antibody specific for dsRNA. In heat shocked worms, we find increased J2 enrichment of downstream of gene transcripts as well as genes involved in translation. To identify altered transcription genome-wide, we developed an algorithm called Dogcatcher that identifies DoG locations, genes that overlap with DoGs on the same or opposite strand, and an optional pipeline to provide differential expression of DoGs.

Results

Heat shock induces nuclear dsRNA foci in C. elegans

While looking for conditions that might induce dsRNA foci besides loss of tdp-1, we found that heat shock robustly induced dsRNA nuclear foci. Upshifting wild type worms to 35ºC or 37ºC induced foci detectable with the J2 dsRNA-specific monoclonal antibody within 30 minutes, primarily visible in intestinal and hypodermal nuclei. To determine if these foci overlapped with previously identified nuclear HSF-1 stress granules, we repeated the heat shock experiment with strain OG497 (drSI13)[296]. This strain has a single copy insertion of hsf-1 with a C-terminal GFP driven by the hsf-1 promoter, and shows nuclear GFP expression that redistributes into granules after a one minute heat shock at 35ºC[296]. Using the J2 antibody for immunohistochemistry, we found J2 dsRNA foci in nuclear regions that partially overlapped with nuclear HSF-1 stress granules when drSI-13 worms were heat shocked for 35ºC for 40 minutes (Fig III-1 A-C). Measuring coincidence of foci over one hour in 10 minute increments, we observe a significant change [Family Wise Error Rate (FWER) < 0.05] for all time points 30-60 minutes compared to 10 minutes (Fig III-1 D).

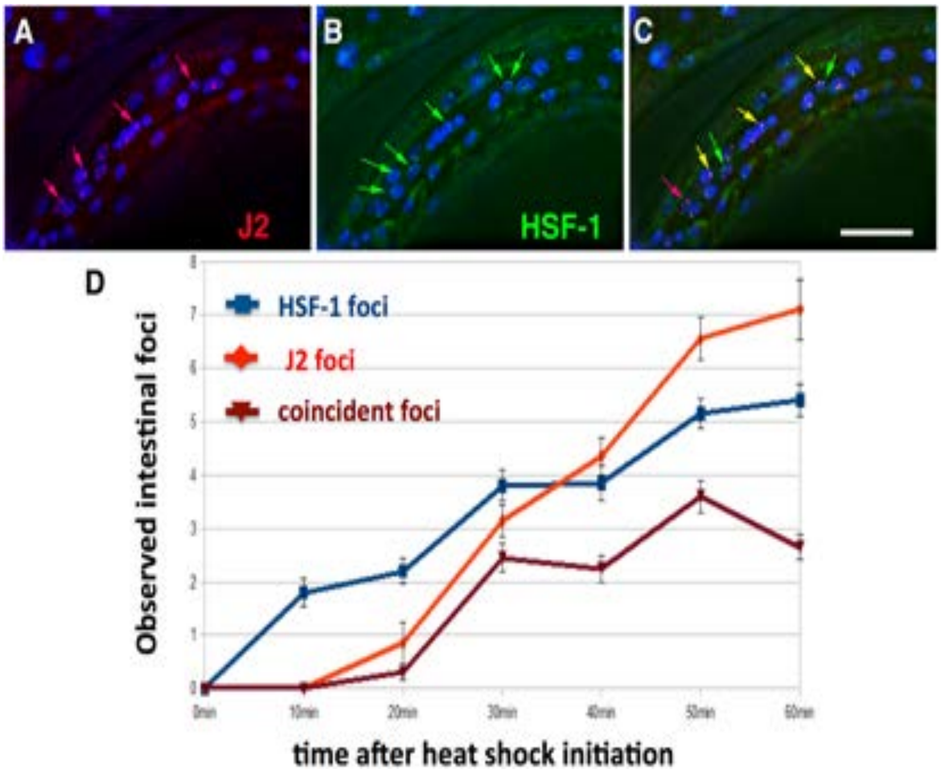

Figure III-1: Heat shock induces nuclear foci detectable with dsRNA-specific antibody J2

Mid-animal intestinal region of 4th larval stage drSI13 worm fixed 40 minutes after heat shock at 35º C. (A) Nuclear J2 foci (red arrows). (B) HSF-1 foci (green arrows). (C) Overlap of J2 foci and HSF-1 foci (orange arrows). White size bar in bottom right corner (20 microns across). DNA stained with DAPI (blue). (D) Quantification of occurrence of HSF-1 and J2 foci over time. 19-20 worms scored per time point with 4 intestinal nuclei scored per worm.

Since dsRNA foci partially overlap with HSF-1 stress granules, we were curious if a HSF-1 partial loss of function mutant *hsf-1(sy441)*[298] would change the amount of dsRNA foci present. We found no significant (FWER < 0.05) differences upon heat shock in the amount of intestinal J2 foci in the *hsf-1(sy44)* mutant compared to heat shocked wild type (Fig III-2). Similar to the *hsf-1(sy441)* mutant, we looked for any effect upon dsRNA formation in a *rde-4(n337)* knockout strain. RDE-4 is a double-stranded RNA binding protein (dsRBP) required for the initiation of RNA interference (RNAi) in C. elegans[299]. We found no significant differences (FWER < 0.05) between heat shocked rde-4(n337) and heat shocked wild type strains, although we found low levels of dsRNA foci in non-heat shocked rde-4 (Fig III-2).

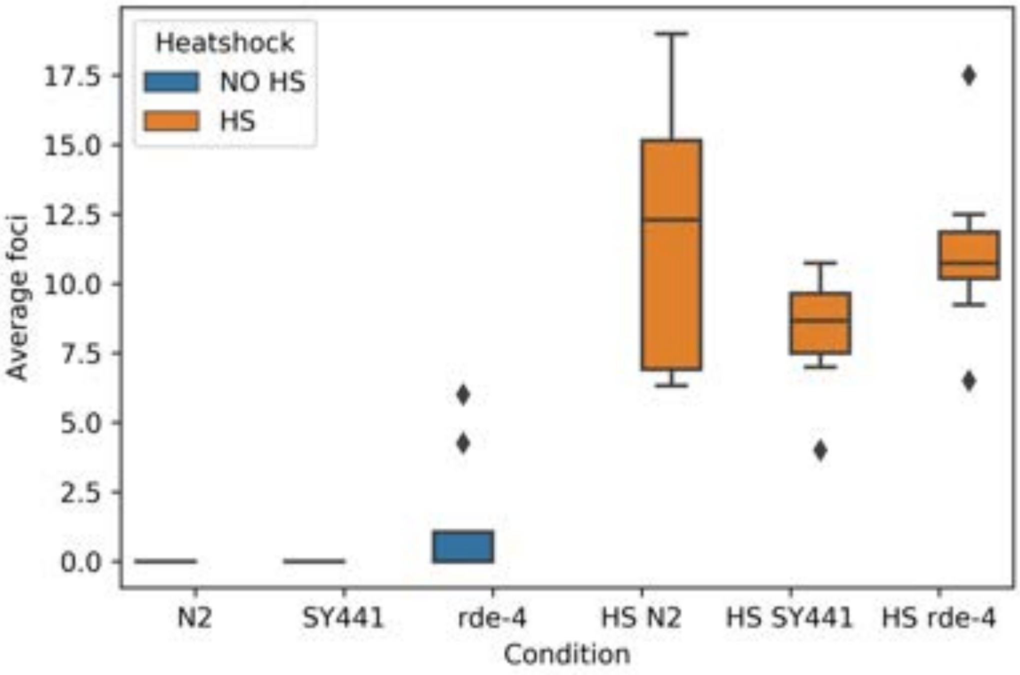

Figure III-2. Quantification of J2 foci with or without heat shock in N2, rde-4, and sy441 mutants.

Box plots of average J2 foci per 2-4 intestinal nuclei scored. 8 worms scored per condition. Analysis of variance (ANOVA) and Tukey HSD post-hoc analysis were used for multiple comparisons between conditions with a significance threshold of < 0.05 (Family Wise Error Rate). Importantly, the null hypothesis is rejected for any heat shocked condition compared to non-heat shocked condition. None of the heat shocked strains were significantly different from one another. None of the non-heat shocked strains were significantly different from one another, although there are some low-level amounts of J2 foci in *rde-4*.

Recovery of dsRNA by J2 immunoprecipitation

In order to identify dsRNA transcripts induced by heat shock, we performed strand-specific RNA sequencing (RNA-seq) and strand-specific RNA immunoprecipitation sequencing (RIP-seq) (Figure III-3). Input RNA and RNA immunoprecipitated with the J2 antibody was extracted and sequenced for heat shocked N2 (wild type) worms (in duplicate) and non-heat shocked worms (in triplicate). The J2 antibody is specific for dsRNA 40bp or more[300] and transcripts from the J2 Immunoprecipitation (IP) could include full length dsRNA transcripts or single stranded RNA (ssRNA) adjacent to 40bp or more sections of dsRNA. dsRNA can occur via base pairing with a different transcript (interstrand) or self-complementarity within the same transcript (intrastrand). Similar to previous experiments, RNA immunoprecipitated samples were normalized to input RNA samples[174,301].

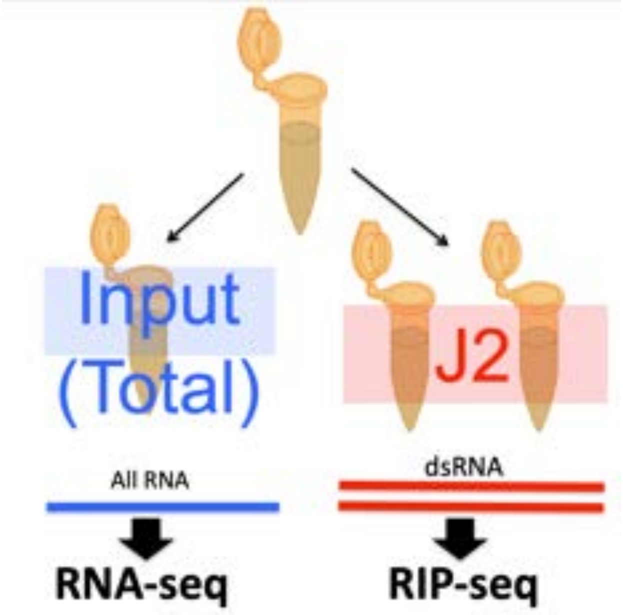

Figure III-3: Schematic of recovery of RNA pools for high throughput sequencing analysis. Control and heat shocked worm populations were recovered and lysed. Worm lysates were then split to recover total input RNA or immunoprecipitated with the J2 antibody.

Measurement of antisense gene transcripts after heat shock

The apparent increase in dsRNA we observed in heat shocked worms [and previously observed in the *tdp-1(ok803)* mutant] could result from an increased accumulation of antisense transcripts. To obtain a global view of antisense levels, we calculated an antisense/sense ratio for genes using the input RNA samples. For a stringent view of fold changes between conditions, genes with a minimum of 20 mean read count (sense and antisense pool) between condition and wild type were used for this analysis. Out of 46,760 worm genes, using this cutoff we scored 11091 genes in heat shock compared to wild type, and 10,831 genes in *tdp-1(ok803)* compared to wild type. When we look at antisense/sense ratios of read counts over genes for each condition compared to wild type (Fig III-4A), we find no difference in the ratio with heat shock (5513/11091 ~49.70%), and an increase in the ratio (7551/10831 ~69.71%) with the *tdp*-1 deletion. Since antisense/sense ratios can increase either through depletion of sense transcripts or increases in antisense transcripts, we examined sense and antisense levels separately in each condition compared to wild type (Fig III-4B). With heat shock, we find no increase [log2 fold change (log2FC) > 0] in sense (4125/11091 ~37.91%) or antisense (3679/11091 ~33.79%) transcripts. With the *tdp*-1 deletion, however, we find noticeably fewer genes with increased sense counts (419/10831 ~3.86%) and no increase in antisense counts (3301/10831 ~30.47%) over genes compared to wild type. Thus, the increase in antisense/sense ratio in the *tdp-1* deletion arises because of lowered accumulation of sense transcripts rather than increased antisense transcript levels. This was not unexpected as TDP-1 plays a role in normal transcription[174].

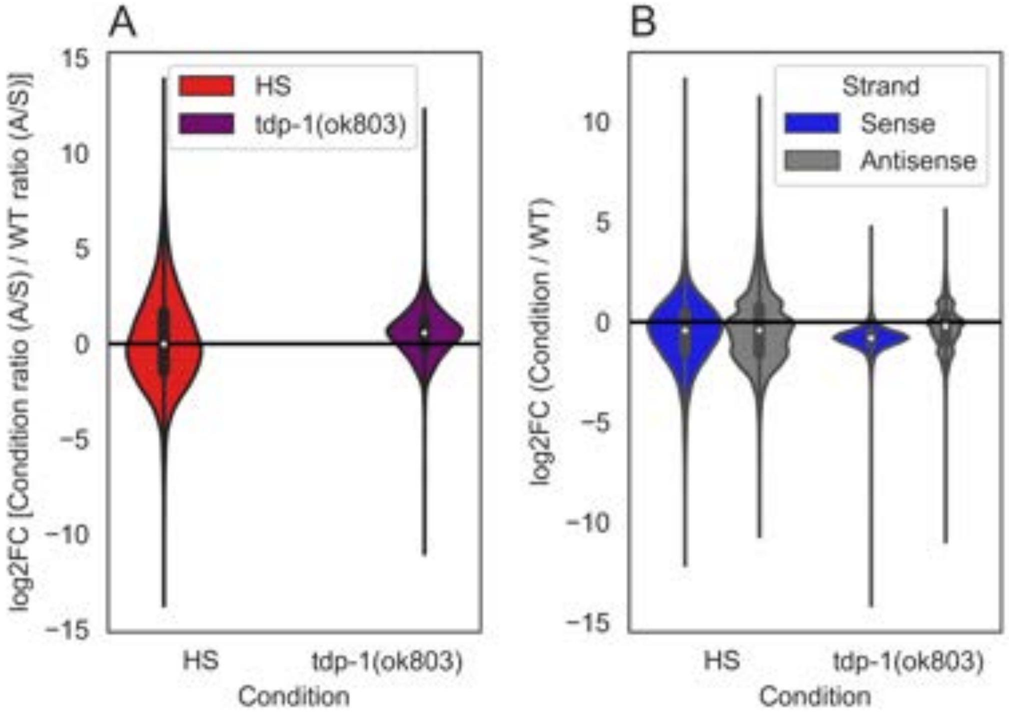

Figure III-4: Quantification of genes with changes in antisense/sense ratios after heat shock or deletion of tdp-1 in input RNA.

Violin plots of the ratio of read counts over gene regions for each condition compared to wild type (WT) [mean >20 for pooled counts (sense and antisense, condition and wild type), n=1]. (A) With heat shock 5513/11091 genes have a higher antisense/sense ratio compared to wild type (log2FC > 0, area of red violin plot above the black line). With *tdp-1(ok803)* 7551/10831 genes have a higher antisense/sense ratio compared to wild type (log2FC > 0, area of purple violin plot above the black line). (B) With heat shock, 4125/11091 sense and 3679/11091 antisense transcripts are upregulated compared to wild type (log2FC > 0, area of violin plot above the black line). With *tdp-1(ok803)* 419/10831 sense and 3301/10831 antisense transcripts are upregulated compared to wild type.

To further validate this result, we analyzed the total RNA-seq data of Brunquell et al[302], who performed similar heat shock experiments in *C. elegans*. In the Brunquell dataset we found a small increase in the amount of genes with significant antisense transcription upon heat shock (28/1818 ~1.54%) compared to no heat shock in wild type worms (Fig III-5). We conclude that in *C. elegans* heat shock does not result in transcriptional dysregulation that leads to a large increase in antisense transcripts.

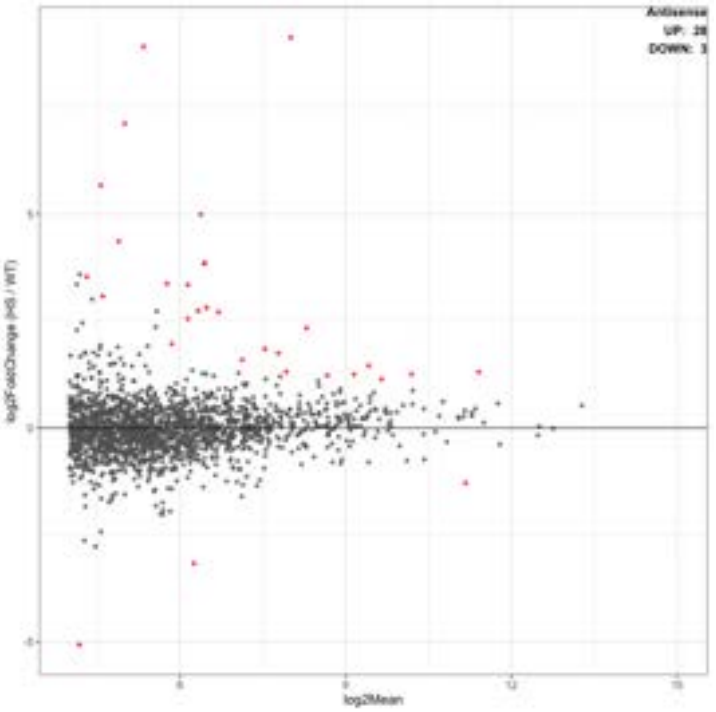

Figure III-5: Comparison of antisense transcripts in worms with heat shock (wild type) vs no heat shock (wild type) from Brunquell et al., 2016

MA plot of significant (FDR <0.05, log2Mean > 4) antisense transcripts. Out of 1818 score-able genes that passed the mean cutoff, 28 were significant in heat shocked worms (red dots above the middle line), and 3 were significant for worms without heat shock (red dots below the middle black line). Data from from Brunquell et al., 2016[303].

Comparison of dsRNAs identified in worms heat shocked or deleted for *tdp-1*

Considering that both heat shock and deletion of the *tdp-1* gene lead to the formation of nuclear dsRNA foci, we sought to determine if this phenotypic similarity also extends to transcripts that are accumulating in the dsRNA pool. After heat shock, we found a large number of significant gene transcripts [false discovery rate (FDR) < 0.05 and a log2 mean expression (log2Mean) > 4]. Specifically, in the pool of RNAs immunoprecipitated by the dsRNA-specific antibody J2 (relative to untreated worms), we found (4774/18737) significantly enriched or (1669/18737) significantly depleted transcripts (Fig. III-6A). We also identified antisense transcripts with significantly altered representation in the J2 IP pool, and found 650/8832 enriched and 477/8832 depleted (Fig. III-6B). A minority of genes had both sense and antisense transcripts significantly enriched (180) or depleted (48) in the heat shock J2 IP pool (Fig. III-6 A-B). In *tdp-1(ok803)* significant (FDR <0.05, log2Mean > 4) gene transcripts, we found a smaller number of sense enriched (418/13223) and depleted (59/13223) (Fig. III-6C), as well as antisense enriched (245/2343) and depleted (14/2343) genes (Fig. III-6D). Similar to heat shock, *tdp-1(ok803)* had relatively fewer genes with both sense and antisense transcripts significantly enriched (6) and depleted (1) (Fig. III-6C-D). We found a significant [P < 1 x 10^{-30}, hypergeometric distribution (hgd)] overlap of J2 enriched gene transcripts between the heat shock and *tdp-1*(*ok803*) populations in both sense (Fig. III-6E) and antisense (Fig. III-6F), suggesting that there might be some similarities between the dsRNA accumulation induced by heat shock and deletion of *tdp-1*. However, with J2 depleted transcripts, we found no significant overlap in (P = 0.165, hgd) in sense transcripts and no significant overlap in antisense transcripts (P = 0.187, hgd).

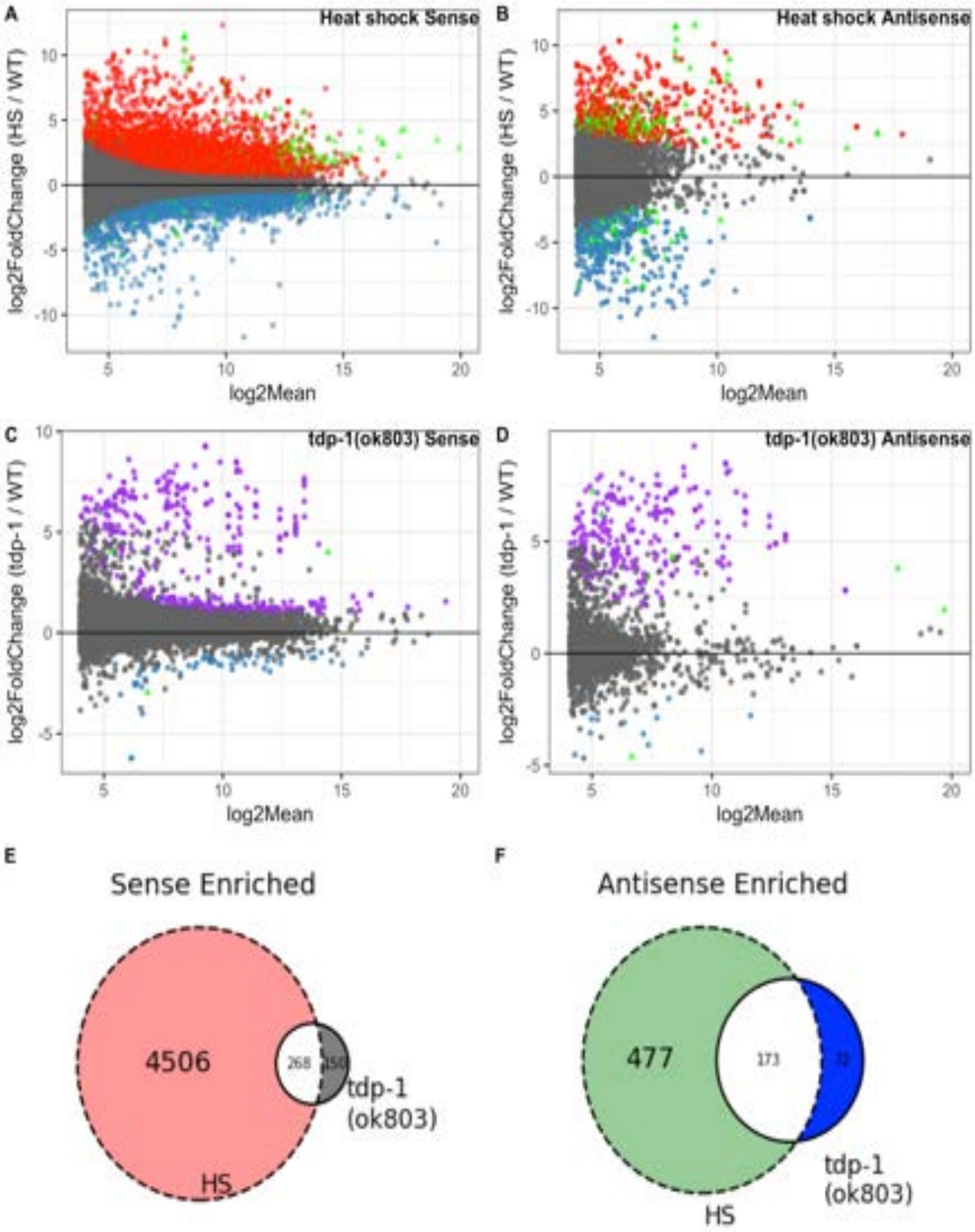

Figure III-6: Comparison of J2 enriched sense and antisense transcripts in heat shock and *tdp-1(ok803)* worms.

MA plots [“M” (log2FC) on y-axis and “A” (log2Mean) on x-axis] of significant (FDR <0.05) dsRNA enrichment for sense and antisense transcripts (analyzed independently) along with Venn diagrams of enrichment for enriched sense and antisense [n=2 for heat shock J2 samples, n=3 for wild type, n=3 for *tdp-1(ok803)*]. (A) Heat shock over wild type J2 enriched sense transcripts with 4774 enriched (red) and 1669 depleted (blue). (B) Heat shock over wild type J2 enriched antisense transcripts with 650 enriched (red) and 477 depleted (blue). (A-B) enriched (180) and depleted (48) heat shock vs wild type transcripts found in both sense and antisense (green triangles). C: *tdp-1(ok803)* over wild type significant J2 enriched sense transcripts with 418 enriched (purple) and 59 depleted (blue). (D) *tdp-1(ok803)* over wild type significant J2 enriched antisense transcripts with 245 enriched (purple) and 14 depleted (blue). (C-D) enriched (6) and depleted (1) *tdp-1(ok803)* vs wild type transcripts found in both sense and antisense (green triangles). (E) Overlap of genes with significantly J2 enriched sense transcripts in both conditions compared to wild type worms. (F) Overlap of genes with significantly J2 enriched antisense transcripts in both conditions compared to wild type worms.

We next sought to examine whether the dsRNAs arising in heat shock or *tdp-1(ok803)* showed enrichment for similar pathways. Using GOATOOLS[304], we found that many Gene ontology (GO) terms related to translation were significantly enriched (FDR < 0.05) in both the heat shock and *tdp-1*(*ok803*) J2 IP pools. Out of 330 translation related genes classified by GOATOOLS, in sense J2 enriched transcripts, we find 234 translation related genes with heat shock, 27 translation related genes in *tdp-1(ok803)*, and 19 translation related genes in the overlap. In the J2 depleted sense pool, only heat shocked worms contained translation related genes (30 total) with none in *tdp-1(ok803)* pool. In J2 enriched antisense transcripts, only heat shocked worms had 33 translation related genes. There were no translation related genes found in J2 depleted antisense transcripts. Thus, the dsRNA recovered by J2 immunoprecipitation is enriched for translation related pathways under both conditions, but there are distinct transcripts in heat shock compared to *tdp-1(ok803)* worms.

J2 Enrichment of Repetitive elements

Previous work in our lab has shown increased repetitive elements (RE) enriched for dsRNA in *tdp-1(ok803)*[174]. We wanted to see if heat shock would also show changes in repetitive elements. Using Repenrich2, REs are organized in clades with class at the top, followed by family, down to fraction. In the worm genome, the majority of repetitive element families are in the DNA class. Within the DNA class we found all significant (FDR <0.05) families were depleted in heat shock J2 (Fig. III-7 A).

Among the DNA class, we found that 11 out of the top 15 most significantly (FDR <0.05) depleted REs were in superfamilies (TcMar, hAT, Piggyback) of Terminal Inverted Repeats of DNA class II transposons. In other non-DNA class families, only three families were significantly (FDR <0.05) enriched in heat shock J2 compared to wild type (Fig. III-7 B). When looking at the fraction level for these three families, we found that the fraction with the highest enrichment belonged to a satellite repeat RCD1. The R2 LINE element Nematode Spliced Leader-1 (NeSL-1), was the next highest J2 enriched fraction which also had the greatest mean out of all of the non-DNA class fractions. The last enriched RE (CELE45) is a type II Sine/tRNA element of unknown function.

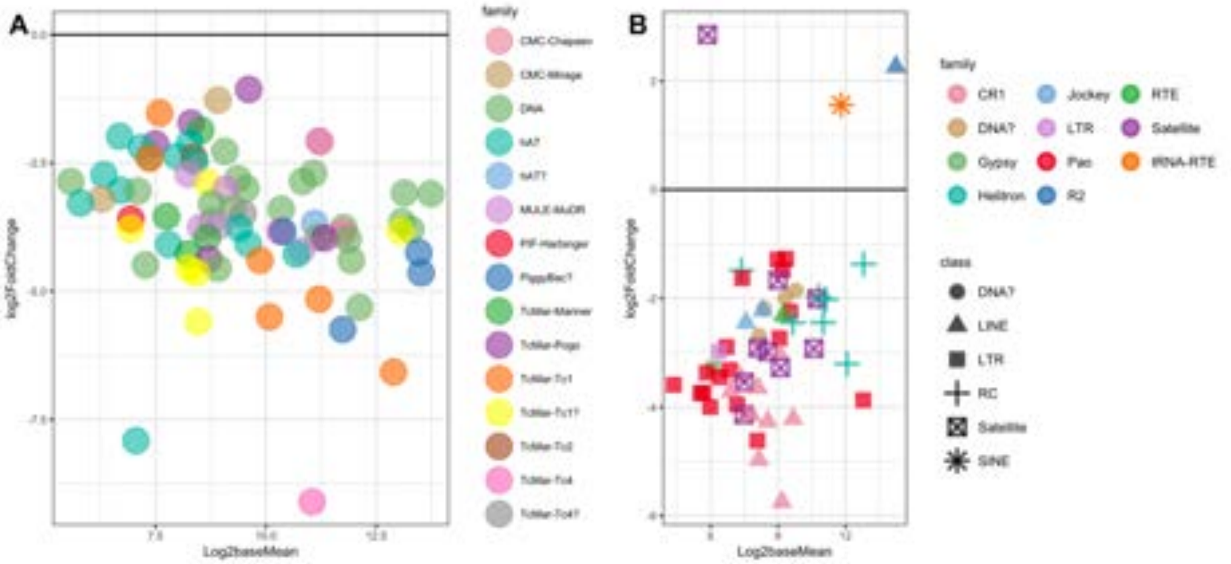

Figure III-7: Heatshock vs wild type J2 enrichment of repetitive elements

Heatshock vs wild type J2 enrichment of Repetitive elements (A) MA plots of heat shock over wild type J2 enrichment from all DNA class families of repetitive elements. (B) MA plot of heat shock over wild type from non-DNA class families. Shape and color correspond to class and family in Repenrich2.

Enrichment of transcripts downstream of genes in the J2 pool

While examining the transcription of known heat shock inducible genes, we noted in heat shocked populations an accumulation of read through transcripts downstream of annotated genes (see example in Fig. III-8). Interestingly, some of these downstream of gene transcripts (DoGs) were also highly enriched in the J2 IP pool. While previous work has characterized the accumulation of downstream of gene transcripts in heat shocked cells from human[190] and mice[191], the phenomenon has not previously been associated with dsRNA accumulation. To annotate read through regions across the whole genome, we created an algorithm called Dogcatcher. Dogcatcher uses a sliding window approach (100 bps) to annotate read through regions. In addition to Dogcatcher, we established an optional wrapper for quantifying differential expression through Rsubread and DESeq2.

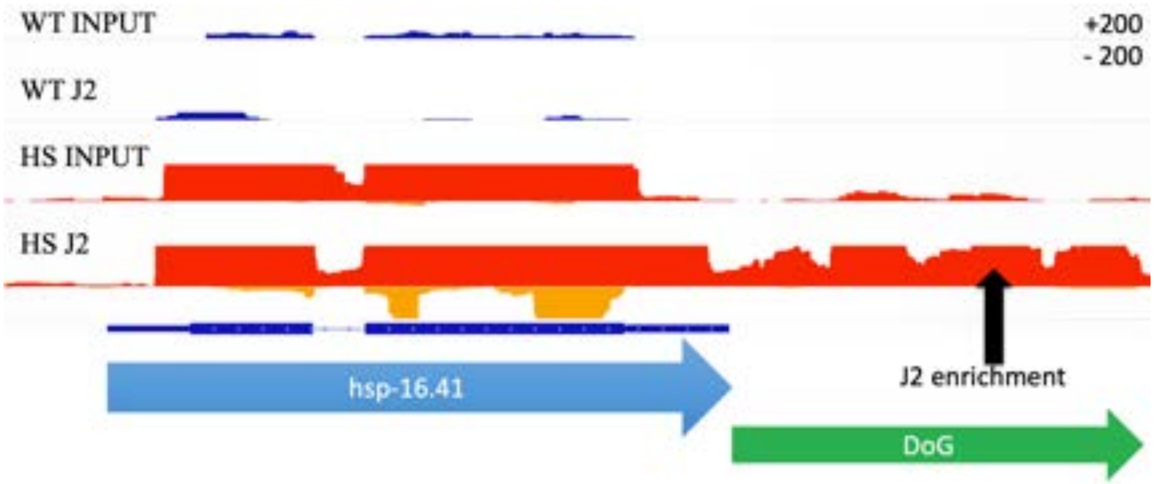

Figure III-8: Aberrant transcription past the end of heat shock family genes showed enrichment in heat shock J2

Normalized histogram from the Integrative Genomics Viewer (IGV). On each track, the sense strand is on the top part of the histogram and antisense is on the bottom (Max read depth +/- 200). Wild type (WT) sense (dark blue) and antisense (light blue), heat shock (HS) sense (red) and antisense (orange). Gene transcription continues past the 3' end of gene (blue arrow) in heat shock, leading to an annotated downstream of gene transcript (DoG) (green arrow).

Using the Dogcatcher algorithm in conjunction with the *C. elegans* genome annotation to identify DoGs *de novo,* we were able to quantify downstream of gene transcripts in the J2 IP pool that would be missed using the standard *C. elegans* genome annotation. Differential expression of transcription in downstream of gene regions can occur via novel DoGs that are in one sample and not another, or by varying levels of transcription of a DoG that is expressed in both samples. Our analysis suggests that both mechanisms may be involved. When we compare heat shocked worms to wild type in the J2 IP and input RNA, we find the majority of annotated DoGs (272/490 ~56%) come from the J2 IP from heat shocked worms (Fig. III-9A). When we compare *tdp-1(ok803)* worms to wild type in the J2 IP and input RNA, we find that the largest fraction comes from input RNA from *tdp-1(ok803)* worms (85/300 ~28%) followed closely by DoGs that are shared between J2 IP and input RNA for both wild type and *tdp-*

1(ok803) worms (52/300 ~17%) (Fig. III-9B). This suggests that with heat shock the majority of DoGs are novel dsRNA enriched regions, in contrast to the *tdp-1* deletion which has a bigger overlap with DoGs in wild type samples.

To quantify differential expression, we used the Dogcatcher optional differential expression wrapper. After heat shock, more read through sections were significantly (FDR <0.05, log2Mean > 4) enriched in the J2 IP pool than depleted (84 vs. 25 out of 421) (Fig. III-10A). Of the 84 DoGs significantly increased in the J2 IP pool with heat shock, the largest group of DoGs (35/84 ~41.66%) are only present in the J2 IP (Fig. III-9C). Of the 25 DoGs significantly decreased in the J2 IP with heat shock, we find a fairly even split between groups (Fig. III-9D).

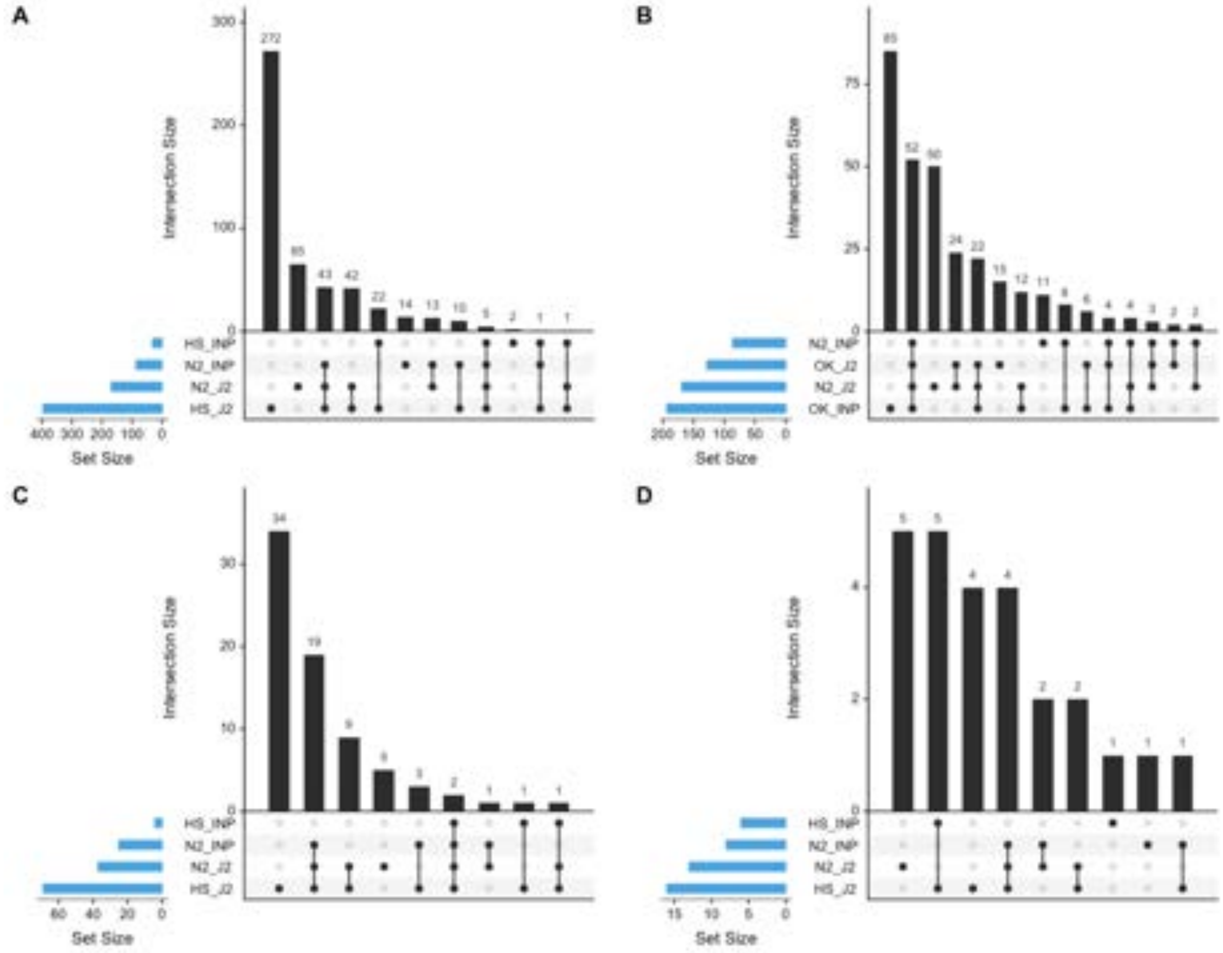

Figure III-9: DoGs and DoGs significant with heat shock or *tdp-1(ok803)* mutation

Upset plots are venn diagram-like plots. Each set is on a row with total amount in a set as a blue bar plot on the left. The black histogram on top shows the counts that are in the intersection of sets (a single dot for one set or connected dots for multiple sets). DOGs that overlap with operons on the same strand have been removed. (A) Upset plot of all DoGs that are found with heat shock (HS) or wild type worms (WT) for the J2 pulldown or input RNA (INP). Heat shock J2 shows the most novel DoGs. (B) Upset plot of all DoGs that are found with the *tdp-1* deletion (OK) or wild type worms for the J2 pulldown or input RNA. the *tdp-1* deletion input sample shows the most amount of novel DoGs. The second highest amount is shared by all four samples. (C) DoGs significantly enriched with heat shock in the dsRNA pull down. Notably 35 of these are only found in the heat shock J2 pull down. (D) DoGs significantly depleted with heat shock in the dsRNA pull down.

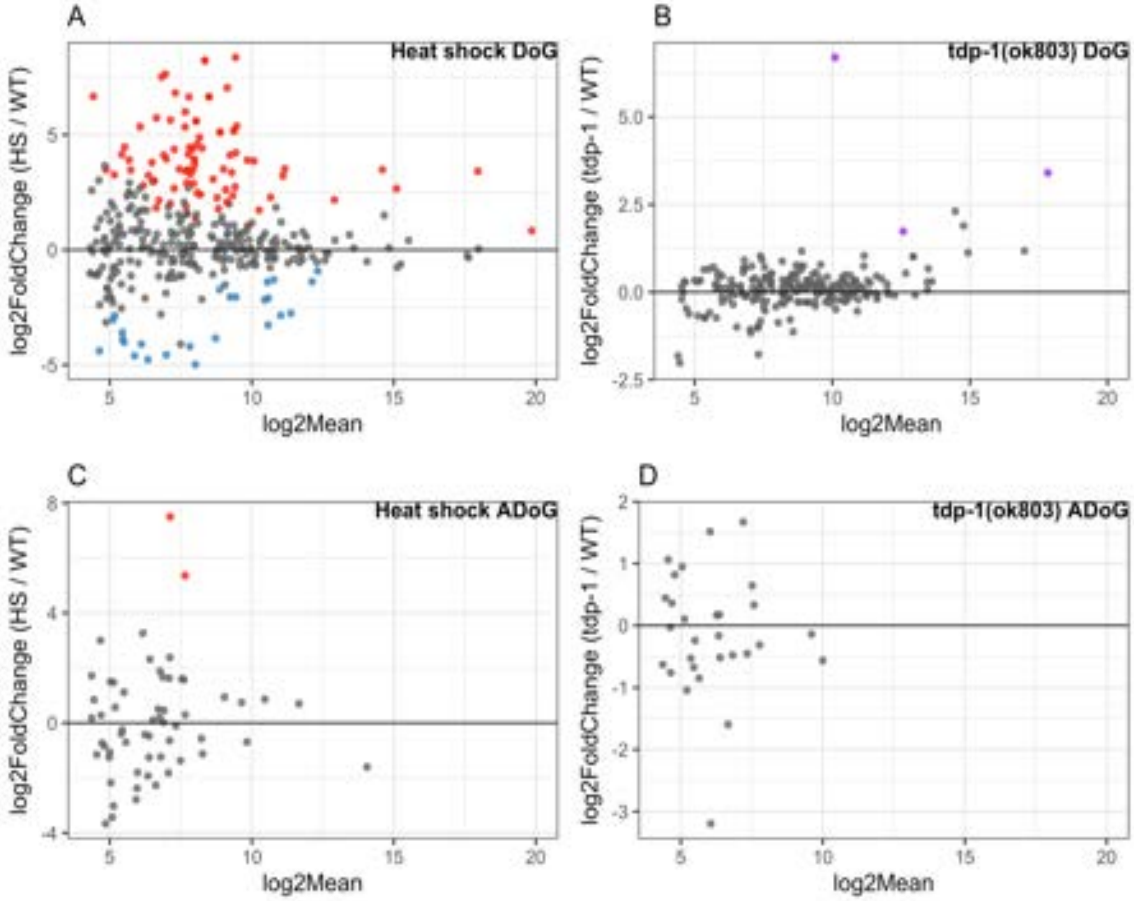

Figure III-10: enrichment of DoGs and ADoGs in heat shock and *tdp-1(ok803)* worms

MA plots of significant (FDR <0.05) dsRNA enrichment for DoGs and ADoGs. Annotated genes that were not significantly changing were added in with DoGs or ADoGs for DESeq2 normalization but were taken out of the plots for clarity [n=2 for heat shock J2 samples, n=3 for wild type, n=3 for *tdp-1(ok803)*]. (A) Heat shock over wild type J2 enriched read through sense transcripts with 84 enriched (red) and 25 depleted (blue) out of 421 scored DoGs. (B) *tdp-1(ok803)* over wild type significant J2 enriched read through sense transcripts with 3 enriched (purple) out of 265 scored DoGs. (C) Heat shock over wild type J2 enriched read through antisense transcripts with 2 enriched (red) out of 70 scored ADoGs. (D) No significant *tdp-1(ok803)* over wild type J2 enriched read through antisense transcripts out of 43 scored ADoGs.

We found that for DoGs enriched in the J2 IP pool after heat shock, the majority corresponds to protein coding genes (60%), followed by non-coding RNA (ncRNA) (20%), pseudogenes (9%), and small nucleolar RNA (snoRNA) (9%). When we looked at significant sense genes with corresponding significant DoGs upon heat shock, we found 62 out of 84 enriched and 11 out of 25 depleted DoGs in the J2 IP have corresponding significant genes (Fig. III-11A-B). Consistent with our results, we found a small increase in the amount of significant (FDR <0.05, log2Mean > 4) DoGs upon heat shock (23/488 ~4.7%) compared to wild type (2/488 ~0.4%) in the Brunquell et al[302] dataset (Fig. III-12A).

We found far fewer significantly (FDR <0.05, log2Mean > 4) J2 enriched DoGs from *tdp-1(ok803)* (3 out of 265) with no regions being depleted (Fig. III-10B). Interestingly, 2 out of the 3 DoGs in *tdp-1(ok803)* were also enriched in the heat shock J2 pool. When we looked at significant sense genes with corresponding significant DoGs upon *tdp-1* deletion, we found 1 out of 3 DoGs enriched in the J2 IP have corresponding significant genes (Fig. III-11C).

From the significantly enriched GO terms of DoGs in heat shock and *tdp-1(ok803)* worms, only heat shocked worms had any significantly enriched GO terms, which primarily consisted of histone genes. As a possible explanation for the formation of dsRNA at downstream of gene regions, we found DoGs to be enriched in terminal repeat sequences compared to a random intergenic downstream background. (Fig III-13).

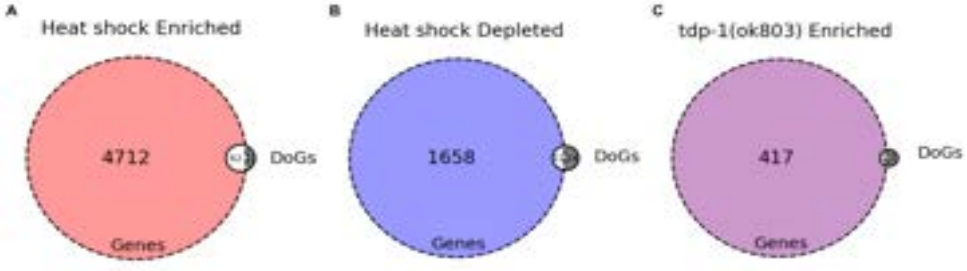

Figure III-11: Venn Diagrams of overlap between significant genes and DoGs

(A) Out of 4774 genes and 84 DoGs enriched in the J2-IP with heat shock, 62 overlap. (B) Out of 1667 genes and 25 DoGs depleted in the J2-IP with heat shock, 11 overlap. (C) Out of 418 genes and 25 DoGs enriched in the J2-IP with the *tdp-1* deletion, 1 overlap. Since there was no significant DoGs depleted with the *tdp-1* deletion there was no overlap.

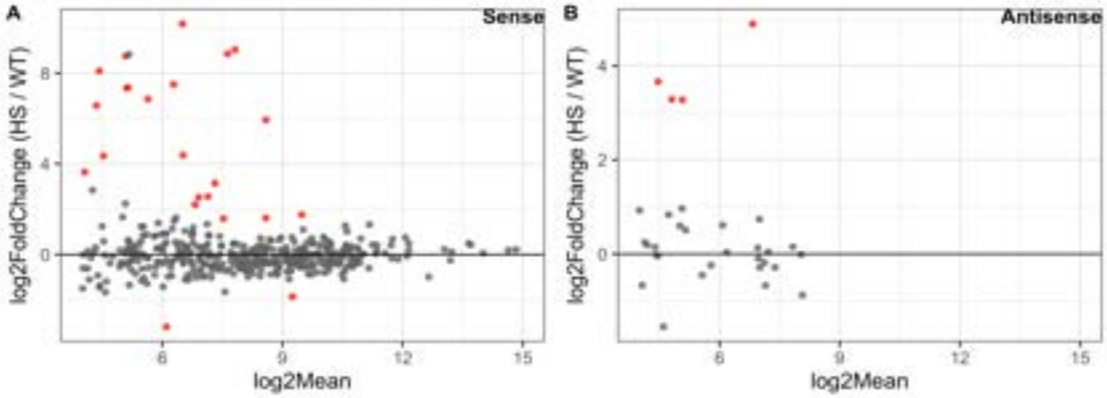

Figure III-12: Comparison of DoGs and ADoGs in worms with heat shock (wild type) vs no heat shock (wild type) from Brunquell et al., 2016.

MA plots of significant (FDR <0.05, log2Mean > 4) DoGs and ADoGs. (A) out of 488 DoGs, 23 DoGs were significantly up with heat shock (red dots above the middle line) compared to 2 down (red dots below the middle line. (B) Out of 51 ADoGs, 4 ADoGs were significantly up with heat shock.

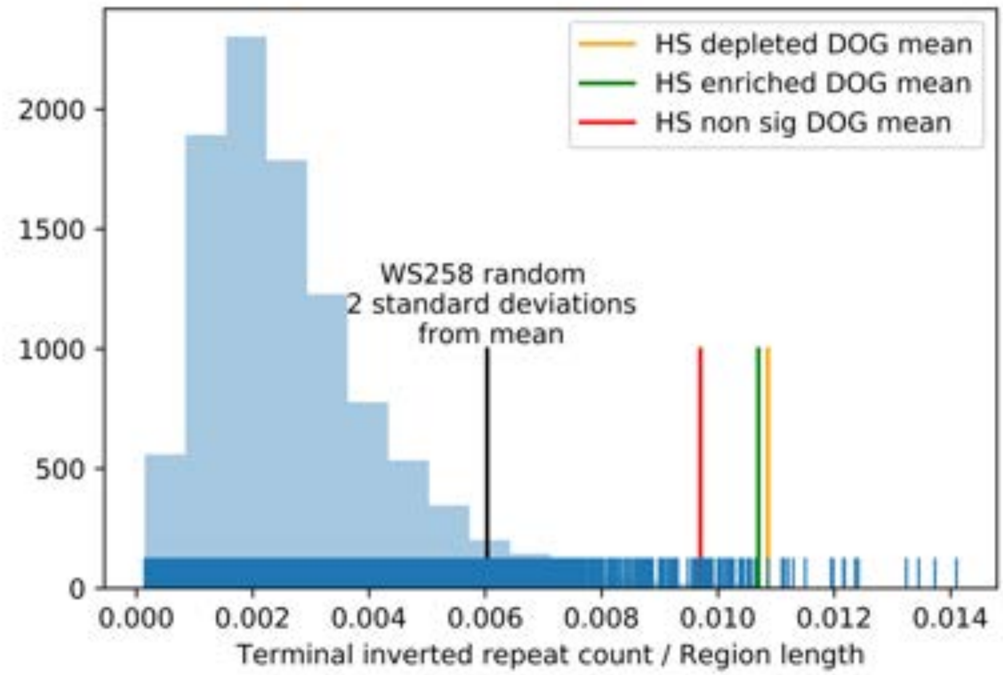

Figure III-13: Number of Terminal Inverted Repeats (TIR) overlapping downstream regions.

DoGs have a significantly different percentage of terminal inverted repeat counts compared to a random downstream intergenic background. Heat shock enriched (green), depleted (yellow), and non-significant (red) means of normalized TIR counts. Histogram was created with 10,000 means of random downstream intergenic background regions with rugplot counts for each mean (blue) and two standard deviations from the mean marked (black).

Additional non-annotated transcripts are minimally enriched in the J2 pool after heat shock or tdp-1 deletion

Next, we were curious if other sections around genes would show aberrant transcription in heat shock or *tdp-1(ok803)* worms. Expanding on the DoG nomenclature, the terms we use for the three other types of transcription flanking an annotated gene are as follows: regions downstream of genes with antisense reads (ADoGs), sense reads in regions previous of the gene (PoGs), and antisense reads in regions previous of the gene (APoGs) (Fig. III-14). Importantly, novel areas of intergenic transcription are obtained by filtering out PoGs with any overlap to DoGs on the same strand, as well as ADoGs or APoGs with any overlap to DoGs (or genes) on the opposite strand (Fig. III-14 and Fig. III-15). We did not find any significantly (FDR <0.05, log2Mean > 4) J2 enriched PoGs or APoGs in either condition compared to wild type. We found a small amount of significant (FDR <0.05, log2Mean > 4) J2 enrichment in heat shock ADoGs (2 out of 70) (Fig. III-10C) and no ADoGs enriched in *tdp-1(ok803)* out of 43 scored ADoGs (Fig. III-10D).

Consistent with our heat shock results, in our analysis of Brunquell et al., 2016, we found a small increase in the amount of significant (FDR <0.05, log2Mean > 4) ADoGs upon heat shock in wild type worms (3/51 ~5.88%) and no significant ADoGs without heat shock. (Fig. III-12B).

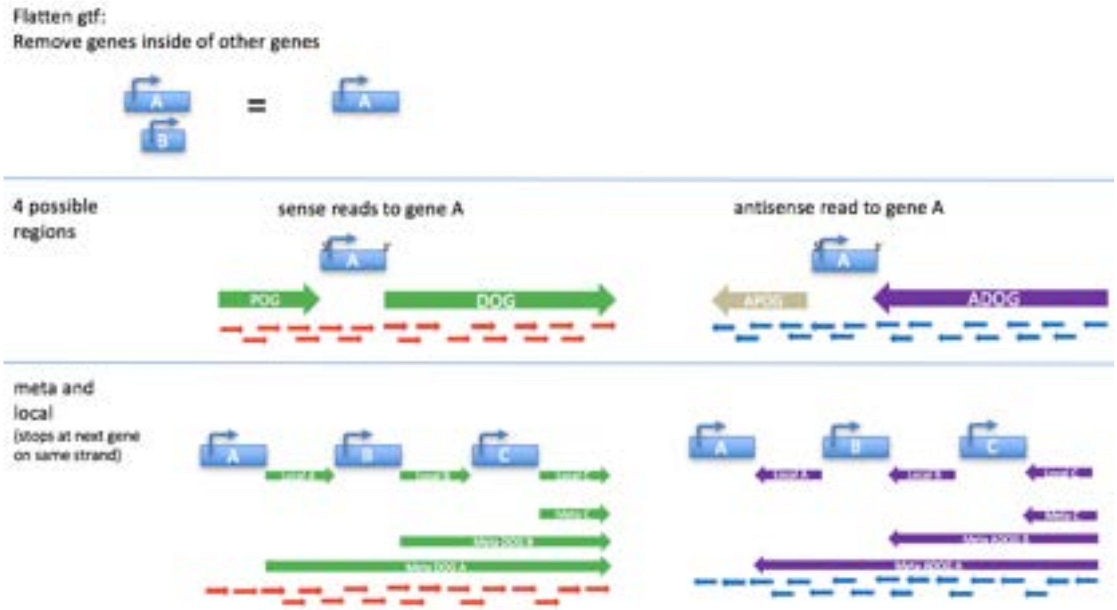

Figure III-14: Dogcatcher flattening and nomenclature.

First, genes were removed that were inside of other genes. With sense and antisense reads and two regions flanking a gene, four types of classification can be made. Downstream of gene transcription with sense reads (DoGs), downstream of gene transcription with antisense reads (ADoGs), previous of gene transcription with sense reads (PoGs), and previous of gene transcription with antisense reads (APoGs).

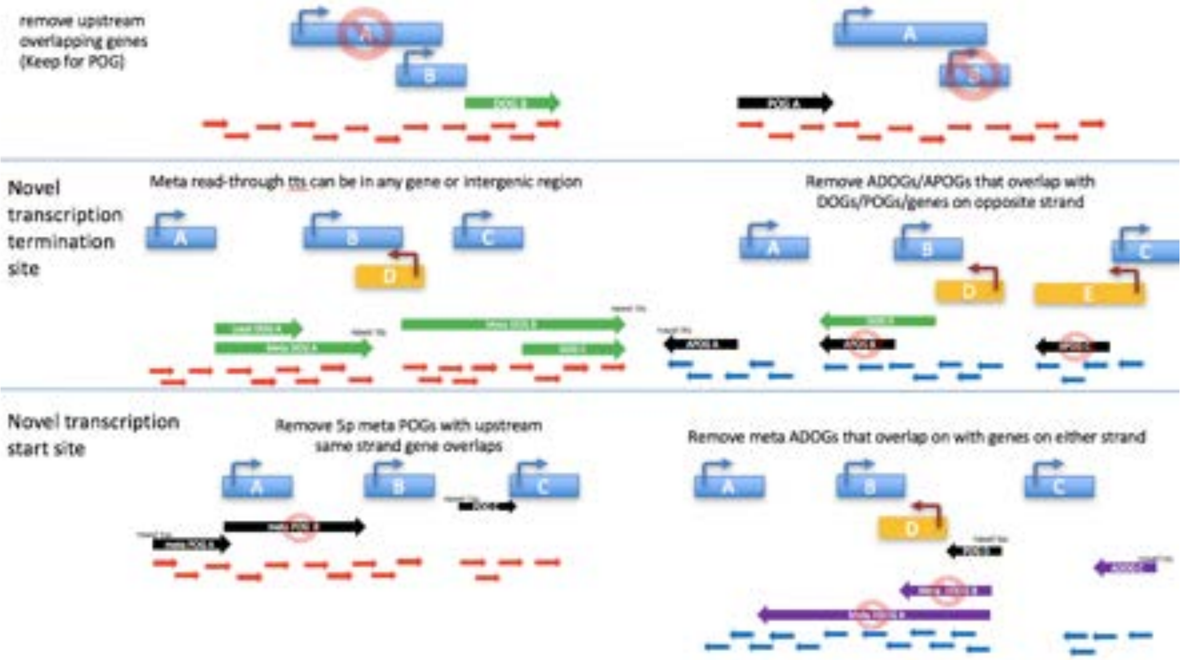
remove upstream overlapping genes (Keep for POG)
A
B
A
B
Novel transcription termination site
Meta read-through tts can be in any gene or intergenic region
A
B
C
D
Remove ADOGs/APOGs that overlap with DOGs/POGs/genes on opposite strand
A
B
C
D
E
Novel transcription start site
Remove 5p meta POGs with upstream same strand gene overlaps
A
B
C
Remove meta ADOGs that overlap on with genes on either strand
A
B
C
D

Figure III-15: Dogcatcher additional filtering.

When searching for DoGs, we removed any genes that had overlap with genes downstream on the same strand. For PoGs we removed genes with any upstream overlap. Novel transcription termination sites can be found downstream of genes with DoGs. Novel transcription start sites can be found with PoGs. ADoGs or APoGs with overlap to genes or DoGs on the opposite strand are removed.

Increased antisense transcription over genes associated with DoGS and ADoGS

Next, we were curious if any aberrant read through transcription might overlap genes and contribute to increased antisense reads within the gene. We define an overlapped gene as any gene that has an ADoG associated with it or an opposite strand DoG with any overlap to the gene. We next define a significant overlapped gene as any gene that has significant (FDR <0.05, log2Mean > 4) antisense transcript levels as well as an associated significant (FDR <0.05, log2Mean > 4) ADoG or opposite strand DoG (overlapping the gene). From our significant overlapped genes, we found 17 enriched and 5 depleted with heat shock, and only 4 enriched and no depleted in *tdp-1(ok803)* worms. We did not find any overlapped genes that were significantly enriched for GO terms related to translation.

Antisense read through into *eif-3.B* in heat shocked worms

Visual inspection of DoG transcripts identified one transcript downstream of the ncRNA *W01D2.8* (*doW01D2.8*) that ran into the gene *eif-3.B* on the opposite strand (Fig. III-16) (Since *doW01D2.8* is inside of the gene W01D2.3 on the same strand, annotation of this DoG starts at the end of the W01D2.3). *eif-3.B* is an ortholog of human EIF-3.B (eukaryotic translation initiation factor 3 subunit B) and is involved in translation initiation. As the *doW01D2.8* transcript was strongly increased by heat shock in both the input and J2 IP pools, we chose to target this transcript to confirm our RNA-seq data. Fluorescent *in situ* hybridization (FISH) was used as this could both demonstrate the accumulation of the *doW01D2.8* transcript and determine its cellular and subcellular (i.e., possible colocalization with J2 foci) distribution.

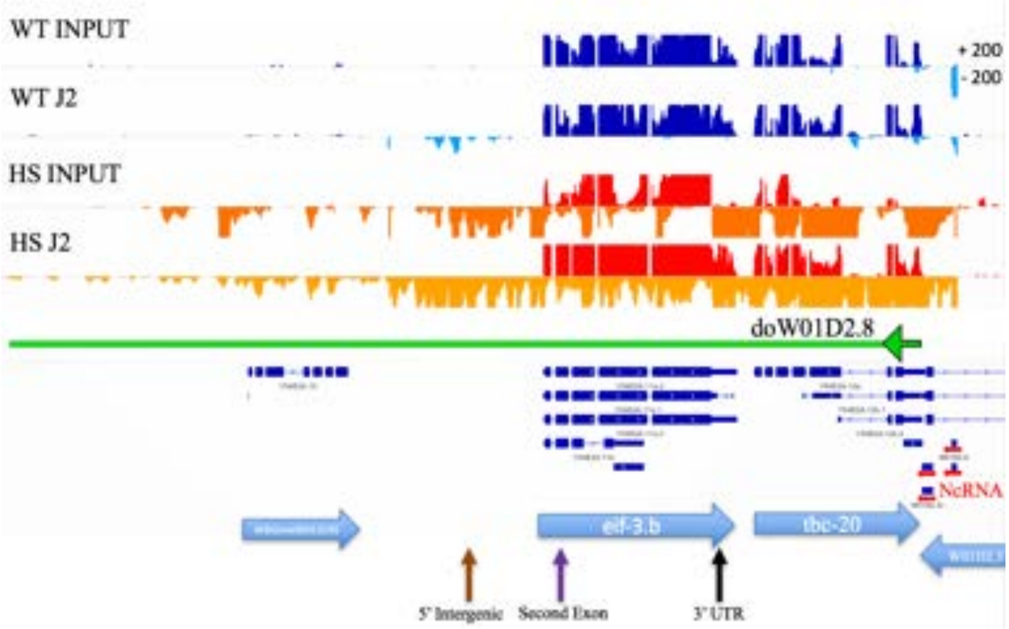

Figure III-16: Heat shock induces transcripts antisense to the eif-3B locus.

IGV view of *eif-3*.B. Normalized tracks with the sense strand on the top part of the histogram and antisense on the bottom with a max read depth of 200 for sense or antisense. Wild type (WT) sense (dark blue), WT antisense (light blue), heat shock (HS) sense (red), heat shock antisense (orange). Horizontal blue arrows indicated genes and gene direction 5' to 3'. Horizontal red arrows on the right show a cluster of ncRNAs including *W01D2.8* and transcription downstream of *W01D2.8 (doW01D2.8)* into *eif-3.B* (green arrow going left). Arrows on the bottom correspond to locations of probes for FISH (brown: 5' Intergenic, purple: Second exon, black: 3' UTR).

Three strand-specific fish probes at the 5' intergenic region (5' INT) (antisense), first 3 exons (sense), and the last exon along with the 3' UTR (LE 3'UTR) (antisense) of *eif-3.B* (Fig. III-17D) were designed. First, we performed immunohistochemistry with the J2 antibody along with FISH for antisense transcripts that contain the last exon and 3' UTR (Fig. III-17A). We find that *doW01D2.8* is transcribed in this region with heat shock and commonly forms two foci per nucleus, but does not colocalize with dsRNA foci. The 5' and 3’ doW01D2.8 probes do strongly colocalize in the nuclear foci (Fig. III-17B), consistent with a single transcript spanning this region. Unfortunately, when probing for the 5’ intergenic region (antisense) and first three exons (sense), the sense probe was undetectable in young adults. However, in embryos the sense probe was detectable, and we did not find the sense probes colocalizing to the antisense foci (Fig. III-18). To inquire if the *eif-3B* antisense foci were a general site of transcript accumulation, we probed for *C30E1.9*, a long ncRNA that is highly expressed, forms nuclear foci, but is not induced in heat shock. We observed that this transcript does not overlap with the *eif-3B* antisense foci (Fig. III-17C). Lastly, we wanted to see if deletion of *tdp-1,* which does not lead to accumulation of *eif-3B* antisense transcripts, would alter heat shock induced accumulation of these transcripts. We found that the *tdp-1* deletion did not alter the formation of *eif-3B* antisense transcripts (Fig. III-17E).

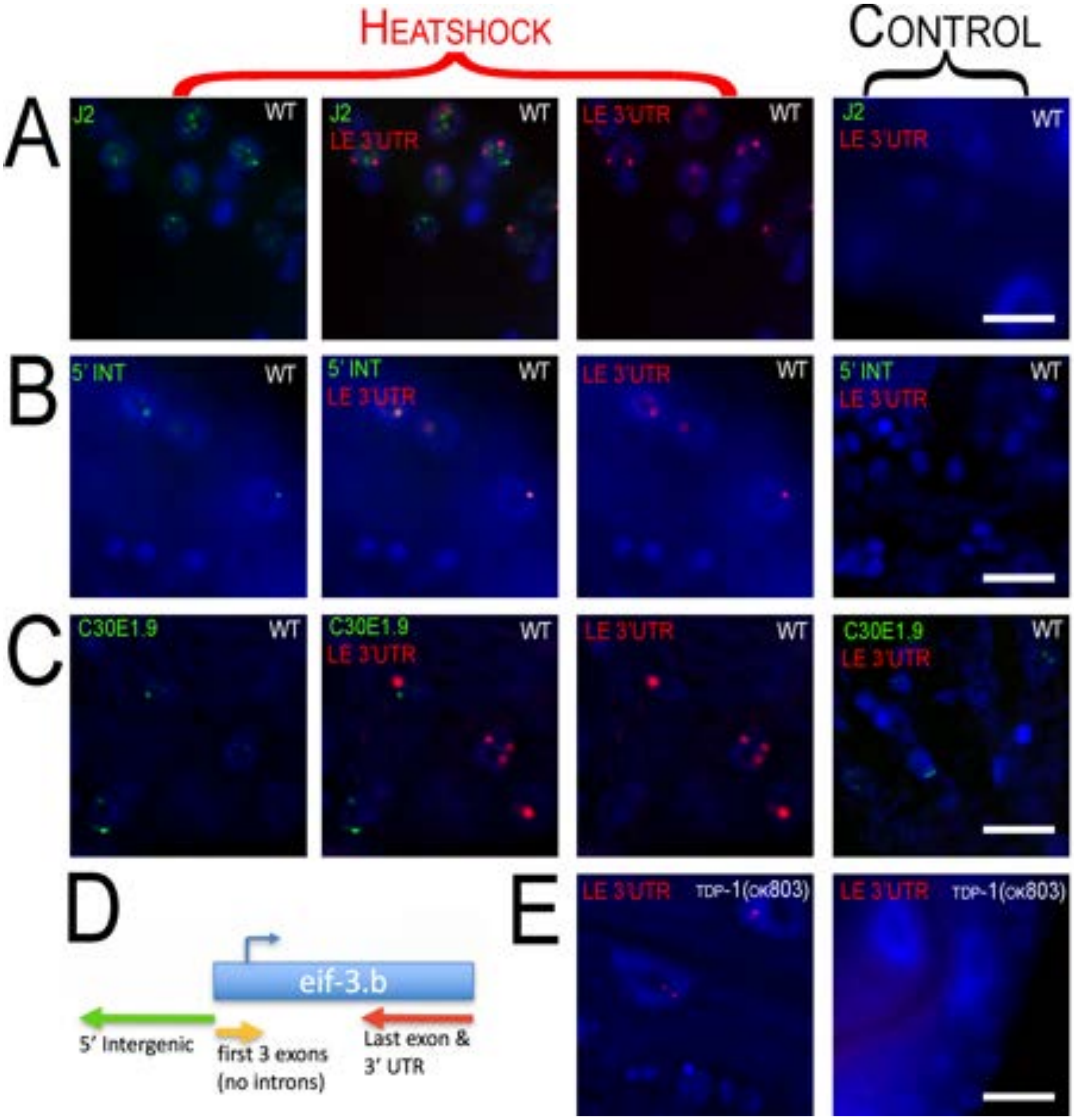

Figure III-17: Fluorescence in situ Hybridization (FISH) of eif-3.B regions

100x oil immersion images of worm hypodermal and neuronal cells. Heat shock panels are in the three columns to the left (merged channel in the middle column). Control panels show exposure from every channel (right column). Row (A) Immunohistochemistry with J2 antibody (green) along with FISH of *doW01D2.8* antisense to the last exon and 3' UTR (LE 3' UTR) (red) of *eif-3.B*. dsRNA and the antisense LE 3'UTR transcript aggregate into nuclear foci with heat shock and do not appear to colocalize. Row (B) FISH of *doW01D2.8* in two regions antisense to the 5' intergenic region (5' INT) (green) and last exon and 3' UTR (LE 3'UTR) (red) of *eif-3.B*. Row (C) FISH of *doW01D2.8* antisense to the last exon and 3' UTR (LE 3'UTR) (red) of *eif-3.B* and sense probe of ncRNA C3DE1.9 (green). Probing of C3DE1.9 is not affected by heat shock and C3DE1.9 is not induced by heat shock. C3DE1.9 and LE 3'UTR show no overlap. (D) Diagram of *eif-3.b* gene with FISH probe locations and orientation. (E) Heat shock of *tdp-1(ok803)* induces nuclear foci from probes antisense to the last exon and 3’ UTR (LE 3’UTR) of *eif-3.B* (left panel) and is not visible with no heat shock (right).

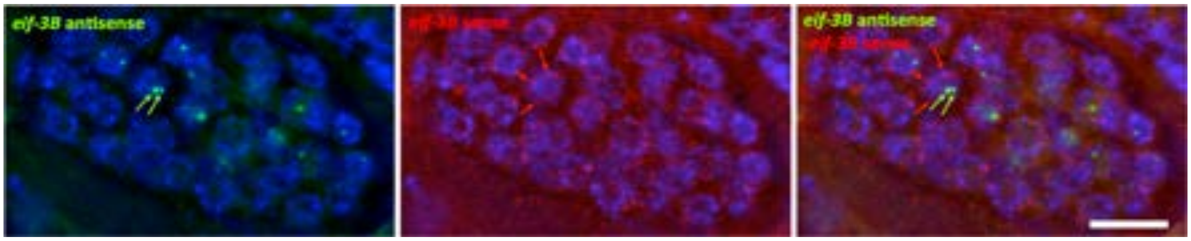

Figure III-18: Sense and antisense eif-3B transcripts do not colocalize.

Shown is a ~ 30 cell embryo fixed and probed for *eif-3B* antisense (left panel) and sense (middle panel) transcripts by fluorescence *in situ* hybridization (FISH) (white size bar is 10 microns across). Note that the *eif-3B* antisense transcripts localize to distinct nuclear foci (green arrows) which do not colocalize with the *eif-3B* sense transcripts (red arrows).

Discussion

We show that heat shock induces nuclear dsRNA foci that partially overlap with HSF-1 nuclear stress granules. A loss of function mutation in *hsf-1* does not block the formation of heat shock induced dsRNA foci, although because this is a hypomorphic mutation we cannot exclude the possibility that HSF-1 plays some role in the formation of heat shock induced dsRNA foci. After heat shock, we find a general increase in the amount of dsRNA and expression levels of transcripts with dsRNA structure, assayed using the dsRNA-specific monoclonal antibody J2. The dsRNA transcripts recovered by J2 immunoprecipitation after heat shock partially overlap with J2 transcripts previously identified in *C. elegans* worms deleted for *tdp-1*. This result suggests that while heat shock does not directly mimic the effects of loss of *tdp-1*, these two conditions likely share some overlapping biological processes. In addition, we find that heat shock induces accumulation of novel downstream of gene transcripts. To our knowledge this is the first time heat shock has been shown to lead to the accumulation of these abnormal transcripts in an *in vivo* model. In addition to global depletion of J2 immunoprecipitated transcripts from LTR's, we find that all DNA class elements are depleted globally and the majority of significantly depleted superfamilies (TcMar, hAT, Piggyback) are from Terminal Inverted Repeats. Among J2 enriched repetitive transcripts in heat shock, we find one RE (NeSL-1) with a potential mechanism to reduce transcription by inserting into SL-1 genes and causing aberrant trans splicing. How this relates to humans is harder to answer, as none of the splice leader mechanisms exists in humans.

Double-stranded RNA can form intrastrand or interstrand base-pairing. Our data suggest that both types of dsRNA may be contributing to the dsRNA pool induced by heat shock. We find that novel downstream of gene transcripts are enriched in the J2 IP pool. These novel transcripts are enriched in inverted repeat sequences, which may be contributing to the formation of intrastrand (hairpin) dsRNA. Downstream of gene transcripts also have the potential to generate transcripts antisense to neighboring genes on the other strand. This has been reported in the heat shock study by Vilborg et al[191], and we have noted similar examples in our data. Using our new Dogcatcher algorithm, we have also documented novel transcripts originating in intergenic regions, which also have the potential to generate antisense transcripts. Indeed, we observe that antisense transcripts are enriched in the J2 IP, supporting the formation of interstrand dsRNA. We note that the J2 antibody immunoprecipitation protocol used in our study will recover transcripts that have only partial (at least 40 nucleotides) dsRNA structure, thus it is feasible that some transcriptional regions we recover after J2 IP are single stranded extensions of double stranded regions.

The accumulation of dsRNA transcripts after heat shock could be the result of altered RNA production and changes in RNA stability or turnover. Further studies will be required to definitively determine the relative contribution of these cellular processes. Published studies demonstrate that loci susceptible to heat shock induced downstream of gene transcription are marked by open chromatin before heat shock[191] and are depleted of the transcriptional termination factor CPSF-73 after heat shock[305]. These results suggest that altered transcriptional processing itself leads to the altered transcript accumulation after heat shock. However, the significant overlap of transcripts enriched in the J2 pool resulting from heat shock and from deletion of the *tdp-1* gene suggest that changes in RNA stability may be also contributing to transcript accumulation. TDP-1 is orthologous to mammalian TDP-43, and we have previously shown that human TDP-43 can act as an RNA chaperone in an *in vitro* assay [174]. Conceivably, heat shock could inhibit the function of TDP-1 or other similar RNA binding proteins, leading to the formation of more dsRNA structure in existing transcripts.

We employed fluorescence in situ hybridization (FISH) to confirm heat shock induced expression of DoG and antisense transcripts in the *eif-3.B* region, and to examine their subcellular localization. These novel transcripts were found in nuclear foci that did not overlap with the J2 dsRNA foci, and were typically limited to two spots in each nuclei. This two foci distribution is very similar to the FISH characterization of DoG transcripts described by Vilborg et al, and strongly suggest that the *eif-3.B* loci transcripts are associated in *cis* with their site of production. These antisense transcripts clearly did not contribute to the foci detected by J2 immunostaining, and may reflect a general dysregulation of transcription at the *eif-3.B* locus.

Identification of the dsRNA species present in the J2 foci induced by heat shock may require development of a protocol to purify these RNA granules, as we have identified thousands of transcripts enriched in the J2 pool, and have no additional insight as to which ones might be found specifically in the J2 foci.

A critical issue is whether the accumulation of novel transcripts and dsRNA after heat shock have a biological function. By characterizing transcriptional changes induced by a variety of stresses, Vilborg et al concluded that transcriptional read through was not a random failure, and suggested it might have a functional role in stress responses. We have characterized the accumulation of dsRNA after heat shock, and by gene ontology analysis find that the sense and antisense transcripts in this pool (as well as the J2 IP pool in *tdp-1* deletion mutants) are enriched in genes involved in translation. Given that we find significant J2 IP enrichment of both sense and antisense transcripts from genes related to translation, it is tempting to speculate that the formation of interstrand dsRNA might reduce the translation of these "translation related transcripts", leading to a down regulation of global translation, a protective event against most cellular stress insults including heat shock. While we have no direct evidence that dsRNA dependent translational downregulation happens after heat shock in *C. elegans*, we note that deletion of *tdp-1* has been reported to protect against proteotoxicity and increase lifespan [306]. Translational downregulation would presumably be protective against proteotoxicity, and post developmental knockdown of translation initiation factors strongly increases lifespan in *C. elegans*[307].

Materials and methods

Caenorhabditis elegans culturing and strains

Hermaphrodites from each strain were kept at 16 ºC on Nematode Growth Media (NGM) plates seeded with Escherichia coli strain OP50 as a food source according to standard practices [308]. To obtain age synchronized worms, we used alkaline hypochlorite bleach on gravid adults to obtain eggs that were hatched overnight in S-basal buffer[309]. Worms were then allowed to grow to 1 day old adults (approximately 80h at 16ºC).

Heat stress treatment

Heat stress treatment was performed in an air incubator set to 35 ºC for 3 hours for the RNA-seq experiments. After stress, populations were washed off with S-basal buffer and immediately fixed for immunohistochemistry or fluorescence in situ hybridization (FISH), flash frozen in liquid nitrogen for quantitative reverse transcriptase polymerase chain reaction (qRT-PCR), or crude extracts were created with subsequent J2 Immunoprecipitation (J2 IP) as previously described[174].

RNA isolation, cDNA library preparation, and RNA Sequencing

Total RNA was extracted from worms using TRIzol (Invitrogen #15596026) extraction and used as input RNA. Chloroform was used to solubilize proteins and TURBO DNAase (Invitrogen) was used to remove DNA. For input RNA libraries, 5 µg of RNA was ran through a

RiboZero column (Epicenter, #R2C1046) to remove ribosomal RNA. Libraries were created using Illumina TruSeq kits (RS-122-2001). RNA recovered by immunoprecipitation with the J2 antibody of young adult worms as well as input material (as a loading control) was converted into strand-specific total RNA libraries using V2 Scriptseq (Epicenter #SSV21106) kits following manufacturer's instructions, except reverse transcription was done with SuperScript III (Invitrogen #18080 044) using incrementally increasing temperatures from 42 to 59 °C to allow for transcription though structured RNAs. rRNA was not removed from J2 IP RNA samples. Libraries were sequenced on an Illumina HiSeq 2000 platform at the Genomics Core at the University of Colorado, Denver. Data were deposited under GEO accession number GSE120949.

Immunohistochemistry and Fluorescence in situ Hybridization (FISH)

For immunohistochemistry, all washes used a constant volume of 1ml sterile S-basal buffer unless otherwise noted. Worms were first washed off plates, spun down into a pellet, and fixed in 4% paraformaldehyde. Worms were then resuspended in 1ml of Tris-Triton buffer with 5% beta-mercaptoethanol and incubated in a rocker for two days at 37°C. After two days, worms were washed two times and put into collagenase buffer. Next, worms were placed into a 1:1 dilution of 1mg/ml type IV collagenase (Sigma) and S-basal buffer for 45 minutes at 37 ºC with rocking. Worms were checked under the microscope to ensure cuticle breakage then quenched in cold Antibody buffer A (1X Phosphate buffered saline, 0.1% Bovine Serum Albumin, 0.5% Triton X-100, 0.05% Sodium Azide). Worms were then washed, pelleted, and primary antibodies were added for 16 hours at 4°C. Next, worms were washed twice in Antibody buffer B (same as Antibody buffer A except using 1% Bovine Serum Albumin), pelleted, and secondary antibodies were added with subsequent incubation for 2 hours at room temperature. Finally, worms were washed twice in Antibody buffer B and then placed in 50ul of Antibody buffer A. Permeabilized worms were probed with the primary J2 antibody (English and Scientific Consulting Lot: J2-1102 and J2-1103) at 4μg/mL and secondary antibody Alexa dye-conjugated goat anti-mouse at 4μg/mL. DAPI nuclear stain was added along with secondary antibodies at 5μg/mL to visualize nuclei.

Stellaris FISH probes (Biosearch technologies) [310] were custom designed using the Stellaris RNA FISH probe designer. Three regions were chosen for probing, and each probe was tested against the *C. elegans* genome using BLAST to identify any complementarity to non-target sequences. A probe was excluded if it was in an intron, had a highly repetitive sequence outside of the region, or matched other regions up to 18nt long with high transcriptomic expression viewed in the Integrative Genome Viewer (IGV) [311].

For FISH probing and storage, the Stellaris protocol for *C. elegans* was followed using RNAase OUT (Invitrogen) when applicable. Briefly, worms were washed off plates using nuclease-free water and fixed for 45 minutes at room temperature in a fixation buffer (1:1:8 of 37% formaldehyde, 10X RNAase-free phosphate buffered saline (PBS), nuclease-free water). Worms were then washed twice with 1X RNAase-free PBS and permeabilized in 70% ethanol overnight at 4°C. Worms were then incubated at room temperature in Stellaris Wash Buffer A, pelleted, and incubated for 16 hours in a 37 ºC water bath in the dark with 100μl of the hybridization buffer (9:1 of μl Stellaris RNA FISH hybridization buffer, deionized formamide with a 100:1 Hybridization buffer, FISH probe). Next, 1mL of Stellaris Wash Buffer A was added with 30 more minutes of incubation in the dark 37 ºC water bath. Stellaris Wash Buffer A was then aspirated and incubated with DAPI (1:1000 of 5µg/ml DAPI, Stellaris Wash Buffer A) for 30 more minutes of the dark 37 ºC water bath. Lastly, the DAPI buffer was aspirated and 1mL of Stellaris Wash Buffer B was added with a 5 minute room temperature incubation.

A modification of the immunohistochemistry protocol was used when doing immunohistochemistry and FISH. The immunohistochemistry protocol was the same except all washes were done using RNAase-free PBS or water and RNAase-free reagents (Tris-Triton buffer, collagenase buffer, collagenase, Antibody Buffer A, Antibody Buffer B) were created by adding RNAase OUT (2:10000 of RNAase OUT, reagent). After antibody staining, the FISH protocol was started at the hybridization step.

Microscopy

Images were acquired with a Zeiss Axiophot microscope equipped with digital deconvolution optics (Intelligent Imaging Innovations). Image brightness and contrast were digitally adjusted in Photoshop.

Quantification of coincidence of J2 and HSF-1 foci over time and J2 foci in mutant strains

For the quantification of coincidence of foci over time, intestinal nuclei of the worms were isolated from the rest of the image and the Foci Picker3D plug-in was used to count foci. The FITC channel of the image was converted to 16 bit and analyzed. Foci Picker3D settings were changed from default by changing the Minlsetting to 0.25 and the ToleranceSetting to 20 before running analysis. 19-20 worms were selected for each time point. Analysis of variance (ANOVA) and Tukey honestly significant difference (HSD) post-hoc analysis were used for multiple comparisons between conditions with a significance threshold of < 0.05 (Family Wise Error Rate). The script for the ANOVA and post-hoc is available at https://github.com/Senorelegans/heatshock_and_tdp-1_dsRNA_scripts/tree/master/fig1_coincidents_of_foci/fig_supplemental_graph_foci

Quantification of J2 foci in mutant strains with or without heat shock was done blindly. 8 worms per condition were counted and the average number of foci from 2-4 intestinal nuclei were used to make the box plots. Analysis of variance (ANOVA) and Tukey HSD post-hoc analysis were used for multiple comparisons between conditions with a significance threshold of < 0.05 (Family Wise Error Rate). The script for creating boxplots and ANOVA and post-hoc is available at https://github.com/Senorelegans/heatshock_and_tdp-1_dsRNA_scripts/tree/master/fig_supplemental_graph_foci

Sequencing data analysis

Reads were checked for quality with FastQC v0.11.7 [312], adapters were trimmed using Trimmomatic-0.36[313], and reads were aligned to the worm genome WS258 using STAR-2.5.2b [314]. Genes and DoGs (identified by Dogcatcher, described below) were assigned counts using Rsubread v1.28.1 featureCounts [315] and were rRNA normalized according to the rRNA subtraction ratio (RSR) (described in supplemental). Differential expression was obtained using DESeq2 v1.20.0 and the likelihood ratio test (LRT) set with input and J2 groups treated as separate variables within the condition[20].

We created an algorithm called Dogcatcher to identify and analyze DoGs. Briefly, Dogcatcher uses a sliding window approach to identify contiguous regions of transcription above a defined threshold. If the sliding window runs into a gene on the same strand it will either continue (meta read through) or stop (local read through). Dogcatcher outputs bedfiles, gtf's and dataframes of all DoGs and antisense DoGs identified within a sample along with differential expression and genes overlapping DoGs. For improved normalization in DESeq2, non-significant genes are added when calculating differential expression and removed for visualization. The Dogcatcher algorithm and README is available at https://github.com/Senorelegans/Dogcatcher. For processing J2 enrichment, a modified version of Dogcatcher was used that applies the likelihood ratio test from DESeq2 (available at https://github.com/Senorelegans/heatshock_and_tdp-1_dsRNA_scripts/J2_enrichment_Dogcatcher).

DoGs identified by Dogcatcher that overlap operons on the same strand were removed.
All of the scripts used to process the data and create figures can be found at https://github.com/Dogcatcher/heatshock_and_tdp-1_dsRNA_scripts

IV. Mystery Miner: Application of a bioinformatic pipeline to RNA-seq data identifies novel virus-like sequence in human blood

Contributions from fellow researchers

I would like to thank Patrick Gonzales for implementing the LAST alignment and protein BLAST step. I would like to thank Thomas J. LaRocca for consultation on the early development of the pipeline. I would like to thank the researchers from the Mayo Clinic (Yuping Song, Joanne Wuu, Michael Benatar, Björn Oskarsson, Leonard Petrucelli, Mercedes Prudencio) for performing all of the wet lab experiments including study design, sample collection for sequencing, and RT-PCR.

Citation

All of this chapter was published as Marko Melnick, Patrick Gonzales, Thomas J. LaRocca, Yuping Song, Joanne Wuu, Michael Benatar, Björn Oskarsson, Leonard Petrucelli, Robin D. Dowell, Christopher D. Link, Mercedes Prudencio in Application of a bioinformatic pipeline to RNA-seq data identifies novel virus-like sequence in human blood. *G3 Genes|Genomes|Genetics*. April 2021. doi:10.1093/g3journal/jkab141

Introduction

Numerous reports suggest that microbes could play a role in neurodegenerative diseases. Microbial sequences are routinely identified in human RNA sequencing (RNA-seq) data[316], which is typically acquired to assay gene expression. The origins of these microbial sequences are generally unknown, although in theory disease-relevant microbes could be identified if their sequences are significantly enriched in patients compared to controls. We therefore sought to develop a bioinformatic pipeline that could identify microbial sequences over-represented in RNA-seq data from patients compared to controls. Importantly, our pipeline can recover both known and novel microbial sequences.

Infection has been proposed to play a role in multiple neurodegenerative diseases[317], including amyotrophic lateral sclerosis (ALS)[318]. ALS is the most common motor neuron disease in adults, with the majority of individuals dying within 3-5 years of symptom onset. The disease is defined by the degeneration and death of motor neurons in the brain and spinal cord, resulting in progressive weakness and eventually death, typically from respiratory muscle weakness[319]. Around 10% of ALS patients have a family history that suggests an autosomal dominant inheritance which is classified as familial ALS (fALS), with the remaining 90% of patients classified as having sporadic ALS (sALS)[165]. After decades of study, the etiology of sALS remains a mystery, although suspected risk factors for ALS include exposure to heavy metals, pesticides, chemical solvents, cigarette smoke, and unidentified factors related to US military service[320–323]. Along with these environmental risk factors, there has been a long history, with variable success, in the search for pathogens that might contribute to ALS[246–250] and other neurodegenerative diseases such as Alzheimer's disease (AD)[239–241], Parkinson's disease (PD)[242–244], and multiple sclerosis (MS)[245].

Studies on ALS primarily come from European populations and within these populations four genes [TAR DNA-binding protein 43 (TDP-43), fused in sarcoma/translocated liposarcoma

(FUS), superoxide dismutase 1 (SOD1), *chromosome 9 open reading frame 72 (C9ORF72)*] account for 70% of fALS[166]. Of these four genes, *C9ORF72* accounts for up to 30-50% of cases in fALS and 7% of sALS (in all populations)[165]. In *C9ORF72*-associated ALS (c9ALS), a hexanucleotide repeat expansion (HRE) occurs that can form RNA with highly stable parallel G-quadruplex structures (G4 RNA). How neurodegeneration occurs from HRE in c9ALS is not well understood, but putative mechanisms include reduction of *C9ORF72* expression, production of poly-dipeptides as a result of Repeat Associated Non-AUG (RAN) translation of repeat sequences, and the formation of RNA foci that may sequester RNA binding proteins[176,177]. Identifying disease modifiers is of significant translational interest, as it is currently unknown how patients with c9ALS (sporadic or familial) progress from asymptomatic to symptomatic states. Evidence is mounting that persistent immune activation can play a causative role in disease progression, and some recent treatments focus on reducing the elevated neuroinflammation that occurs in patients with the HRE[324]. Indeed, one study showed that a lower abundance of immune-stimulating bacteria contributes to reduced inflammation and protection from premature mortality in a *C9orf72* loss-of-function mouse model[325].

Numerous studies have looked for biomarkers of ALS[326] using metabolomics[327,328], neuroinflammation[329,330], DNA methylation[331,332], gene expression[333], microRNA expression[334,335] and our previous study which analyzed protein levels of poly(GP) in c9ALS[336]. The search for pathogens using sequencing data from blood samples in ALS patients has been conducted by others [337–340], but previous efforts have not looked for novel pathogens. Next-generation sequencing (NGS) technologies have shown broad detection of pathogens in a target-independent unbiased fashion[341–344], however, designing a microbial detection experiment is non-trivial considering the variety of methods[345] and algorithms[346] that can be applied. Our primary goal when designing a new pipeline was to be conservative and unbiased with regards to discovery and quantification of novel pathogens. Furthermore, our intention was not to "reinvent the wheel" for microbiota classification, and instead opt to provide an end-to-end pipeline that leverages data across samples to obtain biologically significant fold changes of microbiota between diseased and healthy subjects.

While other pipelines have used reads that do not map to the host genome (unmapped reads) for microbial identification and quantification, these pipelines cannot be used for discovery as they rely on existing databases of microbial genomes[234–237]. One popular pipeline for viral classification that uses non-host reads includes ViromeScan[347], which utilizes a database of reference viral sequences to assign reads to taxonomic categories, but is "blind" to viral sequences not closely related to those in the database. Thus, we opted for de-novo assembly of unmapped reads into contigs, similar to the strategy employed by Kraken[348] and MetaShot[349]. Additionally, we use a hierarchical method to assemble unmapped reads into contigs (single samples, group, all) to increase the chance of assembling a correct contig from partial sequences that are present in multiple samples, and to remove outlier contigs present in single samples that are unlikely to contribute to the statistical analysis.

Where MetaShot stops at providing reads assigned to taxonomical categories, we map reads back to contigs and provide proper library normalization for statistical quantification. A similar pipeline known as IMSA[238] also maps reads back to contigs, but disregards contigs that might be identified by translated amino acid sequence similarity using BLASTX (a set we call the "dark biome") as well as subsequent contigs with no BLASTN or BLASTX hit (a set we call the "double dark biome").

We validated our pipeline by using datasets (synthetic and real) with known bacterial or viral infections. We also examined the differences in microbial identification between polyA and total RNA recovery in multiple tissues, and investigated the effects of globin depletion of blood samples. We then used our pipeline on a novel ALS blood dataset (termed "Our Study") as well as on five other published ALS datasets from blood or spinal cord samples, analyzed each dataset individually, and analyzed across datasets for changes in microbiota. While we did not identify any microbes enriched in the blood of ALS patients, we did identify and validate a novel virus-like sequence, demonstrating the potential of the bioinformatic pipeline we have established.

Results

Pipeline description

Our novel pipeline, Mystery Miner, is written as a Nextflow pipeline. Below is a short overview of the Mystery Miner pipeline (Fig. IV-1). A more in-depth explanation, list of software and versions used, and all of the code used in this manuscript can be found at https://github.com/Senorelegans/MysteryMiner.

Raw reads were first checked for quality using FastQC then trimmed to remove both adaptor contamination and low quality basecalls using Trimmomatic. Trimmed reads were then mapped to the host genome using STAR for a fast first-pass followed by a 2nd pass with bowtie2 for sensitivity. Unmapped reads were retained for contig assembly. Filtering out host reads made downstream assembly faster and required less memory. We assembled contigs from unmapped reads with the SPAdes assembler (with "-rna" setting). This assembler was chosen for its recent use in studies of microbial diversity[350] and proven robustness to biological and technical variation[351]. The species each contig belongs to was identified with BLASTN using default settings, and the top hit for each contig was retained (a set we call "regular biome"). Contigs with no BLASTN hits were then filtered to remove highly repetitive regions (DUST). Next, contigs were retained if they had a greater than 60% pairwise alignment (LAST) between contigs assembled from a single sample, group/condition, or all samples (for example; contigs from groups that match singles are retained, we then use this new set to match with contigs from the all assembly).

We then identified contigs that lacked detectable nucleotide similarity to any GenBank entry but showed similarity at the amino acid level using BLASTX ("dark biome"). Furthermore, contigs with no BLASTN or BLASTX hits were labelled as "double dark biome" (also filtered by DUST and LAST). Every contig in the "regular biome" and "dark biome" were then queried against the Joint Genome Institute Server for additional taxonomic information. As Mystery Miner is an agnostic tool, it cannot distinguish between true tissue or cell-associated microbes and experimentally introduced contamination.

For quantification, we mapped the non-host reads using Bowtie2 to the contigs obtained from SPAdes. Next, we mapped reads to contigs using samtools mpileup (default mapq score) to calculate the amount of reads over each base pair in a contig. We then calculated coverage on the contigs by summing all of the counts for each base pair in a contig and dividing by the length of the contig. We then calculated normalized coverage by library size using the number of mapped reads to the host genome. This gave us normalized coverage (NC) for a contig or binned normalized coverage (BNC) for multiple contigs within a species/genus, etc. To assess statistical differences between conditions, a Student's *t*-test was calculated through NC or BNC, using the number of contigs or genus/species to obtain an FDR corrected adjusted p-value (q-value) using statsmodels in Python.

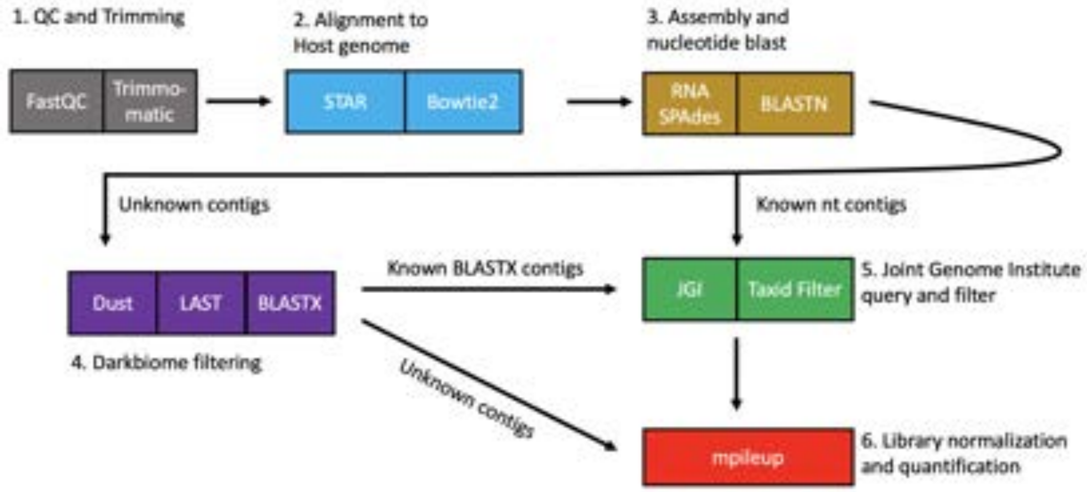

Figure IV-1: Diagram of Mystery Miner Pipeline

Reads were first checked with FastQC and trimmed using Trimmomatic (1. grey). Reads were then aligned to the host genome using various aligners (2. blue). Non-host (unmapped) reads were assembled into contigs with RNA SPAdes and regular biome contigs were identified with BLASTN (3. yellow). Unidentified contigs were filtered for repetitive sequences with Dust, filter by single, group or all with LAST, and dark biome contigs were identified with BLASTX. Double dark biome unidentified BLASTX contigs were sent directly to quantification (4. purple). Dark biome and regular biome contigs were assigned complete taxonomy using the JGI server and filtered one last time to remove mammalian/host genome contigs (5. Green). Non-host reads were then mapped to all contigs and normalized coverage was calculated for subsequent statistical analysis.

Validating Mystery Miner on datasets with known bacterial or viral infection

To confirm that Mystery Miner is able to recover and quantify known bacterial infections from sequencing data, we utilized an *in vitro* model of *Chlamydia trachomatis* infection from (Humphrys et al., 2016)[352]. In this study, epithelial cell monolayers were infected with *Chlamydia trachomatis*; and polyA RNA (6 samples) and total RNA (6 samples) were sequenced 1 hour and 24 hours post infection (hpi). Using the Mystery Miner pipeline, out of 5.32 X 10^6 reads from all of the samples, 6.04 X 10^5 reads remained unmapped (~11.34%) after trimming and mapping to the host genome. From these non-host reads, 3,257 contigs were assembled and 1,199 of these contigs were identified as regular biome. An additional 27 contigs had no BLASTN hit. Of these, we identified 2 dark biome (BLASTX identified) and no double dark biome (no BLASTX or BLASTN hit) contigs.

Using Mystery Miner we successfully identified, and found significantly elevated levels, of *Chlamydia trachomatis* (BNC by species) in 24 hours post infection (hpi) samples compared to 1 hpi samples in both polyA (q = 0.004) and total RNA (q = 0.0005). In addition to *Chlamydia trachomatis,* we identified 6 additional bacterial species and one viral species (Alphapapillomavirus 7) in the samples (Fig IV-2A), including significantly elevated levels of *Mycoplasma hyorhinis* contigs in total RNA samples. No significant differences were observed in the dark or double dark contigs.

To confirm that the pipeline can detect known viral infections, we ran Mystery Miner on a total RNA dataset from an *in vitro* model of severe acute respiratory syndrome coronavirus (SARS-CoV) 1 or 2 infection (Emanuel et al., 2020[353]). In this study human epithelial Calu3 cells were infected with SARS-CoV-1 or SARS-CoV-2 (4, 12, or 24 hours), mock (4 or 24 hours), or untreated (4 hours).

Out of the 2.81 X 10^8 reads obtained from all of the samples, 8.23 X 10^7 reads remained unmapped (~29%) after trimming and mapping to the host genome. From these non-host reads, 42,816 contigs were assembled, of which 1,346 regular biome, 27 dark biome, and 7 double dark biome contigs passed the filtering steps.

Mystery Miner successfully identified both SARS-CoV-2 and SARS-CoV-1 isolates and found significantly elevated levels of each virus compared to controls (Fig IV-2B). Hereafter we refer to SARS-CoV-1 or SARS-CoV-2 infected cells as COV1 or COV2 to avoid confusion with recovered names of contigs. Consistent with the viruses being similar, we identified both SARS-CoV-2 and SARS-CoV-1 in both the COV1-24hr and COV2-24hr samples when compared to mock-24hr. However, when we compared COV2-24hr to COV1-24hr, we were able to distinguish SARS-CoV-1 isolates from SARS-CoV-2 in the appropriate samples (i.e., SARS-CoV-2 was significantly elevated in COV2). Similar results were seen in the 12 hour samples but the 4 hour samples did not have sufficient viral reads to detect either SARS-CoV virus. To simulate a novel virus, we ran Mystery Miner on versions of the BLASTN and BLASTX databases obtained before SARS-CoV-2 was discovered and were able to properly identify SARS-CoV-2 as a bat related coronavirus[354].

Collectively, these data show that Mystery Miner is able to identify potential bacterial and viral infections, properly identify infected samples using quantification, and detect significant differences between infected samples and controls for bacteria, viruses, and isolates of a virus.

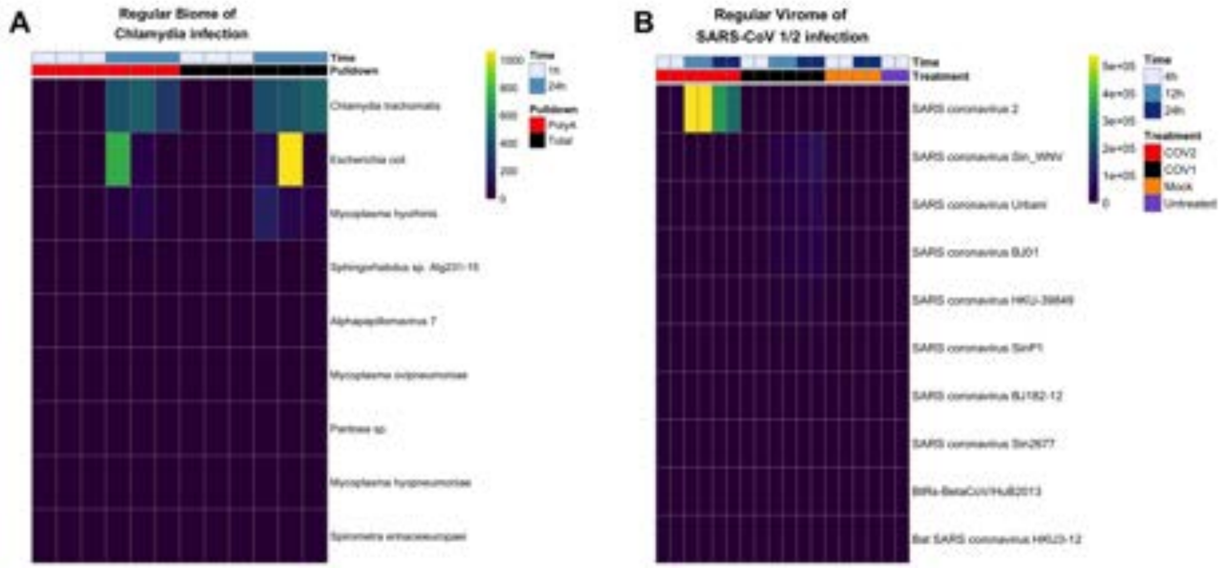

Figure IV-2: Heatmap of binned normalized coverage for bacterial or viral infected datasets

(A) Regular biome contigs binned by species from Humphrys et al., 2016. Time refers to 1or 24 hours post infection (hpi) of epithelial cell monolayers with *Chlamydia trachomatis* (blue). Pulldown refers to library enrichment for polyA RNA (red) or total RNA (black). (B) Regular virome of contigs binned by name from Emanuel et al., 2020 for SARS-CoV-2 infected cells (COV2) (red), or SARS-CoV-1infected cells (COV1) (black), mock virus (orange), or untreated sample (purple). Time refers 4,12, or 24 hpi of Calu3 cells with indicated virus (blue). Top 10 hits per experiment shown for brevity.

Validating Mystery Miner on a synthetic minibiome

We next looked at the detection and quantification limits of Mystery Miner using generated read data to create a synthetic minibiome. We used Polyester[355] to generate paired end read data (100bp read size) at various coverage levels and various fold change differences between two groups (Group A, Group B) with 10 samples each (20 samples total). Our synthetic minibiome consists of 10 human sequences and 10 sequences from non-human organisms (4 pathogenic, 6 commensal). The first four organisms in the synthetic minibiome are SARS-CoV-1, SARS-CoV-2, *Chlamydia trachomatis*, and *Chlamydia pneumoniae*. The next 6 (*Mageeibacillus indolicus*, *Prevotella melaninogenica*, *Filifactor alocis, Mobiluncus curtisii, Rothia dentocariosa, Aeromicrobium marinum*) are commensals that are part of the representative bacteria list from the Human microbiome project[356].

For the human sequences, we first generated a pool of human reads using the first 10kb of 10 scaffolds from chromosome 22 (default value for human read generation in Polyester) at 1000x coverage with no fold change differences between groups. For non-human organisms, we took the first 10kb of the nucleotide sequence for the organism and generated reads at coverage levels of 1000x, 100x, 10x, 1x, 0.1x, and 0.01x. Lastly, we combined the 1000x coverage human reads separately with each level of coverage for non-human organisms and ran Mystery Miner (6 pipeline runs in total).

We found sequences below 1x coverage did not assemble, suggesting that this is our limit of detection (all further data omits 0.1x and 0.01x coverage). For the SARS strains, we successfully identified both strains at 1000x coverage but found that with lower coverage levels, SARS-CoV-1 was identified as a SARS-related coronavirus. This ambiguity is likely due to the 73% nucleotide sequence identity (aligned with CLUSTAL OMEGA[357]) between the first 10kb of SARS-CoV-1 and SARS-CoV-2. For the selected *Chlamydia species* (59% sequence identity of the first 10kb) and the rest of the commensal bacteria, we were able to successfully assemble and correctly identify each species at every level of coverage.

Along with identification, we looked at Mystery Miners ability to quantify fold change differences between groups (A and B) using the synthetic minibiome. For the four pathogenic organisms, we selected one sequence from each kingdom to have a 2 fold difference (SARS-CoV-2, *Chlamydia trachomatis).* For the 6 commensals, we chose the first three species to have fold change differences of 1.8, 1.5, and 1.3 (*Mageeibacillus indolicus*, *Prevotella melaninogenica*, *Filifactor alocis)*. For SARS, we found that at 1x coverage, the 2 fold difference of SARS-CoV-2 was correctly called significant (q = 5.14 e^{-10}), but the ambiguously identified SARS-related coronavirus contig was not called significant (q = 0.489). At 1000x coverage, we found that the correctly identified SARS-CoV-1 contig was falsely called significant (q = 0.0028), this is likely due to ambiguous read mapping from the closely related SARS-CoV-2 sequence, as mentioned above. We found similar results for each coverage level (from 1x to 1000x) for the rest of the organisms and will subsequently use values from 1x coverage as that is the lowest level of detection. For *Chlamydia,* we found Mystery Miner successfully called *Chlamydia trachomatis* significant (q = 3.57 e^{-10}) and *Chlamydia pneumoniae* not significant (q = 0.709). For the commensals with FC differences, we successfully called each one significant [*Mageeibacillus indolicus (q = 6.92 e^{-7}*), *Prevotella melaninogenica (q = 4.91 e^{-5}*), *Filifactor alocis (q = 0.017*)] (Fig. IV-3). Using synthetic data, we conclude that Mystery Miner is able to identify organisms down to the species level and correctly call significant fold changes at low levels of coverage but has difficulty from ambiguity when reads come from highly similar sequences (72% >).

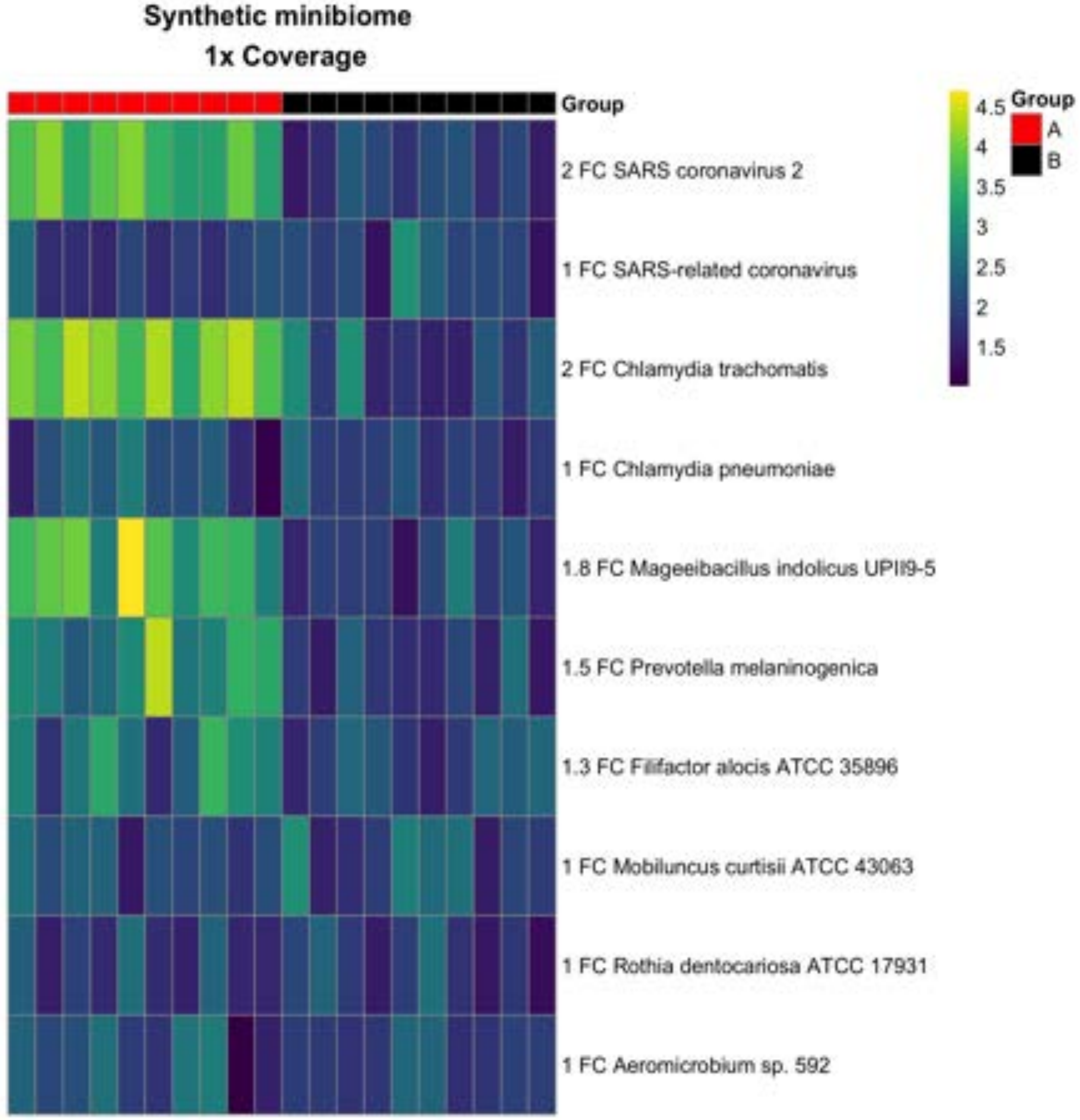

Figure IV-3: Heatmap of coverage of synthetic minibiome (1x coverage).

Heatmap of coverage for synthetic minibiome at 1x coverage. Fold change (FC) in the row name refers to group A (red) over group B (black). The first four rows are pathogenic organisms, the next 6 rows are commensals identified from the human microbiome project.

Effects of library pulldown or globin depletion in RNA-seq datasets

In order to compare effects of library enrichment or depletion, we compared recovered pathogens in a dataset that has polyA enrichment or rRNA depleted total RNA from blood or colonic tissue (VonSchack et al., 2018)[358]. When we compared polyA RNA vs total RNA and looked at BNC by superkingdom of bacteria we found significantly more reads map to bacteria for total RNA than polyA RNA (q = 0.0349), in blood but not in colon (q = 0.11709) (Fig. IV-4). We found similar amounts of significant BNC by species for polyA RNA vs total RNA in blood (29) and in colon (26). We then looked at significant BNC by genus and found double the amount in blood (14) compared to colon (7), with only one significant genus (*Actinomyces*) found in both comparisons. We did not find any significant differences in coverage when we looked at the species, genus or superkingdom level for viruses. We conclude that library enrichment with total RNA compared to polyA RNA increases bacterial recovery and diversity in blood.

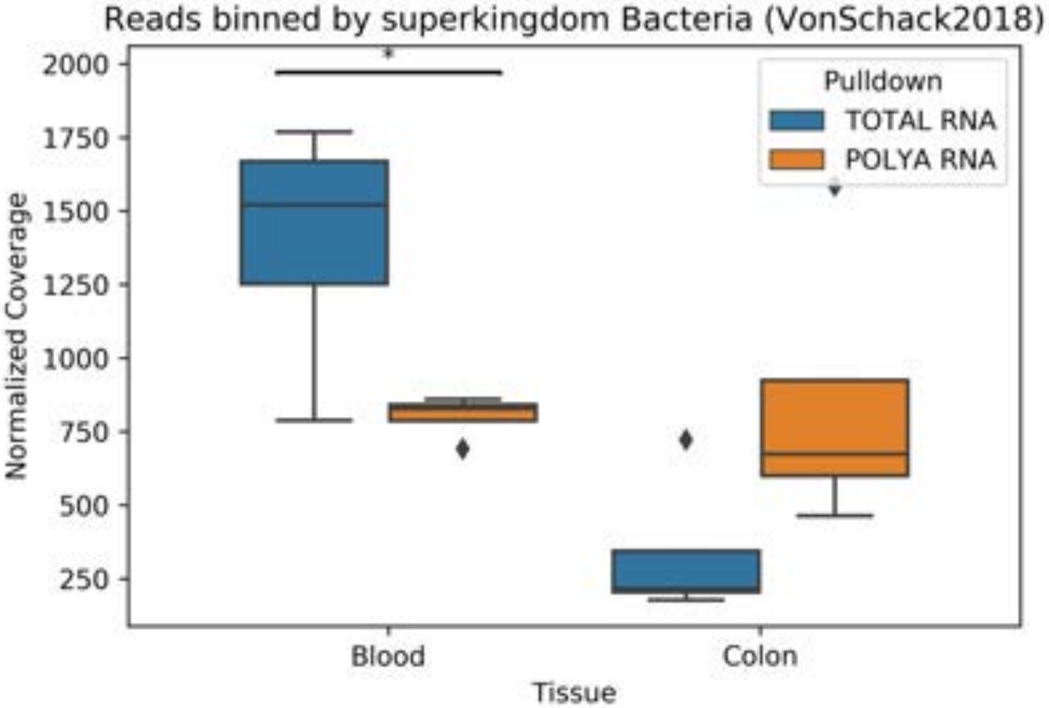

Figure IV-4: Boxplot of normalized coverage for superkingdom Bacteria in VonSchack et al., 2018

Boxplot of normalized coverage of regular biome contigs binned by superkingdom Bacteria. Blood shows significantly more reads in total RNA vs polyA RNA compared to Colon tissue.

We next looked at a RNA-seq dataset from whole blood with globin depleted (GD) vs non-globin depleted (NGD) total RNA (Shin et al., 2014[359]). With BNC by superkingdom, we found significantly increased levels in globin depleted vs. not-depleted samples for both bacteria (q = 0.004) (Fig. IV-5) and viruses (q = 0.030) (Fig. IV-6). We also found significant differences in BNC by species or genus primarily from *E. coli* with elevated levels in globin-depleted blood RNA. We did not find any significant differences when we looked for viruses at the species or genus level.

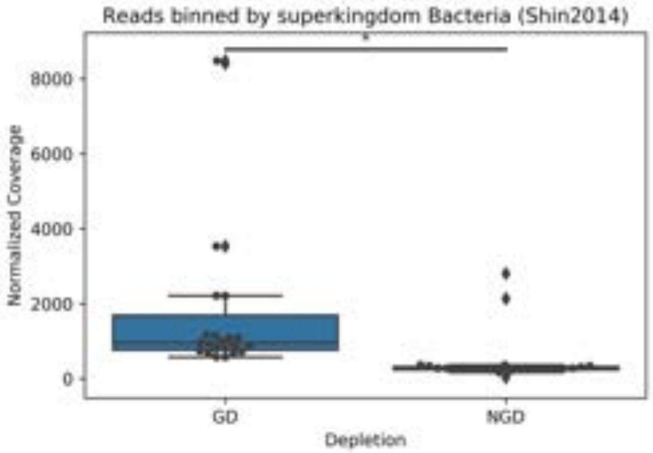

Figure IV-5: Boxplot of normalized coverage for superkingdom Bacteria in Shin et al., 2014

Boxplot of normalized coverage of regular biome contigs binned by superkingdom Bacteria. Globin depletion (GD) has significantly more coverage than non-globin depleted (NGD) blood.

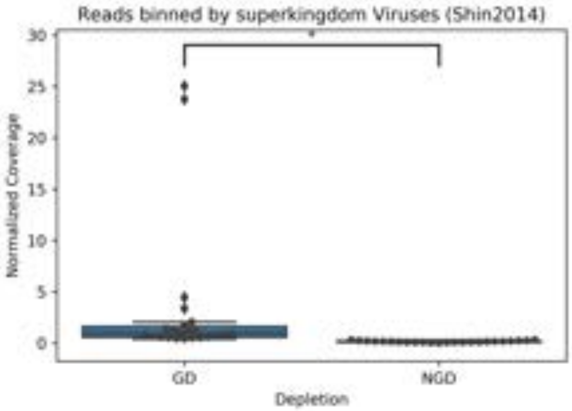

Figure IV-6: Boxplot of normalized coverage for superkingdom Viruses in Shin et al., 2014

Boxplot of normalized coverage of regular biome contigs binned by superkingdom Viruses. Globin depletion (GD) has significantly more coverage than non-globin depleted (NGD) blood.

Analysis of Our Study

We used Mystery Miner on our novel RNA-seq dataset of globin depleted and rRNA depleted total blood RNA from 120 individuals. These samples were from four subject groups including healthy control participants (CTL), ALS symptomatic *C9ORF72* negative patients (SYM), *C9ORF72* positive ALS symptomatic patients (C9S) and *C9ORF72* positive asymptomatic individuals (C9A).

The entire dataset contains a combined 8.64 X 10^9 reads. Approximately 2.7% (2.34 X 10^8) of the reads did not map to the human genome. From these non-host reads 2,976,988 contigs were assembled and 17,047 BLASTN contigs (regular biome) were identified. A total of 25,815 contigs had no BLASTN hit and after filtering we identified 2,980 dark biome (BLASTX identified) and 859 double dark biome (no BLASTX or BLASTN hit) contigs.

In general, we found a modest positive correlation between library size and number of bacterial contigs assembled, species detected (Fig. IV-7), and genera detected for each sample as well as a homogenous mixture of samples with respect to disease status. No specific taxonomy or contig sequence correlated with participant class within the dataset. By pooling bacterial read counts across all of the samples, we found *alpha proteo-bacteria*, *Actinobacteria, Firmicutes,* and *Bacteroidetes* as the most highly represented taxonomies, consistent with other blood biome studies[360] (Fig. IV-8). Most of the bacterial genera (~65%) assigned to the dark biome contigs were also found in the regular biome, however this was not the case in the viral sets, as only 5% (4/78) of dark viral contigs were observed in the regular biome. This observation suggested that our pipeline might be identifying novel viral sequences.

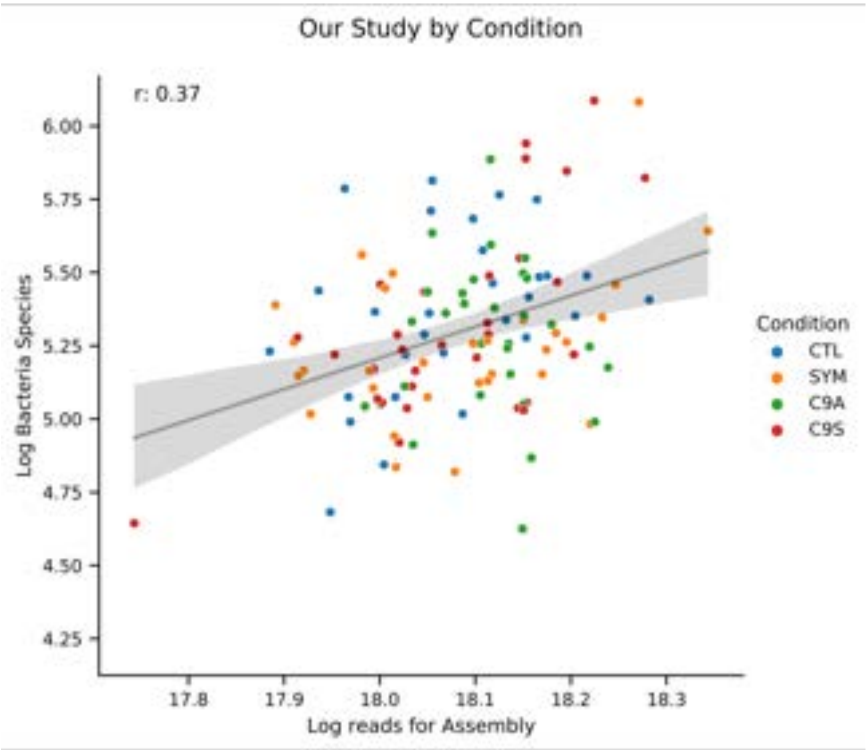

Figure IV-7: Log number of bacterial species vs Log reads for Assembly in Our Study

Scatterplot where each dot is a sample from a dataset with log number of bacterial contigs assembled on the Y-axis and Log reads used in SPAdes on the X-axis. Samples show a modest correlation (Pearson's r=0.37) between library size and bacterial species recovered. Data fit with a regression (black line) and 95% confidence interval (gray area).

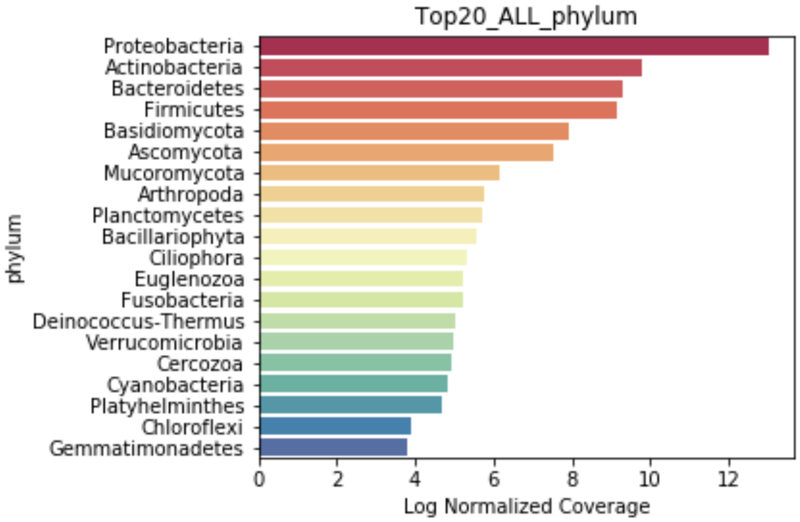

Figure IV-8: Log normalized coverage binned by phylum from our ALS dataset

Log normalized coverage is summed for all of the samples and *alpha proteo-bacteria*, *Actinobacteria, Firmicutes,* and *Bacteroidetes* are the most highly represented.

Within the dark biome contigs, we noted numerous contigs with a region of protein sequence similarity to RNA-dependent RNA polymerase (RdRP) from multiple RNA viruses, including the velvet tobacco mottle virus (first row in heatmap of Fig. IV-9). Our attention was drawn to the largest (~5 kb) dark biome contig hereafter labeled as “RDRP contig”. This large contig showed no nucleotide sequence similarity to any sequence in GENBANK, and no protein sequence similarity except for a long open reading frame with significant similarity to viral RDRPs (BLASTX P ~ 1e^{-26}). A phylogeny based solely on viral RDRP protein sequences places the RDRP contig closest to single-stranded (+) viruses of the *Barnavirus*, *Sobemovirus*, and *Polerovirus* genera (Fig. IV-10). However, given the absence of detectable similarity in this contig to other (non-RDRP) viral proteins of these genera, the relationship of the contig sequence to other virus groups is unclear, which supports the view that this contig represents a novel viral sequence.

To confirm the presence of the RDRP contig in the original samples, we designed primers to the RDRP contig and performed reverse transcriptase polymerase chain reaction (RT-PCR) on seven samples, four of which had high coverage (predicted present) and three with zero coverage (predicted absent). We found typical levels for detection of a virus[75] in the samples with high coverage and detected no signal above background in samples with zero coverage (Table IV-1). We conclude that Mystery Miner can recover true novel sequences that could represent previously unknown pathogens.

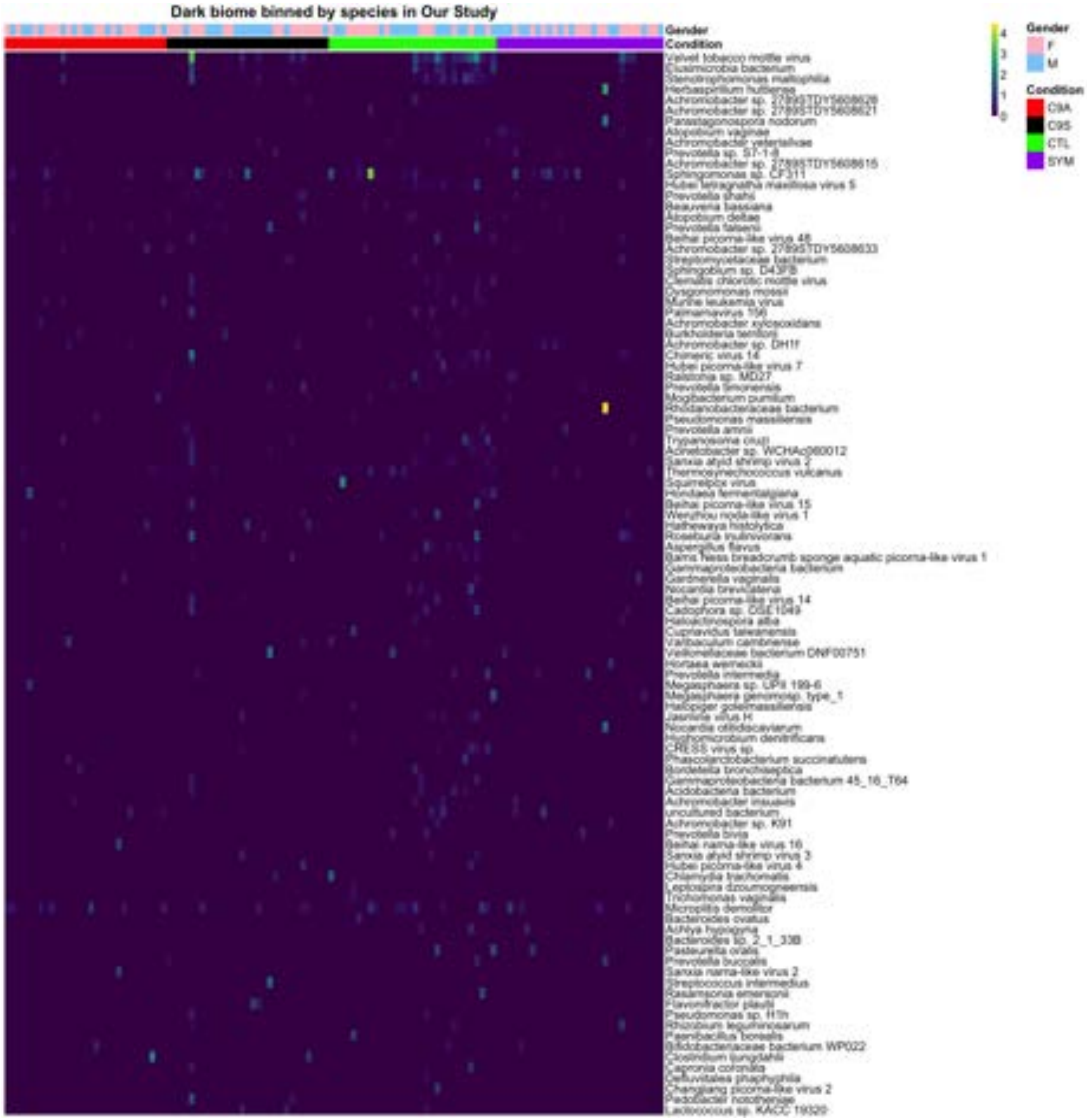

Figure IV-9: Heatmap of dark biome contigs binned by species in Our Study

Heatmap of normalized coverage of dark biome contigs binned by species. The highest coverage belongs to contigs that show high similarity to velvet tobacco mottle virus. Zero coverage is dark blue and goes to yellow with increasing values. These samples were from four subject groups including healthy controls [(CTL) green], *C9ORF72* negative ALS symptomatic [(SYM) purple], *C9ORF72* positive ALS symptomatic [(C9S) blue] and *C9ORF72* positive asymptomatic [(C9A) red] patients. Sex indicated as light blue (male) and pink (female). Top 100 species sorted by binned normalized coverage shown for brevity.

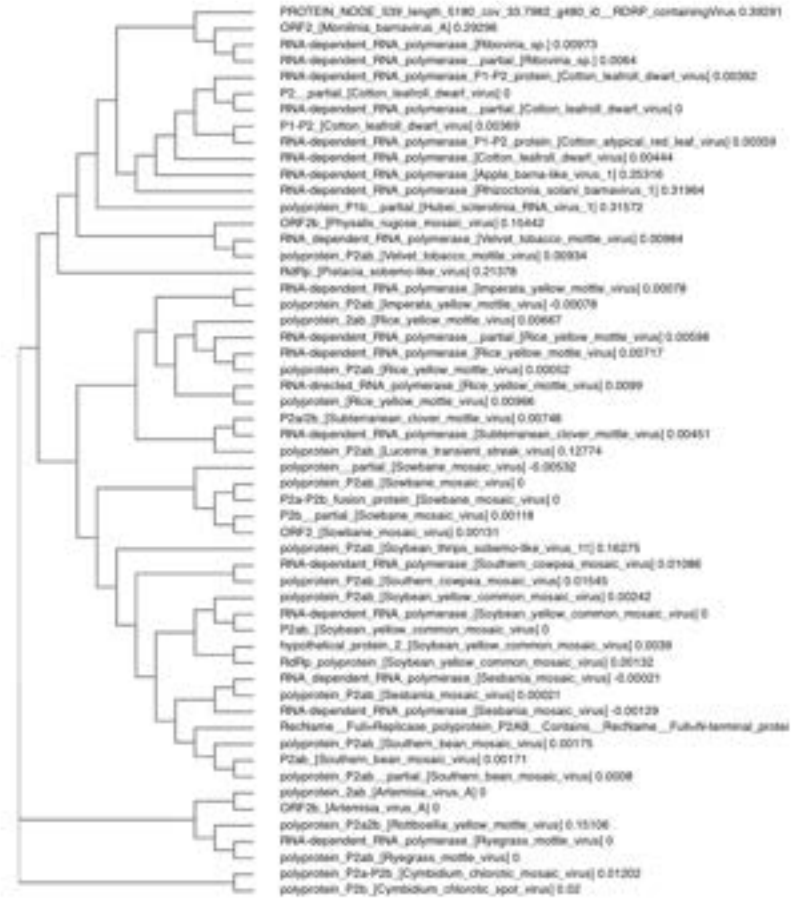
PROTEIN_NODE_539_length_5180_cov_33.7982_g480_i0__RDRP_containingVirus 0.39291
ORF2_[Monilinia_barnavirus_A] 0.29296
RNA-dependent_RNA_polymerase_[Ribovirus_sp.] 0.00973
RNA-dependent_RNA_polymerase__partial_[Ribovirus_sp.] 0.0064
RNA-dependent_RNA_polymerase_P1-P2_protein_[Cotton_leafroll_dwarf_virus] 0.00392
P2__partial_[Cotton_leafroll_dwarf_virus] 0
RNA-dependent_RNA_polymerase__partial_[Cotton_leafroll_dwarf_virus] 0
P1-P2_[Cotton_leafroll_dwarf_virus] 0.00369
RNA-dependent_RNA_polymerase_P1-P2_protein_[Cotton_atypical_red_leaf_virus] 0.00059
RNA-dependent_RNA_polymerase_[Cotton_leafroll_dwarf_virus] 0.00444
RNA-dependent_RNA_polymerase_[Apple_barna-like_virus_1] 0.25316
RNA-dependent_RNA_polymerase_[Rhizoctonia_solani_barnavirus_1] 0.31964
polyprotein_P1b__partial_[Hubei_sclerotinia_RNA_virus_1] 0.31572
ORF2b_[Physalis_rugose_mosaic_virus] 0.10442
RNA_dependent_RNA_polymerase_[Velvet_tobacco_mottle_virus] 0.00984
polyprotein_P2ab_[Velvet_tobacco_mottle_virus] 0.00934
RdRp_[Pistacia_sobemo-like_virus] 0.21378
RNA-dependent_RNA_polymerase_[Imperata_yellow_mottle_virus] 0.00078
polyprotein_P2ab_[Imperata_yellow_mottle_virus] -0.00078
polyprotein_2ab_[Rice_yellow_mottle_virus] 0.00667
RNA-dependent_RNA_polymerase__partial_[Rice_yellow_mottle_virus] 0.00598
RNA-dependent_RNA_polymerase_[Rice_yellow_mottle_virus] 0.00717
polyprotein_P2ab_[Rice_yellow_mottle_virus] 0.00052
RNA-directed_RNA_polymerase_[Rice_yellow_mottle_virus] 0.0099
polyprotein_[Rice_yellow_mottle_virus] 0.00966
P2a/2b_[Subterranean_clover_mottle_virus] 0.00748
RNA-dependent_RNA_polymerase_[Subterranean_clover_mottle_virus] 0.00451
polyprotein_P2ab_[Lucerne_transient_streak_virus] 0.12774
polyprotein__partial_[Sowbane_mosaic_virus] -0.00532
polyprotein_P2ab_[Sowbane_mosaic_virus] 0
P2a-P2b_fusion_protein_[Sowbane_mosaic_virus] 0
P2b__partial_[Sowbane_mosaic_virus] 0.00118
ORF2_[Sowbane_mosaic_virus] 0.00131
polyprotein_P2ab_[Soybean_thrips_sobemo-like_virus_11] 0.16275
RNA-dependant_RNA_polymerase_[Southern_cowpea_mosaic_virus] 0.01086
polyprotein_P2ab_[Southern_cowpea_mosaic_virus] 0.01545
polyprotein_P2ab_[Soybean_yellow_common_mosaic_virus] 0.00242
RNA-dependent_RNA_polymerase_[Soybean_yellow_common_mosaic_virus] 0
P2ab_[Soybean_yellow_common_mosaic_virus] 0
hypothetical_protein_2_[Soybean_yellow_common_mosaic_virus] 0.0039
RdRp_polyprotein_[Soybean_yellow_common_mosaic_virus] 0.00132
RNA_dependent_RNA_polymerase_[Sesbania_mosaic_virus] -0.00021
polyprotein_P2ab_[Sesbania_mosaic_virus] 0.00021
RNA-dependent_RNA_polymerase_[Sesbania_mosaic_virus] -0.00129
RecName__Full=Replicase_polyprotein_P2AB__Contains__RecName__Full=N-terminal_protei
polyprotein_P2ab_[Southern_bean_mosaic_virus] 0.00175
P2ab_[Southern_bean_mosaic_virus] 0.00171
polyprotein_P2ab__partial_[Southern_bean_mosaic_virus] 0.0008
polyprotein_2ab_[Artemisia_virus_A] 0
ORF2b_[Artemisia_virus_A] 0
polyprotein_P2a2b_[Rottboellia_yellow_mottle_virus] 0.15106
RNA-dependent_RNA_polymerase_[Ryegrass_mottle_virus] 0
polyprotein_P2ab_[Ryegrass_mottle_virus] 0
polyprotein_P2a-P2b_[Cymbidium_chlorotic_mosaic_virus] 0.01202
polyprotein_P2b_[Cymbidium_chlorotic_spot_virus] 0.02

Figure IV-10: Protein BLAST phylogeny of closest hits to our RDRP contig

A protein BLAST phylogeny of closest hits to our RDRP contig (top row) aligned with CLUSTALW2 and built using Simple Phylogeny (both default settings). The RDRP contig closest to single-stranded (+) viruses of the *Barnavirus*, *Sobemovirus*, and *Polerovirus* genera.

Condition	Sample	GAPDH RT-PCR Ct Value	RDRP RT-PCR Ct Value	RDRP RNA-seq Normalized Coverage
SYM	LP00274	20.562019	36.401	1.56
C9S	LP00041	20.783213	36.346	3.39

C9S	**LP00192**	**20.899612**	**35.636**	**0.67**
C9A	**LP000180**	**19.982108**	**34.832**	**1.11**
C9S	**LP000183**	**20.176418**	**undetermined**	**0**
C9S	**LP000197**	**20.125161**	**undetermined**	**0**
C9A	**LP000157**	**20.062433**	**undetermined**	**0**

TABLE IV-1. RT-PCR and normalized coverage for RDP contig

Quantitative RT-PCR and normalized coverage results from the 5180 bp RDRP contig. For the RDRP contig positive samples (top 4) with high normalized coverage and detectable Ct values and negative samples (bottom 3) with no normalized coverage and undetectable Ct values. GAPDH was used as a positive control for qRT-PCR and shows comparable levels for all samples. These samples were from three conditions *C9ORF72* negative ALS symptomatic patients (SYM), *C9ORF72* positive ALS symptomatic patients (C9S) and *C9ORF72* positive asymptomatic individuals (C9A).

Analysis of published ALS datasets

We next sought to explore whether similar results would be obtained from other ALS datasets. To this end, we examined five other publicly available ALS datasets, consisting of two that used total RNA from blood (Linsley et al., 2014[361], Gagliardi et al., 2018[337]), and three datasets from spinal cord (Brohawn et al., 2016[362], Ladd et al., 2017[363], Brohawn et al., 2019[364]). We have provided a summary table of datasets (Table IV-2). As we observed in our study, we first noted that increased library size correlated with an increased number of bacterial contigs assembled, species detected, and genera detected (Figure IV-11).

Name	Groups	# Samples	Tissue	Pulldown
Humphrys2016	1- or 24-hours post infection with *Chlamydia trachomatis*	12	Cultured epithelial cell monolayers	PolyA Total RNA
VonSchack2018	PolyA or Total RNA from blood or colon	16	Whole Blood Colon	PolyA RNA Total RNA
Shin2014	Globin depleted Not globin depleted	24	Whole Blood	Total RNA
Emanuel2020	Severe acute respiratory syndrome coronavirus 1 or 2 infection Controls	18	Calu3 cells	Total RNA
Our Study	*C9ORF72* negative ALS, *C9ORF72* positive and symptomatic ALS, *C9ORF72* positive asymptomatic participants Controls	120	Whole Blood	Total RNA hemoglobin and rRNA depleted
Linsley2014	ALS type 1 diabetes, sepsis, multiple sclerosis patients before and 24 hours after the first treatment with IFN-beta Controls	134	Whole blood	Total RNA
Gagliardi2018	Sporadic ALS, ALS with mutations in *FUS, SOD1, TARDBP* Controls	20	Peripheral blood mononuclear cells	Total RNA

Brohawn2016	**ALS** **Controls**	**15**	**Cervical spinal cord**	**Total RNA** **rRNA depleted**
Ladd2017	**ALS** **Controls**	**10**	**Laser capture microdissection (LCM) to isolate cervical spinal cord motor neurons**	**Total RNA**
Brohawn2019	**ALS, Alzheimer's disease (AD), Parkinson's disease (PD)** **Controls**	**53**	**Cervical spinal cord**	**Total RNA**

TABLE IV-2. Studies used in the Mystery Miner analysis

The first three studies are only used to validate our pipeline. The six subsequent studies are ALS related from both blood and spinal cord.

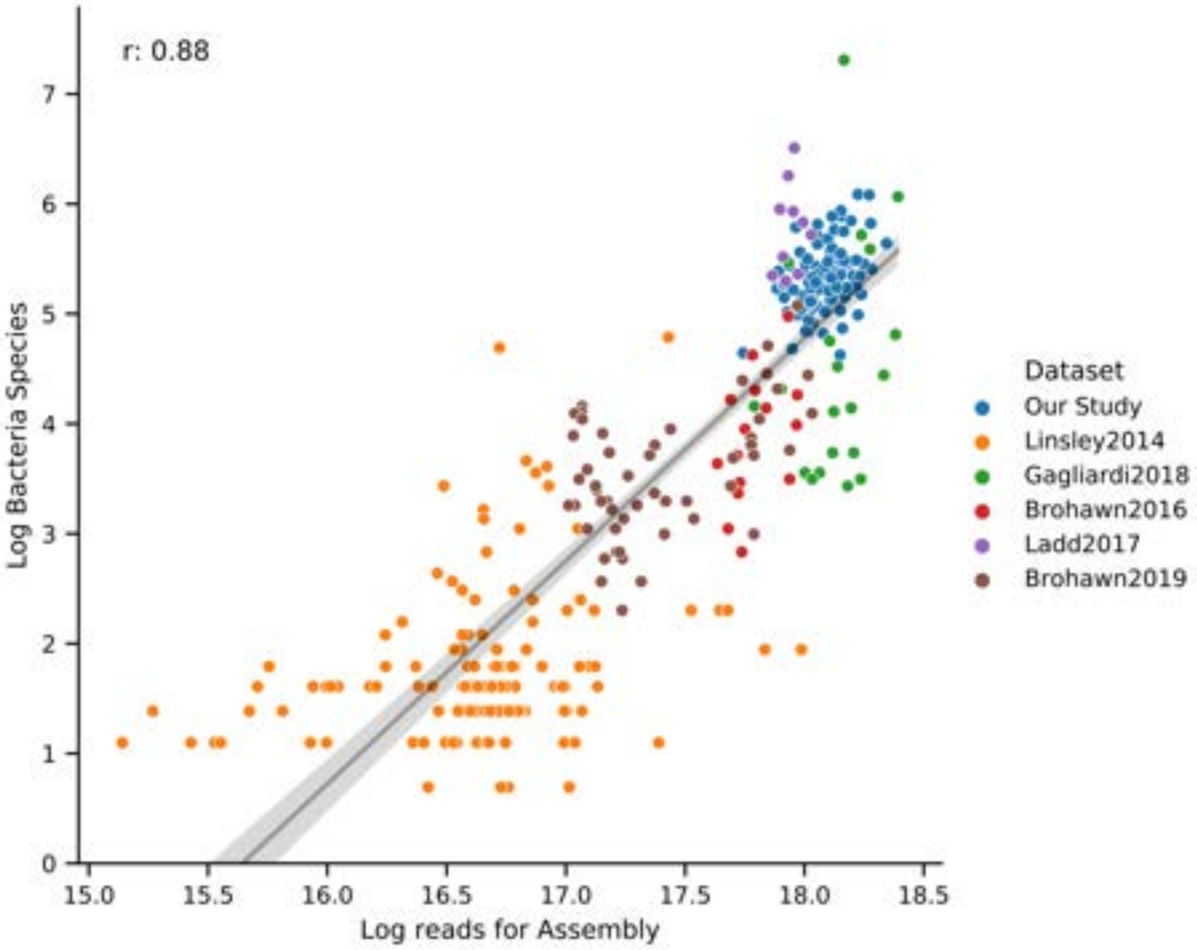

Figure IV-11: Log number of bacterial species vs Log reads for Assembly for ALS Datasets

Scatterplot where each dot is a sample from a dataset with log number of bacterial contigs assembled on the Y-axis and Log reads used in SPAdes on the X-axis. ALS datasets show a high correlation (Pearson's r = 0.88) between library size and bacterial species recovered. Data fit with a regression (black line) and 95% confidence interval (gray area).

We then looked at the total overlap of genus or species to determine if there are similarities in recovered microbial or viral sequences between datasets. For genus in the regular bacteriome, our dataset had the highest number of unique genus (185), followed by Ladd et al., 2017 (117), and Gagliardi et al., 2018 (38). The highest number of overlapping bacterial genus was between our dataset and Ladd et al., 2017 (133) followed by the intersection between our dataset, Ladd et al., 2017 and Gagliardi et al., 2018 (61) and there was a modest overlap (24) for

9/10 datasets (Fig. IV-12). We observed roughly the same trend in the regular bacterial biome at the species level and in the dark bacterial biome. In contrast, the regular virome of each dataset was relatively unique with very low amounts of overlap (<= 3) between datasets (species and genus shows a similar pattern). Interestingly, the highest overlap for species in the dark virome was between our dataset and Ladd et al., 2017 (13), one of which is similar to RDRP viruses, although the contigs in Ladd's data were not similar to the velvet tobacco mottle virus in our dataset.

In addition to comparing datasets using taxonomy, we also compared contigs between datasets for nucleotide similarity (> 70%) using LAST. We found that in general, datasets in the regular biome with the largest amount of contigs have the most overlap. Unsurprisingly, in the dark biome we observed less overlap by nucleotide similarity and that our RDRP contig does not share nucleotide similarity with contigs from any dataset. In addition, we also compared the nucleotide similarity of double dark biome contigs and found there is not a large percentage of similar contigs between datasets.

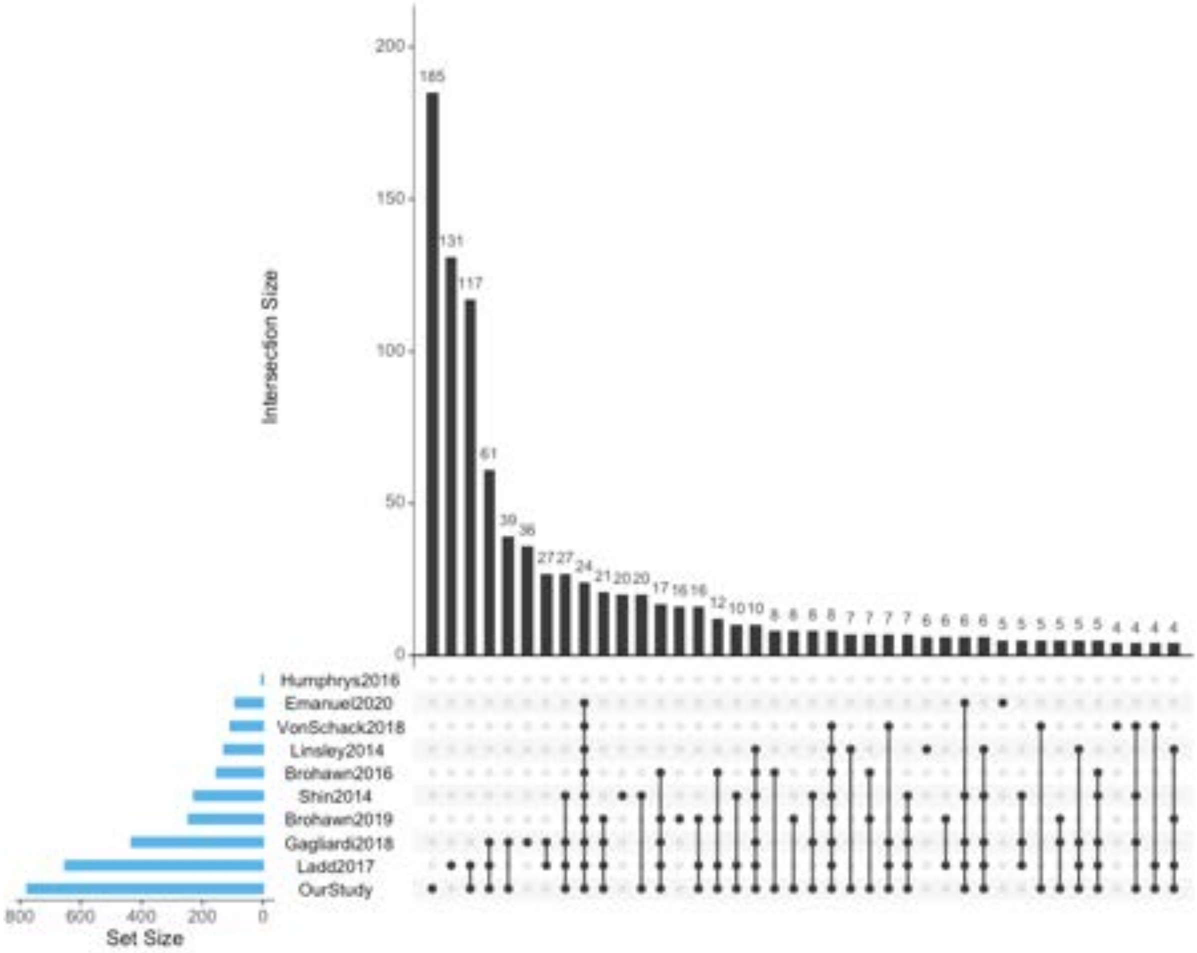

Figure IV-12: Upset plots of overlapping genus in the regular bacteriome between datasets

Upset plots are Venn diagram-like plots. A set refers to a dataset used in this study and each set is on a row with total amounts in a set as a blue bar plot on the left (ordered by set size). The black histogram on top shows the counts that are in the intersection of sets (a single dot for one dataset or connected dots for overlap of multiple datasets). Intersections less than 4 are removed for visualization purposes.

Meta-analysis between datasets

Since our dataset and many others had few to no significant comparisons for ALS vs control groups within each dataset, a meta-analysis between datasets using this criteria would be difficult. As a second pass analysis we created a less stringent filtering method in order to compare the presence of microbes for each group between datasets (ALS vs. ALS; or controls vs. controls) (Fig. IV-13). We assigned a contig to a condition if ≥ 2 samples from that condition contain at least 90% of the summed normalized coverage (from all samples) to the contig. This filtering greatly reduced the number of comparable genus/species for each dataset and, for example, reduced the genus of the regular bacteriome in our dataset from 305 for all samples to 33 (SYM:8, C9S:6, C9A:2, CTL:17).

When we looked at ALS or control contigs in the regular bacteriome, the highest number of unique genus or species was from Ladd et al., 2017, and in general there was a small amount of overlap between datasets (≤1 for ALS or ≤ 8 for controls) (Fig. IV-13). When we looked at genus in the dark bacteriome we saw no overlap for ALS contigs and low overlap (≤ 1) between control conditions (species was similar). In the regular virome there was no overlap between datasets and only our study (one contig from ALS) and Ladd et al., 2017 (three from ALS, five from controls) had contigs that passed the filter (similar values for species). When we looked in the dark virome by genus there was no overlap between datasets, and our dataset had only one genus (*Sobemovirus* from controls*)* with the rest coming from Ladd et al., 2017 (18 from controls, 5 from ALS). In conclusion, despite our conservative and loose approaches, we did not find any convincing evidence in ALS samples that the presence (or lack of presence) of an organism (or multiple organisms) was different compared to control samples.

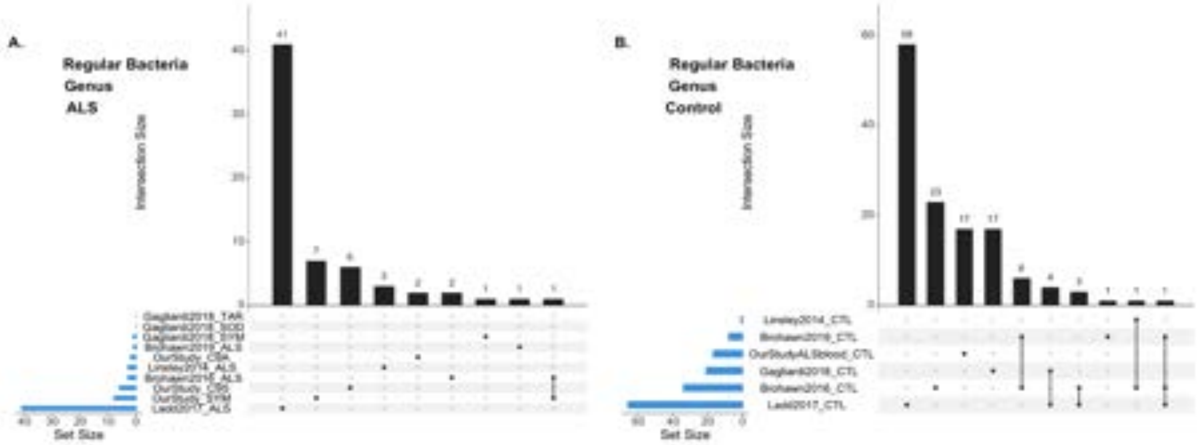

Figure IV-13: Upset plots of overlapping genus between datasets in the regular biome for ALS or controls.

Upset plots are Venn diagram-like plots. A set refers to a contig that was assigned to a condition from a dataset. Each set is on a row with total amounts in a set as a blue bar plot on the left (ordered by set size). The black histogram on top shows the counts that are in the intersection of sets (a single dot for one dataset or connected dots for overlap of multiple datasets). A. ALS contigs in the regular bacteriome. B. Control contigs from the regular bacteriome.

Discussion

We have created Mystery Miner to search for and quantify known and unknown microbes in RNA-seq datasets as a tool for researchers to study dysbiosis and identify infectious agents. We validated the pipeline by recovering and quantifying *Chlamydia* and SARS-CoV reads from RNA-seq datasets from intentionally infected cells. Interestingly, we also identified *Mycoplasma* reads in the *Chlamydia* dataset, suggesting that Mystery Miner may also be able to identify unsuspected bacterial infections or contamination. Next, we created a synthetic minibiome of two different *Chlamydia* species and SARS strains, along with 6 representative bacteria from the human microbiome to investigate the sensitivity of Mystery Miner with regards to species and strain detection and quantification of small fold changes at low coverage. We find that the pipeline is able to recover and quantify significant fold changes for the bacterial species but has difficulty distinguishing reads that come from highly related sequences. We also use published data to investigate the difference of polyA vs total RNA recovery of bacterial species in multiple tissues. Perhaps surprisingly, we did not see a consistent difference in the recovery of bacterial reads between the two types of RNA-seq libraries, considering that bacterial transcripts are not expected to be polyadenylated. However, it is well-recognized that polyA selection is imperfect, and libraries constructed from polyA-selected RNA routinely contain transcripts thought not to be polyadenylated (e.g., rRNA). We also found increased recovery of bacterial species with globin RNA depletion in blood. This could be the result of an effective increase in read depth for bacteria when not sequencing globin, or an increase in contamination from the globin depletion step. We stress that our bioinformatic approach alone cannot distinguish between contamination and the true existence of microbial sequences in human tissue.

We then used Mystery Miner on a novel ALS blood dataset (Our Study) consisting of 8.64 X 10^9 reads. This dataset was generated from whole blood total RNA that was depleted for both ribosomal and globin transcripts. It encompasses samples from controls, participants with a *C9ORF72* hexanucleotide expansion (symptomatic and pre-symptomatic), and *C9ORF72* negative ALS patients. We found no statistical difference in microbial sequence read coverage between controls and any class of ALS patients, examining either individual contigs or using various taxonomical binning approaches. We also did not detect any batch effects or obvious age- or sex-biases in the recovery of microbial reads. Overall, we found no evidence of blood microbial sequences contributing to either *C9ORF72* negative ALS or symptomatic patients harboring the

C9ORF72 hexanucleotide expansion. However, ALS is a CNS disease, so our findings in these blood samples do not necessarily preclude a role for microbes in this disease.

A unique aspect of Mystery Miner is that it tracks non-human reads that do not have significant BLASTN hits in GenBank. We were intrigued by the identification of a large contig (>5kb) in the dark biome of our ALS dataset that showed protein sequence similarity to RNA-dependent RNA polymerases, the essential replicase of RNA viruses. To validate that this virus-like sequence was not an artifact of contig assembly or a contaminant introduced during library construction or sequencing, we used RT-PCR of the original patient samples to demonstrate that this sequence was present in positive samples identified through the RNA-seq analysis and not detectable in negative samples. We cannot extrapolate from this specific example to determine what fraction of the "dark" and "double dark" sequences represent true novel microbial sequences present in human blood, but we note that analysis of human cell free blood DNA has reported the identification of thousands of novel bacterial sequences[365]. We suggest that Mystery Miner is a generally useful tool for the identification of novel microbial sequences in RNA-seq data.

To extend our analysis we processed publicly available blood and spinal cord ALS datasets through our pipeline. As observed in our dataset, library size generally correlated with number of bacterial contigs assembled and number of bacterial genera/species recovered. When the microbial sequences we found in our dataset were compared to the other datasets we found similar genera/species and, not surprisingly, larger datasets generally had greater overlap. One dataset (Ladd et al.,2017) yielded comparable recovery of bacteria and viruses for the regular biome but a far greater recovery bacteria and viruses in the dark biome compared to all the other datasets. This study used laser capture microdissection (LCM) to isolate cervical spinal cord motor neurons and had comparable read amounts per sample to other studies and was conducted in the same laboratory as two other studies (Brohawn et al., 2016, Brohawn et al., 2019). We are unsure why this dataset yielded a much larger dark biome compared to the other datasets. Potentially these differences are a byproduct of using LCM to acquire samples.

We then analyzed several publicly available ALS datasets for statistically significant differences between recovered microbial sequences in ALS and control samples. Only one dataset (Gagliardi et al., 2018) had any significant microbial sequence differences between control and ALS samples, specifically ALS patients with *FUS* or *SOD1* mutations. However, the sample number in this sub-study was small (N = 2-3), and in the case of the *SOD1* patients the excess microbial reads were in the control samples. In the absence of additional information (e.g., batch assignments for the samples) it is difficult to conclude that these sequence/sample correlations are meaningful. Finally, we compared identified microbial sequences in the control and ALS samples across the datasets and did not identify any genera/species that were preferentially recovered in either sample type.

Using our bioinformatic analysis pipeline Mystery Miner, we have not identified an association between ALS pathology and sequences corresponding to known or unknown microbial species. However, there are intrinsic limitations in using "repurposed" RNA-seq data to assay tissue-associated microbial sequences, including the relatively small number of non-human reads (<1% of total) upon which the analysis is based. This limited sequence signal could preclude identification of rarer microbes. Perhaps more problematic is the possibility that contaminating sequences obscure true tissue-associated microbial sequences. Any candidate microbes identified using Mystery Miner that correlate with human phenotypes will necessarily require independent validation. Despite these limitations, we believe Mystery Miner will be a useful tool for future researchers investigating known and unknown microbes that could contribute to disease, as our analyses have shown it to be sensitive to bacterial/viral agents in sequencing data.

Materials and methods

Blood Collection and RNA Extraction

A total of 120 RNA whole blood samples constitutes Our Study, which included 30 healthy controls (from general population that do not have blood relatives suffering from ALS, CTL), 30 pre-symptomatic *C9ORF72* mutant carriers (C9A), 30 symptomatic *C9ORF72* ALS cases (C9S), and 30 symptomatic *C9ORF72*-negative ALS cases (SYM). PAXgene blood RNA tubes were collected at Mayo Clinic Jacksonville and at University of Miami. All 120 RNA samples selected for RNA-seq were received and processed at Mayo Clinic Jacksonville using PAXgene blood RNA kit following manufacturer's recommendations (Qiagen). Blood RNA was of high quality, assessed in an Agilent Bioanalyzer (Agilent), with RNA integrity values ranging from 7.4 to 9.8, with a median value of 8.7. RNA samples were then sent to The Jackson Laboratory for globin depletion, library preparation and sequencing of total blood RNA.

Globin Depletion

Due to the abundance of large haemoglobin RNA transcripts present in the blood, a globin depletion step, using the Ambion GLOBINclear kit (AM1980), was performed before sequencing of the blood RNA samples in order maximize coverage on non-globin genes. In brief, one microgram of total RNA was used as starting material, and specific biotinylated oligos were used to capture globin mRNA transcripts. The capture oligos were hybridized with total RNA samples at 50°C for 30 min. Streptavidin magnetic beads were then used to bind to the biotinylated capture oligos hybridized to globin mRNA by incubating at 50°C for 30 min. The magnetic streptavidin beads-biotin complex were then captured to the side of the tubes by a magnet, and the resulting supernatant is free of globin mRNA. The globin depleted RNA was further purified by RNA binding beads and finally eluted in elution buffer. The resulting RNA free of >95% globin mRNA transcripts was then processed for next generation sequencing. Of note, to assess the efficiency of the globin RNA depletion, 10% of the samples processed were selected randomly and amplified using a Target-Amp Nano labeling kit (Epicentre). Samples were normalized to 100 ng input and reverse transcribed. First strand cDNA was generated by incubating at 50°C for 30 min with first strand premix and Superscript III. This was followed by second strand cDNA synthesis through DNA polymerase by incubating at 65°C for 10 min and at 80°C for 3 min. In-vitro transcription was then performed at 42°C for 4 hours followed by purification using RNeasy mini kit (Qiagen).

Due to the large number of samples, the globin depletion step was performed in two batches. We provided guidelines on how samples would be divided among the batches and also for how samples would be grouped in the sequencing runs in order to minimize technical variability. The Jackson Laboratory personnel were blinded to the identity of the samples.

RNA-seq of total blood RNA (globin and ribosomal RNA depleted) was performed in an Illumina HiSeq4000 with >70 million read pairs per sample (100bp read lengths). Raw reads were then sent back to us for bioinformatics analyses.

Quantitative RT-PCR for blood RNA samples

A total of 500 ng of total blood RNA was used for reverse transcription polymerase chain reaction (RT-PCR), using the High Capacity cDNA Transcription Kit with random primers (Applied Biosystems). Quantitative real-time PCR (qRT-PCR) was performed using SYBR GreenER qPCR SuperMix (Invitrogen). Samples were run in triplicate, and qRT-PCRs were run on a QuantStudio 7 Flex Real-Time system (Applied Biosystems).

List of primers and their sequences in this study:

Primers targeting the novel RDRP contig from our study

RDRP forward 5’-GCTGTCAAATCGGTTTCCAAC-3’;

RDRP reverse 5’-CTGCCTTCGTCATCTTGGAG-3’;

Primers targeting highly expressed control regions

GAPDH forward 5’-GTTCGACAGTCAGCCGCATC-3’;

GAPDH reverse 5’-GGAATTTGCCATGGGTGGA-3’.

Transcriptomics

For downloading the pipeline and detailed instruction for running the pipeline please read the README at https://github.com/Senorelegans/MysteryMiner. All data in this study were processed identically using the pipeline.

Statistical Analysis

To assess statistical differences between conditions, a two tailed Student's *t*-test was calculated using normalized coverage for contigs or binned normalized coverage for species/genus, etc. The number of contigs or genus/species is used to obtain an False discovery rate corrected (using the Benjamini/Hochberg method) adjusted p-value (q-value) via statsmodels in Python. Cutoff for statistical significance is less than an q-value of 0.05 unless otherwise stated.

Data availability

Raw sequencing data for Our Study dataset is available in the NCBI Sequence Read Archive under the accession number PRJNA715316.

All other datasets are publicly available, and all of the code used in this manuscript is available at https://github.com/Senorelegans/MysteryMiner. Supplemental material available at https://figshare.com/s/71d8bbdd30c72f6557d2.

V. MaDDoG

Introduction

The methods to accurately discover and quantify Downstream of Gene (DoGs) expression are still in their infancy. Initial efforts used a sliding window approach that looked at expression levels in the last 1kb of a gene compared to subsequent downstream windows stopping when coverage was less than 1%[190]. The first program dedicated to DoG detection was DoGFinder and used a similar sliding window approach along with down sampling of BAM files for calculating differential expression of genes[366]. Dogcatcher followed soon after and uses a similar approach to DoGFinder but also quantifies antisense downstream of gene (ADoGs), previous of gene (PoG), and antisense previous of gene (APoG) transcripts[367]. The most recent algorithm is automatic readthrough transcription detection (ARTDeco), which uses arbitrary lengths (15kb in humans) to define upstream or downstream regions in order to compare their expression levels to the expression levels in the gene body. In addition, ARTDeco removes read-in counts (reads likely coming from transcripts originating outside of the gene body) from genes which might falsely be called as differentially expressed[368].

However, various improvements that can still be made to these algorithms. One could improve the estimation of DoG lengths by using a Hidden Markov Models that doesn't depend on arbitrary window lengths[369]. Additionally, Bayesian approaches such as Bayesian online changepoint (BOC) detection could be used to differentiate noise from true transcription in DoG regions[370,371]. BOC seeks to identify abrupt changes (usually of the mean) in sequential data, e.g. change points (or switch points), typically with a known number of latent states (Fig.

V-1). Modifying this Bayesian approach, I have created an algorithm called Multi analysis of differential expression of downstream of gene regions (MaDDoG), that partitions downstream of gene regions based on changes in read-depth (using a rolling average from a group of samples) with subsequent differential expression analysis of each section (between groups). Importantly, MaDDoG is a tool that can be applied to any area of the genome but is built to primarily classify DoG regions.

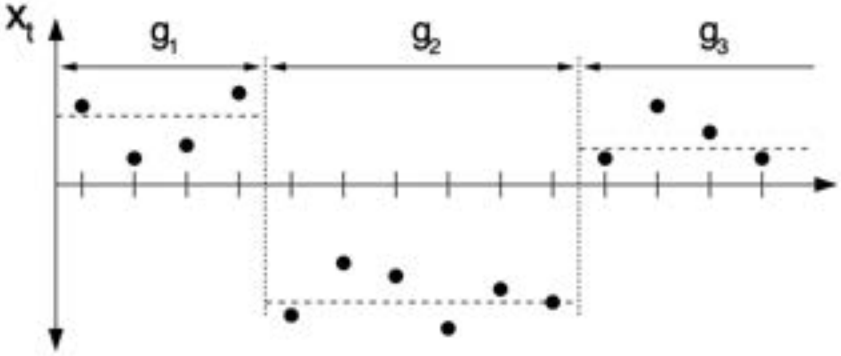

Figure V-1: Example of Bayesian online changepoint detection

Example with synthetic data (x_t is an observation in sequence, typically read depth) cut into three segments (g1, g2, g3) with two latent states, described by horizontal dashed lines (state 1 = g1, g3 ; state 2 = g2) by two changepoints (dashed vertical lines).

Results

Pipeline description

MaDDoG seeks to identify the latent states within previously identified DoGs, on the presumption that within a latent state read count depth is variable but that the identified segments will correspond to interpretable transcripts. Furthermore, MaDDoG then uses these regions to inquire about differential signal, with the overall goal of classifying DoGs based on their changes observed between two conditions. In implementation, MaDDoG is written as a series of python scripts with an optional Nextflow pipeline to parallelize across chromosomes. MaDDog can be run on any section of the genome but a typical application takes DoG regions as input, typically obtained from Dogcatcher. Below is a short overview of the MaDDoG pipeline (Fig. IV-1). On a per sample basis reads are binned across the entire genome (default 50bp windows) then grouped by condition taking the average at each window. Dogcatcher is then applied, this consists of flattening the annotation to remove overlapping genes, selecting a minimum read depth per gene, and using sliding windows with cutoffs for the end of a gene (default stops when DoG is 1% read depth of gene). The result of Dogcatcher is the identification of candidate DoG regions which are then used as input to MaDDoG. MaDDoG first applies a convolution window (rolling average) to smooth reads (default 500bp windows), as smoothing reduces the impact of towers and other outlier anomalies often seen in sequencing data. MaDDoG then applies multiple Bayesian change point analysis models (e.g. different numbers of latent states) and Bayesian Information Criteria (BIC) to select a preferred model (explained in further detail below). To determine whether regions identified change between two conditions, MaDDoG adds all changepoints from both experiments and clusters the change points, and finally takes the resulting regions sectioned by change points and applies differential expression analysis (default DESeq2). I will now use a toy example of synthetic data to clarify the methodology, focusing on the Bayesian change point model, BIC model selection, and inference of differential signal. I will then illustrate the use of MaDDoG on real data to clarify each step of the pipeline, using visualizations where appropriate.

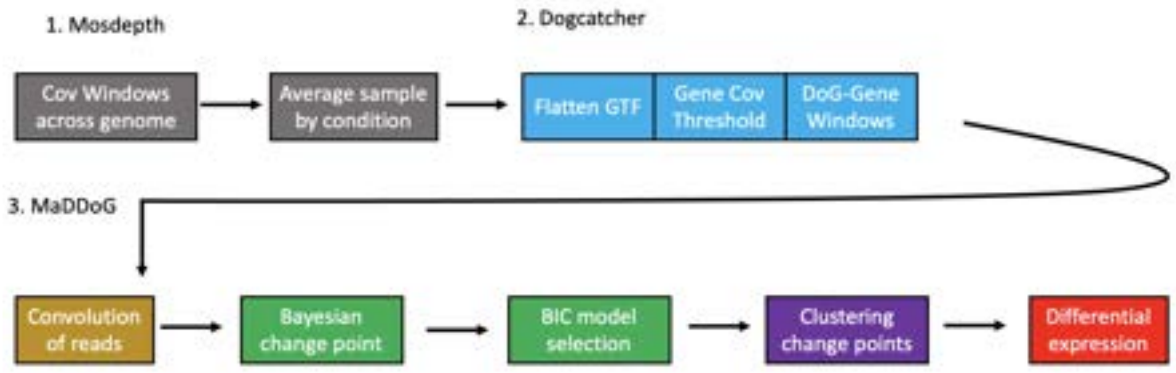

Figure V-2: Diagram of MaDDoG pipeline

On a per sample basis reads are binned across the entire genome then grouped by condition taking the average at each window (1. grey). Dogcatcher is then applied which consists of flattening the annotation to remove overlapping genes, selecting a minimum read depth per gene, and using sliding windows with cutoffs for the end of a gene (2. blue). After input regions are obtained, MaDDoG first applies a convolution window to smooth reads (brown), applies multiple Bayesian change point analysis models with Bayesian Information Criteria (BIC) to select a model (green), adds all changepoints from both groups and clusters the points (purple), and finally takes the regions sectioned by change points and applies differential expression analysis (red).

Toy example of Bayesian change point and BIC model selection

To further clarify the Bayesian change point model implemented in MaDDoG, I will walk through a synthetic toy data example – with three latent states and signal in four regions of variable length (e.g. three switch points). I first generate synthetic data with different mean signal lasting a different number of durations (variable lengths). To do this I use the Poisson distribution (rate = mean) to generate observed values (e.g. pseudo read counts) using three different means (40, 5, 20) over four durations (10, 20, 30, 40) (e.g. positions 1-10 at 40, positions 11-30 at 5, positions 31-60 at 20, and finally positions 61-100 at 40 again) (Fig. V-3).

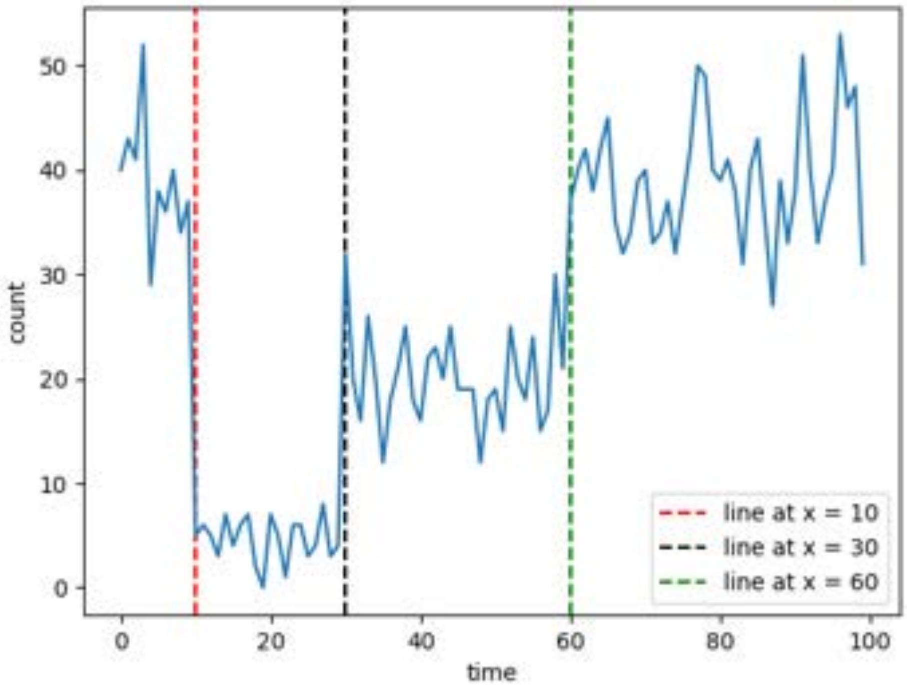

Figure V-3: Toy example of generated data

Data generated using the Poisson process with three different means (40, 5, 20, 40 again) over four durations (10, 20, 30, 40) (or 1-10, 11-30, 31-60, 61-100). Note the fourth duration reuses the first latent state (e.g. it's at 40 read depth). Vertical colored dotted lines indicate location when the mean (or rate) changes.

Assuming the number of latent states is known (often unrealistic), I built a hidden Markov model (HMM) using the Poisson rates (the mean in our case) as emissions. Essentially this imbeds the Poisson process into a simple HMM and allows us to predict the true mean values. I will not go over the HMM model in detail here and instead refer the reader to this great review on their applications in biology[369]. The model is initialized with uniform probabilities for the starting state, uniform probabilities of transitioning from one state to another, and a log normal distribution for choosing initial means. The model is then run using an optimizer (default Adam[372] with learning rate = 0.01) to compute the maximum a posteriori fit to the observed count data to predict the true means. Once I have fit the model (predicted the means), I can then predict the hidden latent state at each time step by using the forward-backward algorithm (a common step in an HMM) to obtain posterior probabilities for each state at each time point (Fig. V-4). I then select the best state for a time point by choosing the latent state with the highest posterior probability at that time point (Fig. V-5).

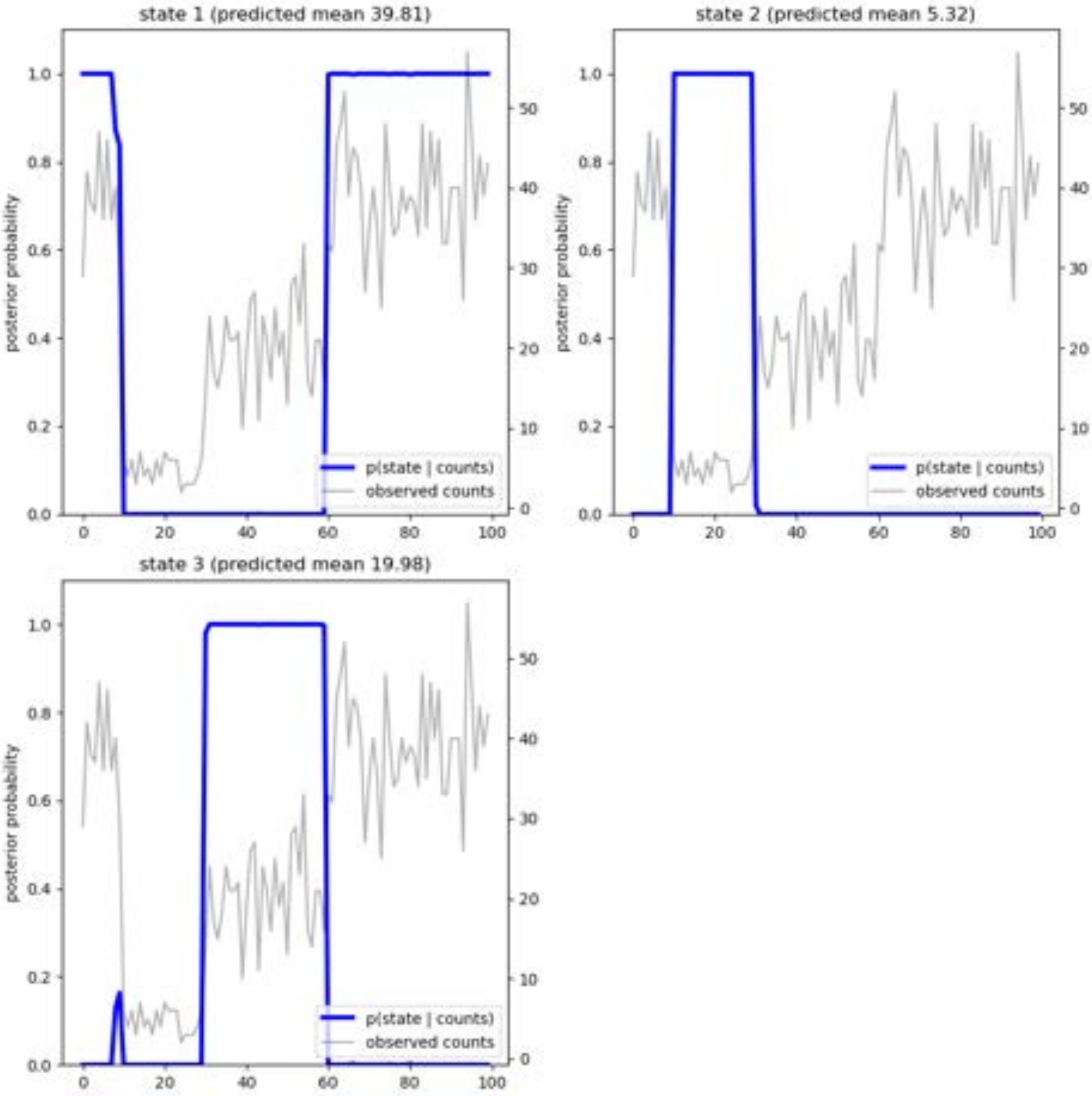

Figure V-4: Posterior probabilities of toy data

Posterior probabilities of the toy data example. State 1 has a high probability of occurring early and later in the sequence. Notice the horizontal line at posterior probability 1.0 runs from positions 0 to 9 and then drops through position 11 before dropping to zero until position 61 where it returns to a high posterior (1.0). State 2 is predicted to occur after the first drop, e.g. between positions 12 and 30. State 3 occurs near the middle of the sequence (positions 31 to 60), but also has a low probability of occurring at positions 9-11 – the same positions where the posterior probability of state 1 were reduced. Thus, the best state path (shown in Figure V-5) most supported is state 1 (positions 0-11), state 2 (positions 12-30, state 3 (positions 31-60), followed by a return to state 1 (61 to end of region).

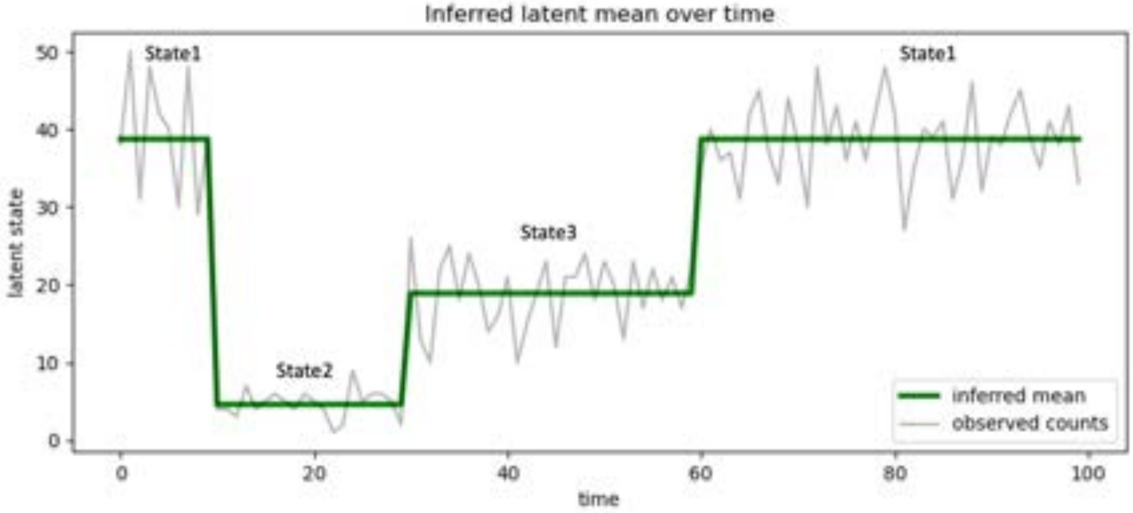

Figure V-5: Inferred latent mean over time

The inferred latent mean (or rate) over time is chosen by selecting the latent state with the highest probability for that time point.

Typically, the true number of latent states is unknown. Theoretically, the number of latent states could be infinite (because there is no upper bound on counts), but with increased number of latent states you will essentially overfit the model to the data i.e., each count level eventually becomes a latent state. Therefore, I need a method for selecting the best number of latent states that strikes a balance between accuracy and overfitting.

Here we simply generate all possible latent states (within some range) to evaluate as possible best models. Since the number of latent states is unknown, I must first decide on the max number of latent states (upper bound) to generate models (each model is fit with an increasing number of latent states), then fit parameters for each model simultaneously and sum over the priors for each model to compute the marginal likelihood for all models (Fig. V-6).

The most practical way to choose the maximum number of latent states comes from domain knowledge and empirical testing. In practice when using MaDDoG on actual count data, I find that the marginal likelihood plateaus around 6 (default max in MaDDoG), making it a good choice for max latent states for this application.

In general, the best model is the simplest one (lowest number of states) that still has a reasonable marginal likelihood compared to more complex models. For speed and practicality, I select the best model using Bayesian information criteria (BIC), specifically using Bayes Factor[373], which is a ratio of the marginal likelihood of one model over the marginal likelihood of another model. I start with the least complex model then use the arbitrary value 1.3 as our threshold for Bayes Factor when deciding to select a more complex model (generally, Bayes Factor values of 1 to 2 are strong evidence to favor another model)[373]. In our toy example, the Bayes Factor is 1.33 at model 3 and does not reach above .98 on more complex models, so model 3 is chosen as our best model and agrees with the true number of hidden latent states for our toy example (Fig. V-7).

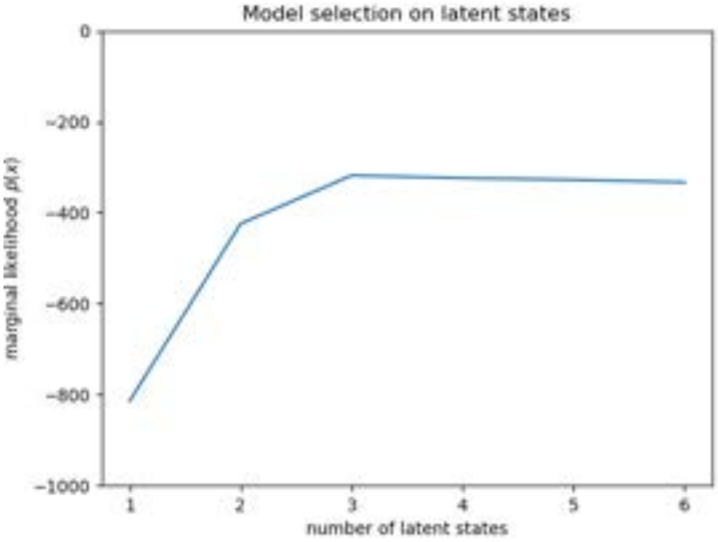

Figure V-6: Marginal likelihood for 6 latent states of toy data

Plot showing the marginal likelihood (y-axis) for the multiple models, differing in number of latent states (x-axis), fit using the toy data. Six models are built (default maximum is 6), with each number indicating the number of latent states in that model.

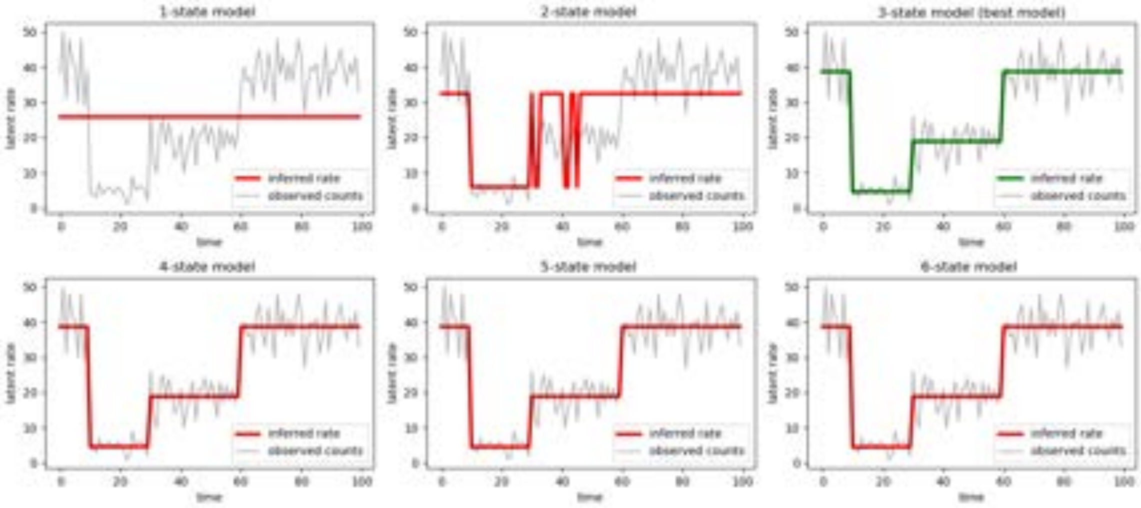

Figure V-7: Model selection for toy data

Plot showing the inferred levels (horizontal lines) for each fit model. To select the best model (model 3, in green), a Bayes Factor is implemented to select the simplest model above the chosen threshold (default 1.3). Note: State models 4-6 appear identical in the figure to the 3-state model but have multiple latent states that are very close in proximity (i.e., 4-state model has latent states 39, 5, 20, 40).

After choosing the best model and the corresponding switchpoints, switchpoints from the treatment and controls are combined for clustering. To illustrate clustering between conditions, I have kept the toy data used above (now labelled “control”) and generated additional synthetic data, labeled “treatment” that extends past where the control sample stopped, but has shorter durations to the 3rd, 4th, and 5th switch points (see Fig. V-8 A). For clustering, I use DBSCAN[374] that does not require the user to specify the number of clusters a priori (compared to K-means). The most important parameter of DBSCAN is called “eps”, which is the minimum number of points required to form a dense region (e.g. what is the minimum width of a latent state), I set this to 0.3 which is slightly more conservative than the default (0.5). In our toy example, I see that the 2nd and 3rd switchpoints from each sample form two clusters (Fig. V-8 B). Next, I take the mean of each cluster to obtain our final switchpoints (Fig. V-8 C). Finally, I use these last switchpoints to partition the region and input the segments into DESeq2 for differential expression analysis.

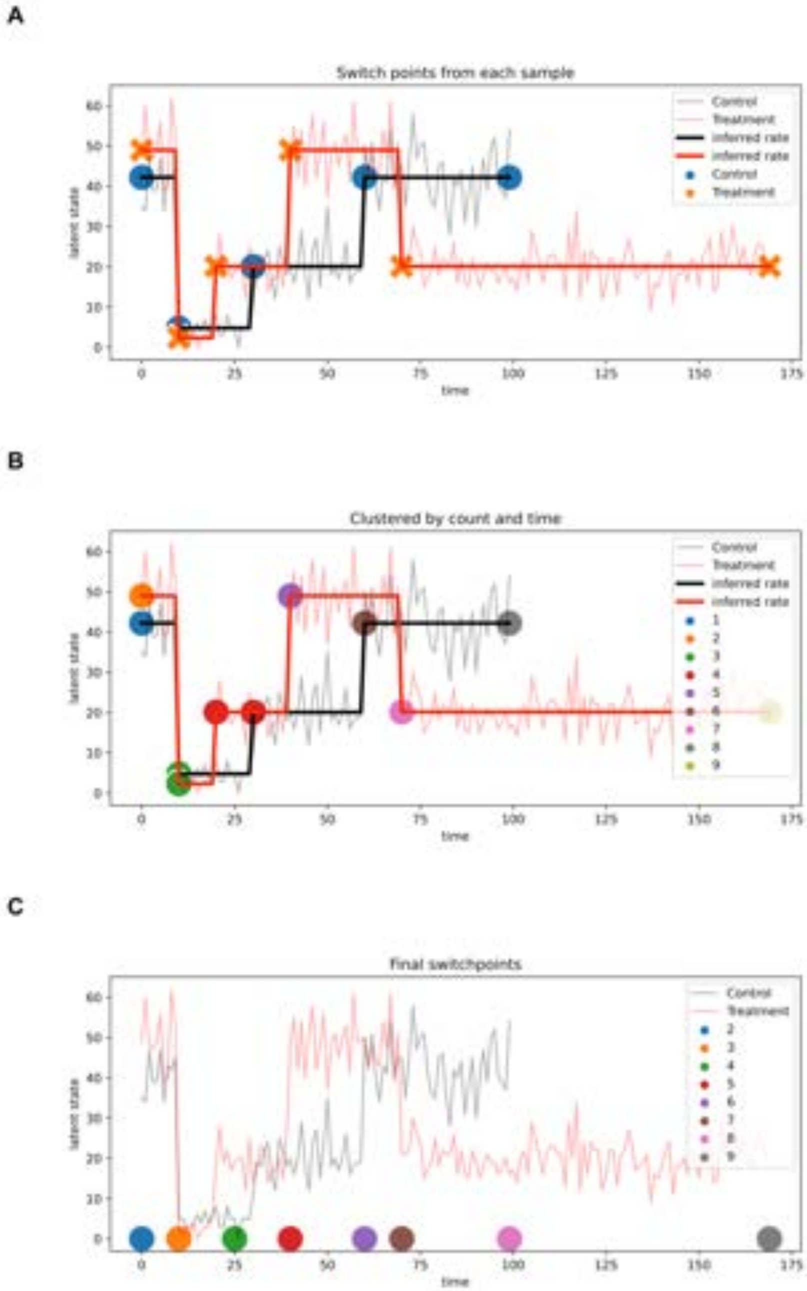

Figure V-8: Clustering switchpoints

(A) Toy data generated for “control”, (in grey) and “treatment” (in red) that extends past where the other sample stopped, but has shorter durations to get to the 3rd, 4th, and 5th switch point. (B) For clustering, I use DBSCAN, which makes clusters from the 2nd and 3rd switchpoints from each sample (each color is a cluster). (C) Finally, I take the mean of each cluster to obtain the final switchpoints. Switchpoints are moved to the X-axis and the inferred rate (mean) is removed for visualization purposes.

Application of MaDDoG to select genes from Vilborg et al., 2015

To identify candidate loci for evaluation of MaDDoG, I first ran Dogcatcher with very loose settings (settings in methods below) to overcall DoG regions. I did this because MaDDoG cannot extend the length of input DoGs but is able to "shorten" DoGs because regions of transcriptional noise between two samples will not be differentially expressed. Here I describe the manual evaluation of the results of MaDDoG on three hand selected loci.

The first gene I looked at is CXXC4, which is also an example DoG from Vilborg et al., 2015 (Fig. V-9 A). At this locus, I identify seven switch point regions (Fig. V-9 B) and all of these regions are significantly differentially expressed (Padj < 0.05) in the KCL treated samples compared to controls. In the KCL sample, the signal is much stronger in the DoG region and there are clearly multiple "levels" of transcription (by manual inspection). However, after collapsing neighboring segments that are consistently differentially expressed (e.g. called by DEseq2 in the same direction), the entire region would be considered differentially expressed. Therefore, the most likely explanation for this region is that there is not a DoG transcribed for this gene in the control sample, as every section is differentially expressed. Why the treatment (KCL) DoG has numerous apparent levels is unclear, but could arise from either biological or technical sources.

Next, I looked at NAT8L (Fig. V-9 C), which appears to have a DoG in both the treatment and control samples and identify only two switch points at the end of either the treatment or control sample (Fig. V-9 D). This highlights the algorithms flexibility, as it chooses the simplest partitions when there is low variance in the mean across the DoG. In addition, I see that for NAT8L, there is no differential expression in the first region, suggesting that the DoG is at a comparable expression level in both the control and treatment, but likely the control sample DoG ends 500bp after the end of the gene. Thus, the data suggests that at this gene the DoG is merely lengthened in the KCL sample.

Finally, I looked at SYNPO2 (Fig. V-9 E). Here I see that the first two sections are not differentially expressed, followed by two differentially expressed sections unique to the treatment, ending with a non-differentially expressed section. I also note that the expression pattern in these DoGs (especially the treatment DoG) appear to show exon-intron boundaries which are being called as switch points (Fig. V-9 F), suggesting the possibility of treatment specific splicing that is not currently annotated. Given that the last section is not differentially expressed, this suggests the region is merely transcriptional noise and I posit that the true ending of the treatment DoG happens at the end of the fourth section.

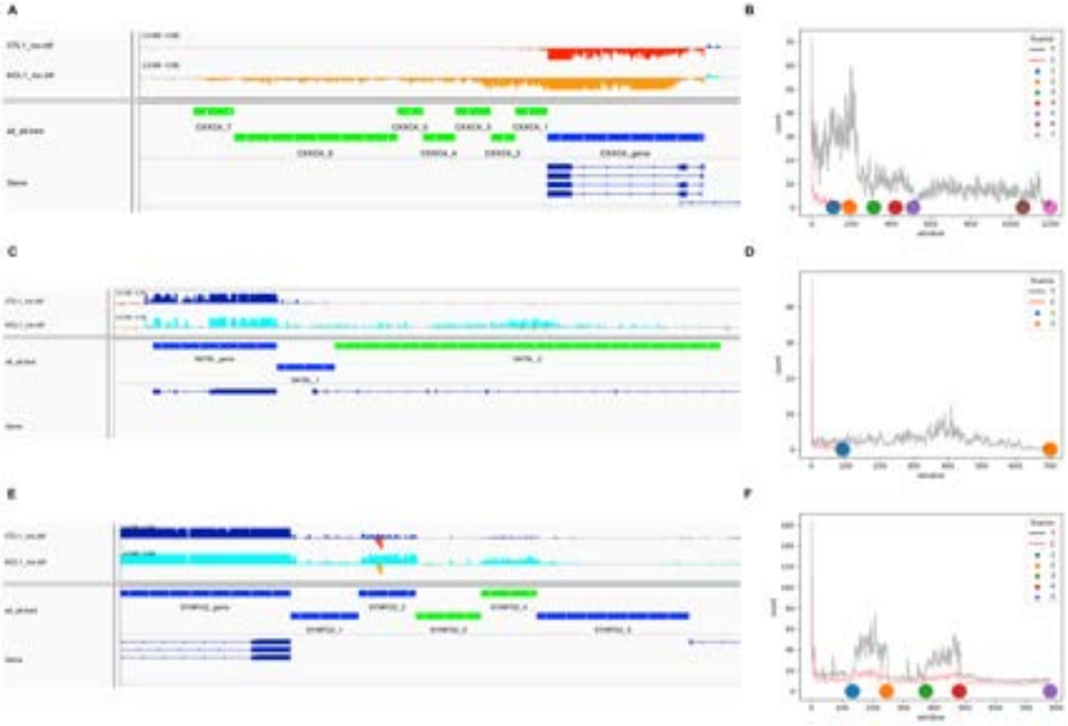

Figure V-9: Select DoGs from Vilborg2015

Three examples showing the CXXC4 locus (A,B), the NAT8L locus (C,D) and the SYNPO2 locus (E,F). IGV (version 2.6.3, and hg38 genome) plots (A,C,E) on left show tracts of data with top two tracks: histograms of normalized reads for one of the control samples (red: minus strand; blue: plus strand) and one of the KCL treated samples (orange: minus strand; light blue: plus strand). The bars on the third track indicate segments that are statistically differentially expressed (green) or not differentially expressed (blue), as called by MaDDoG. The fourth (bottom) track shows the gene annotation (blue). Switch point plots (B,D,F) on the right show the observed counts (treatment: grey; control : red) and various random colors for each switch point. Note: all switch point plots show the gene 5' to 3' from left to right regardless of strand depicted in the IGV plots. Windows are 50bps in size.

Using MaDDoG for intron retention or alternative exon discovery

Although MaDDoG is primarily built to handle downstream of gene regions, one other potential application of MaDDoG is to run it on gene regions to look for intron retention, alternative exon usage, or incorrect annotation. As most exons are well annotated, this also serves as a secondary validation of the capabilities of the method. To this end, I hand selected one sample from Vilborg et al., 2015, and then chose three genes that appeared the cleanest with regards to exon-intron boundaries (KAT5, DRAP1, POLA1). For all genes I chose to skip the first exon because of transcriptional noise.

The first gene I looked at was KAT5, as I found this gene to have the cleanest exon-intron boundaries up until the 8th exon (before annotated alternative exons). I chose only to analyze a subregion of the gene (green bar in Fig. V-10 A), to see if our algorithm can call switchpoints accurately at exon boundaries. MaDDoG selected two latent states, although the three-state model is probabilistically similar and highlights exon 4 (around window 20) as an alternative exon (and possibly exon 2) (Fig. V-10 B). Next, I looked at DRAP1 (Fig V-11 A), where MaDDoG selects a three-state model as the best model, most likely because there is alternative exon usage (or incomplete annotation) before exon 4 (around window 15). For both KAT5 and DRAP1, the switch point boundaries correspond well to the annotated exons, suggesting that MaDDoG is finding reasonable transitions. Finally, I looked at POL1A, which is highly transcribed, has a high quantity of short exons, and is a much longer region compared to the first two genes (~30kb vs 2kb for KAT5) (Fig V-12 A). I find that the Bayesian switch point model has low confidence in any model being correct (Marginal likelihood for all POLA1 models are less than -5000, compared to the lowest being -500 for KAT5), likely due to the added complexity from variance in transcriptional levels of exons, length, and transcriptional noise in intronic regions (Fig V-12 B).

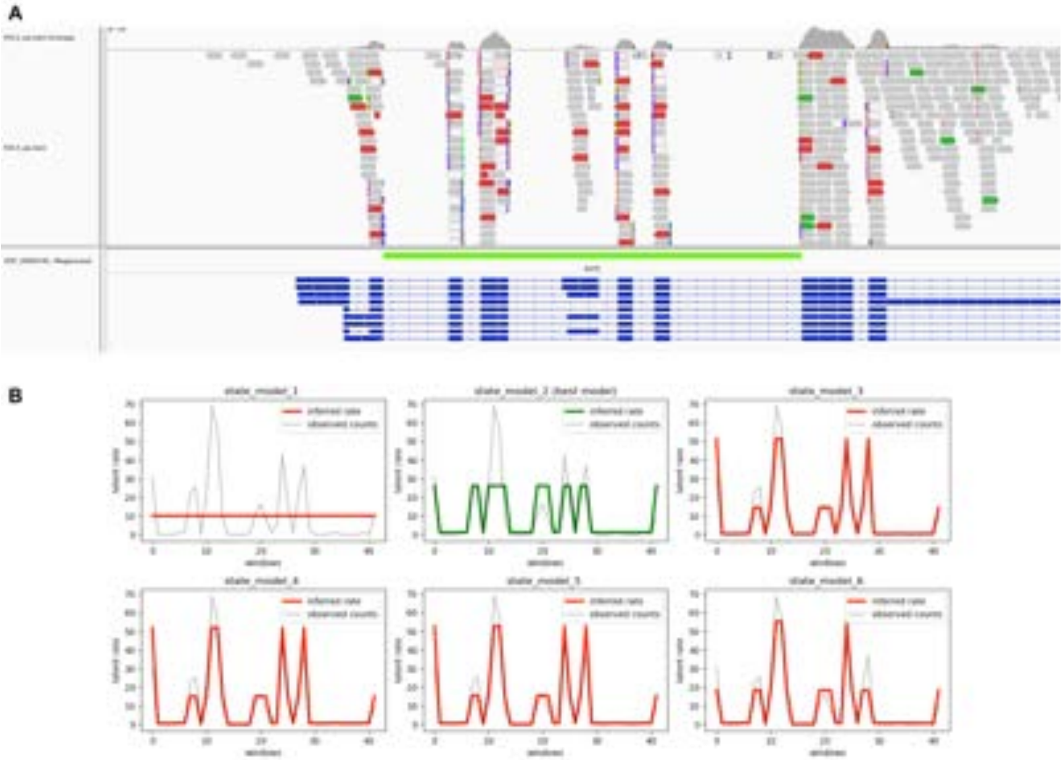

Figure V-10: Bayesian switchpoints of KAT5

(A) IGV screenshot of KAT5 (going from left to right). Green bar indicates the region on which MaDDoG was applied. (B) State model selection showing the two latent state model (in green) as the best model.

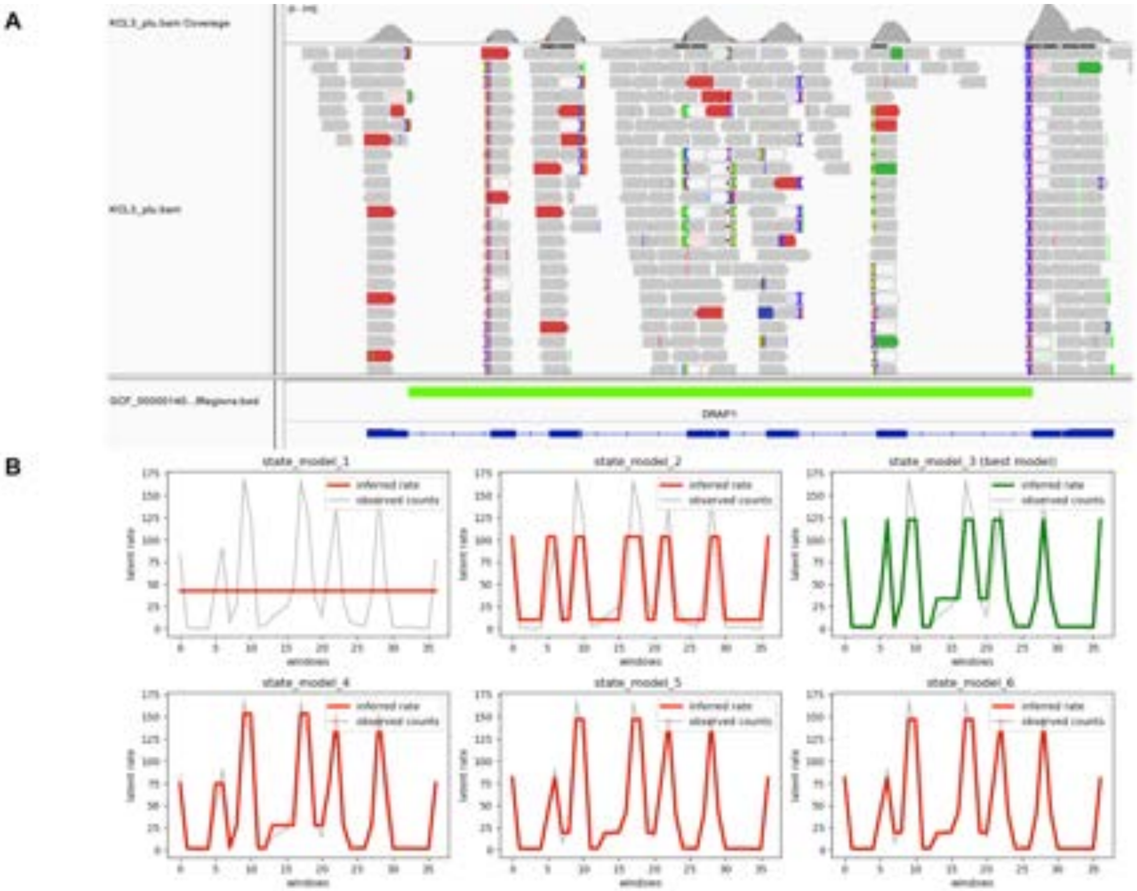

Figure V-11: Bayesian switchpoints of DRAP1

(A) IGV screenshot of DRAP1. Green bar indicates the region on which MaDDoG was applied. (B) State model selection showing the three latent state model (in green) as the best model.

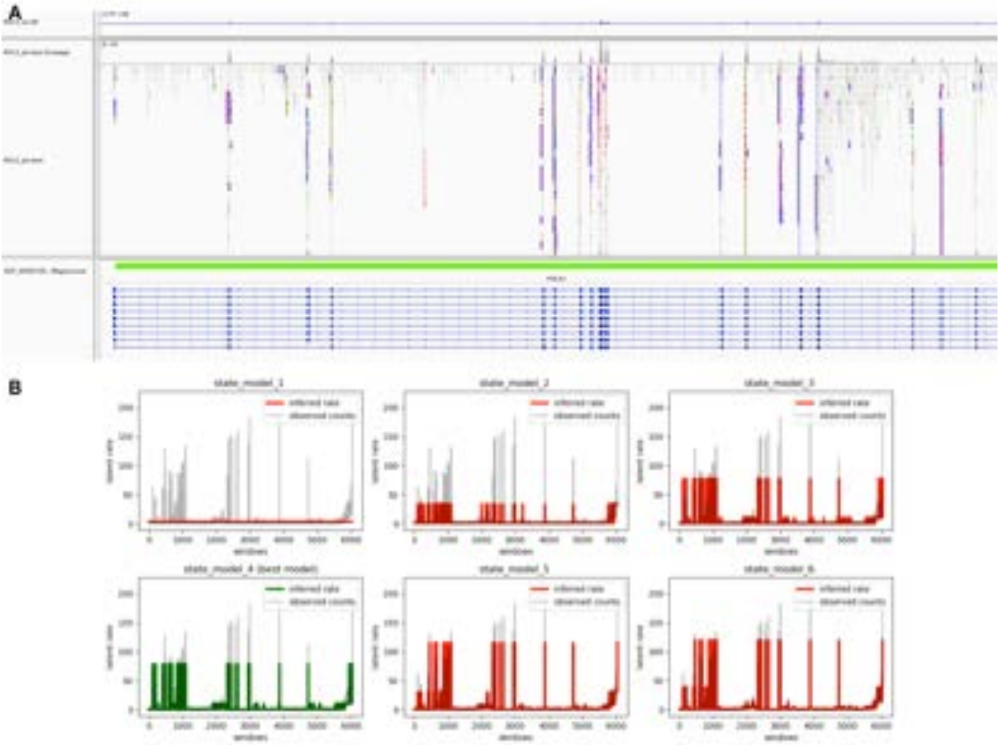

Figure V-12: Bayesian switchpoints of POLA1

(A) Partial IGV screenshot of POLA1. Green bar shows the region over which MaDDoG was applied. (B) State model selection showing the four latent state model (in green) as the best model.

Global results of MaDDoG from Vilborg et al., 2015

Subsequently, I applied MaDDoG to all DoGs called by Dogcatcher in the Vilborg et al., 2015 dataset, primarily to evaluate the runtime feasibility of whole genome analysis. Since the amount of switchpoints in a DoG is not biologically interpretable for thousands of DoGs, I decided to focus on DoGs that have all segments significant for the same condition (indicating the whole region is significantly different), DoGs with or without a significant first segment (indicating both conditions have DoGs of different lengths), and DoGs with or without a significant last segment (indicating that the longest DoG is overcalled and likely transcriptional noise). Out of 4224 DoGs with at least one significant segment, I removed 53 DoGs that had segments significant for both treatment and control as these require individual biological interpretation. Out of the remaining 4171 DoGs, unsurprisingly, I find the majority of DoGs have significant segments from the KCL treatment (treated: 3262; control 909). Next, I filtered for all DoGs that have every segment significant in a condition across the entire loci (similar CXXC4 in Fig. V-9 A), and find that most DoGs do not have all segments significant (total: 2779; treatment: 2248; control 531) (Fig. V-13 A). I next filtered for DoGs that have a significant first segment, I find that there is a similar number of DoGs with a significant first segment (total: 2116; treatment: 1375; control 741) and without (total: 2055; treatment: 1887; control 168) (Fig. V-13 B). Finally, I filtered for DoGs that have a significant last segment, I find that the majority of DoGs do have a significant last segment (total: 3121; treatment: 2651; control 611) than those without (total: 1050; treatment: 470; control 439) (Fig. V-13 C).

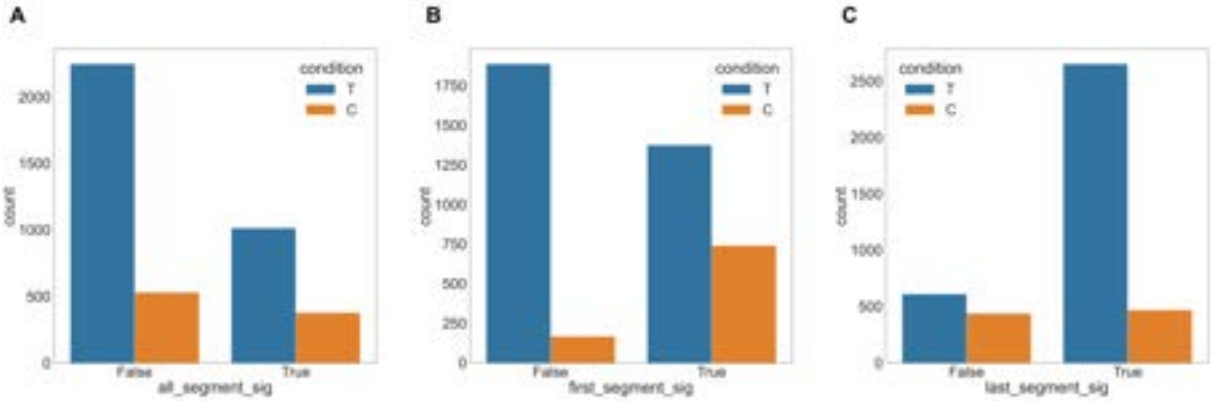

Figure V-13: DoGs with significant segments from Vilborg et al., 2015

Bar plots displaying the counts of DoGs with significant segments. Note: each graph is out of 4171 DoGs as 53 DoGs have been removed because they contain significant segments from both conditions. (A) Count plot showing DoGs with all segments significant (total: 1392; treatment: 1014; control 378) and without (total: 2779; treatment: 2248; control 531). (B) DoGs with a significant first segment (total: 2116; treatment: 1375; control 741) and without (total: 2055; treatment: 1887; control 168). (C) DoGs with a significant last segment (total: 3121; treatment: 2651; control 611) and those without (total: 1050; treatment: 470; control 439).

Discussion

This chapter presented the MaDDoG algorithm and its preliminary application to several genomic regions. I created MaDDoG to partition downstream of gene regions and call differential expression on these partitioned sections. MaDDoG can be run on any region of the genome but is expected to take DoGs as input, which may come from DoGFinder, ARTDeco, or Dogcatcher. The ultimate goal of MaDDoG is to distinguish regions of true transcription from transcriptional noise through patterns of differential transcription.

Using Bayesian changepoints analysis with clustering of switchpoints I have found many promising applications of MaDDoG. Firstly, although MaDDoG does not find DoG regions, it can be used to trim end-regions of transcriptional noise from overcalled DoGs. This trimming approach reduces the windowing issues that the previously mentioned algorithms suffer from and is a simple way to apply MaDDoG to DoGs genome-wide (as running MaDDoG on all genic and intergenic regions would be computationally infeasible). I used three example genes from Vilborg et al., 2015 to show that MaDDoG can delineate genes that have only a treatment DoG, have a treatment and control DoG, and a “trimmed” treatment DoG. Next, I use three example genic regions and apply Bayesian changepoint analysis for discovering intron retention or alternative/unannotated exon usage, as well as highlight where the algorithm struggles on very long regions that have high transcriptional variability and noise (POLA1). Finally, I apply MaDDoG genome-wide and show that most DoGs do not have significant segments across the entire loci, justifying the fine-grained approach of MaDDoG. Next, I show that there is a similar amount of significant and non-significant first segments indicating in the lower expressed condition, there is likely an even split between genes that have no DoG or a truncated DoG. Finally, I show that the majority of DoGs have significant last segments, although there is a fair number of DoGs (1050) that benefit from end trimming.

Moving forward, extensive quantitative performance evaluation is needed and many potential improvements to MaDDoG can be made. For example, the accuracy of exon boundaries could be quantified based on window level accuracy, rather than the more manual inspection approach shown here. In the improvement direction, MaDDoG is built using the Poisson process (i.e., the mean in our case) which is a one parameter distribution, it should be possible to instead use the Negative Binomial which can take two parameters as input (mean and variance). This Negative Binomial addition would likely improve the modelling of counts between samples within a group, as currently the variance is not used. I admit to trying to implement this method but ran into trouble because it changes not only the underlying HMM model but also training, which becomes much more complex especially when training multiple models in parallel. In addition to changing the Bayesian changepoint analysis method, selecting an alternative clustering algorithm to DBscan may improve speed without sacrificing clustering performance[375].

Materials and methods

MaDDoG was created using a modified version of Dogcatcher, tensor flow probability, and DBscan in scikit learn. Using a virtual environment to install tensor flow probability is highly recommended. We have provided a requirements.txt file (in extra scripts on GITHUB) that will install identical versions of the packages. When first using MaDDoG, going through the README on GITHUB is highly recommended. Briefly, a Nextflow pipeline has been provided that will trim, map, and process bams with mosdepth. Next, run the normalizeMosdepth jupyter notebook to average and normalize samples per condition across windows. After that, run the dogcatcher_nextflow.sh script that runs a modified version of Dogcatcher, then runs a nextflow pipeline that runs MaDDoG on each chromosome in parallel. Finally, run the FilterDESeq2 jupyter notebook which aggregates the output of each chromosome and provides Bed files and data frames of significant segments.

All code for MaDDoG can be accessed at https://github.com/Senorelegans/MaDDoG

VI. Summary and Conclusion

Summary of major findings and future experiments

In this thesis, I described two major project directions: (1) the development of Mystery Miner[376], a tool for quantifying pathogen presence within RNA-seq data from the set of reads that do not map to the reference human genome; and (2) the development of tools for the study of downstream of gene (DoG) transcripts, which resulted in two tools: Dogcatcher[367] and MaDDoG (unpublished, described in Chapter V). My application domain has been in neurodegenerative disease; for example, I applied Mystery Miner to a large cohort of patients with Amyotrophic Lateral Sclerosis. The DoG work was initially driven by observations of dsRNA foci in *tdp-1* knockouts, but also focuses on conditions that give rise to stress granules, which are observed in Alzheimer's disease (though Alzheimer's was not directly assayed in this work).

Mystery Miner and potential pathogens in ALS

How Amyotrophic Lateral Sclerosis, Alzheimer's disease, and numerous other neurodegenerative diseases develop largely remains a mystery. The hunt for factors that might contribute to the formation of these diseases has a storied history and include but are not limited to inheritance of alleles or mutations of certain genes, exposure to environmental factors such as heavy metals, and pathogens.

In the search for potential pathogens that might contribute to these neurodegenerative diseases, we developed Mystery Miner, an algorithm that utilizes non-host reads to assemble sequences and identify potential pathogens or contamination. We then apply Mystery Miner to our large novel ALS-related dataset. While our work showed Mystery Miner was effective on both real and synthetic datasets (Chapter IV), the application to a large ALS patient dataset had mixed results. While on the one hand, we recovered a viral RNA-dependent RNA polymerase, this RDRP was not observed in other ALS-related datasets. Lastly, we perform a meta-analysis of other comparable datasets in the ALS field and accumulate additional evidence that a single microbe or set of microbes do not contribute to this neurodegenerative disease.

Despite not finding convincing evidence for the contribution of pathogens to neurodegeneration in our datasets or other datasets analyzed, we are optimistic that Mystery Miner will be a valuable tool for researchers analyzing publicly available datasets. Additionally, with the ever-decreasing cost of sequencing, we hope Mystery Miner will be used on datasets yet to be released. Given that RNA-seq is less than 15 years old, our novel dataset is tremendously valuable for the field because of the number of individuals (most datasets are 3-4 samples per condition) in multiple patient classes. Along with the search for pathogens, this dataset is ripe for other researchers to apply different types of analysis such as repetitive element analysis, alternative splicing, or RNA editing.

There are many avenues of research that can be pursued using Mystery Miner. Since many neurodegenerative diseases have overlapping pathology, it would be interesting to compare recovered microbes from individuals with a variety of these diseases. Additionally, it is possible that older individuals that already have the disease suffered insult from pathogens many years before that are longer resident. As biome sequencing of healthy individuals becomes standard medical practice, we might discover that specific biomes in younger healthy individuals predispose them for certain neurodegenerative diseases which may lead to preventative treatments.

Aside from our dataset, we hope that Mystery Miner will be a useful tool for other researchers that are looking for potential pathogens that might contribute to any disease or to quality check potential contamination of samples. Once RNA-seq becomes more commonplace, applying Mystery Miner in a clinical setting would help physicians link disease phenotypes to known or unknown pathogens. In addition, Mystery Miner might help alleviate the current problem of over-prescribing antibiotics. For example, if a patient has a viral infection identified by Mystery Miner, it would be pointless to prescribe bacterial antibiotics.

Heat shock in *C. elegans,* Dogcatcher, and MaDDoG

For Heat shock in *C elegans*, our work builds on a large body of work that characterizes the effects of this stress in the worm. We have shown that with heat shock in *C. elegans*, many previously unknown changes occur. We first identify the presence of nuclear dsRNA foci and quantify formation of foci over time. Next, we note increased transcription of downstream of gene regions and develop an algorithm called Dogcatcher to identify and quantify these regions genome-wide. We show that these regions are enriched in RNA immunoprecipitated using a dsRNA specific antibody and biologically validate a region identified using Dogcatcher. Given our current knowledge of *C. elegans*, nuclear dsRNA foci has only been identified after heat shock or with *tdp-1* knockout. I believe that discovering how, why, and what these foci are composed of will bring significant biological insights to the field.

With regards to dsRNA in worms, there are many unresolved questions that can be further explored. Although we used J2 IP for RIP-seq, we still have no idea if any of the enriched transcripts are in the dsRNA foci. The foci could be composed of a large amount of highly similar transcripts (repetitive elements) or composed of many separate transcripts. Additionally, if the dsRNA foci are strictly composed of RNA or also contain protein remains unresolved. Stress induced nuclear granules (STING) form in worms under salt stress, oxidative stress, but not with heat shock, although STING formation can be inhibited by a brief pre-exposure heat shock that is dependent upon HSF-1[377]. It would be a worthwhile follow up to look for co-localization of STINGs and nuclear dsRNA foci. Finally, we have only looked at dsRNA foci from worms that have a *tdp*-1 deletion or have been heat shocked, it is likely that there are other conditions that form dsRNA, such as osmotic stress, salt stress, or starvation.

Additionally, identifying and quantifying downstream of gene transcripts is a young field and we are hopeful that there are many important discoveries to be made. As an improvement to Dogcatcher, we developed MaDDoG. MaDDoG is built to partition DoGs into segments based on changepoints of the mean and to quantify these segments between groups of samples. We first use synthetic data to validate our approach, then apply it genome-wide on a real dataset. We then highlight MaDDoGs ability to identify genes with or without truncated control DoGs, remove transcriptional noise from overcalled DoGs, and lastly run it on genic regions where it can be used to look for intron retention and alternative exon use.

Given that we have no idea what DoGs even do, we hope our DoG algorithms will be used by researchers to answer many questions yet to come. As we learn more about the pathways of DoG formation and mechanisms of action, it is likely the field will grow in both number of researchers and complexity of the data and algorithms. To our knowledge, Bayesian change point analysis has not been applied to DoGs before and shows great potential for obtaining true biological insights.

Additionally, Biological validation of MaDDoG using RT-PCR or FISH in segments that we believe are true DoGs or transcriptional noise would be a tremendously beneficial experiment to prove the usefulness of the algorithm. This same approach can also be used on genic regions where we believe there is intron retention or unannotated exons that only appear in treatment conditions.

Conclusion

In this thesis, we have shown that often discarded data can be a treasure trove of knowledge for researchers that go the extra mile. With the current ever-increasing stream of new data, it will be vital that researchers correctly utilize new approaches so biological findings don’t slip through the cracks. We hope that any biological insights gained and developed algorithms will be beneficial to patients suffering from neurodegenerative diseases.

VII. Additional Research

Introduction

There is a complex interplay between repetitive element transcription and activation of the immune system in both viral infection[378,379] and neurodegeneration[73,380]. Due to the recent pandemic and our familiarity with aberrant repetitive element expression in neurodegenerative diseases[174], we sought to apply this analysis to datasets of coronavirus infection [SARS-CoV-1, SARS-CoV-2, Middle East Respiratory Syndrome (MERS)] to look for any differences in host response. To our knowledge no meta-analysis has been performed that analyzes repetitive element expression between these various coronaviruses. We analyzed three publicly available datasets which are described below.

The first dataset has already been used in chapter IV, which has (SARS-CoV) 1 or 2 infection (Emanuel et al., 2020[353]). In the first study human epithelial Calu3 cells were infected with SARS-CoV-1 or SARS-CoV-2 (4, 12, or 24 hours), mock (4 or 24 hours), or untreated (4 hours). The second dataset from Blanco-Melo et al.,2020[381] uses Calu 3 cells infected with SARS-CoV-1 for 24 hours with mock controls. The third dataset from Zhang et al., 2020[382] uses Calu 3 cells infected with MERS for 24 hours with mock controls.

Results

Using Repenrich2[383], we found statistically significant differential repetitive element expression (run at the class level) in virus infected cells compared to mock in all datasets (Fig. VII-1). MERS infection showed the greatest difference with 332 enriched and 396 depleted, with every SINE element depleted (Fig. VII-1 A). Next, we looked at SARS-CoV-2 from BlancoMelo et al., 2020 and found 113 enriched and 106 depleted with the majority coming from LTR and DNA families (Fig. VII-1 B). Lastly, we looked at SARS-CoV-1 or SARS-CoV-2 in Emanuel et al., 2020, we found that for SARS-CoV-2 compared to SARS-CoV-1, there is increased differential expression for both enriched (155 vs 70) and depleted (150 vs 56) transcripts (Fig. VII-1 C, D).

Additionally, since Emanuel et al., 2020 had SARS-CoV-1 or SARS-CoV-2 infected cells, we were able to compare these strains directly. We found that the main difference in RE expression from SARS-CoV-2 vs SARS-CoV-1 is increased expression of LTRs with SARS-CoV-2 (Fig. VII-2 A) and that the majority of these LTRs are from Endogenous retroviral elements (Fig. VII-2 B).

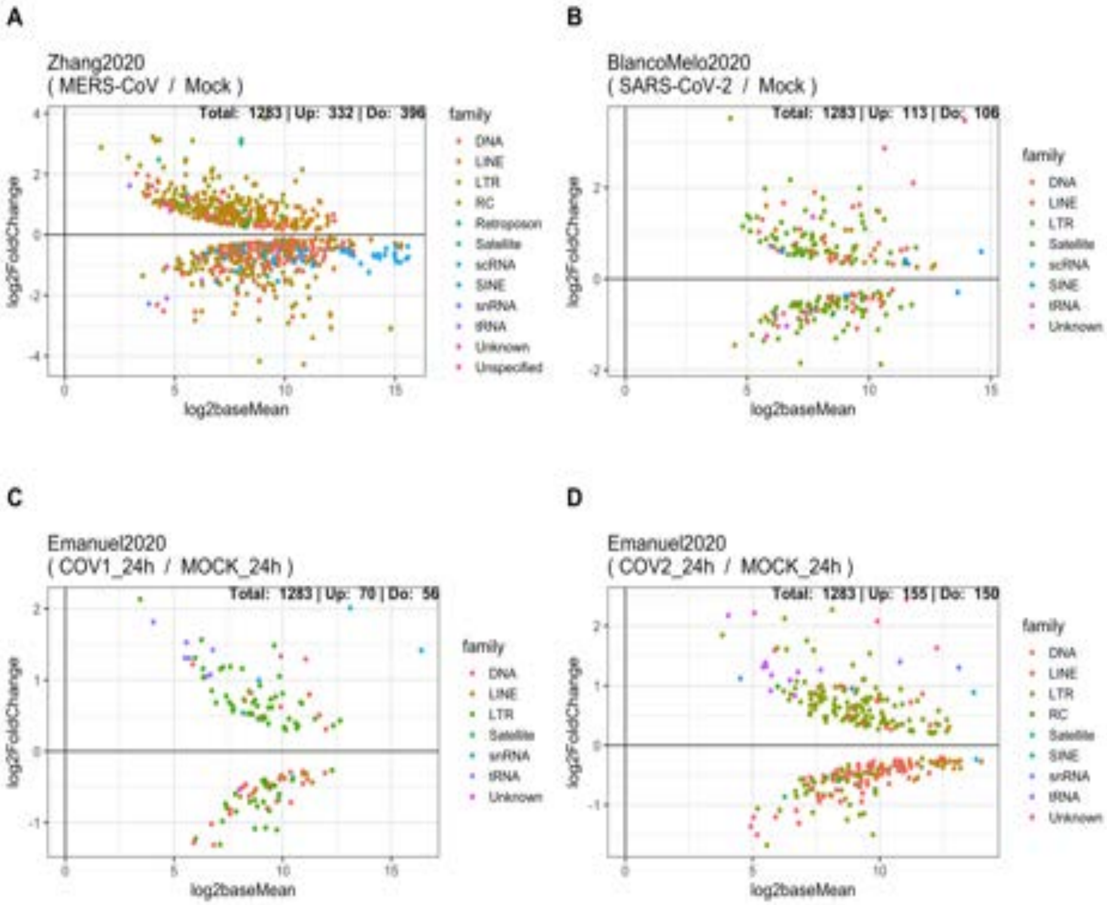

Figure VII-1: Repetitive element expression in coronavirus datasets

MA plot of Calu-3 infected cells from various coronaviruses showing statistically significant differences in repetitive element expression. (A) MERS infection showed the greatest difference with 332 enriched and 396 depleted, with every SINE element depleted. (B) In the first SARS-CoV-1 data, we found 113 enriched and 106 depleted with the majority coming from LTR and DNA families. For SARS-CoV-2 compared to SARS-CoV-1, there is increased differential expression for both enriched (155 vs 70) and depleted (150 vs 56) transcripts (C, D). All plots colored by family.

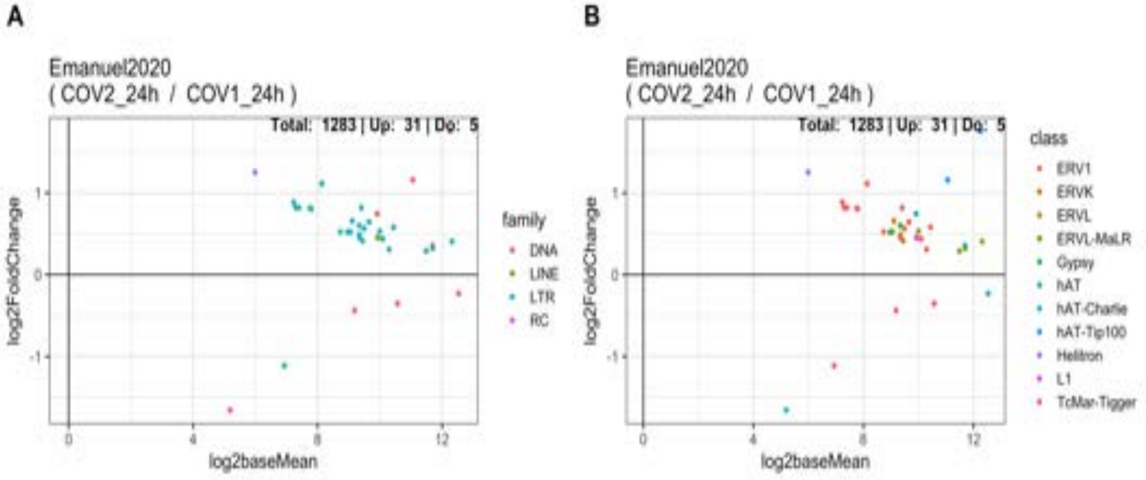

Figure VII-2: SARS-CoV-2 vs SARS-CoV-1 in Emanuel et al., 2020

MA plot of Calu-3 infected cells for SARS-CoV-2 vs SARS-CoV-1 in Emanuel et al., 2020.
(A) Comparing SARS-CoV-2 vs SARS-CoV-1 directly we found increased expression of LTRs with SARS-CoV-2 infection. (B) When we identify elements by class, we find the majority of these LTRs are from Endogenous retroviral elements.

Discussion

To our knowledge, we have performed the first meta-analysis of differential repetitive element expression from coronavirus infected cell lines. We found broad changes in repetitive element expression for each study and show that MERS infection has the most changes, followed by SARS-CoV-2, and finally SARS-CoV-1 infection. We also show that a large difference of RE expression from SARS-CoV-2 compared to SARS-CoV-1 comes from human endogenous retroviral elements (HERV). HERVs are ancient infections that for the most part lay dormant in the genome, although basal expression confers some ability to modulate the immune system[384]. Numerous other viruses such as HIV-1, Influenza A virus, and herpesviruses can induce HERV activation which can contribute to development of viral disease and virus-associated tumors[385]. Additional follow up work should be conducted on HERV activation in SARS-CoV-2 so we may better understand this pathogen and create potential treatments.

Materials and methods

All data was processed with Repenrich2 (downloaded May 5th 2020, default settings). Differential expression was called using DESeq2 (version 1.30.1, default settings) with a statistical significance value cutoff (P adj < 0.05). All data used in this analysis is publicly available.

IX. Bibliography

1. Alwine JC, Kemp DJ, Stark GR. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc Natl Acad Sci U S A*. 1977. doi:10.1073/pnas.74.12.5350

2. VanGuilder HD, Vrana KE, Freeman WM. Twenty-five years of quantitative PCR for gene expression analysis. *Biotechniques*. 2008. doi:10.2144/000112776

3. Wang Z, Gerstein M, Snyder M. RNA-Seq: A revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;10(1):57-63. doi:10.1038/nrg2484

4. Emrich SJ, Barbazuk WB, Li L, Schnable PS. Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res*. 2007;17(1):69-73. doi:10.1101/gr.5145806

5. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*. 2008;18(9):1509-1517. doi:10.1101/gr.079558.108

6. Sultan M, Schulz MH, Richard H, et al. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science (80-)*. 2008;321(5891):956-960. doi:10.1126/science.1160342

7. Goya R, Sun MGF, Morin RD, et al. SNVMix: Predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics*. 2010;26(6):730-736. doi:10.1093/bioinformatics/btq040

8. Djebali S, Davis CA, Merkel A, et al. Landscape of transcription in human cells. *Nature*. 2012;489(7414):101-108. doi:10.1038/nature11233

9. Morris K V., Mattick JS. The rise of regulatory RNA. *Nat Rev Genet*. 2014;15(6):423-437. doi:10.1038/nrg3722

10. Arnold PR, Wells AD, Li XC. Diversity and Emerging Roles of Enhancer RNA in Regulation of Gene Expression and Cell Fate. *Front Cell Dev Biol*. 2020;7:377. doi:10.3389/fcell.2019.00377

11. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. *PLoS Comput Biol*. 2017. doi:10.1371/journal.pcbi.1005457

12. Kukurba KR, Montgomery SB. RNA sequencing and analysis. *Cold Spring Harb Protoc*. 2015. doi:10.1101/pdb.top084970

13. For all you seq. https://www.illumina.com/content/dam/illumina-marketing/documents/applications/ngs-library-prep/ForAllYouSeqMethods.pdf. Published 2020. Accessed October 14, 2020.

14. Wang J, Zhao Y, Zhou X, Hiebert SW, Liu Q, Shyr Y. Nascent RNA sequencing analysis provides insights into enhancer-mediated gene regulation. *BMC Genomics*. 2018. doi:10.1186/s12864-018-5016-z

15. Gong J, Shao D, Xu K, et al. RISE: A database of RNA interactome from sequencing experiments. *Nucleic Acids Res*. 2018. doi:10.1093/nar/gkx864

16. Chen G, Ning B, Shi T. Single-cell RNA-seq technologies and related computational data analysis. *Front Genet*. 2019. doi:10.3389/fgene.2019.00317

17. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol*. 2010. doi:10.1186/gb-2010-11-10-r106

18. Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*. 2007. doi:10.1093/bioinformatics/btm453

19. Chen Y, Lun ATL, Smyth GK. Differential Expression Analysis of Complex RNA-seq Experiments Using edgeR. In: *Statistical Analysis of Next Generation Sequencing Data*. ; 2014. doi:10.1007/978-3-319-07212-8_3

20. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):1-21. doi:10.1186/s13059-014-0550-8

21. Huang HC, Niu Y, Qin LX. Differential expression analysis for RNA-Seq: An overview of statistical methods and computational software. *Cancer Inform*. 2015. doi:10.4137/CIN.S21631

22. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B*. 1995. doi:10.1111/j.2517-6161.1995.tb02031.x

23. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*. 2013. doi:10.1186/1471-2105-14-91

24. Rapaport F, Khanin R, Liang Y, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol*. 2013. doi:10.1186/gb-2013-14-9-r95

25. Ren X, Kuan PF. Negative Binomial Additive Model for RNA-Seq Data Analysis. *bioRxiv*. 2019. doi:10.1101/599811

26. Law CW, Chen Y, Shi W, Smyth GK. Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol*. 2014. doi:10.1186/gb-2014-15-2-r29

27. Hawinkel S, Rayner JCW, Bijnens L, Thas O. Sequence count data are poorly fit by the negative binomial distribution. *PLoS One*. 2020. doi:10.1371/journal.pone.0224909

28. Bottomly D, Walter NAR, Hunter JE, et al. Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. *PLoS One*. 2011. doi:10.1371/journal.pone.0017820

29. Mehmood A, Laiho A, Venäläinen MS, McGlinchey AJ, Wang N, Elo LL. Systematic evaluation of differential splicing tools for RNA-seq studies. *Brief Bioinform*. 2019.

doi:10.1093/bib/bbz126

30. Wilkinson ME, Charenton C, Nagai K. RNA Splicing by the Spliceosome. *Annu Rev Biochem*. 2020. doi:10.1146/annurev-biochem-091719-064225

31. Breitbart RE, Andreadis A, Nadal-Ginard B. Alternative splicing: a ubiquitous mechanism for the generation of multiple protein isoforms from single genes. *Annu Rev Biochem*. 1987. doi:10.1146/annurev.bi.56.070187.002343

32. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*. 2008. doi:10.1038/ng.259

33. Frankiw L, Baltimore D, Li G. Alternative mRNA splicing in cancer immunotherapy. *Nat Rev Immunol*. 2019. doi:10.1038/s41577-019-0195-7

34. Wang ET, Sandberg R, Luo S, et al. Alternative isoform regulation in human tissue transcriptomes. *Nature*. 2008;456(7221):470-476. doi:10.1038/nature07509

35. Vuong CK, Black DL, Zheng S. The neurogenetics of alternative splicing. *Nat Rev Neurosci*. 2016. doi:10.1038/nrn.2016.27

36. Xue Y, Ouyang K, Huang J, et al. Direct conversion of fibroblasts to neurons by reprogramming PTB-regulated MicroRNA circuits. *Cell*. 2013. doi:10.1016/j.cell.2012.11.045

37. Scotti MM, Swanson MS. RNA mis-splicing in disease. *Nat Rev Genet*. 2016. doi:10.1038/nrg.2015.3

38. Niblock M, Gallo JM. Tau alternative splicing in familial and sporadic tauopathies. In: *Biochemical Society Transactions*. ; 2012. doi:10.1042/BST20120091

39. Liu MM, Zack DJ. Alternative splicing and retinal degeneration. *Clin Genet*. 2013. doi:10.1111/cge.12181

40. Kole R, Krainer AR, Altman S. RNA therapeutics: Beyond RNA interference and antisense oligonucleotides. *Nat Rev Drug Discov*. 2012. doi:10.1038/nrd3625

41. Ruegg UT. Pharmacological prospects in the treatment of Duchenne muscular dystrophy. *Curr Opin Neurol*. 2013. doi:10.1097/WCO.0b013e328364fbaf

42. Ren X, Zhang K, Deng R, Li J. RNA Splicing Analysis: From In Vitro Testing to Single-Cell Imaging. *Chem*. 2019. doi:10.1016/j.chempr.2019.05.027

43. Furukawa K, Abe H, Tamura Y, et al. Fluorescence detection of intron lariat RNA with reduction-triggered fluorescent probes. *Angew Chemie - Int Ed*. 2011. doi:10.1002/anie.201104425

44. Wang H, Wang H, Duan X, Sun Y, Wang X, Li Z. Highly sensitive and multiplexed quantification of mRNA splice variants by the direct ligation of DNA probes at the exon junction and universal PCR amplification. *Chem Sci*. 2017. doi:10.1039/c7sc00094d

45. Sun L, Yu C, Irudayaraj J. Raman multiplexers for alternative gene splicing. *Anal Chem*. 2008. doi:10.1021/ac702542n

46. Ren X, Deng R, Wang L, Zhang K, Li J. RNA splicing process analysis for identifying antisense oligonucleotide inhibitors with padlock probe-based isothermal amplification. *Chem Sci*. 2017. doi:10.1039/c7sc01336a

47. Trapnell C, Roberts A, Goff L, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*. 2012. doi:10.1038/nprot.2012.016

48. Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods*. 2010. doi:10.1038/nmeth.1528

49. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*. 2016. doi:10.1038/nbt.3519

50. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*. 2017. doi:10.1038/nmeth.4197

51. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2009. doi:10.1093/bioinformatics/btp616

52. Shen S, Park JW, Lu ZX, et al. rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc Natl Acad Sci U S A*. 2014. doi:10.1073/pnas.1419161111

53. Vaquero-Garcia J, Barrera A, Gazzara MR, et al. A new view of transcriptome complexity and regulation through the lens of local splicing variations. *Elife*. 2016. doi:10.7554/eLife.11752

54. Li YI, Knowles DA, Humphrey J, et al. Annotation-free quantification of RNA splicing using LeafCutter. *Nat Genet*. 2018;50(1):151-158. doi:10.1038/s41588-017-0004-9

55. Deschamps-Francoeur G, Simoneau J, Scott MS. Handling multi-mapped reads in RNA-seq. *Comput Struct Biotechnol J*. 2020. doi:10.1016/j.csbj.2020.06.014

56. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001. doi:10.1038/35057062

57. de Koning APJ, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over Two-Thirds of the human genome. *PLoS Genet*. 2011;7(12). doi:10.1371/journal.pgen.1002384

58. Criscione SW, Zhang Y, Thompson W, Sedivy JM, Neretti N. Transcriptional landscape of repetitive elements in normal and cancer human cells. *BMC Genomics*. 2014;15(1):583. doi:10.1186/1471-2164-15-583

59. Boeke JD, Garfinkel DJ, Styles CA, Fink GR. Ty elements transpose through an RNA intermediate. *Cell*. 1985. doi:10.1016/0092-8674(85)90197-7

60. Munoz-Lopez M, Garcia-Perez J. DNA Transposons: Nature and Applications in Genomics. *Curr Genomics*. 2010. doi:10.2174/138920210790886871

61. Mayer J, Meese E. Human endogenous retroviruses in the primate lineage and their influence on host genomes. *Cytogenet Genome Res*. 2005. doi:10.1159/000084977

62. Levin HL, Moran J V. Dynamic interactions between transposable elements and their hosts. *Nat Rev Genet*. 2011. doi:10.1038/nrg3030

63. Louzada S, Lopes M, Ferreira D, et al. Decoding the role of satellite DNA in genome architecture and plasticity—an evolutionary and clinical affair. *Genes (Basel)*. 2020. doi:10.3390/genes11010072

64. Emanuel BS, Shaikh TH. Segmental duplications: An “expanding” role in genomic instability and disease. *Nat Rev Genet*. 2001. doi:10.1038/35093500

65. Usdin K. The biological effects of simple tandem repeats: Lessons from the repeat expansion diseases. *Genome Res*. 2008. doi:10.1101/gr.070409.107

66. Hancks DC, Kazazian HH. Roles for retrotransposon insertions in human disease. *Mob DNA*. 2016. doi:10.1186/s13100-016-0065-9

67. Brouha B, Schustak J, Badge RM, et al. Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci U S A*. 2003. doi:10.1073/pnas.0831042100

68. Larsen PA, Lutz MW, Hunnicutt KE, et al. The Alu neurodegeneration hypothesis: A primate-specific mechanism for neuronal transcription noise, mitochondrial dysfunction, and manifestation of neurodegenerative disease. *Alzheimer’s Dement*. 2017. doi:10.1016/j.jalz.2017.01.017

69. Kuehnen P, Krude H. Alu elements and human common diseases like obesity. *Mob Genet Elements*. 2012. doi:10.4161/mge.21470

70. Payer LM, Burns KH. Transposable elements in human genetic disease. *Nat Rev Genet*. 2019. doi:10.1038/s41576-019-0165-8

71. Treger RS, Pope SD, Kong Y, Tokuyama M, Taura M, Iwasaki A. The Lupus Susceptibility Locus Sgp3 Encodes the Suppressor of Endogenous Retrovirus Expression SNERV. *Immunity*. 2019. doi:10.1016/j.immuni.2018.12.022

72. Perron H, Bernard C, Bertrand JB, et al. Endogenous retroviral genes, Herpesviruses and gender in Multiple Sclerosis. *J Neurol Sci*. 2009. doi:10.1016/j.jns.2009.04.034

73. Li W, Jin Y, Prazak L, Hammell M, Dubnau J. Transposable Elements in TDP-43-Mediated Neurodegenerative Disorders. *PLoS One*. 2012. doi:10.1371/journal.pone.0044099

74. Wang J, Huda A, Lunyak V V., Jordan IK. A gibbs sampling strategy applied to the mapping of ambiguous short-sequence tags. *Bioinformatics*. 2010.

doi:10.1093/bioinformatics/btq460

75. Jeong HH, Yalamanchili HK, Guo C, Shulman JM, Liu Z. An ultra-fast and scalable quantification pipeline for transposable elements from next generation sequencing data. In: *Pacific Symposium on Biocomputing*. ; 2018. doi:10.1142/9789813235533_0016

76. Jin Y, Tam OH, Paniagua E, Hammell M. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics*. 2015;31(22):3593-3599. doi:10.1093/bioinformatics/btv422

77. Billingsley KJ, Lättekivi F, Planken A, et al. Analysis of repetitive element expression in the blood and skin of patients with Parkinson's disease identifies differential expression of satellite elements. *Sci Rep*. 2019. doi:10.1038/s41598-019-40869-z

78. Roth SH, Levanon EY, Eisenberg E. Genome-wide quantification of ADAR adenosine-to-inosine RNA editing activity. *Nat Methods*. 2019. doi:10.1038/s41592-019-0610-9

79. Benne R, Van Den Burg J, Brakenhoff JPJ, Sloof P, Van Boom JH, Tromp MC. Major transcript of the frameshifted coxll gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell*. 1986. doi:10.1016/0092-8674(86)90063-2

80. Powell LM, Wallis SC, Pease RJ, Edwards YH, Knott TJ, Scott J. A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. *Cell*. 1987. doi:10.1016/0092-8674(87)90510-1

81. Bazak L, Haviv A, Barak M, et al. A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. *Genome Res*. 2014. doi:10.1101/gr.164749.113

82. Sommer B, Köhler M, Sprengel R, Seeburg PH. RNA editing in brain controls a determinant of ion flow in glutamate-gated channels. *Cell*. 1991. doi:10.1016/0092-8674(91)90568-J

83. Christofi T, Zaravinos A. RNA editing in the forefront of epitranscriptomics and human health. *J Transl Med*. 2019. doi:10.1186/s12967-019-2071-4

84. Lonsdale J, Thomas J, Salvatore M, et al. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013. doi:10.1038/ng.2653

85. Kim DDY, Kim TTY, Walsh T, et al. Widespread RNA editing of embedded Alu elements in the human transcriptome. *Genome Res*. 2004. doi:10.1101/gr.2855504

86. Weber F, Wagner V, Rasmussen SB, Hartmann R, Paludan SR. Double-Stranded RNA Is Produced by Positive-Strand RNA Viruses and DNA Viruses but Not in Detectable Amounts by Negative-Strand RNA Viruses. *J Virol*. 2006. doi:10.1128/jvi.80.10.5059-5064.2006

87. Feng Q, Hato S V., Langereis MA, et al. MDA5 Detects the Double-Stranded RNA Replicative Form in Picornavirus-Infected Cells. *Cell Rep*. 2012.

doi:10.1016/j.celrep.2012.10.005

88. Mannion NM, Greenwood SM, Young R, et al. The RNA-Editing Enzyme ADAR1 Controls Innate Immune Responses to RNA. *Cell Rep*. 2014. doi:10.1016/j.celrep.2014.10.041

89. Chung H, Calis JJA, Wu X, et al. Human ADAR1 Prevents Endogenous RNA from Triggering Translational Shutdown. *Cell*. 2018;172(4):811-824.e14. doi:10.1016/j.cell.2017.12.038

90. Rice GI, Kasher PR, Forte GMA, et al. Mutations in ADAR1 cause Aicardi-Goutières syndrome associated with a type i interferon signature. *Nat Genet*. 2012. doi:10.1038/ng.2414

91. Orlowski RJ, O'Rourke KS, Olorenshaw I, Hawkins GA, Maas S, Laxminarayana D. Altered editing in cyclic nucleotide phosphodiesterase 8A1 gene transcripts of systemic lupus erythematosus T lymphocytes. *Immunology*. 2008. doi:10.1111/j.1365-2567.2008.02850.x

92. Krestel H, Meier JC. RNA editing and retrotransposons in neurology. *Front Mol Neurosci*. 2018. doi:10.3389/fnmol.2018.00163

93. Maruyama H, Morino H, Ito H, et al. Mutations of optineurin in amyotrophic lateral sclerosis. *Nature*. 2010. doi:10.1038/nature08971

94. Akbarian S, Smith MA, Jones EG. Editing for an AMPA receptor subunit RNA in prefrontal cortex and striatum in Alzheimer's disease, Huntington's disease and schizophrenia. *Brain Res*. 1995. doi:10.1016/0006-8993(95)00922-D

95. Tariq A, Jantsch MF. Transcript diversification in the nervous system: A to I RNA editing in CNS function and disease development. *Front Neurosci*. 2012. doi:10.3389/fnins.2012.00099

96. Gurevich I, Tamir H, Arango V, Dwork AJ, Mann JJ, Schmauss C. Altered editing of serotonin 2C receptor pre-mRNA in the prefrontal cortex of depressed suicide victims. *Neuron*. 2002. doi:10.1016/S0896-6273(02)00660-8

97. Dracheva S, Patel N, Woo DA, Marcus SM, Siever LJ, Haroutunian V. Increased serotonin 2C receptor mRNA editing: A possible risk factor for suicide. *Mol Psychiatry*. 2008. doi:10.1038/sj.mp.4002081

98. Kung C-P, Maggi LB, Weber JD. The Role of RNA Editing in Cancer Development and Metabolic Disorders. *Front Endocrinol (Lausanne)*. 2018. doi:10.3389/fendo.2018.00762

99. Picardi E, D'Erchia AM, Gallo A, Pesole G. Detection of post-transcriptional RNA editing events. *Methods Mol Biol*. 2015. doi:10.1007/978-1-4939-2291-8_12

100. Diroma MA, Ciaccia L, Pesole G, Picardi E. Elucidating the editome: bioinformatics approaches for RNA editing detection. *Brief Bioinform*. 2019;20(2):436-447. doi:10.1093/bib/bbx129

101. Picardi E, Pesole G. REDItools: High-throughput RNA editing detection made easy.

Bioinformatics. 2013. doi:10.1093/bioinformatics/btt287

102. Zhang F, Lu Y, Yan S, Xing Q, Tian W. SPRINT: an SNP-free toolkit for identifying RNA editing sites. *Bioinformatics*. 2017;33(22):3538-3548. doi:10.1093/bioinformatics/btx473

103. Zhang Q, Xiao X. Genome sequence–independent identification of RNA editing sites. *Nat Methods*. 2015;12(4):347-350. doi:10.1038/nmeth.3314

104. Ramaswami G, Lin W, Piskol R, Tan MH, Davis C, Li JB. Accurate identification of human Alu and non-Alu RNA editing sites. *Nat Methods*. 2012. doi:10.1038/nmeth.1982

105. Ramaswami G, Zhang R, Piskol R, et al. Identifying RNA editing sites using RNA sequencing data alone. *Nat Methods*. 2013. doi:10.1038/nmeth.2330

106. Porath HT, Carmi S, Levanon EY. A genome-wide map of hyper-edited RNA reveals numerous new sites. *Nat Commun*. 2014;5(1):4726. doi:10.1038/ncomms5726

107. Zhao HQ, Zhang P, Gao H, et al. Profiling the RNA editomes of wild-type C. elegans and ADAR mutants. *Genome Res*. 2015. doi:10.1101/gr.176107.114

108. Piechotta M, Wyler E, Ohler U, Landthaler M, Dieterich C. JACUSA: Site-specific identification of RNA editing events from replicate sequencing data. *BMC Bioinformatics*. 2017. doi:10.1186/s12859-016-1432-8

109. Wang Z, Lian J, Li Q, et al. RES-Scanner: A software package for genome-wide identification of RNA-editing sites. *Gigascience*. 2016. doi:10.1186/s13742-016-0143-4

110. Picardi E, Manzari C, Mastropasqua F, Aiello I, D’Erchia AM, Pesole G. Profiling RNA editing in human tissues: Towards the inosinome Atlas. *Sci Rep*. 2015. doi:10.1038/srep14941

111. Lo Giudice C, Silvestris DA, Roth SH, et al. Quantifying RNA Editing in Deep Transcriptome Datasets. *Front Genet*. 2020. doi:10.3389/fgene.2020.00194

112. Hurd MD, Martorell P, Delavande A, Mullen KJ, Langa KM. Monetary Costs of Dementia in the United States. *N Engl J Med*. 2013. doi:10.1056/nejmsa1204629

113. Gan L, Cookson MR, Petrucelli L, La Spada AR. Converging pathways in neurodegeneration, from genetics to mechanisms. *Nat Neurosci*. 2018. doi:10.1038/s41593-018-0237-7

114. Taylor JP, Hardy J, Fischbeck KH. Toxic proteins in neurodegenerative disease. *Science (80-)*. 2002. doi:10.1126/science.1067122

115. Wolozin B, Ivanov P. Stress granules and neurodegeneration. *Nat Rev Neurosci*. 2019. doi:10.1038/s41583-019-0222-5

116. Guzman-Martinez L, Maccioni RB, Andrade V, Navarrete LP, Pastor MG, Ramos-Escobar N. Neuroinflammation as a common feature of neurodegenerative disorders. *Front Pharmacol*. 2019. doi:10.3389/fphar.2019.01008

117. Nik S, Bowman T V. Splicing and neurodegeneration: Insights and mechanisms. *Wiley Interdiscip Rev RNA*. 2019. doi:10.1002/wrna.1532

118. Leija-Salazar M, Piette C, Proukakis C. Review: Somatic mutations in neurodegeneration. *Neuropathol Appl Neurobiol*. 2018. doi:10.1111/nan.12465

119. Diaz-Ortiz ME, Chen-Plotkin AS. Omics in Neurodegenerative Disease: Hope or Hype? *Trends Genet*. 2020. doi:10.1016/j.tig.2019.12.002

120. Diekstra FP, Van Deerlin VM, Van Swieten JC, et al. C9orf72 and UNC13A are shared risk loci for amyotrophic lateral sclerosis and frontotemporal dementia: A genome-wide meta-analysis. *Ann Neurol*. 2014. doi:10.1002/ana.24198

121. Marioni RE, Harris SE, Zhang Q, et al. GWAS on family history of Alzheimer's disease. *Transl Psychiatry*. 2018. doi:10.1038/s41398-018-0150-6

122. Lambert JC, Ibrahim-Verbaas CA, Harold D, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet*. 2013. doi:10.1038/ng.2802

123. Nalls MA, Blauwendraat C, Vallerga CL, et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol*. 2019. doi:10.1016/S1474-4422(19)30320-5

124. Chang D, Nalls MA, Hallgrímsdóttir IB, et al. A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nat Genet*. 2017. doi:10.1038/ng.3955

125. Hamza TH, Zabetian CP, Tenesa A, et al. Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease. *Nat Genet*. 2010. doi:10.1038/ng.642

126. Gratten J, Zhao Q, Benyamin B, et al. Whole-exome sequencing in amyotrophic lateral sclerosis suggests NEK1 is a risk gene in Chinese. *Genome Med*. 2017. doi:10.1186/s13073-017-0487-0

127. MacDonald ME, Ambrose CM, Duyao MP, et al. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*. 1993. doi:10.1016/0092-8674(93)90585-E

128. Raux G, Guyant-Maréchal L, Martin C, et al. Molecular diagnosis of autosomal dominant early onset Alzheimer's disease: An update. *J Med Genet*. 2005. doi:10.1136/jmg.2005.033456

129. Pihlstrøm L, Wiethoff S, Houlden H. Genetics of neurodegenerative diseases: an overview. In: *Handbook of Clinical Neurology*. ; 2018. doi:10.1016/B978-0-12-802395-2.00022-5

130. DeJesus-Hernandez M, Mackenzie IR, Boeve BF, et al. Expanded GGGGCC Hexanucleotide Repeat in Noncoding Region of C9ORF72 Causes Chromosome 9p-Linked FTD and ALS.

Neuron. 2011. doi:10.1016/j.neuron.2011.09.011

131. Cirulli ET, Lasseigne BN, Petrovski S, et al. Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and pathways. *Science (80-)*. 2015. doi:10.1126/science.aaa3650

132. Ling SC, Polymenidou M, Cleveland DW. Converging mechanisms in als and FTD: Disrupted RNA and protein homeostasis. *Neuron*. 2013. doi:10.1016/j.neuron.2013.07.033

133. Bondi MW, Edmonds EC, Salmon DP. Alzheimer's disease: Past, present, and future. *J Int Neuropsychol Soc*. 2017. doi:10.1017/S135561771700100X

134. Katzman R. The Prevalence and Malignancy of Alzheimer Disease. A Major Killer. *Alzheimer's Dement*. 2008. doi:10.1016/j.jalz.2008.10.003

135. Strittmatter WJ, Saunders AM, Schmechel D, et al. Apolipoprotein E: High-avidity binding to β-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proc Natl Acad Sci U S A*. 1993. doi:10.1073/pnas.90.5.1977

136. Katzman R, Kawas C. The epidemiology of dementia and Alzheimer disease. In: *Alzheimer Disease*. ; 1994.

137. Budson AE, Solomon PR. New criteria for Alzheimer disease and mild cognitive impairment: Implications for the practicing clinician. *Neurologist*. 2012. doi:10.1097/NRL.0b013e31826a998d

138. Cummings J, Lee G, Ritter A, Sabbagh M, Zhong K. Alzheimer's disease drug development pipeline: 2020. *Alzheimer's Dement Transl Res Clin Interv*. 2020. doi:10.1002/trc2.12050

139. O'Brien RJ, Wong PC. Amyloid precursor protein processing and alzheimer's disease. *Annu Rev Neurosci*. 2011. doi:10.1146/annurev-neuro-061010-113613

140. Bibl M, Mollenhauer B, Lewczuk P, et al. Validation of amyloid-β peptides in CSF diagnosis of neurodegenerative dementias. *Mol Psychiatry*. 2007. doi:10.1038/sj.mp.4001967

141. Iwatsubo T. Alzheimer's disease: basic aspects. *Nippon Ronen Igakkai zasshi Japanese J Geriatr*. 2000. doi:10.3143/geriatrics.37.207

142. Eisenberg DS, Sawaya MR. Structural studies of amyloid proteins at the molecular level. *Annu Rev Biochem*. 2017. doi:10.1146/annurev-biochem-061516-045104

143. Julien C, Tomberlin C, Roberts CM, et al. In vivo induction of membrane damage by β-amyloid peptide oligomers. *Acta Neuropathol Commun*. 2018. doi:10.1186/s40478-018-0634-x

144. Tracy TE, Gan L. Tau-mediated synaptic and neuronal dysfunction in neurodegenerative disease. *Curr Opin Neurobiol*. 2018. doi:10.1016/j.conb.2018.04.027

145. Hirokawa N, Funakoshi T, Sato-Harada R, Kanai Y. Selective stabilization of tau in axons

and microtubule-associated protein 2C in cell bodies and dendrites contributes to polarized localization of cytoskeletal proteins in mature neurons. *J Cell Biol*. 1996. doi:10.1083/jcb.132.4.667

146. Kellogg EH, Hejab NMA, Poepsel S, Downing KH, DiMaio F, Nogales E. Near-atomic model of microtubule-tau interactions. *Science (80-)*. 2018. doi:10.1126/science.aat1780

147. Goedert M, Spillantini MG, Jakes R, Rutherford D, Crowther RA. Multiple isoforms of human microtubule-associated protein tau: sequences and localization in neurofibrillary tangles of Alzheimer’s disease. *Neuron*. 1989. doi:10.1016/0896-6273(89)90210-9

148. Tapia-Rojas C, Cabezas-Opazo F, Deaton CA, Vergara EH, Johnson GVW, Quintanilla RA. It’s all about tau. *Prog Neurobiol*. 2019. doi:10.1016/j.pneurobio.2018.12.005

149. Sohn PD, Tracy TE, Son HI, et al. Acetylated tau destabilizes the cytoskeleton in the axon initial segment and is mislocalized to the somatodendritic compartment. *Mol Neurodegener*. 2016. doi:10.1186/s13024-016-0109-0

150. Wang JZ, Grundke-Iqbal I, Iqbal K. Glycosylation of microtubule-associated protein tau: An abnormal posttranslational modification in Alzheimer’s disease. *Nat Med*. 1996. doi:10.1038/nm0896-871

151. Fiandaca MS, Kapogiannis D, Mapstone M, et al. Identification of preclinical Alzheimer’s disease by a profile of pathogenic proteins in neurally derived blood exosomes: A case-control study. *Alzheimer’s Dement*. 2015. doi:10.1016/j.jalz.2014.06.008

152. Hutton M, Lendon CL, Rizzu P, et al. Association of missense and 5’-splice-site mutations in tau with the inherited dementia FTDP-17. *Nature*. 1998. doi:10.1038/31508

153. Karran E, Hardy J. A critique of the drug discovery and phase 3 clinical programs targeting the amyloid hypothesis for Alzheimer disease. *Ann Neurol*. 2014. doi:10.1002/ana.24188

154. Bloom GS. Amyloid-β and tau: The trigger and bullet in Alzheimer disease pathogenesis. *JAMA Neurol*. 2014. doi:10.1001/jamaneurol.2013.5847

155. Oddo S, Caccamo A, Tran L, et al. Temporal profile of amyloid-β (Aβ) oligomerization in an in vivo model of Alzheimer disease: A link between Aβ and tau pathology. *J Biol Chem*. 2006. doi:10.1074/jbc.M507892200

156. Guo JP, Arai T, Miklossy J, McGeer PL. Aβ and tau form soluble complexes that may promote self aggregation of both into the insoluble forms in Alzheimer’s diseases. *Proc Natl Acad Sci U S A*. 2006. doi:10.1073/pnas.0509386103

157. Terrill-Usery SE, Mohan MJ, Nichols MR. Amyloid-β(1-42) protofibrils stimulate a quantum of secreted IL-1β despite significant intracellular IL-1β accumulation in microglia. *Biochim Biophys Acta - Mol Basis Dis*. 2014. doi:10.1016/j.bbadis.2014.08.001

158. Matousek SB, Ghosh S, Shaftel SS, Kyrkanides S, Olschowka JA, O’Banion MK. Chronic IL-1β-mediated neuroinflammation mitigates amyloid pathology in a mouse model of alzheimer’s disease without inducing overt neurodegeneration. *J Neuroimmune*

Pharmacol. 2012. doi:10.1007/s11481-011-9331-2

159. Ikeda K, Akiyama H, Kondo H, et al. Thorn-shaped astrocytes: possibly secondarily induced tau-positive glial fibrillary tangles. *Acta Neuropathol*. 1995. doi:10.1007/BF00318575

160. Barres BA. The Mystery and Magic of Glia: A Perspective on Their Roles in Health and Disease. *Neuron*. 2008. doi:10.1016/j.neuron.2008.10.013

161. Kahlson MA, Colodner KJ. Glial tau pathology in tauopathies: Functional consequences. *J Exp Neurosci*. 2015. doi:10.4137/JEN.S25515

162. Guo T, Zhang D, Zeng Y, Huang TY, Xu H, Zhao Y. Molecular and cellular mechanisms underlying the pathogenesis of Alzheimer's disease. *Mol Neurodegener*. 2020. doi:10.1186/s13024-020-00391-7

163. Stroke. NI of ND and. Amyotrophic Lateral Sclerosis (ALS) Fact Sheet. NIH.

164. Siddique T, Dellefave L. Amyotrophic lateral sclerosis. In: *Neurogenetics: Scientific and Clinical Advances*. ; 2005. doi:10.5005/jp/books/12672_132

165. Masrori P, Van Damme P. Amyotrophic lateral sclerosis: a clinical review. *Eur J Neurol*. 2020. doi:10.1111/ene.14393

166. Kiernan MC, Vucic S, Cheah BC, et al. Amyotrophic lateral sclerosis. In: *The Lancet*. ; 2011. doi:10.1016/S0140-6736(10)61156-7

167. Wang MD, Little J, Gomes J, Cashman NR, Krewski D. Identification of risk factors associated with onset and progression of amyotrophic lateral sclerosis using systematic review and meta-analysis. *Neurotoxicology*. 2017. doi:10.1016/j.neuro.2016.06.015

168. Santamaria N, Alhothali M, Alfonso MH, Breydo L, Uversky VN. Intrinsic disorder in proteins involved in amyotrophic lateral sclerosis. *Cell Mol Life Sci*. 2017. doi:10.1007/s00018-016-2416-6

169. Zhang Y-J, Xu Y-F, Cook C, et al. Aberrant cleavage of TDP-43 enhances aggregation and cellular toxicity. *Proc Natl Acad Sci*. 2009;106(18):7607-7612. doi:10.1073/pnas.0900688106

170. Higashi S, Iseki E, Yamamoto R, et al. Concurrence of TDP-43, tau and α-synuclein pathology in brains of Alzheimer's disease and dementia with Lewy bodies. *Brain Res*. 2007. doi:10.1016/j.brainres.2007.09.048

171. Chang XL, Tan MS, Tan L, Yu JT. The Role of TDP-43 in Alzheimer's Disease. *Mol Neurobiol*. 2016. doi:10.1007/s12035-015-9264-5

172. Prasad A, Bharathi V, Sivalingam V, Girdhar A, Patel BK. Molecular mechanisms of TDP-43 misfolding and pathology in amyotrophic lateral sclerosis. *Front Mol Neurosci*. 2019. doi:10.3389/fnmol.2019.00025

173. Xiao S, Sanelli T, Dib S, et al. RNA targets of TDP-43 identified by UV-CLIP are deregulated

in ALS. *Mol Cell Neurosci*. 2011. doi:10.1016/j.mcn.2011.02.013

174. Saldi TK, Ash PE, Wilson G, et al. TDP-1, the Caenorhabditis elegans ortholog of TDP-43, limits the accumulation of double-stranded RNA. *EMBO J*. 2014;33(24):2947-2966. doi:10.15252/embj.201488740

175. Saldi TK, Gonzales PK, LaRocca TJ, Link CD. Neurodegeneration, Heterochromatin, and Double-Stranded RNA. *J Exp Neurosci*. 2019. doi:10.1177/1179069519830697

176. Reddy K, Zamiri B, Stanley SYR, Macgregor RB, Pearson CE. The disease-associated r(GGGGCC)n repeat from the C9orf72 gene forms tract length-dependent uni- and multimolecular RNA G-quadruplex structures. *J Biol Chem*. 2013. doi:10.1074/jbc.C113.452532

177. Tang X, Toro A, Sahana TG, et al. Divergence, Convergence, and Therapeutic Implications: A Cell Biology Perspective of C9ORF72-ALS/FTD. *Mol Neurodegener*. 2020. doi:10.1186/s13024-020-00383-7

178. Ishiguro A, Kimura N, Watanabe Y, Watanabe S, Ishihama A. TDP-43 binds and transports G-quadruplex-containing mRNAs into neurites for local translation. *Genes to Cells*. 2016. doi:10.1111/gtc.12352

179. Prudencio M, Gonzales PK, Cook CN, et al. Repetitive element transcripts are elevated in the brain of C9orf72 ALS/FTLD patients. *Hum Mol Genet*. 2017;26(17):3421-3431. doi:10.1093/hmg/ddx233

180. Zhang K, Donnelly CJ, Haeusler AR, et al. The C9orf72 repeat expansion disrupts nucleocytoplasmic transport. *Nature*. 2015. doi:10.1038/nature14973

181. Hurlimann T, Jaitovich Groisman I, Godard B. Exploring neurologists' perspectives on the return of next generation sequencing results to their patients: A needed step in the development of guidelines 06 Biological Sciences 0604 Genetics. *BMC Med Ethics*. 2018. doi:10.1186/s12910-018-0320-3

182. Tian X, Liang WC, Feng Y, et al. Expanding genotype/phenotype of neuromuscular diseases by comprehensive target capture/NGS. *Neurol Genet*. 2015. doi:10.1212/NXG.0000000000000015

183. Farwell KD, Shahmirzadi L, El-Khechen D, et al. Enhanced utility of family-centered diagnostic exome sequencing with inheritance model-based analysis: Results from 500 unselected families with undiagnosed genetic conditions. *Genet Med*. 2015. doi:10.1038/gim.2014.154

184. Westra D, Schouten MI, Stunnenberg BC, et al. Panel-based exome sequencing for neuromuscular disorders as a diagnostic service. *J Neuromuscul Dis*. 2019. doi:10.3233/JND-180376

185. Novarino G, Fenstermaker AG, Zaki MS, et al. Exome sequencing links corticospinal motor neuron disease to common neurodegenerative disorders. *Science (80-)*. 2014.

doi:10.1126/science.1247363

186. Klein CJ, Foroud TM. Neurology Individualized Medicine: When to Use Next-Generation Sequencing Panels. *Mayo Clin Proc*. 2017. doi:10.1016/j.mayocp.2016.09.008

187. Shademan B, Biray Avci C, Nikanfar M, Nourazarian A. Application of Next-Generation Sequencing in Neurodegenerative Diseases: Opportunities and Challenges. *NeuroMolecular Med*. 2020. doi:10.1007/s12017-020-08601-7

188. Proudfoot NJ. Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science (80-)*. 2016;352(6291):aad9926. doi:10.1126/science.aad9926

189. Eaton JD, West S. Termination of Transcription by RNA Polymerase II: BOOM! *Trends Genet*. 2020. doi:10.1016/j.tig.2020.05.008

190. Vilborg A, Passarelli MC, Yario TA, Tycowski KT, Steitz JA. Widespread Inducible Transcription Downstream of Human Genes. *Mol Cell*. 2015;59(3):449-461. doi:10.1016/j.molcel.2015.06.016

191. Vilborg A, Sabath N, Wiesel Y, et al. Comparative analysis reveals genomic features of stress-induced transcriptional readthrough. *Proc Natl Acad Sci*. 2017;114(40):E8362-E8371. doi:10.1073/pnas.1711120114

192. Rutkowski AJ, Erhard F, L'Hernault A, et al. Widespread disruption of host transcription termination in HSV-1 infection. *Nat Commun*. 2015;6(1):7126. doi:10.1038/ncomms8126

193. Grosso AR, Leite AP, Carvalho S, et al. Pervasive transcription read-through promotes aberrant expression of oncogenes and RNA chimeras in renal carcinoma. *Elife*. 2015;4(NOVEMBER2015):1-16. doi:10.7554/eLife.09214

194. Muniz L, Deb MK, Aguirrebengoa M, Lazorthes S, Trouche D, Nicolas E. Control of Gene Expression in Senescence through Transcriptional Read-Through of Convergent Protein-Coding Genes. *Cell Rep*. 2017;21(9):2433-2446. doi:10.1016/j.celrep.2017.11.006

195. Vilborg A, Steitz JA. Readthrough transcription: How are DoGs made and what do they do? *RNA Biol*. 2017. doi:10.1080/15476286.2016.1149680

196. Wang X, Hennig T, Whisnant AW, et al. Herpes simplex virus blocks host transcription termination via the bimodal activities of ICP27. *Nat Commun*. 2020. doi:10.1038/s41467-019-14109-x

197. Hennig T, Michalski M, Rutkowski AJ, et al. HSV-1-induced disruption of transcription termination resembles a cellular stress response but selectively increases chromatin accessibility downstream of genes. *PLoS Pathog*. 2018. doi:10.1371/journal.ppat.1006954

198. Rosa-Mercado NA, Zimmer JT, Apostolidi M, Rinehart J, Simon MD, Steitz JA. Hyperosmotic stress induces downstream-of-gene transcription and alters the RNA Polymerase II interactome despite widespread transcriptional repression. *bioRxiv*. January 2020:2020.06.30.178103. doi:10.1101/2020.06.30.178103

199. He Y, Vogelstein B, Velculescu VE, Papadopoulos N, Kinzler KW. The antisense transcriptomes of human cells. *Science (80-)*. 2008. doi:10.1126/science.1163853

200. Wight M, Werner A. Europe PMC Funders Group The functions of natural antisense transcripts. 2015:91-101. doi:10.1042/bse0540091.The

201. Dauber B, Saffran HA, Smiley JR. The herpes simplex virus host shutoff (vhs) RNase limits accumulation of double stranded RNA in infected cells: Evidence for accelerated decay of duplex RNA. *PLoS Pathog*. 2019. doi:10.1371/journal.ppat.1008111

202. Forum I of M (US) F. Study of the Human Microbiome. 2013. https://www.ncbi.nlm.nih.gov/books/NBK154091/. Accessed November 14, 2020.

203. Sender R, Fuchs S, Milo R. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS Biol*. 2016. doi:10.1371/journal.pbio.1002533

204. Tierney BT, Yang Z, Luber JM, et al. The Landscape of Genetic Content in the Gut and Oral Human Microbiome. *Cell Host Microbe*. 2019. doi:10.1016/j.chom.2019.07.008

205. Hadrich D. Microbiome research is becoming the key to better understanding health and nutrition. *Front Genet*. 2018. doi:10.3389/fgene.2018.00212

206. Durack J, Lynch S V. The gut microbiome: Relationships with disease and opportunities for therapy. *J Exp Med*. 2019. doi:10.1084/jem.20180448

207. Nagy E, Boyanova L, Justesen US. How to isolate, identify and determine antimicrobial susceptibility of anaerobic bacteria in routine laboratories. *Clin Microbiol Infect*. 2018. doi:10.1016/j.cmi.2018.02.008

208. Kennedy EA, King KY, Baldridge MT. Mouse microbiota models: Comparing germ-free mice and antibiotics treatment as tools for modifying gut bacteria. *Front Physiol*. 2018. doi:10.3389/fphys.2018.01534

209. Khoruts A, Sadowsky MJ. Understanding the mechanisms of faecal microbiota transplantation. *Nat Rev Gastroenterol Hepatol*. 2016. doi:10.1038/nrgastro.2016.98

210. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The Human Microbiome Project. *Nature*. 2007. doi:10.1038/nature06244

211. Fournier PE, Drancourt M, Colson P, Rolain JM, Scola B La, Raoult D. Modern clinical microbiology: New challenges and solutions. *Nat Rev Microbiol*. 2013. doi:10.1038/nrmicro3068

212. Houpikian P, Raoult D. Traditional and molecular techniques for the study of emerging bacterial diseases: One laboratory's perspective. *Emerg Infect Dis*. 2002. doi:10.3201/eid0802.010141

213. Versalovick J, Carroll K, Funke G, Jorgense J, Landry M, Warnock D. *Manual of Clinical Microbiology (Manual of Clinical Microbiology)*.; 2015.

214. Milne JLS, Borgnia MJ, Bartesaghi A, et al. Cryo-electron microscopy - A primer for the

non-microscopist. *FEBS J*. 2013. doi:10.1111/febs.12078

215. Pace B, Campbell LL. Homology of ribosomal ribonucleic acid diverse bacterial species with Escherichia coli and Bacillus stearothermophilus. *J Bacteriol*. 1971. doi:10.1128/jb.107.2.543-547.1971

216. Higuchi R, Dollinger G, Sean Walsh P, Griffith R. Simultaneous amplification and detection of specific DNA sequences. *Bio/Technology*. 1992. doi:10.1038/nbt0492-413

217. Klein D. Quantification using real-time PCR technology: Applications and limitations. *Trends Mol Med*. 2002. doi:10.1016/S1471-4914(02)02355-9

218. Rodríguez-Lázaro D, Cook N, Hernández M. Real-time PCR in food science: PCR diagnostics. *Curr Issues Mol Biol*. 2013. doi:10.21775/cimb.015.039

219. A review of 10 years of human microbiome research activities at the US National Institutes of Health, Fiscal Years 2007-2016. *Microbiome*. 2019. doi:10.1186/s40168-019-0620-y

220. Konstantinidis KT, Tiedje JM. Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci U S A*. 2005. doi:10.1073/pnas.0409727102

221. Chaudhary N, Sharma AK, Agarwal P, Gupta A, Sharma VK. 16S classifier: A tool for fast and accurate taxonomic classification of 16S rRNA hypervariable regions in metagenomic datasets. *PLoS One*. 2015. doi:10.1371/journal.pone.0116106

222. Fox GE, Wisotzkey JD, Jurtshuk P. How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *Int J Syst Bacteriol*. 1992. doi:10.1099/00207713-42-1-166

223. Eren AM, Maignien L, Sul WJ, et al. Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol Evol*. 2013. doi:10.1111/2041-210X.12114

224. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J*. 2017. doi:10.1038/ismej.2017.119

225. Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochem Biophys Res Commun*. 2016;469(4):967-977. doi:10.1016/j.bbrc.2015.12.083

226. Wallace RJ, Rooke JA, McKain N, et al. The rumen microbial metagenome associated with high methane production in cattle. *BMC Genomics*. 2015. doi:10.1186/s12864-015-2032-0

227. Sunagawa S, Coelho LP, Chaffron S, et al. Structure and function of the global ocean microbiome. *Science (80-)*. 2015. doi:10.1126/science.1261359

228. Nagarajan N, Pop M. Sequence assembly demystified. *Nat Rev Genet*. 2013.

doi:10.1038/nrg3367

229. Grabherr MG, Haas BJ, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 2011. doi:10.1038/nbt.1883

230. Robertson G, Schein J, Chiu R, et al. De novo assembly and analysis of RNA-seq data. *Nat Methods*. 2010. doi:10.1038/nmeth.1517

231. Bankevich A, Nurk S, Antipov D, et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 2012. doi:10.1089/cmb.2012.0021

232. Li S, Tighe SW, Nicolet CM, et al. Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nat Biotechnol*. 2014. doi:10.1038/nbt.2972

233. Su Z, Łabaj PP, Li S, et al. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat Biotechnol*. 2014. doi:10.1038/nbt.2957

234. Mangul S, Yang HT, Strauli N, et al. ROP: dumpster diving in RNA-sequencing to find the source of 1 trillion reads across diverse adult human tissues. *Genome Biol*. 2018;19(1):36. doi:10.1186/s13059-018-1403-7

235. Cavadas B, Ferreira J, Camacho R, Fonseca NA, Pereira L. QmihR: Pipeline for Quantification of Microbiome in Human RNA-seq. In: Springer, Cham; 2017:173-179. doi:10.1007/978-3-319-60816-7_21

236. Simon LM, Karg S, Westermann AJ, et al. MetaMap: an atlas of metatranscriptomic reads in human disease-related RNA-seq data. *Gigascience*. 2018;7(6). doi:10.1093/gigascience/giy070

237. Gihawi A, Rallapalli G, Hurst R, Cooper CS, Leggett RM, Brewer DS. SEPATH: benchmarking the search for pathogens in human tissue whole genome sequence data leads to template pipelines. *Genome Biol*. 2019;20(1):208. doi:10.1186/s13059-019-1819-8

238. Cox JW, Ballweg RA, Taft DH, Velayutham P, Haslam DB, Porollo A. A fast and robust protocol for metataxonomic analysis using RNAseq data. *Microbiome*. 2017;5(1):7. doi:10.1186/s40168-016-0219-5

239. Deutsch SI, Mohs RC, Davis KL. A rationale for studying the transmissibility of Alzheimer's disease. *Neurobiol Aging*. 1982;3(2):145-147. doi:10.1016/0197-4580(82)90011-2

240. Taylor GR, Crow TJ, Markakis DA, Lofthouse R, Neeley S, Carter GI. Herpes simplex virus and Alzheimer's disease: A search for virus DNA by spot hybridisation. *J Neurol Neurosurg Psychiatry*. 1984;47(10):1061-1065. doi:10.1136/jnnp.47.10.1061

241. Sochocka M, Zwolińska K, Leszek J. The Infectious Etiology of Alzheimer's Disease. *Curr Neuropharmacol*. 2017;15(7). doi:10.2174/1570159x15666170313122937

242. Irkeç C. [Virologic and immunologic considerations in Parkinson's disease]. *Mikrobiyol Bul*. 1982;16(4):293-296. http://www.ncbi.nlm.nih.gov/pubmed/6304477. Accessed December 9, 2019.

243. Abushouk AI, El-Husseny MWA, Magdy M, et al. Evidence for association between hepatitis C virus and Parkinson's disease. *Neurol Sci*. 2017;38(11):1913-1920. doi:10.1007/s10072-017-3077-4

244. Parashar A, Udayabanu M. Gut microbiota: Implications in Parkinson's disease. *Park Relat Disord*. 2017;38:1-7. doi:10.1016/j.parkreldis.2017.02.002

245. Libbey JE, Cusick MF, Fujinami RS. Role of pathogens in multiple sclerosis. *Int Rev Immunol*. 2014;33(4):266-283. doi:10.3109/08830185.2013.823422

246. Kohne DE, Gibbs CJ, White L, Tracy SM, Meinke W, Smith RA. Virus detection by nucleic acid hybridization: Examination of normal and ALS tissues for the presence of poliovirus. *J Gen Virol*. 1981;56(2):223-233. doi:10.1099/0022-1317-56-2-223

247. Pertschuk LP, Broome JD, Brigati DJ, et al. JEJUNAL IMMUNOPATHOLOGY IN AMYOTROPHIC LATERAL SCLEROSIS AND MULTIPLE SCLEROSIS IDENTIFICATION OF VIRAL ANTIGENS BY IMMUNOFLUORESCENCE. *Lancet*. 1977;309(8022):1119-1123. doi:10.1016/S0140-6736(77)92382-0

248. Xue YC, Feuer R, Cashman N, Luo H. Enteroviral Infection: The Forgotten Link to Amyotrophic Lateral Sclerosis? *Front Mol Neurosci*. 2018;11:63. doi:10.3389/fnmol.2018.00063

249. Alonso R, Pisa D, Fernández-Fernández AM, Rábano A, Carrasco L. Fungal infection in neural tissue of patients with amyotrophic lateral sclerosis. *Neurobiol Dis*. 2017;108:249-260. doi:10.1016/j.nbd.2017.09.001

250. Andrade FC, Vergetti V, Cozza G, Falcao MC, Azevedo G. Amyotrophic Lateral Sclerosis-like Syndrome after Chikungunya. *Cureus*. October 2019. doi:10.7759/cureus.5876

251. Foster JA, Lyte M, Meyer E, Cryan JF. Gut microbiota and brain function: An evolving field in neuroscience. *Int J Neuropsychopharmacol*. 2016. doi:10.1093/ijnp/pyv114

252. Dinan TG, Cryan JF. Gut instincts: microbiota as a key regulator of brain development, ageing and neurodegeneration. *J Physiol*. 2017. doi:10.1113/JP273106

253. Pawate S, Sriram S. The role of infections in the pathogenesis and course of multiple sclerosis. *Ann Indian Acad Neurol*. 2010. doi:10.4103/0972-2327.64622

254. Fierz W. Multiple sclerosis: an example of pathogenic viral interaction? *Virol J*. 2017. doi:10.1186/s12985-017-0719-3

255. Engdahl E, Gustafsson R, Huang J, et al. Increased Serological Response Against Human Herpesvirus 6A Is Associated With Risk for Multiple Sclerosis. *Front Immunol*. 2019. doi:10.3389/fimmu.2019.02715

256. Guan Y, Jakimovski D, Ramanathan M, Weinstock-Guttman B, Zivadinov R. The role of Epstein-Barr virus in multiple sclerosis: From molecular pathophysiology to in vivo imaging. *Neural Regen Res*. 2019. doi:10.4103/1673-5374.245462

257. Cossu D, Yokoyama K, Hattori N. Bacteria–host interactions in multiple sclerosis. *Front Microbiol*. 2018. doi:10.3389/fmicb.2018.02966

258. Askarova S, Umbayev B, Masoud AR, et al. The Links Between the Gut Microbiome, Aging, Modern Lifestyle and Alzheimer's Disease. *Front Cell Infect Microbiol*. 2020. doi:10.3389/fcimb.2020.00104

259. Sun J, Zhang S, Zhang X, Zhang X, Dong H, Qian Y. IL-17A is implicated in lipopolysaccharide-induced neuroinflammation and cognitive impairment in aged rats via microglial activation. *J Neuroinflammation*. 2015. doi:10.1186/s12974-015-0394-5

260. Köhler C, Maes M, Slyepchenko A, et al. The Gut-Brain Axis, Including the Microbiome, Leaky Gut and Bacterial Translocation: Mechanisms and Pathophysiological Role in Alzheimer's Disease. *Curr Pharm Des*. 2016. doi:10.2174/1381612822666160907093807

261. Van Gerven N, Van der Verren SE, Reiter DM, Remaut H. The Role of Functional Amyloids in Bacterial Virulence. *J Mol Biol*. 2018. doi:10.1016/j.jmb.2018.07.010

262. Horvath I, Weise CF, Andersson EK, et al. Mechanisms of protein oligomerization: Inhibitor of functional amyloids templates α-synuclein fibrillation. *J Am Chem Soc*. 2012. doi:10.1021/ja209829m

263. Zhan X, Stamova B, Sharp FR. Lipopolysaccharide associates with amyloid plaques, neurons and oligodendrocytes in Alzheimer's disease brain: A review. *Front Aging Neurosci*. 2018. doi:10.3389/fnagi.2018.00042

264. Alonso R, Pisa D, Carrasco L. Searching for Bacteria in Neural Tissue From Amyotrophic Lateral Sclerosis. *Front Neurosci*. 2019;13:171. doi:10.3389/fnins.2019.00171

265. Gil C, González AAS, León IL, et al. Detection of Mycoplasmas in Patients with Amyotrophic Lateral Sclerosis. *Adv Microbiol*. 2014;04(11):712-719. doi:10.4236/aim.2014.411077

266. Alonso R, Pisa D, Marina AI, et al. Evidence for fungal infection in cerebrospinal fluid and brain tissue from patients with amyotrophic lateral sclerosis. *Int J Biol Sci*. 2015;11(5):546-558. doi:10.7150/ijbs.11084

267. Pisa D, Alonso R, Rábano A, Carrasco L. Corpora Amylacea of Brain Tissue from Neurodegenerative Diseases Are Stained with Specific Antifungal Antibodies. *Front Neurosci*. 2016;10:86. doi:10.3389/fnins.2016.00086

268. Cermelli C, Vinceti M, Beretti F, et al. Risk of sporadic amyotrophic lateral sclerosis associated with seropositivity for herpesviruses and echovirus-7. *Eur J Epidemiol*. 2003;18(2):123-127. doi:10.1023/a:1023067728557

269. Berger MM, Kopp N, Vital C, Redl B, Aymard M, Lina B. Detection and cellular localization

of enterovirus RNA sequences in spinal cord of patients with ALS. *Neurology*. 2000;54(1):20-25. doi:10.1212/wnl.54.1.20

270. Vandenberghe N, Leveque N, Corcia P, et al. Cerebrospinal fluid detection of enterovirus genome in ALS: A study of 242 patients and 354 controls. *Amyotroph Lateral Scler*. 2010;11(3):277-282. doi:10.3109/17482960903262083

271. Xue YC, Feuer R, Cashman N, Luo H. Enteroviral infection: The forgotten link to amyotrophic lateral sclerosis? *Front Mol Neurosci*. 2018;11. doi:10.3389/fnmol.2018.00063

272. Giraud P, Beaulieux F, Ono S, Shimizu N, Chazot G, Lina B. Detection of enteroviral sequences from frozen spinal cord samples of Japanese ALS patients. *Neurology*. 2001;56(12):1777-1778. doi:10.1212/wnl.56.12.1777

273. Verma A, Berger JR. ALS syndrome in patients with HIV-1 infection. *J Neurol Sci*. 2006;240(1-2):59-64. doi:10.1016/j.jns.2005.09.005

274. Moodley K, Bill PLA, Bhigjee AI, Patel VB. A comparative study of motor neuron disease in HIV-infected and HIV-uninfected patients. *J Neurol Sci*. 2019;397:96-102. doi:10.1016/J.JNS.2018.12.030

275. Douville R, Liu J, Rothstein J, Nath A. Identification of active loci of a human endogenous retrovirus in neurons of patients with amyotrophic lateral sclerosis. *Ann Neurol*. 2011;69(1):141-151. doi:10.1002/ana.22149

276. Li W, Lee M-H, Henderson L, et al. Human endogenous retrovirus-K contributes to motor neuron disease. *Sci Transl Med*. 2015;7(307):307ra153-307ra153. doi:10.1126/scitranslmed.aac8201

277. Arru G, Mameli G, Deiana GA, et al. Humoral immunity response to human endogenous retroviruses K/W differentiates between amyotrophic lateral sclerosis and other neurological diseases. *Eur J Neurol*. 2018;25(8):1076-e84. doi:10.1111/ene.13648

278. Blacher E, Bashiardes S, Shapiro H, et al. Potential roles of gut microbiome and metabolites in modulating ALS in mice. *Nature*. 2019;572(7770):474-480. doi:10.1038/s41586-019-1443-5

279. Fang X, Wang X, Yang S, et al. Evaluation of the microbial diversity in amyotrophic lateral sclerosis using high-throughput sequencing. *Front Microbiol*. 2016;7(SEP). doi:10.3389/fmicb.2016.01479

280. Sun J, Zhan Y, Mariosa D, et al. Antibiotics use and risk of amyotrophic lateral sclerosis in Sweden. *Eur J Neurol*. 2019;26(11):1355-1361. doi:10.1111/ene.13986

281. Zhang Y guo, Wu S, Yi J, et al. Target Intestinal Microbiota to Alleviate Disease Progression in Amyotrophic Lateral Sclerosis. *Clin Ther*. 2017;39(2):322-336. doi:10.1016/j.clinthera.2016.12.014

282. Obrenovich M, Jaworski H, Tadimalla T, et al. The role of the microbiota–gut–brain axis

and antibiotics in ALS and neurodegenerative diseases. *Microorganisms*. 2020;8(5). doi:10.3390/microorganisms8050784

283. Brenner D, Hiergeist A, Adis C, et al. The fecal microbiome of ALS patients. *Neurobiol Aging*. 2018;61:132-137. doi:10.1016/j.neurobiolaging.2017.09.023

284. Lindquist S. The Heat-Shock Response. *Annu Rev Biochem*. 1986;55(1):1151-1191. doi:10.1146/annurev.bi.55.070186.005443

285. Gomez-Pastor R, Burchfiel ET, Thiele DJ. Regulation of heat shock transcription factors and their roles in physiology and disease. *Nat Rev Mol Cell Biol*. 2018;19(1):4-19. doi:10.1038/nrm.2017.73

286. Brewer-Jensen P, Wilson CB, Abernethy J, Mollison L, Card S, Searles LL. Suppressor of sable [Su(s)] and Wdr82 down-regulate RNA from heat-shock-inducible repetitive elements by a mechanism that involves transcription termination. *RNA*. 2016;22(1):139-154. doi:10.1261/rna.048819.114

287. Shalgi R, Hurt JA, Lindquist S, Burge CB. Widespread inhibition of posttranscriptional splicing shapes the cellular transcriptome following heat shock. *Cell Rep*. 2014;7(5):1362-1370. doi:10.1016/j.celrep.2014.04.044

288. Gidalevitz T, Prahlad V, Morimoto RI. The stress of protein misfolding: From single cells to multicellular organisms. *Cold Spring Harb Perspect Biol*. 2011;3(6):1-18. doi:10.1101/cshperspect.a009704

289. Westwood JT, Clos J, Wu C. Stress-induced oligomerization and chromosomal relocalization of heat-shock factor. *Nature*. 1991;353:822-827. doi:10.1038/353822a0

290. Åkerfelt M, Morimoto RI, Sistonen L. Heat shock factors: Integrators of cell stress, development and lifespan. *Nat Rev Mol Cell Biol*. 2010;11(8):545-555. doi:10.1038/nrm2938

291. Lindquist S. Regulation of protein synthesis during heat shock. *Nature*. 1981;293(5830):311-314. doi:10.1038/293311a0

292. Shalgi R, Hurt JA, Krykbaeva I, Taipale M, Lindquist S, Burge CB. Widespread Regulation of Translation by Elongation Pausing in Heat Shock. *Mol Cell*. 2013;49(3):439-452. doi:10.1016/j.molcel.2012.11.028

293. Sonenberg N, Hinnebusch AG. Regulation of Translation Initiation in Eukaryotes: Mechanisms and Biological Targets. *Cell*. 2009;136(4):731-745. doi:10.1016/j.cell.2009.01.042

294. Smith HL, Li W, Cheetham ME. Molecular chaperones and neuronal proteostasis. *Semin Cell Dev Biol*. 2015;40:142-152. doi:10.1016/j.semcdb.2015.03.003

295. Jolly C, Metz A, Govin J, et al. Stress-induced transcription of satellite III repeats. *J Cell Biol*. 2004;164(1):25-33. doi:10.1083/jcb.200306104

296. Morton EA, Lamitina T. Caenorhabditis elegans HSF-1 is an essential nuclear protein that forms stress granule-like structures following heat shock. *Aging Cell*. 2013;12(1):112-120. doi:10.1111/acel.12024

297. Rutkowski AJ, Erhard F, L'Hernault A, et al. Widespread disruption of host transcription termination in HSV-1 infection. *Nat Commun*. 2015;6(May). doi:10.1038/ncomms8126

298. Morton EA, Lamitina T. Caenorhabditis elegans HSF-1 is an essential nuclear protein that forms stress granule-like structures following heat shock. *Aging Cell*. 2013;12(1):112-120. doi:10.1111/acel.12024

299. Parker GS, Eckert DM, Bass BL. RDE-4 preferentially binds long dsRNA and its dimerization is necessary for cleavage of dsRNA to siRNA. *RNA*. 2006;12(5):807-818. doi:10.1261/rna.2338706

300. Kaneko H, Dridi S, Tarallo V, et al. DICER1 deficit induces Alu RNA toxicity in age-related macular degeneration. *Nature*. 2011;471(7338):325-332. doi:10.1038/nature09830

301. Daume M, Uhl M, Backofen R, Randau L. RIP-seq suggests translational regulation by L7Ae in Archaea. *MBio*. 2017;8(4):e00730-17. doi:10.1128/mBio.00730-17

302. Brunquell J, Morris S, Lu Y, Cheng F, Westerheide SD. The genome-wide role of HSF-1 in the regulation of gene expression in Caenorhabditis elegans. *BMC Genomics*. 2016;17(1):1-18. doi:10.1186/s12864-016-2837-5

303. Brunquell J, Morris S, Lu Y, Cheng F, Westerheide SD. The genome-wide role of HSF-1 in the regulation of gene expression in Caenorhabditis elegans. *BMC Genomics*. 2016;17(1):559. doi:10.1186/s12864-016-2837-5

304. Klopfenstein D V., Zhang L, Pedersen BS, et al. GOATOOLS: A Python library for Gene Ontology analyses. *Sci Rep*. 2018;8(1):10872. doi:10.1038/s41598-018-28948-z

305. Cardiello JF, Goodrich JA, Kugel JF. Heat shock causes a reversible increase in RNA polymerase II occupancy downstream of mRNA genes consistent with a global loss in transcriptional termination. *Mol Cell Biol*. 2018;38(18):MCB.00181-18. doi:10.1128/MCB.00181-18

306. Zhang T, Hwang HY, Hao H, Talbot C, Wang J. Caenorhabditis elegans RNA-processing protein TDP-1 regulates protein homeostasis and life span. *J Biol Chem*. 2012;287(11):8371-8382. doi:10.1074/jbc.M111.311977

307. Curran SP, Ruvkun G. Lifespan regulation by evolutionarily conserved genes essential for viability. *PLoS Genet*. 2007;3(4):e56. doi:10.1371/journal.pgen.0030056

308. Brenner S. The genetics of Caenorhabditis elegans. *Genetics*. 1974;77(1):71-94. doi:10.1002/cbic.200300625

309. Porta-de-la-Riva M, Fontrodona L, Villanueva A, Cerón J. Basic Caenorhabditis elegans methods: synchronization and observation. *J Vis Exp*. 2012;(64):e4019. doi:10.3791/4019

310. Orjalo A, Johansson HE, Ruth JL. Stellaris fluorescence in situ hybridization (FISH) probes: A powerful tool for mRNA detection. *Nat Methods*. 2011;8(10):i-ii. doi:10.1038/nmeth.f.349

311. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*. 2013;14(2):178-192. doi:10.1093/bib/bbs017

312. Andrews S. FastQC: A quality control tool for high throughput sequence data. *available from http//www.bioinformatics.babraham.ac.uk/projects/fastqc/*. 2017:1.

313. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114-2120. doi:10.1093/bioinformatics/btu170

314. Dobin A, Gingeras TR. Mapping RNA-seq Reads with STAR. *Curr Protoc Bioinforma*. 2015;51:11.14.1-11.14.19. doi:10.1002/0471250953.bi1114s51

315. Liao Y, Smyth GK, Shi W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 2014;30(7):923-930. doi:10.1093/bioinformatics/btt656

316. Mangul S, Yang HT, Strauli N, et al. ROP: dumpster diving in RNA-sequencing to find the source of 1 trillion reads across diverse adult human tissues. *Genome Biol*. 2018;19(1):36. doi:10.1186/s13059-018-1403-7

317. Patrick KL, Bell SL, Weindel CG, Watson RO. Exploring the “multiple-hit hypothesis” of neurodegenerative disease: Bacterial infection comes up to bat. *Front Cell Infect Microbiol*. 2019. doi:10.3389/fcimb.2019.00138

318. Castanedo-Vazquez D, Bosque-Varela P, Sainz-Pelayo A, Riancho J. Infectious agents and amyotrophic lateral sclerosis: another piece of the puzzle of motor neuron degeneration. *J Neurol*. 2019. doi:10.1007/s00415-018-8919-3

319. Mehta P, Kaye W, Raymond J, et al. Prevalence of amyotrophic lateral sclerosis — United States, 2015. *Morb Mortal Wkly Rep*. 2018;67(46):1285-1289. doi:10.15585/mmwr.mm6746a1

320. Talbott EO, Malek AM, Lacomis D. The epidemiology of amyotrophic lateral sclerosis. In: *Handbook of Clinical Neurology*. Vol 138. Elsevier B.V.; 2016:225-238. doi:10.1016/B978-0-12-802973-2.00013-6

321. Ingre C, Roos PM, Piehl F, Kamel F, Fang F. Risk factors for amyotrophic lateral sclerosis. *Clin Epidemiol*. 2015. doi:10.2147/CLEP.S37505

322. Zhan Y, Fang F. Smoking and amyotrophic lateral sclerosis: A mendelian randomization study. *Ann Neurol*. 2019. doi:10.1002/ana.25443

323. Opie-Martin S, Wootton RE, Budu-Aggrey A, et al. Relationship between smoking and ALS: Mendelian randomisation interrogation of causality. *J Neurol Neurosurg Psychiatry*. 2020. doi:10.1136/jnnp-2020-323316

324. Trageser KJ, Smith C, Herman FJ, Ono K, Pasinetti GM. Mechanisms of Immune Activation by c9orf72-Expansions in Amyotrophic Lateral Sclerosis and Frontotemporal Dementia. *Front Neurosci*. 2019;13. doi:10.3389/fnins.2019.01298

325. Burberry A, Wells MF, Limone F, et al. C9orf72 suppresses systemic and neural inflammation induced by gut bacteria. *Nature*. 2020;582(7810):89-94. doi:10.1038/s41586-020-2288-7

326. Verber NS, Shepheard SR, Sassani M, et al. Biomarkers in motor neuron disease: A state of the art review. *Front Neurol*. 2019;10(APR):291. doi:10.3389/fneur.2019.00291

327. Blasco H, Corcia P, Moreau C, et al. 1H-NMR-Based metabolomic profiling of CSF in early amyotrophic lateral sclerosis. *PLoS One*. 2010;5(10). doi:10.1371/journal.pone.0013223

328. Blasco H, Veyrat-Durebex C, Bocca C, et al. Lipidomics Reveals Cerebrospinal-Fluid Signatures of ALS. *Sci Rep*. 2017;7(1). doi:10.1038/s41598-017-17389-9

329. Mitchell RM, Freeman WM, Randazzo WT, et al. A CSF biomarker panel for identification of patients with amyotrophic lateral sclerosis. *Neurology*. 2009;72(1):14-19. doi:10.1212/01.wnl.0000333251.36681.a5

330. Guo J, Yang X, Gao L, Zang D. Evaluating the levels of CSF and serum factors in ALS. *Brain Behav*. 2017;7(3). doi:10.1002/brb3.637

331. Young PE, Jew SK, Buckland ME, Pamphlett R, Suter CM. Epigenetic differences between monozygotic twins discordant for amyotrophic lateral sclerosis (ALS) provide clues to disease pathogenesis. *PLoS One*. 2017;12(8). doi:10.1371/journal.pone.0182638

332. Coppedè F, Stoccoro A, Mosca L, et al. Increase in DNA methylation in patients with amyotrophic lateral sclerosis carriers of not fully penetrant SOD1 mutations. *Amyotroph Lateral Scler Front Degener*. 2018;19(1-2):93-101. doi:10.1080/21678421.2017.1367401

333. Swindell WR, Kruse CPS, List EO, Berryman DE, Kopchick JJ. ALS blood expression profiling identifies new biomarkers, patient subgroups, and evidence for neutrophilia and hypoxia. *J Transl Med*. 2019;17(1):170. doi:10.1186/s12967-019-1909-0

334. Waller R, Wyles M, Heath PR, et al. Small RNA sequencing of sporadic amyotrophic lateral sclerosis cerebrospinal fluid reveals differentially expressed miRNAs related to neural and glial activity. *Front Neurosci*. 2018;11(JAN). doi:10.3389/fnins.2017.00731

335. Waller R, Goodall EF, Milo M, et al. Serum miRNAs miR-206, 143-3p and 374b-5p as potential biomarkers for amyotrophic lateral sclerosis (ALS). *Neurobiol Aging*. 2017;55:123-131. doi:10.1016/j.neurobiolaging.2017.03.027

336. Gendron TF, Chew J, Stankowski JN, et al. Poly(GP) proteins are a useful pharmacodynamic marker for C9ORF72-associated amyotrophic lateral sclerosis. *Sci Transl Med*. 2017;9(383):7866. doi:10.1126/scitranslmed.aai7866

337. Gagliardi S, Zucca S, Pandini C, et al. Long non-coding and coding RNAs characterization in Peripheral Blood Mononuclear Cells and Spinal Cord from Amyotrophic Lateral

Sclerosis patients. *Sci Rep*. 2018;8(1):2378. doi:10.1038/s41598-018-20679-5

338. Zucca S, Gagliardi S, Pandini C, et al. RNA-Seq profiling in peripheral blood mononuclear cells of amyotrophic lateral sclerosis patients and controls. *Sci Data*. 2019;6(1):190006. doi:10.1038/sdata.2019.6

339. Rahman MR, Islam T, Huq F, Quinn JMW, Moni MA. Identification of molecular signatures and pathways common to blood cells and brain tissue of amyotrophic lateral sclerosis patients. *Informatics Med Unlocked*. 2019;16:100193. doi:10.1016/J.IMU.2019.100193

340. van Rheenen W, Diekstra FP, Harschnitz O, et al. Whole blood transcriptome analysis in amyotrophic lateral sclerosis: A biomarker study. *PLoS One*. 2018;13(6):e0198874. doi:10.1371/journal.pone.0198874

341. Parker J, Chen J. Application of next generation sequencing for the detection of human viral pathogens in clinical specimens. *J Clin Virol*. 2017;86:20-26. doi:10.1016/j.jcv.2016.11.010

342. Bouquet J, Gardy JL, Brown S, et al. RNA-Seq Analysis of Gene Expression, Viral Pathogen, and B-Cell/T-Cell Receptor Signatures in Complex Chronic Disease. *Clin Infect Dis*. 2017;64(4):476-481. doi:10.1093/cid/ciw767

343. Westermann AJ, Barquist L, Vogel J. Resolving host-pathogen interactions by dual RNA-seq. *PLoS Pathog*. 2017;13(2):e1006033. doi:10.1371/journal.ppat.1006033

344. Moore RA, Warren RL, Freeman JD, et al. The Sensitivity of Massively Parallel Sequencing for Detecting Candidate Infectious Agents Associated with Human Tissue. Jordan IK, ed. *PLoS One*. 2011;6(5):e19838. doi:10.1371/journal.pone.0019838

345. Poussin C, Sierro N, Boué S, et al. Interrogating the microbiome: experimental and computational considerations in support of study reproducibility. *Drug Discov Today*. 2018;23(9):1644-1657. doi:10.1016/j.drudis.2018.06.005

346. Roumpeka DD, Wallace RJ, Escalettes F, Fotheringham I, Watson M. A review of bioinformatics tools for bio-prospecting from metagenomic sequence data. *Front Genet*. 2017;8(MAR). doi:10.3389/fgene.2017.00023

347. Rampelli S, Soverini M, Turroni S, et al. ViromeScan: A new tool for metagenomic viral community profiling. *BMC Genomics*. 2016;17(1):165. doi:10.1186/s12864-016-2446-3

348. Wood DE, Salzberg SL. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol*. 2014;15(3):R46. doi:10.1186/gb-2014-15-3-r46

349. Fosso B, Santamaria M, D'Antonio M, et al. MetaShot: An accurate workflow for taxon classification of host-associated microbiome from shotgun metagenomic data. *Bioinformatics*. 2017;33(11):1730-1732. doi:10.1093/bioinformatics/btx036

350. Almeida A, Mitchell AL, Boland M, et al. A new genomic blueprint of the human gut microbiota. *Nature*. 2019;568(7753):499-504. doi:10.1038/s41586-019-0965-1

351. Papudeshi B, Haggerty JM, Doane M, et al. Optimizing and evaluating the reconstruction of Metagenome-assembled microbial genomes. *BMC Genomics*. 2017;18(1):915. doi:10.1186/s12864-017-4294-1

352. Humphrys MS, Creasy T, Sun Y, et al. Simultaneous transcriptional profiling of bacteria and their host cells. Ramsey K, ed. *PLoS One*. 2013;8(12):e80597. doi:10.1371/journal.pone.0080597

353. Emanuel W, Kirstin M, Vedran F, et al. Bulk and single-cell gene expression profiling of SARS-CoV-2 infected human cell lines identifies molecular targets for therapeutic intervention. *bioRxiv*. May 2020:2020.05.05.079194. doi:10.1101/2020.05.05.079194

354. Boni MF, Lemey P, Jiang X, et al. Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat Microbiol*. July 2020:1-10. doi:10.1038/s41564-020-0771-4

355. Frazee AC, Jaffe AE, Langmead B, Leek JT. *Polyester* : simulating RNA-seq datasets with differential transcript expression. *Bioinformatics*. 2015;31(17):2778-2784. doi:10.1093/bioinformatics/btv272

356. Ribeiro FJ, Przybylski D, Yin S, et al. Finished bacterial genomes from shotgun sequence data. doi:10.1101/gr.141515.112

357. Ninfali P. Clustal Omega : Multiple Sequence Alignment. *Eur Mol Biol Lab*. 2003:2222-2226. https://www.ebi.ac.uk/Tools/msa/clustalo/. Accessed March 29, 2021.

358. Zhao S, Zhang Y, Gamini R, Zhang B, von Schack D. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Sci Rep*. 2018;8(1):4781. doi:10.1038/s41598-018-23226-4

359. Shin H, Shannon CP, Fishbane N, et al. Variation in RNA-Seq Transcriptome Profiles of Peripheral Whole Blood from Healthy Individuals with and without Globin Depletion. Wang K, ed. *PLoS One*. 2014;9(3):e91041. doi:10.1371/journal.pone.0091041

360. Castillo DJ, Rifkin RF, Cowan DA, Potgieter M. The healthy human blood microbiome: Fact or fiction? *Front Cell Infect Microbiol*. 2019;9(MAY):148. doi:10.3389/fcimb.2019.00148

361. Linsley PS, Speake C, Whalen E, Chaussabel D. Copy number loss of the interferon gene cluster in melanomas is linked to reduced T cell infiltrate and poor patient prognosis. *PLoS One*. 2014;9(10). doi:10.1371/journal.pone.0109760

362. Brohawn DG, O'Brien LC, Bennett JP. RNAseq analyses identify tumor necrosis factor-mediated inflammation as a major abnormality in ALS spinal cord. *PLoS One*. 2016;11(8):e0160520. doi:10.1371/journal.pone.0160520

363. C Ladd A, G Brohawn D, P Bennett J. Laser-captured spinal cord motorneurons from ALS subjects show increased gene expression in vacuolar ATPase networks. *J Syst Integr Neurosci*. 2017;3(6). doi:10.15761/jsin.1000182

364. Bennett JP, Keeney PM, Brohawn DG. RNA Sequencing Reveals Small and Variable

Contributions of Infectious Agents to Transcriptomes of Postmortem Nervous Tissues From Amyotrophic Lateral Sclerosis, Alzheimer's Disease and Parkinson's Disease Subjects, and Increased Expression of Genes From Disease-Activated Microglia. *Front Neurosci*. 2019;13. doi:10.3389/fnins.2019.00235

365. Kowarsky M, Camunas-Soler J, Kertesz M, et al. Numerous uncharacterized and highly divergent microbes which colonize humans are revealed by circulating cell-free DNA. *Proc Natl Acad Sci U S A*. 2017;114(36):9623-9628. doi:10.1073/pnas.1707009114

366. Wiesel Y, Sabath N, Shalgi R. DoGFinder: A software for the discovery and quantification of readthrough transcripts from RNA-seq. *BMC Genomics*. 2018. doi:10.1186/s12864-018-4983-4

367. Melnick M, Gonzales P, Cabral J, Allen MA, Dowell RD, Link CD. Heat shock in C. elegans induces downstream of gene transcription and accumulation of double-stranded RNA. *PLoS One*. 2019;14(4). doi:10.1371/journal.pone.0206715

368. Roth SJ, Heinz S, Benner C. ARTDeco: Automatic readthrough transcription detection. *BMC Bioinformatics*. 2020. doi:10.1186/s12859-020-03551-0

369. Yoon B-J. Hidden Markov Models and their Applications in Biological Sequence Analysis. *Curr Genomics*. 2009. doi:10.2174/138920209789177575

370. van den Burg GJJ, Williams CKI. An evaluation of change point detection algorithms. *arXiv*. 2020.

371. Adams RP, MacKay DJC. Bayesian Online Changepoint Detection. October 2007. http://arxiv.org/abs/0710.3742. Accessed April 28, 2021.

372. Kingma DP, Ba JL. Adam: A method for stochastic optimization. In: *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR; 2015. https://arxiv.org/abs/1412.6980v9. Accessed April 30, 2021.

373. Kass RE, Raftery AE. Bayes factors. *J Am Stat Assoc*. 1995;90(430):773-795. doi:10.1080/01621459.1995.10476572

374. Ester M, Ester M, Kriegel H-P, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. 1996:226--231. https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.121.9220. Accessed May 3, 2021.

375. Braune C, Besecke S, Kruse R. Density based clustering: Alternatives to DBSCAN. In: *Partitional Clustering Algorithms*. Springer International Publishing; 2015:193-213. doi:10.1007/978-3-319-09259-1_6

376. Melnick M, Gonzales P, LaRocca TJ, et al. Application of a bioinformatic pipeline to RNA-seq data identifies novel viruslike sequence in human blood. *G3 Genes|Genomes|Genetics*. April 2021. doi:10.1093/g3journal/jkab141

377. Sampuda KM, Riley M, Boyd L. Stress induced nuclear granules form in response to accumulation of misfolded proteins in Caenorhabditis elegans. *BMC Cell Biol*. 2017;18(1):1-18. doi:10.1186/s12860-017-0136-x

378. Roy M, Viginier B, Saint-Michel É, Arnaud F, Ratinier M, Fablet M. Viral infection impacts transposable element transcript amounts in Drosophila. *Proc Natl Acad Sci U S A*. 2020;117(22):12249-12257. doi:10.1073/pnas.2006106117

379. Broecker F, Moelling K. Evolution of immune systems from viruses and transposable elements. *Front Microbiol*. 2019;10(JAN). doi:10.3389/fmicb.2019.00051

380. Saleh A, Macia A, Muotri AR. Transposable elements, inflammation, and neurological disease. *Front Neurol*. 2019;10(AUG):894. doi:10.3389/fneur.2019.00894

381. Blanco-Melo D, Nilsson-Payant BE, Liu WC, et al. Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19. *Cell*. 2020;181(5):1036-1045.e9. doi:10.1016/j.cell.2020.04.026

382. Zhang X, Chu H, Wen L, et al. Competing endogenous RNA network profiling reveals novel host dependency factors required for MERS-CoV propagation. *Emerg Microbes Infect*. 2020;9(1):733-746. doi:10.1080/22221751.2020.1738277

383. Criscione SW, Zhang Y, Thompson W, Sedivy JM, Neretti N. Transcriptional landscape of repetitive elements in normal and cancer human cells. *BMC Genomics*. 2014;15(1):583. doi:10.1186/1471-2164-15-583

384. Grandi N, Tramontano E. Human endogenous retroviruses are ancient acquired elements still shaping innate immune responses. *Front Immunol*. 2018;9(SEP):2039. doi:10.3389/fimmu.2018.02039

385. Chen J, Foroozesh M, Qin Z. Transactivation of human endogenous retroviruses by tumor viruses and their functions in virus-associated malignancies. *Oncogenesis*. 2019;8(1):1-9. doi:10.1038/s41389-018-0114-y