EXAMINATION OF DYNAMIC LONG RNAS

A Dissertation Presented

By

Christian K. Roy

Submitted to the Faculty of the University of the
Massachusetts Graduate School of Biomedical Sciences,
Worcester
in partial filfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

MAY 21st 2014

BIOCHEMISTRY

EXAMINATION OF DYNAMIC LONG RNAS

A Dissertation Presented

By

Christian K. Roy

The signatures of the Dissertation Defense Committee signify completion and approval as to style and content of the Dissertation

---

Melissa J. Moore, Co-Thesis Advisor

Phillip D. Zamore, Co-Thesis Advisor

Scot Wolfe, Member of Committee

Job Dekker, Member of Committee

The signature of the Chair of the Committee signifies that the written dissertation meets the requirements of the Dissertation Committee

---

Zhiping Weng, Chair of Committee

The signature of the Dean of the Graduate School of Biomedical Sciences signifies that the student has met all graduation requirements of the school.

Anthony Carruthers, Ph.D.,

Dean of the Graduate School of Biomedical Sciences

Biochemistry and Molecular Pharmacology

MAY 21st 2014

UNIVERSITY NAME (IN BLOCK CAPITALS)

Abstract

Faculty Name
BIOCHEMISTRY AND MOLECULAR PHARMACOLOGY

Doctor of Philosophy

Examination of dynamic long RNAs

by Christian Knauf ROY

The Thesis Abstract is written here (and usually kept to just this page). The page is kept centered vertically so can expand into the blank space above the title too...

Acknowledgements

First I would like to acknowledge and thank Laura Geagan and Rebecca Sendak at Genzyme. They were my supervisors while I was a research associate there, and assisted and encouraged my transition back to graduate school. Without the confidence they instilled in my abilities as a young scientist in, I doubt I would have ever signed up for more school.

Next I'd like to thank Melissa. During my 1st year retreat at Wood's Hole I first learned that Melissa is a fantastic communicator of interesting and important science. When she brought out her rope representing the unspliced pre-mRNA of dystrophin—a rope that reached to the back of a rather large auditorium—and then dramatically held up a No.2 pencil representing the final mRNA product, both to scale, I knew the that I wanted to do my graduate research in her lab. I have never once doubted the decision to join Melissa's lab, and have learned so much from the broad and interconnected approach she takes to important scientific questions. Thank you so much for teaching me to always consider the big picture, go for the answer, and to just ask when I need help.

Soon after joining Melissa's lab, and a project going well, it was proposed to me that I be a joint student between Melissa and Phil. It was not difficult to not jump at the opportunity to be advised by two Howard Hughes Investigators, and I also haven't regretted the decision. Over the past few years, I have been continually amazed at the depth of Phil's knowledge, in scientific and general topics. He is a careful, meticulous, quantitative, and calculating mentor. While I feel that I clicked 'on the level' with Melissa, interacting with Phil forced me to think and act outside my comfort zone, something I always tell myself is a critical aspect of change and growth. Thank you Phil for everything I've learned.

My committee has also been very supportive throughput my PhD. I hardly believed the ease with which I passed my qualifying exam, and took it as a big confidence boost. The following years of TRAC meetings confirmed that I was not thinking way off-base. The one-on-one meetings just prior to my QE were

especially helpful. Thanks to Scot, Job, and especially Zhiping for all the guidance.

- Lab Members; Aaron; Alper; Amrit
- Eric and Erin
- Collaborators
- Dave Weaver
- Muro
- Graveley
- Heinrich
- Anna
- Ogo
- Family
- Jul Owen

Contents

List of Figures

List of Tables

List of Abreviations

AS	Alternative Splicing
DNA	Deoxyribonucleic acid
ssDNA	Single-stranded DNA
RNA	Ribonucleic acid
ssRNA	Single-stranded RNA
ATP	Adenosine triphosphate
NAD	Nicotinamide adenine dinucleotide
ChIP-Seq	Chromatin Immunoprecipitation followed by sequencing
HTS	High-throughput sequencing (see also NGS)
NGS	Next-generation sequencing
nt	A nucleotide of either DNA or RNA
bp	A base pair of DNA
SRE	Splicing Regulatory Element
IRE	Intron Recognition Element
CNS	Central Nervous System
TSS	{Transcription or Translation} Start Site
TTS	{Transcription or Translation} Termination Site
SAGE	Serial Analysis of Gene Expression

List of Symbols

5′ The 5 prime end of a DNA or RNA molecule
3′ The 5 prime end of a DNA or RNA molecule
μ Micro. A value of $1x10^{-6}$ standard units

Definitions

RNA-Seq

A technology wherein RNA is fragmented, converted to DNA, and analyzed on a high-throughput sequencing instrument

A 'Read'

The sequence of nucleotides produced from each spot on a high-throughput sequencing machine

Insert

The RNA molecule captured between two cloning sequences in a high-throughput sequencing library preparation workflow

Read length

The number of nucleotides for each given 'read'

Read depth

The number of reads obtained from each high-throughput sequencing analysis

Coverage

A measure of the number of times each nt of a genome is sequenced. E.g. 100 million reads of a 10 million nt genome = 10X coverage, assuming uniform distribution of the 'reads'

Paired-end

oach1995a When both sides of a DNA insert or template are sequenced, utilizing the original length of DNA between the reads to facilitate mapping (Roach et al. [1995]).

Scaffold or contig

A draft sequence of nucleotides, meant to represent the actual biological sequence as closely as possible, examples include unassembled fragments of chromosomes or fragments of mRNA transcripts.

I would like to dedicate this Doctoral dissertation to my grandfather, George Knauf. My grandfather passed away on September 23rd, 2011, just one week shy of his 82nd birthday. I find it difficult to articulate how much I miss him. He spoke carefully and never without purpose or conviction. While I hear from others that he was proud of me, he rarely, if ever, betrayed that type of emotion directly. It is my goal to build as solid a life as he, founded on hard work, playing the long game, responsibility, and maintaining friendships. These are just a few of the personality traits that I observed and try to emulate. The fact that he passed before he could meet our son Owen is one of my biggest regrets. Of all the possessions he left behind, it is the memory of our time together that I will cherish the most. Rest in peace, Grump. I did it.

Preface

The work reported in this dissertation has been published in the following articles. Chapter III has been published previously as Li, X. Z. Z., Roy, C. K. K., Dong, X., Bolcun-Filas, E., Wang, J., Han, B. W. W., ... Zamore, P. D. D. (2013). An Ancient Transcription Factor Initiates the Burst of piRNA Production during Early Meiosis in Mouse Testes. Molecular Cell, 50(1), 1–15. doi:10.1016/j.molcel.2013.02.016

Some contents of Chapter I are included in an accepted for publication:

Chapter 1

Introduction

1.1 On the importance of gene expression

The Old Testament chapter Exodus tells of the liberation of the Israelite people from Egyptian slavery. Their humble and reluctant leader Moses, acting under the direction of God, forces the Pharaoh Ramses to release the people of Israel through a series of 10 plagues. Pharaoh is stalwart and stubborn as he watches water turn to blood. As frogs, lice, and flies flood the city streets, he refuses to free the Israelites. When Egyptian livestock fell dead from disease, people and animals both were covered in boils, and land burned in storms of fire, Pharaoh did not bend. The 8th plague was a swarm of Locusts, described in Exodus 10: 14–15:

> *[14] And the locusts went up over all the land of Egypt, and rested in all the coasts of Egypt: very grievous were they; before them there were no such locusts as they, neither after them shall be such.*
> *[15] For they covered the face of the whole earth, so that the land was darkened; and they did eat every herb of the land, and all the fruit of the trees which the hail had left: and there remained not any green thing in the trees, or in the herbs of the field, through all the land of Egypt.*

The desolation left by the locust plague was not enough to persuade Ramses. Nor was three days of darkness. Only the death of all first-born Egyptians, included Ramses own son, was enough to persuade Pharaoh to let the Israelites leave Egypt.

The power of a locust swarm is not just a fanciful biblical story, and is perhaps the most *believable* of the 10 plagues. In current times, the United Nations' (UN) Food and Agriculture division maintains a Locust watch website providing weekly updates on potential locust swarms in northern Africa and middle east. The locust has long been, and continues to be, a powerful and feared force of Nature.

Unlike fire and brimstone from the heavens, locusts are something we can hold and study. Surely science can help us understand what triggers them to swam and cause massive destruction. We know that the desert locust, Schistocerca gregaria, is the main species of about 10 that swarms in vast numbers and causes extensive crop damage. They are members of the insect

Order Orthoptera, whose other famous members include crickets and katydids. Orthoptern members make sound known as stridulation by vigorously rubbing their wings. They also undergo incomplete metamorphosis (formally Hemimetabolism), and do not have a pupal stage during development.

- While only 2–2.5 inches and weighing 0.05–0.07 oz, can consume its own weight in food per day
- Can fly 60 miles in 5–8 hours
- Thought to be separate species from solitary form until 1921

The power and destruction this animal can inflict makes it difficult to believe that it is nothing more than a grasshopper. It is nothing more than a grasshopper not just by analogy, but by actual Taxonomy. The infamous desert locust is actually the *gregarious* form of *Schistocerca gregaria* (See Figure 1.1), while the more familiar and docile looking Grasshopper is the *solitary form*. Scientists are just now beginning to understand how it is possible for such a dichotomy to exist within the same organism, or more specifically, within the same *genome*.

Schistocerca gregaria is a polyphenic organism. Grasshoppers become locusts by going through a phase transition. Polyphenism is a general feature of insects, often stark in transformation. For example, pea aphids (*Acyrthosiphon pisum*), which usually exist in an asexually reproducing, wingless female form, responds

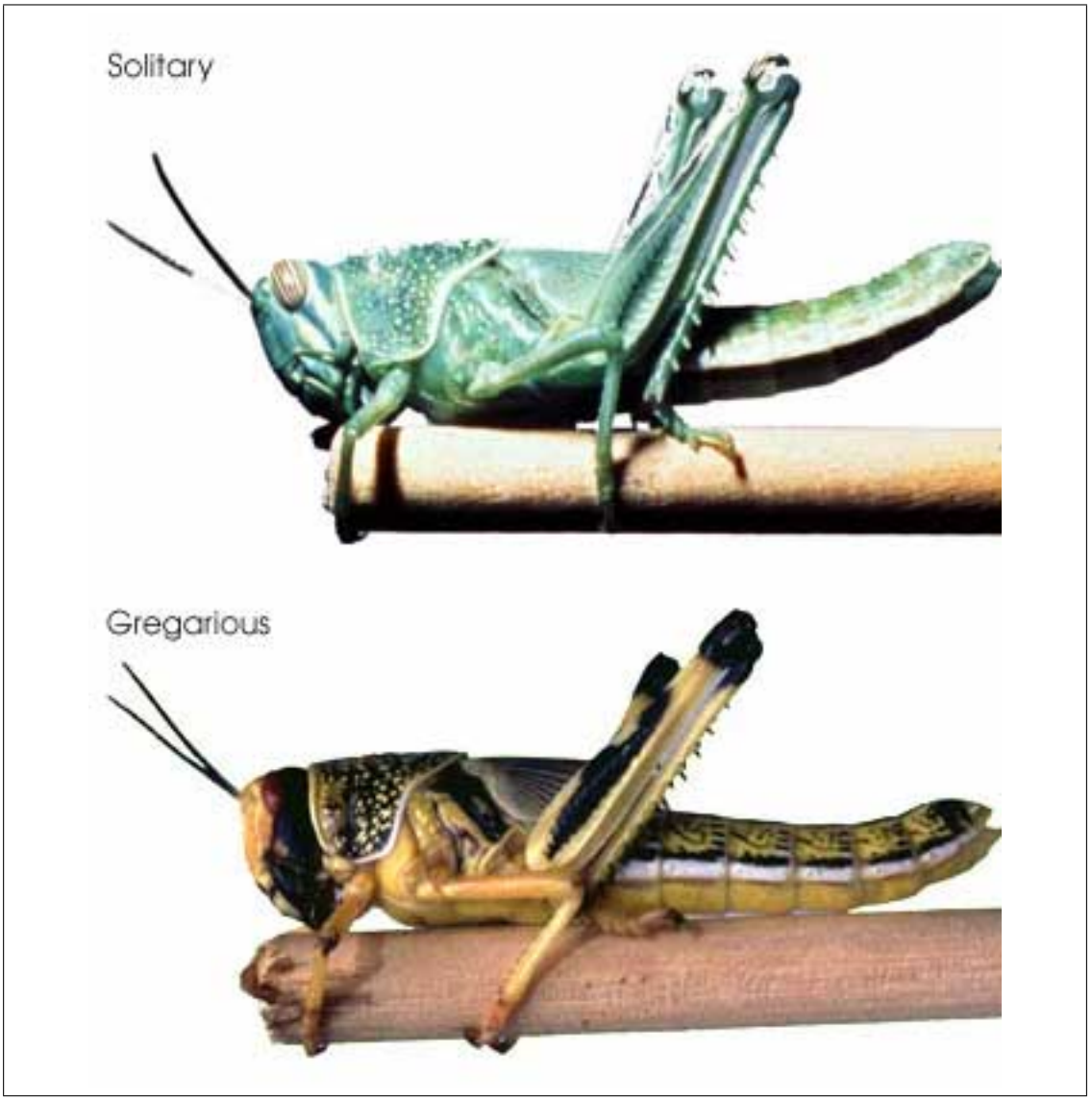

FIGURE 1.1: The Solitary and Gregarious forms of *Schistocerca gregaria*

The two phenotypic forms of Schistocerca gregaria appear very different. The Solitary form is green an generally larger, while its "gregarious" form is more brightly colored, smaller, and capable of swarming in vast numbers, destroying crops vegetation. Photo from Wikicommons.

to overcrowding (often as a result of dwindling food supply) by producing winged offspring that travel to new sources of food [Purandare et al., 2014, Shingleton et al., 2003].

1.2 DNA Sequencing History

Soon after it was realized that DNA is the source of genetic information in all living organisms [Watson and Crick, 1953], and the "pretty" and "elegant" arrangement of complementary, antiparrallel DNA strands was known [Watson et al., 2012], the ability to determine the specific arrangement, or "sequence", of nucleotide bases in a given length of DNA was seen as a critical missing piece of technology. It took 25 years after the nature of DNA's architecture to be able to determine the specific arrangement of nucleotides in the polymer—to sequence it. By 1977, two completely different methods developed by Sanger [Sanger and Coulson, 1975, Sanger et al., 1977] and Maxam-Gilbert [Maxam and Gilbert, 1992] were reported. These sequencing technologies, from then on referred to eponymously as 'Sanger' or 'Maxam-Gilbert' sequencing, were used to determine the specific order of a small piece of DNA (200–300 nt). Sanger sequencing soon dominated most sequencing reactions, likely due to the conceptually more intuitive nature of the technology, and over the past 35 years,

DNA sequences have been slowly cloned, sequenced, analyzed, and dutifully cataloged into knowledge.

During the late 1970's and throughout the 1980's, DNA sequences were typically communicated in important publications [Bell et al., 1980, Sanger et al., 1978]. The birth of the Internet in the 1990's made essential publically-funded repositories for sequence information easily available [Benson et al., 2011]. However, it was the human genome project [Lander, 2011, Venter et al., 2001], that provided the important activation energy that brought DNA sequencing from a hard-to-perform, but necessary, analysis, to an organized large-scale effort of assembling the complete genetic material complex genomes. An often criticized, but undeniably disrupting force in the human genome project was the competing efforts of the privately-owned company Celera [Venter, 2007]. Taking a higher-throughput and centralized approach to determining the sequence of the human genome, Celera fundamentally changed the landscape of genome assembly. Instead of assigning specific sections of the genome to be worked out by individual labs, Celera parallelized the effort, by collecting many of the best "high-throughput" Sanger-sequencing devices from Agilent (ABI 3700 DNA Analyzer). Using "shotgun" approach [Staden, 1979], sequenced pairwise [Roach et al., 1995], and combined with sequence scaffolds made available by the publicly-funded project, Celera was able to assemble high-quality genomic sequences very quickly. Arguably, this was the first deep sequencing effort, and

changed the landscape of molecular and biochemical research, coincident with the beginning of a new millennium.

1.3 History of High-throughput Sequencing

Sequencing DNA by Sanger's technology remains a valuable and critical tool in every biological scientist's arsenal. However, the technology has a practical throughput limit. Each DNA molecule to be sequenced must be isolated and clonally amplified, typically using bacteria. Given that the human genome [Consortium, 2004] comprises > 3 billion nt (on just one strand), and that each Sanger reaction will provide 800nt of quality sequence, we need at least 4 million individual reactions to determine the sequence of the human genome, assuming that all of our reads are of sufficient quality, length, and do not overlap by even 1 nt. Even the best practical improvements to work-flows could not bring the Sanger approach to DNA sequencing in-line with aspirations of analyzing genomes of many different species or individual organisms.

In the early 2000's, efforts to change the approach to DNA sequencing, first using MPSS [Brenner et al., 2000], but perhaps more importantly, by Pyrosequencing [Ronaghi et al., 1998] and Polony sequencing [Shendure et al., 2005]. Both of the latter methods utilize emulsion PCR [Nakano et al., 2003] for clonal amplification prior to sequencing, removing the bottleneck of bacterial cloning.

In contrast to Sanger sequencing, where the signal is from fluorescence of the last incorporated chain-terminating nucleotide, Pyrosequencing visualizes light given off by luciferase as it reacts with ATP generated from the pyrophosphate (PPi) by-product of nucleotide addition. Pyrosequencing has been commercialized by 454 technologies. Polony sequencing involves a more complicated sequencing-by-ligation method, eventually commercialized by Applied Biosystems and branded as SOLiD sequencing. While both of these technologies provided valuable, high-throughput sequences, neither has been as successful as the approach commercialized by Solexa, eventually purchased and now known as Illumina. Illumina uses a sequencing-by-synthesis approach using fluorescent nucleotides after clonal amplification of DNA on a slide surface [Bentley et al., 2008]. Since 2006, iterations of the Illumina platform (eg. GE, GE-II(x), Hi-Seq, Hi-Seq 2500) have demonstrated a steady and impressive increases in both read depth and length. On February 15th 2012, Illumina announced on its Basespace blog, that they had sequenced a HapMap sample at 40X coverage, using the HiSeq 2500 platform and paired-end 100 nt reads in a single run. This announcement demonstrated that in a single analysis attempt (but certainly not the day claimed by the title), analysis and assembly of a human genome is no longer the monumental endeavor it once was, and that completely new experimental possibilities are a reality for life science research.

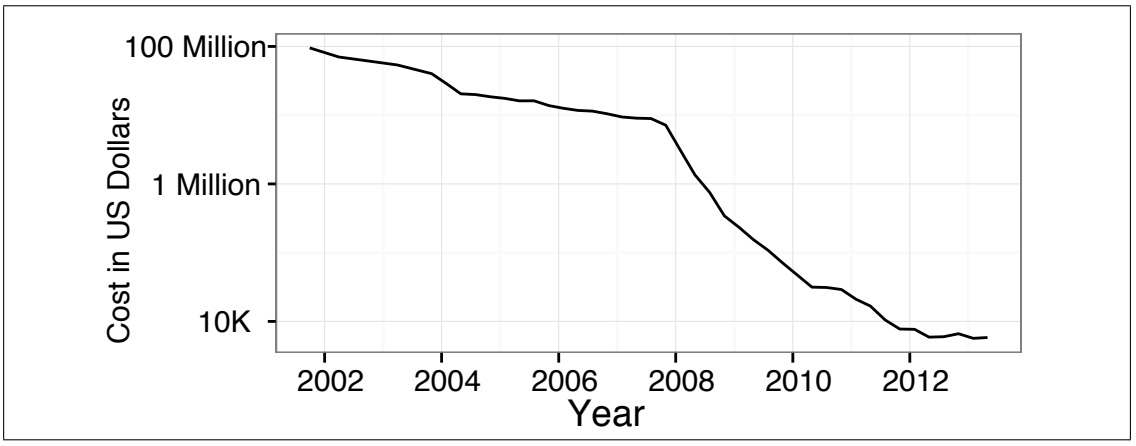

FIGURE 1.2: Cost of sequencing the human genome over time

The costs of sequencing the human genome has decreased on a log scale over a roughly 10 year period thanks to major improvements in high-throughput sequencing. Data from Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcosts. Accessed 2013-09-03).

1.4 Deep-sequencing RNA methodologies

The first widely-accepted method for measuring gene expression via sequencing by proxy of cDNA molecules was Serial Analysis of Gene Expression (SAGE) [Velculescu et al., 1995]. While the importance of microarrays in the measurement of gene expression via cannot be overstated [Marioni et al., 2008, Shendure and Ji, 2008] the technologies limited ability to investigate novel sequences, and analogue signal, makes their relevance to this section somewhat off-topic. However, **SAGE**, (similar to the previously discussed MPSS technique) produces a digital output of gene expression using a cleaver procedure of cleaving cDNA molecules using restriction endonucleases that leaves a sticky

end. After cleavage, these molecules are ligated and concatenated together to form longer DNA fragments. Fragments are cloned into a vector, amplified, and Sanger sequenced. Using known sequences incorporated during concatenation, the number of sequenced 'fragments' that align to a given gene is related to the abundance of the original mRNA molecule. While SAGE was a cleaver molecular trick allowing researches to dip into the 5-log range of expression typically seen in mRNA expression, it is still limited by read lengths and practical read depth of Sanger sequencing. Not long after the Solexa/Illumina platform produced read lengths of sufficient length of depth to consider measuring gene expression were the first RNA-Seq papers published [Lister et al., 2008, Mortazavi et al., 2008, Nagalakshmi et al., 2008]. These papers gave a powerful glimpse into the future of molecular biology. Indeed, in the years since, analysis by RNA-Seq has quickly overtaken other forms of gene expression analysis, as demonstrated by the number of accessions deposited in GEO per year [Barrett et al., 2013]. RNA-Seq allows for digital quantification of RNA expression across physiologically-relevant ranges [Blencowe et al., 2009]. While simultaneously measuring gene expression, the data can be used for novel sequence discovery, measuring RNA-editing [Li et al., 2011], transcript assembly [Trapnell et al., 2010]. By modifying the basic protocol or performing additional biochemical steps, RNA-Seq can be used to investigate many aspects of RNA biology (see 1.3).

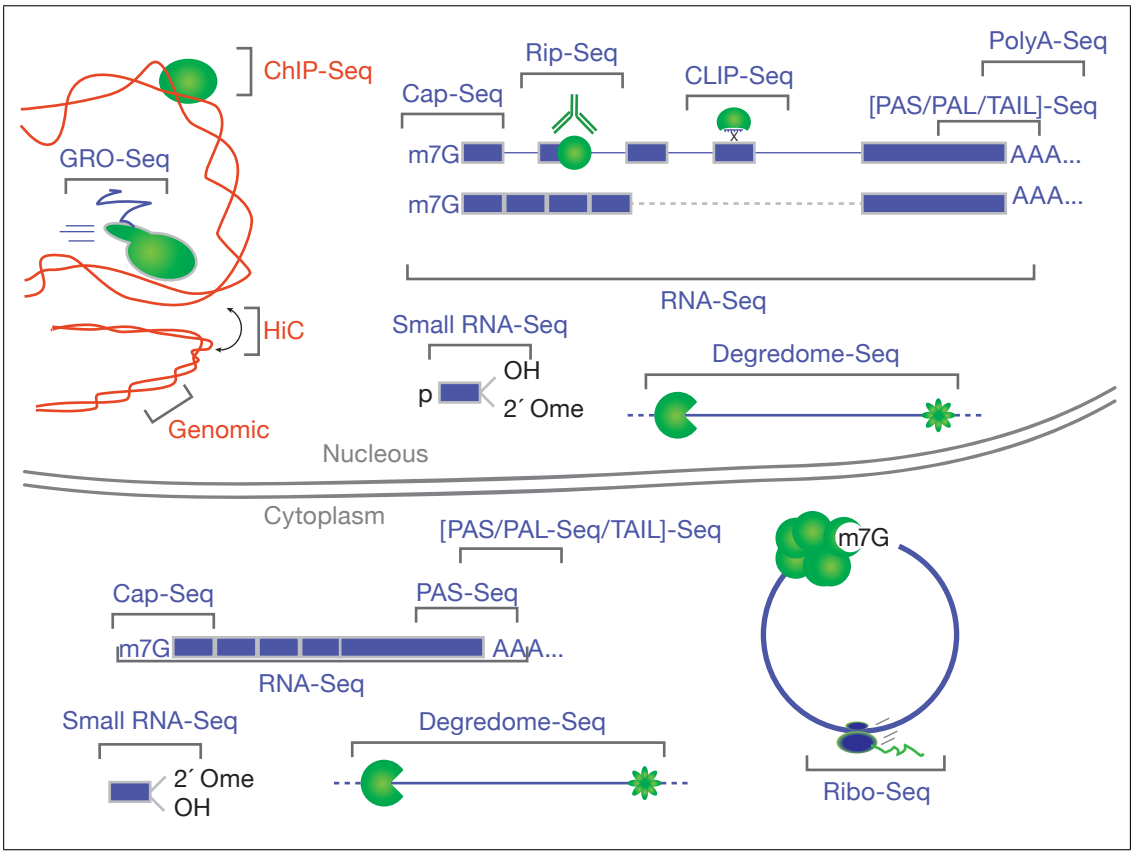

FIGURE 1.3: Methods for High-throughput sequencing of RNA

In the short years since the first report of RNA-Seq, many variations have been reported. The figure above provides an incomplete graphical illustration of some of these variations. A more complete list of *Seq applications is maintained on this blog.

RNA processing begins the moment the nascent RNA is exposed from the polymerase exit channel. Many methodologies have been developed that enrich RNA-Seq libraries for RNA molecules. For example, measurement of nascently transcribed RNA can be performed via GRO-Seq [Core et al., 2008]. Measuring the extremely complicated process of RNA turnover (referring to the rates at which RNAs both are produced and degraded) [Ghosh and Jacobson, 2010],

can be done using XXX-Seq after incorporation of XX nucleotides or a biochemical handle such as biotin. RNA::protein interactions can be measured with or without cross-linking the protein to the RNA, via CLIP or RIP, respectively. Once an RNA has been fully transcribed, known processing steps such as Cap formation and poly(A) tail formation can be measured using any of the Cap-Seq/CAGE methodologies (Shiraki et al. 2003), or PAS-Seq (Shepard et al. 2011). With appropriate size-selection steps, small RNAs [Ghildiyal et al., 2008] can also be captured into a sequencing library. Finally, traditional RNA-Seq, can effectively capture fragments of all of the above mentioned libraries, even though it is mainly associated with measurement or analysis of traditional mRNAs.

RNA-Seq and its associated flavors are also traditionally associated with measuring gene expression in tissue culture cells, or RNA extracted from particular tissues. Recently, efforts to measure the RNA expression occurring in individual cells has gained attention [Shapiro et al., 2013]. Perhaps the most interesting concept when thinking about measurement of gene expression in a single cell is the biological uncertainty principle, wherein it is possible to either know, or change —but not both —the RNA composition of a single cell. The name borrows from Heisenberg's uncertainty principle [Kennard, 1927] and is often confused with the more appropriate 'observed effect' [Riley and Steitz, 2013]. Leaving that issue aside, measuring the unique transcriptome of a given

cell amoung cells of a common tissue is an exciting and informative endeavor [Shalek et al., 2013, Wills et al., 2013]. Compared to DNA, the diversity of RNA synthesis within living cells is potentially much more complicated [Shendure and Aiden, 2012], and the ability to accurately measure RNA dynamics should allow us to make much more informative observations concerning biology then is currently possible [Djebali et al., 2012].

1.5 RNA Expression

The encode project revealed that most of the genome is transcribed into RNA. This was done in cancerous cell lines, and while it revealed the potential for transcription, it did not reveal much biology beyond cells in culture simply perpetuating their existence.

1.6 Alternative Splicing

Soon after the discovery of introns, it was reasoned that genes could be arranged in different combinations, greatly increasing the coding potential of a genome [Gilbert, 1978]. The process of rearranging genes, now known as alternative splicing (AS), has proven to be an integral phase of gene expression in

most eukaryotes. In just 15 years, the number of genes estimated to be alternatively spliced has grown considerably. Phillip Sharp, Co-Nobel-prize winner for the discovery of splicing, stated that: “Approximately, one of every twenty genes is expressed by alternative pathways of RNA splicing in different cell types or growth states” Sharp [2014]. Not long after the assembly of the first human genome, a number of groups combed through Expressed Sequence Tag (EST) databases to increase that estimate to 35%-59% [Modrek and Lee, 2002]. Soon after, analysis using specially designed microarrays resulted in an increased estimate of 74% [Johnson et al., 2003]. However, in late 2008, three groups utilizing high-throughput sequencing (HTS) of cDNA (referred to as RNA-Seq) demonstrated that between 86% and 95% of human multi-exon genes are subject to AS [Pan et al., 2008, Sultan et al., 2008, Wang et al., 2008]. Not only did they demonstrate that almost all genes are alternatively spliced, they also showed that AS often occurs in a tissue- and cell type-specific manner. In combination with regulation of transcription itself, the study of AS is critical to our understanding of the connections between the comparably static genomic DNA sequence and the highly flexible and adaptive abilities of organisms.

1.7 Deciphering a splicing code

A gene is alternatively spliced when, as a result of transcription and processing, there are at least two unique transcripts produced from one genomic sequence. Beyond counting observed isoforms, one major area of effort is to decode sequence regulatory elements (SREs) contained in pre-mRNA that define AS site selection [Wang et al., 2008]. In contrast to the core splicing signals, we have limited knowledge of the SREs that serve to increase, or decrease, the strength of a particular splice site, often within a sea of other potential sites. Through a variety of mechanisms, these elements serve as cis-acting sequences and binding sites for trans-acting factors. Some of the best-studied SREs include Exon Splicing Enhancers and Silencers (ESEs and ESSs). Members of the Serine-Arginine (SR) protein family typically bind to ESEs located in an exon, promoting its definition and thereby increasing the probability that the exon will be included in the final transcript [Graveley, 2000, Long and Caceres, 2009]. Meanwhile, ESSs serve to squelch inclusion, often through binding trans-acting heterogeneous ribonucleoprotein particles (hnRNPs) [Martinez-Contreras et al., 2007]. Therefore, binding of these trans-acting factors to their appropriate SREs can either promote or inhibit interactions between the splicing machinery and the pre-mRNA. The current working hypothesis is that a finely tuned combination of

these binding events determines the final exon content of each isoform [House and Lynch, 2008].

Sequence motifs that compose the AS code have been teased out [Barash et al., 2010, Ladd and Cooper, 2002]. Additionally, assignment of the binding motifs to tissue-specific trans-acting factors has also progressed [Jin et al., 2003, Licatalosi et al., 2008, Ule et al., 2005]. Many of these binding motifs were identified using combined computational and biochemical approaches. Computational approaches usually involve searching for a comparative enrichment of sequences near splice sites. Biochemical approaches typically include gel shift, SELEX, and cross-linking. Many of these approaches are performed in vitro and disregard the importance of cellular context on binding affinities. However, with the increasing accessibility of deep sequencing, many groups are extracting physiologically relevant, high-resolution data from traditional biochemical techniques [Ingolia et al., 2009, 2011]. Deep-sequencing approaches are also being applied to questions involving mechanisms of AS. In addition to the RNA-Seq experiments, High-Throughput Sequencing [following] Cross-Linking Immunoprecipitation (HTS-CLIP) has confirmed SRE motif data predicted from computational and microarray experiments [Hafner et al., 2010, Licatalosi et al., 2008]. Using this approach, researchers can now enrich their samples for sequences that bind trans-acting factors of interest.

1.8 The Isoform Problem

As with many areas of basic research, the field of AS relies on large-scale (aka – global, genome-wide, high-throughput) techniques. Two of the most widely applied technologies employed for large-scale analysis of gene expression are microarrays and '2nd generation' HTS sequencing. Unfortunately, both of these techniques have fundamental limitations, with the major issues being probe specificity for the former and read length for the latter.

Microarrays rely on hybridization of a target sequence to a known probe averaging 25 to 100 nt in length [Southern, 2001]. Therefore, microarrays indicate only the presence of short sequences in the target sample and do not provide adequate linkage information of these sequences. A hypothetical scenario can be used to describe it another way. Say we are investigating a transcript known to display two different regions of AS (See Figure 1.4). Probes targeting these two regions demonstrated an increase in signal for both AS events. Unfortunately, we could not determine if we observed an increase in unique transcripts, each containing only one region of AS, or an increase in production of a single transcript containing both regions [Calarco et al., 2007a]. This binary analysis is the heart of the "connectivity problem." Microarrays have proven extremely

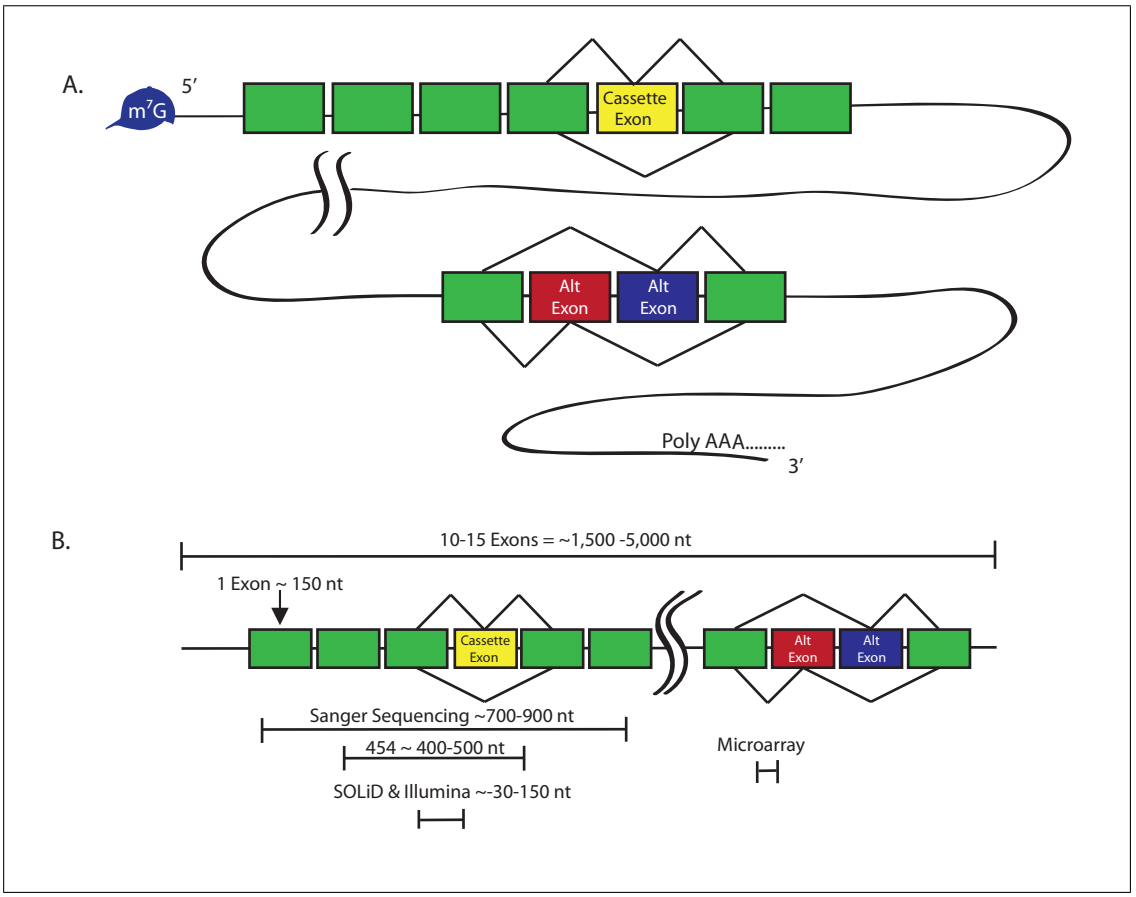

FIGURE 1.4: HTS read lengths are not sufficient to maintain AS connectivity

A) Long RNAs may have multiple sites of AS, separated by 1000's of nt; B) Most mRNAs have 10 exons of 150 nt each. Some have many more (and longer) exons. Read lengths of current sequencing technologies do not maintain connectivity between distant sites.

informative and will likely continue to do so in more targeted applications. However, this issue, combined with concerns of cross-hybridization, reproducibility, and a comparably small dynamic range, will likely hasten microarray displacement by RNA-Seq as the preferred method for comprehensive analysis of gene expression [Shendure and Ji, 2008].

Many researchers are turning toward 2nd generation HTS methodologies for

comprehensive transcriptome analysis. This sequencing approach has significance advantages over microarrays. Specifically, it allows de novo identification of isoforms, over a larger dynamic range, in a quantitative fashion [Mortazavi et al., 2008]. Additionally, techniques exist to enrich samples for low-abundance isoforms, making the complete cataloging of AS events a possibility [Djebali et al., 2008, Salehi-Ashtiani et al., 2008]. Unfortunately, the current read-length abilities (depicted in Figure 1-1,B) of all sequencing platforms do not solve the connectivity problem. Excluding single-molecule HTS read lengths of sufficient length [Shendure et al., 2004], other approaches proposed to solve the connectivity problem include traditional cloning and sequencing or hybridization of query oligos to single-molecule transcripts [Calarco et al., 2007a, Emerick et al., 2007, Zhu et al., 2003]. While these approaches can determine exon sequence connectivity, they scale poorly and are not feasible for large-scale applications. Clearly, AS is an essential regulatory mechanism involved in the control of human gene expression. Its combinatorial nature could potentially answer many questions, such as a physical explanation of what separates us from our closest evolutionary ancestor, the chimpanzee [Calarco et al., 2007b]. Additionally, the influence of AS on disease and cancer is slowly coming to light [Tazi et al., 2009]. Unfortunately, because of the limitations of methods currently used for the large-scale analysis of isoform expression we fail to obtain the complete picture of AS. One specific missing element of that picture is the prevalence

of coordination between different regions of AS separated by large spans of sequence. An efficient, large-scale, single-molecule technique that maintains isoform sequence connectivity is required to complete the complicated picture of AS.

1.9 Coordination in splicing

Identification of proximally acting SREs is progressing at a rapid pace. New and traditional biochemical methods, coupled with HTS, will undoubtedly fuel this progress. Unfortunately, a critical component of AS regulation currently neglected by the field is that of SREs acting across a considerable distance (>800 nt). One observation that may lead to the identification of long-range SREs is intramolecular coordination between distal splicing decisions. Figure 1.4 a model transcript that may exhibit coordinated distal regions of AS. In this model, the 5′ region of AS contains a cassette exon, which may or may not be included. This region is separated from the 3′ region of AS by many thousands of nucleotides. Does the decision to include the cassette exon have an effect on which of the mutually exclusive exons is included? This type of AS regulation may represent a general and pervasive phenomenon.

There is precedence in the literature for genes known to display coordinated regions of AS. One of the clearest examples is mouse Fibronectin *Fn1* (Figure

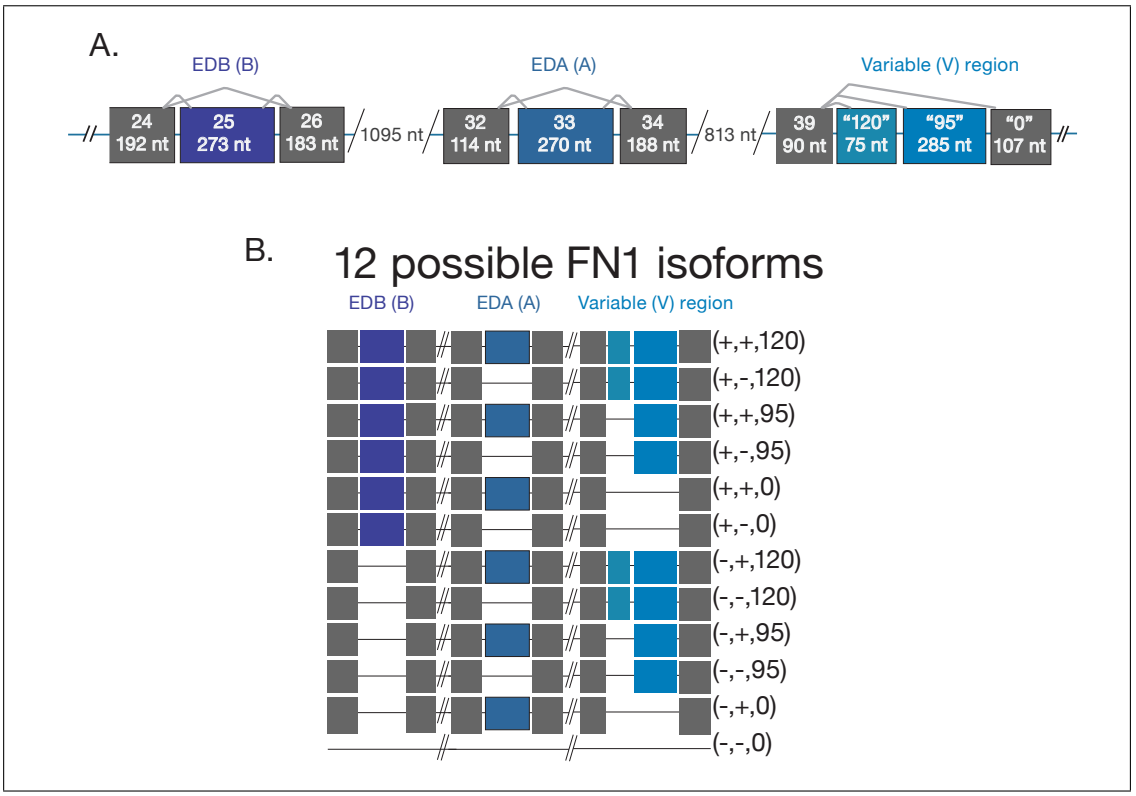

FIGURE 1.5: Mouse *Fn1* contains multiple sites of Alternative Splicing

A) There are three highly-studied regions of AS in mouse *Fn1*: The cassette exons EDB and EDA, and the Variable(V)-region (AKA the IIICS) exon, which displays multiple 3′ splice sites. Each of these sites is separated by multiple constitutive exons.; B) Considering simplistic splicing of these three exons, there are 12 different isoforms of mouse *Fn1*.

1.5) [Schwarzbauer et al., 1983, White and Muro, 2011]. In this gene, inclusion of the alternatively spliced Extra Domain A (EDI or EDA) region promotes splicing from one of three alternative 3′ Splice Site (3′ SS) in the type III homology connecting segment (IIICS) region, resulting in more frequent production of shorter transcripts [Fededa et al., 2005]. This effect occurs over six constitutively expressed exons and 800 nt of sequence (5400 nt if introns are considered). [Fededa et al., 2005] also analyzed EST databases, concluding that

approximately 25% of human genes contain multiple regions of AS. How many of these regions could show a coordinated effect, similar to that observed in Fibronectin? Providing some insight into this question, Fagnani et al used microarrays designed to report on inclusion levels of cassette exons in mammalian central nervous system tissues [Fagnani et al., 2007]. The results produced a set of 38 pairs of exons mapping to the same gene that showed a coordinated increase or decrease of inclusion levels.

There have been a few studies that investigate forms of splicing coordination between adjacent exons present in mRNA. The vertebrate genes 4.1B and 4.1R, members of the protein 4.1 family encoding for cytoskeletal adaptor proteins, both undergo splicing of upstream 5′ first exons to distal 3′ second exons, skipping a stronger proximal 3′ second exon [Parra et al., 2008, 2012]. This is accomplished through 'intrasplicing' involving an intronic sequence element (the 'intraexon') only present when transcription begins at the upstream 5′ exon, allowing the exon to ligate do the weaker distal 3′ second exon via an intermediate splicing event. Importantly, this type of splicing would be similar, but different from recursive splicing seen in drosophila [**?**]. Another example of the importance of intron sequence elements on AS is observed in the equine β-casin gene, where the authors propose a model involving an intronic splicing enhancer bound to the exit channel of the elongating polymerase, promoting inclusion of downstream cassette exons [Lenasi et al., 2006]. Taking a

more genome-wide approach Peng et al. examined human and mouse EST data looking for correlations between adjacent AS cassette exons [Peng et al., 2008]. The authors note that positively correlated pairs of adjacent cassette exons typically resemble constitutive exons in splice strength, whereas negatively, or weakly correlated pairs are likely to be newly emerging exons, whose strength of splicing has not evolved enough to be constitutively included.

The last, most current, and thorough study of intra-gene splicing coordination involves the *Caenorhabditis elegans* gene *slo1* [Glauser et al., 2011, Johnson et al., 2011]. *slo1* is the C.elegans orthologue of the human BK channel gene *Kcnma*, also known to undergo extensive alternative splicing [Nilsen and Graveley, 2010] via 13 cassette exons, potentially coding for over 1,000 different isoforms. *Kcnma* is highly developmentally, spatially, and tissue regulated. It is involved in a diverse range of cellular processes, including hearing, circadian rhythms, urinary function, and vasoregulation [Fodor and Aldrich, 2009]. While the gene is highly conserved, as organism complexity grows, so does the apparent transcriptional diversity of this gene. In worms, *slo1* can produce up to 12 different isoforms. Glauser et al. used QPCR to demonstrate individual, AS region inclusion frequencies do not correspond to complete isoform frequencies, when measured via a TaqMan probe approach. They go on to describe a interdependent-splicing model that best fits the data, and support interdependence via mutations at one sight altering splicing at both upstream

and downstream sites of AS, separated by atleast one other splicing event. After measuring the biophysical properties of the isoforms [Johnson et al., 2011], they conclude that coordinated AS is critical for proper BK channel function in vivo. It is interesting to note that this study also identified an intronic sequence element that displayed some type of coordinated, or co-regulated effect on AS.

1.10 Many isoforms per gene

It is easy to think of AS as a binary process. Isoform A or B is produced based upon picking either exon A or B. What quickly becomes evident, and is far too real for researchers building transcriptome assembly algorithms, is that the combinatorial nature of AS makes it both a power means of generating isoform diversity and a difficult problem to study [Trapnell et al., 2012].

One of the most recent attempts to investigate the breath of combinations produced by AS is the already mentioned ENCODE project [Djebali et al., 2012]. ENCODE performed extremely in depth analysis of 15 cell lines, and find that each isoform produces 10 transcripts per gene, with a broad distribution in terms of isoforms expressed per sample.

The ENCODE project clearly demonstrated that most human genes can under

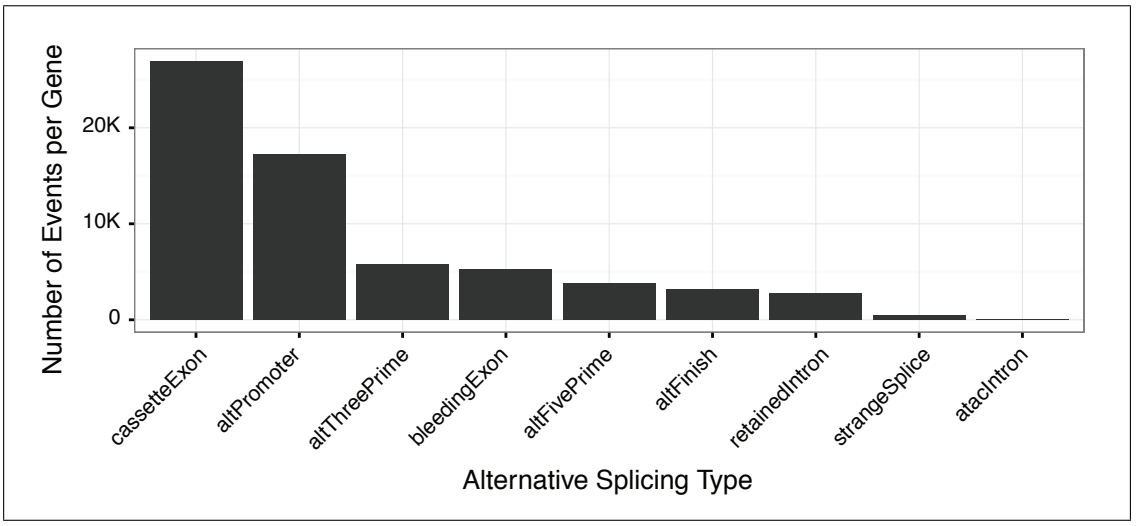

FIGURE 1.6: Number of hg19 Alternative Event types per gene

Alternative Event types per gene. RefSeq on 2014-03-24

AS in many more ways than previously appreciated. Most genes could be considered as undergoing 'complex' AS, with numerous forms of AS (See figure 1.6). Despite the prevalence of complex alternative spliced genes, just a few genes are used as examples to illustrate numerical possibilities and biological significance. For example the human immune system relies heavily on AS to be plastic toward antigen recognition and response [Lynch, 2004]. Modulation of extracellular signaling proteins such as *CD44* and cellular adhesion protein *CD45* have been well-studied [Ponta et al., 2003, Zikherman and Weiss, 2008]. Alternative splicing in humans, however, does not seem to produce the number of unique possible combinations as AS of genes in simpler organisms, such as fruit flies, perhaps due to specialization of genes, or different genes that work in combination or complexes, as oppose to utilizing unique gene isoforms [Park

and Graveley, 2007]. For example, the fruit fly gene muscle myosin heavy chain (*Mhc*) can produce up to 480 different isoforms through AS of 17 different cassette exons [Bernstein et al., 1983].

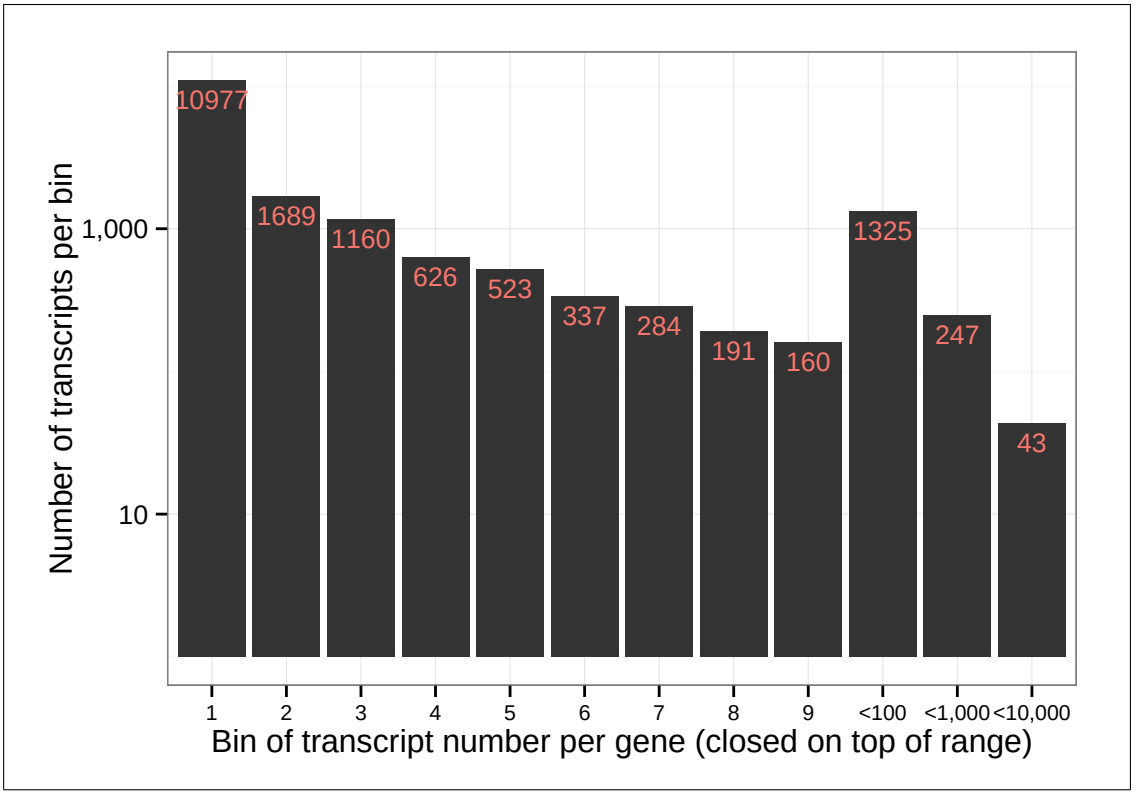

FIGURE 1.7: Number of transcripts per *Drosophila melanogaster* gene

Data from [Brown et al., 2014], Supplemental Table 3. Number of transcript per bin, with bin sizes 'closed' on the upper part of range.

1.11 Drosophila melanogaster *Dscam1*

Unquestionably, the gene most frequently used to demonstrate the combinatorial power of AS is fly *Dscam1*. The 'architecture' of *Dscam1* is rather unique among other organisms, but as we saw in Section 1.10, contain some genes

Gene Name	# Introns	# Transcripts	# Proteins
Mhc	60	2040	511
slo	49	2070	279
ps	30	2099	27
rg	45	2178	23
shot	60	2478	886
scrib	53	2555	259
heph	75	2876	52
CG42748	26	2876	51
rdgA	35	3003	89
Mbs	39	3080	119
CaMKI	41	3992	7
par-1	48	4410	142
GluClalpha	27	4945	188
Sap47	24	5011	49
Patronin	50	5615	590
CG17838	37	8333	147
unc-13	52	8391	279
A2bp1	29	9055	58
Imp	33	9131	12
pan	38	9432	72
Sh	40	15995	66
gish	48	18972	142

TABLE 1.1: Fly genes with >2,000 assembled transcripts according to [Brown et al., 2014].

that generate tremendous isoform diversity from a single genetic locus [Brown et al., 2014]. The basic structure of *Dscam1* is shown in Figure 1.8.

Human *Dscam*, for which *Dscam1* was named, was identified while looking for genes on chromosome 21, specifically band 21q22, where extra copies expressed in Down syndrome patients, a trisomy 21 disorder, may be causative for disease [Yamakawa et al., 1998]. *Dscam* (Down Syndrome Cellular Adhesion Molecule) was named according to this association, and its membership in the immunoglobulin super family of proteins with extracellular adhesion functions. Human *Dscam* does undergo some alternatively splicing and broadly expressed in the developing nervous system. However, it does not contain the

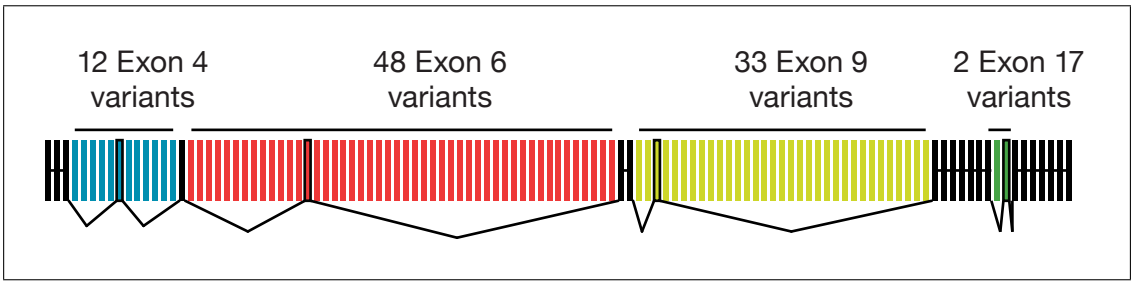

FIGURE 1.8: The architecture of the *Drosophila melanogaster* gene *Dscam1*

Dscam1 has three 'clusters' or 'banks' of alternative cassette exons that are splicing out in a mutually-exclusive manner. The first bank, 'Exon 4', contains 12 different variants, of which one one is ever included into the mRNA. Similarly, banks 6 9 each contain 48 and 33 different variants, respectively. These three banks code for extracellular IgG domains, while the final region of AS, exon 17, encodes two different trans-membrane domains, again of which only one is included in the final mRNA.

architecture of cassette exon banks as *Dscam1*.

Complex AS of*Dscam1* was first noticed by the Zipursky lab in 2000 [Schmucker et al., 2000]. While looking for proteins associated with *dock* and *pak*, two proteins important for neuronal growth cone guidance, they biochemically co-purified DSCAM1. Sequencing of *Dscam1* clones revealed that virtually all contained different combinations of exons 4,6, and 9. In fact, these three exons are chosen from three clusters of mutually-exclusive cassette exons, containing 12, 48, and 33 different options each(see Figure 1.8). The initial report kicked off an exciting period of research into *Dscam1* structure and function. The functional significance of *Dscam1* AS was a major goal of multiple labs.

Before the highlights of *Dscam1* research are reviewed, it is illustrative to discuss some basic *Drosophila melanogaster* anatomy. There are 4 main regions

where *Dscam1* expression has been highly-studied.

- Hemocyte cells of the immune system
- Larva Class IV da Neurons
- Pupal Mushroom-body neurons in the developing brain
- Tetrad synapses of the eye

Dscam1 involvement/expression in three of these four biologically important roles is shown in Figure 1.9. First, *Dscam1* expression in hemocyte cells of the immune system is important for recognition of foreign antigens [Watson et al., 2005]. During larval development, *Dscam1* is expressed in the da neurons of the larval body wall, these neurons create a uniform sensory feed, allowing the larva to respond to mechanical stimulus. In the developing brain, *Dscam1* is expressed in both axonal projections of neurons extend from their Kenyon cell bodies and bifurcate into two different mushroom body lobs [Zhan et al., 2004].

1.12 RNA Sequence investigation by ligation

In the late 1960's and early 1970's, the Lehman and Richardson labs characterized two workhorse-enzymes of modern molecular biology. Robert Lehman and colleagues, working at Standford Medical School, first described the activity of 'polynucleotide-joining enzyme' from *Escherichia coli* (now known as *E.*

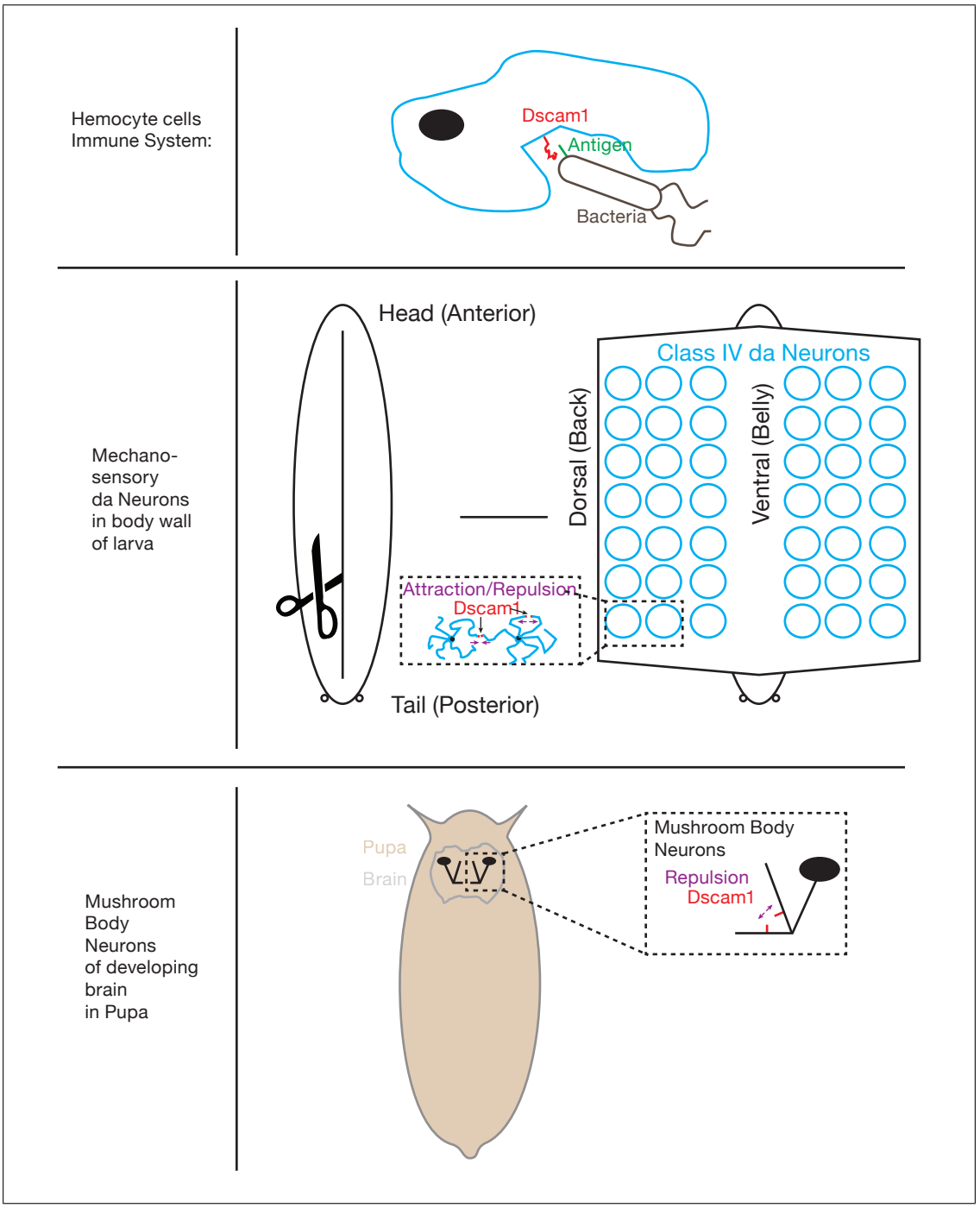

FIGURE 1.9: Important *Dscam1* expression during *Drosophila melanogaster life* cycle

Dscam1 has been high-studied in four different regions/cell types. 1) Hemocytes of the immune system, where DSCAM1 is involved in antigen recognition; 2) In Class IV da neurons, which sense mechanical stimulation of the larval body wall; 3) In mushroom body neurons of the pupal developing brain; and 4) (not shown) in Tetrad neurons of the eyes.

Coli DNA Ligase) [Olivera and Lehman, 1967]. Work on this enzyme paralleled that from the Richardson lab at Harvard Medical School, where they focused on 'polynucleotide ligase' from *Escherichia coli* infected with T4 bacteriophage (now known as T4 DNA ligase) [Weiss and Richardson, 1967]. It became clear that while these two enzyme's shared a common mechanism—later elucidated by [Modrich et al., 1973]—they had important differences. First, T4 DNA ligase required ATP as a cofactor, which *E. Coli* DNA Ligase did not (though it was later discovered that DNA ligase required NAD as a cofactor). Second, only T4 DNA ligase could catalyze ligation of blunt-ended DNA [Tabor, 1987].

The general mechanism of ligation, shown in Figure 1.10, involves three steps: Step 1 (A) involves the ϵ-amino group from the active site lysine performs a nucleophilic attack on the α-phosphate of ATP in solution. B) The ligase is now charged with AMP and inorganic phosphate (PPi) is freed into solution. C) Step 2: Nucleophilic attack by the 5 DNA phosphate on the $3'$ side of the nick to the AMP:ligase phosphate. D) 'Adenylated' DNA is now competent for DNA ligation. E) Step 3: the $3'$ OH on the $5'$ side of the nick performs a nucleophilic attack on the $5'$ PO4 across the DNA nick, liberating AMP into solution. F) Sealed nick resulting in: Ligase; AMP; and intact dsDNA.

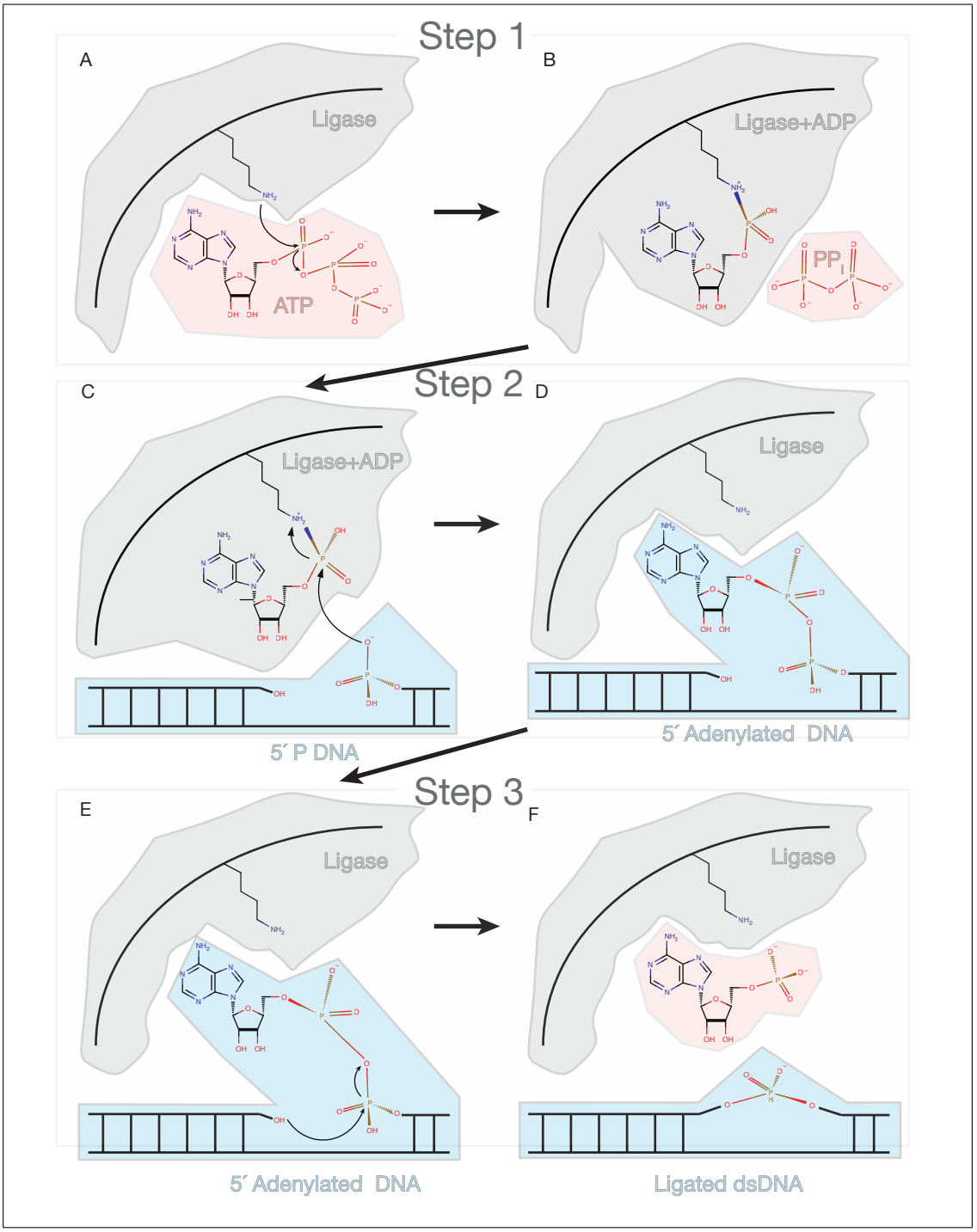

FIGURE 1.10: Mechanism of ATP-dependent ligation

Adapted from [Nandakumar et al., 2006] and specifically for that of T4 RNA ligase 2.

In addition to elucidated the general mechanism of ligation, it was also discovered that T4 DNA ligase lacks a preference for terminal polynucleotide structures. The Khorana and Richardson labs both reported the activity of this enzyme on combinations of RNA and DNA duplexes [Fareed et al., 1971, Kleppe et al., 1970]. Both of these papers describe an activity on T4 DNA ligase, RNA-templated DNA to DNA ligation, that is of particular relevance to this thesis work. Unlike T4 DNA ligase, *E. Coli* DNA Ligase, will not join DNA strands on an RNA template [Bullard and Bowater, 2006]. Soon after demonstrating these activities in vitro, the Khorana lab reported detection of organism-generated DNA [Besmer et al., 1972], setting up an orthogonal field (respective to PCR) of nucleic acid sequence characterization [Conze et al., 2009].

An enzyme that can catalyze an RNA-templated DNA:DNA ligation is a very useful molecular biology tool for two main reasons. First, using RNA as a ligation guide means no modification is made to the template molecule. This contrasts cDNA analysis, where the RNA has been enzymatically converted by reverse transcription, potentially losing valuable RNA-coded information, such as modified bases. Second, synthesis of the DNA probes used in ligation is inherently easier and cheaper compared to synthesis of RNA probes. In addition to being cheaper, synthesis of DNA probes has become high-throughput since the adoption of microarrays as a standard gene expression measurement tool [Schena et al., 1995]. A pair of papers from the Landegren lab first reported

the utility of RNA-templated DNA:DNA ligation for analysis of RNA transcripts [Nilsson et al., 2000, 2001]. The Fu lab applied this approach in a multiplex experimental design in collaboration with Illumina [Li et al., 2012, Yeakley et al., 2002], while Mats Nilsson and Ulf Landegren developed a single molecule application [Conze et al., 2010]. It is important to note that *all* of these studies used T4 DNA ligase. Clearly, there is interest and utility in analyzing RNA in both high-throughput and multiplex experimental designs, using cheap DNA probes, and without cDNA conversion.

For more than 40 years after its first description, T4 DNA ligase was the only choice for RNA-templated DNA:DNA ligation. However, a recent publication from New England Biolabs (NEB) describes this activity by another well-studied ligase, Chlorella Virus PBCV-1 DNA ligase (herein Chlorella DNA ligase) [Lohman et al., 2013]. Chlorella DNA ligase is a long-studied enzyme and had been reported to *not* display RNA-templated DNA:DNA ligation activity [Ho et al., 1997, Sriskanda and Shuman, 1998]. However, at high enough concentrations and under special buffer conditions (specifically a critical concentration of ATP), Lohman et al have shown that Chlorella DNA ligase will join two DNA strands hybridized to an RNA template [Lohman et al., 2013]. They further demonstrated that it performs no worse in this activity than traditional T4 DNA ligase [Nilsson et al., 2001, Yeakley et al., 2002].

Building on the list of available enzymes that join hybrid polymer substrates Chapter **??** presents data supporting RNA-templated DNA:DNA ligation activity for another enzyme, T4 RNA Ligase 2.

1.13 T4 RNA Ligase 2 (Rnl2)

Proteins of the T4 and T7 bacteriophages have been a boon for molecular biology. Without enzymes like polynucleotide kinase [Richardson, 1965], T7 RNA polymerase [Summers and Siegel, 1970], and T4 DNA ligase [Weiss and Richardson, 1967], many essential manipulations of nucleic acids would have been impossible for decades. Obviously, these enzymes also have essential phage functions. T7 RNA polymerase is responsible for late stage replication of T7 phage transcripts, while T4 PNK works in concert with T4 DNA and RNA ligases to repair cleaved nucleic acids resulting from bacterial pathogens defense systems [Wang et al., 2002]. Specifically, T4 RNA ligase 1 (herein "Rnl1", also known as *gene 63* maintains phage replication by repairing tRNAs cleaved by an anticodon nuclease produces from the *prr* locus [Amitsur et al., 1987].

Given the utility and importance of these enzymes, novel enzyme discovery is a fruitful area of research. The Shuman lab has a distinguished record of discovering and characterizing numerous such enzymes, including any involved in nucleic acid synthesis, modification, and repair. Through a blast search looking

for novel ligases with sequences related to *Trypanosoma brucei* RNA-editing ligases TbMP52 and TbMP48 [Ho and Shuman, 2002], they identified motifs in correct arrangement, spacing, and number indicative of an RNA ligase. The gene, identified as *gp24.1*, has quickly become an essential tool in the era of modern genomics.

Initial biochemical purification and characterization of *gp24.1* [Ho and Shuman, 2002] revealed that it indeed codes for an RNA ligase, which was renamed T4 RNA ligase 2 (herein "Rnl2"). Rnl2 is a 374 amino acid monomeric protein composed of 2 distinct domains initially purified as a 42-kDA His-tagged recombinant protein. The N-terminal domain (1–243) is responsible for steps (1) and (3) of the general ligation mechanisms (see Figure 1.10), while the C-terminal domain (244–329) is responsible for adenylation of the 5′ PO4 on the 5′ residue at the 3′ side of the nick, as shown in step (2). Additionally, Rnl2 is routinely purified as a pre-adenylated and immediately poised for its first ligation. In contrast to the N-terminal domain, which is composed of motifs typical to main ligases, the C-terminal domain is significantly different from all other DNA ligases and has no structural homologue. While the biological function of Rnl1 is known, the biological function of Rnl2 remains a mystery, more than 12 years after its discovery [Chauleau and Shuman, 2013]. However, there is some speculation that the flurry of research into bacterial CRISPR phage defense may reveal a role for Rnl2 [Barrangou et al., 2007, Chauleau and Shuman, 2013].

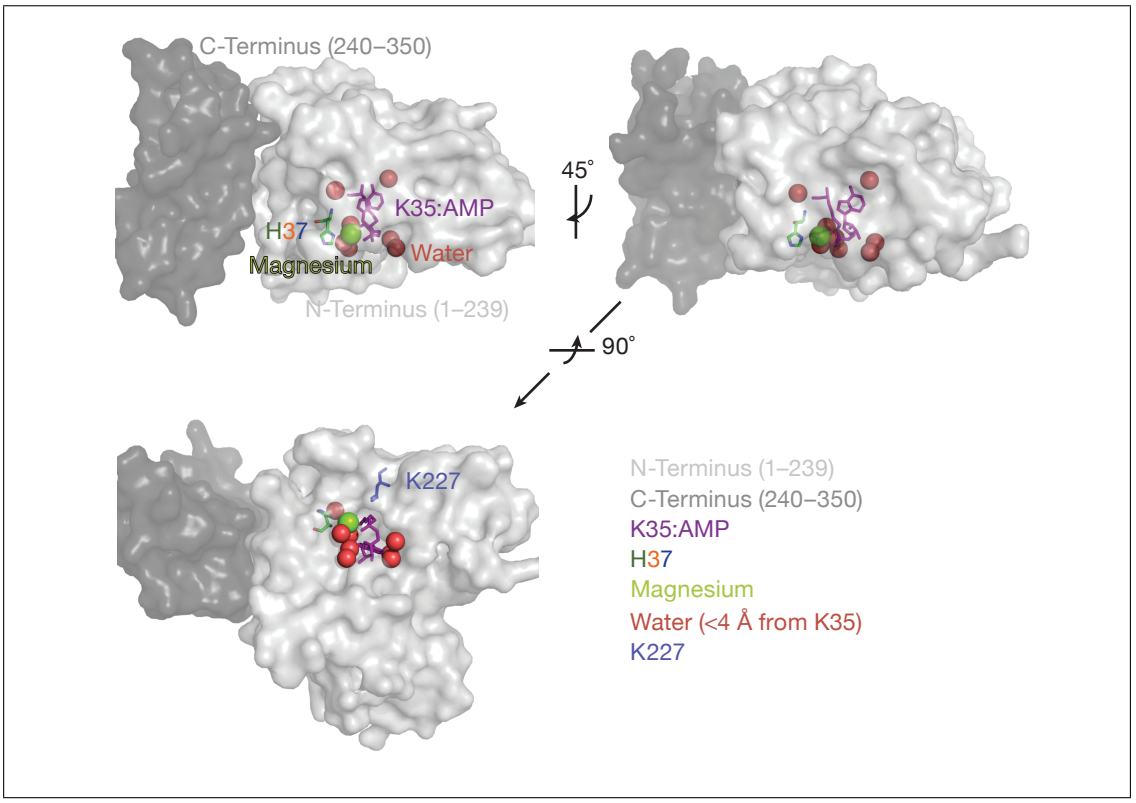

FIGURE 1.11: Structure and active site of pre-adenylated of Rnl2

Rnl2 as crystalized and described by [Nandakumar et al., 2006]. Structures from PDB:2HVQ were generated with PyMol. Top left) Rnl2 is composed of a C-terminal and N-terminal domain. Top Right) The active site of Rnl2 is highlighted. Bottom left) Active site of Rnl2 as shown from 'bottom.' This face interacts with substrate. Residue numbering refers to that of the crystal structure.

Mutational analysis of Rnl2, and later a crystal structure of the enzyme, have identified key functional residues [Ho et al., 2004, Nandakumar et al., 2004, 2006, Yin et al., 2003]. The lysine residue at position 35 (K35) receives the AMP in Step 1. The K227 residue in the C-terminal domain is essential for both forward and reverse adenylation of the 5′ PO_4 at the nick [Viollet et al., 2011]. Mutation of H37 results in an 102 reduced ligation rate, and therefore

indicates the essential nature of this residue. Finally, T39 has been shown to interact with the 2′ OH on the 3′ side of the nick, preferring a C3′ endo sugar pucker confirmation (see Figure 1.12). Rnl2 has a minimal footprint of 13nt, centered on the nick, and only requires magnesium for transfer of AMP to the 5′ phosphate. Work done in the Shuman lab [**?**] observed that 2′ deoxyribose residues on the 5′ side of the nick (i.e. DNA) adopt an RNA-like sugar pucker, leading to the correct orientation of the 3′ OH relative to the AMP leaving group and resulting in ligation. This conformation is of particular importance to this results presented in Chapter **??**.

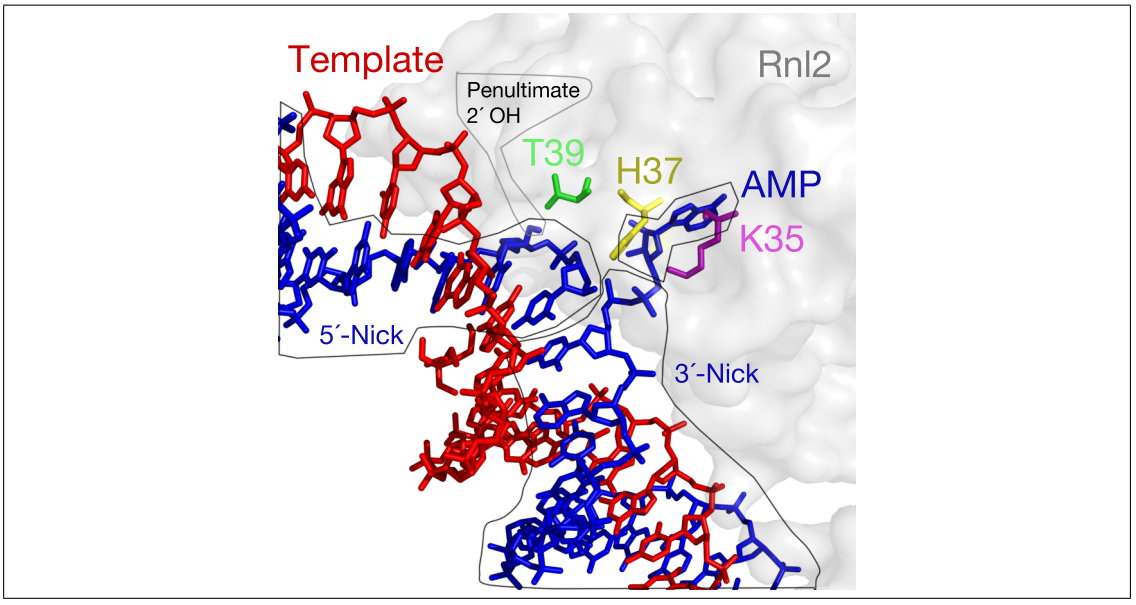

FIGURE 1.12: Structure and active site of pre-adenylated of Rnl2

Rnl2 complexed with nicked dsDNA as crystalized and described by [Nandakumar et al., 2006]. Structures from PDB:2HVR and generated with PyMol

While Rnl2 is extremely efficient at high concentration, displaying little or no

reversible chemistry, a modified version of the enzyme containing only the N-terminal domain and a K227A point mutation ("Truncated mutant") has no adenyltransferase activity. In this case, adenyltransferase refers to the ligase transferring AMP from an adenylated substrate to itself; reverse chemistry of step 2 in Figure 1.10). This mutant has been used in specialized cloning applications [Ghildiyal et al., 2008, Hafner et al., 2008, Viollet et al., 2011] that take advantage of this activity. In these reactions, the use of pre-adenylated 3′ DNA adaptors allows for selective ligation among already phosphorylated species by limiting the enzyme-catalyzed transfer of AMP from the adaptor to other phosphorylated species. Use of this truncated mutant to create a hybrid RNA/DNA molecule has greatly improved many high-throughput sequencing work flows.

Ligation of hybrid substrates (eg. DNA-templated RNA:DNA vs DNA-templated DNA:DNA) have revealed general substrate preferences. DNA ligases appear to prefer the residue bearing the 5′ phosphate on the 3′ side of the nick to be 2′ deoxyribose, and have a relaxed requirement for the sugar on the 5′ side of the nick. RNA ligases have the reverse preference, demonstrating higher activities when the 5′ strand, 3′ OH residue also bears a 2′OH. Rnl2 has an additional preference for an RNA residue at the penultimate 3′ side of a nick [Ho and Shuman, 2002, Ho et al., 2004, Nandakumar et al., 2004, **?**]. The two base requirement for RNA at the 5′ side of the double stranded nick biases Rnl2 to join RNA:[RNA/DNA] strands. Independent labs have measured this preference

and have reported that the RNA-templated DNA:DNA joining activity of Rnl2 is below assay limits of detection [Bullard and Bowater, 2006]. However, results discussed in this work clearly show that with enough enzyme, and sensitive downstream measurements, Rnl2 will catalyze RNA-templated DNA:DNA ligation (see Chapter **??**. Previous reports of Rnl2 lacking this activity are likely due to a single turnover mechanism in this reaction, owning to the poor dissociation rate of nucleic acid-interacting enzymes.

Appendix A

Appendix - Misc Information

A.1 Equations

A.1.1 Determining [RNA] from ^{32}P-α-UTP used during vitro transcription

$$\mu\text{M} = \left(\frac{\text{pmol}}{\mu\text{L}}\right) = \left(\frac{\text{cpm after purification} \times \text{dilution factor}}{\text{cpm before purification} \times \text{dilution factor}}\right) \times \left(\frac{\text{mol UTP in original reaction}}{\text{Reaction Volume}}\right) \times \left(\frac{1}{\text{Number UTPs in transcript}}\right) \times 10^{-12}$$

A.1.2 Determining [RNA] based on A_{260}

$$\text{[RNA in M]} = \left(\frac{\text{A}_{260} \times \text{Dilution Factor}}{10,313 < \text{note 1} > \times \text{nucleotides in message}} \right)$$

note 1: This value represents an average RNA extinction (ϵ) coefficient value

A.1.3 Normalize oxidized small RNA libraries size to time-matched unoxidized library

NB: this equation assumes calibration against a specific time-point , in this case data obtained from '6wk' tests.

$$\text{unox } \tau \text{ norm}_1 = \left(\frac{\left(\frac{\sum \text{miRNA reads } \tau}{\sum \text{miRNA reads 6wk}} \right) \times \text{ depth 6wk}}{1,000,000} \right)$$

$$\text{ox } \tau \text{ norm}_1 = \text{unox } \tau \text{ norm}_1 \times \left(\frac{\sum \text{oxidized shared } \geq 23 \text{ nt reads}}{\sum \text{unoxidized shared } \geq 23 \text{ nt reads}} \right)$$

Bibliography

M Amitsur, R Levitz, and G Kaufmann. Bacteriophage T4 anticodon nuclease, polynucleotide kinase and RNA ligase reprocess the host lysine tRNA. *The EMBO journal*, 6(8):2499–503, August 1987. ISSN 0261-4189. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=553660&tool=pmcentrez&rendertype=abstract.

Yoseph Barash, John A. Calarco, Weijun Gao, Qun Pan, Xinchen Wang, Ofer Shai, Benjamin J. Blencowe, and Brendan J. Frey. Deciphering the splicing code. *Nature*, 465(7294):53–59, May 2010. ISSN 0028-0836. doi: 10.1038/nature09000. URL http://www.nature.com/doifinder/10.1038/nature09000.

Rodolphe Barrangou, Christophe Fremaux, Hélène Deveau, Melissa Richards, Patrick Boyaval, Sylvain Moineau, Dennis a Romero, and Philippe Horvath. CRISPR provides acquired resistance against viruses in prokaryotes. *Science (New York, N.Y.)*, 315(5819):1709–12, March 2007. ISSN 1095-9203. doi: 10.1126/science.1138140. URL http://www.ncbi.nlm.nih.gov/pubmed/17379808.

Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly a Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva. NCBI GEO: archive for functional genomics data sets–update. *Nucleic acids research*, 41(Database issue):D991–5, January 2013. ISSN 1362-4962. doi: 10.1093/nar/gks1193. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531084&tool=pmcentrez&rendertype=abstract.

GI I Bell, RL L Pictet, WJ J Rutter, Barbara Cordell, Edmund Tischer, and Howard M. Goodman. Sequence of the human insulin gene. *Nature*, 284(5751):26–32, March 1980. ISSN 0028-0836. URL http://www2.biglobe.ne.jp/~ashida/bioinfo/data/insdata.pdfhttp://www.ncbi.nlm.nih.gov/pubmed/6243748.

Dennis a Benson, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. GenBank. *Nucleic acids research*, 39(Database issue):D32–7, January 2011. ISSN 1362-4962. doi: 10.1093/nar/gkq1079. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3013681&tool=pmcentrez&rendertype=abstract.

David R Bentley, Shankar Balasubramanian, Harold P Swerdlow, Geoffrey P Smith, John Milton, Clive G Brown, Kevin P Hall, Dirk J Evers, Colin L Barnes, Helen R Bignell, Jonathan M Boutell, Jason Bryant, Richard J Carter, R Keira Cheetham, Anthony J Cox, Darren J Ellis, Michael R Flatbush, Niall a Gormley, Sean J Humphray, Leslie J Irving, Mirian S Karbelashvili, Scott M Kirk, Heng Li, Xiaohai Liu, Klaus S Maisinger, Lisa J Murray, Bojan Obradovic, Tobias Ost, Michael L Parkinson, Mark R Pratt, Isabelle M J Rasolonjatovo, Mark T Reed, Roberto Rigatti, Chiara Rodighiero, Mark T Ross, Andrea Sabot, Subramanian V Sankar, Aylwyn Scally, Gary P Schroth, Mark E Smith, Vincent P Smith, Anastassia Spiridou, Peta E Torrance, Svilen S Tzonev, Eric H Vermaas, Klaudia Walter, Xiaolin Wu, Lu Zhang, Mohammed D Alam, Carole Anastasi, Ify C Aniebo, David M D Bailey, Iain R Bancarz, Saibal Banerjee, Selena G Barbour, Primo a Baybayan, Vincent a Benoit, Kevin F Benson, Claire Bevis, Phillip J Black, Asha Boodhun, Joe S Brennan, John a Bridgham, Rob C Brown, Andrew a Brown, Dale H Buermann, Abass a Bundu, James C Burrows, Nigel P Carter, Nestor Castillo, Maria Chiara E Catenazzi, Simon Chang, R Neil Cooley, Natasha R Crake, Olubunmi O Dada, Konstantinos D Diakoumakos, Belen Dominguez-Fernandez, David J Earnshaw, Ugonna C Egbujor, David W Elmore, Sergey S Etchin, Mark R Ewan, Milan Fedurco, Louise J Fraser, Karin V Fuentes Fajardo, W Scott Furey, David George, Kimberley J Gietzen, Colin P Goddard, George S Golda, Philip a Granieri, David E Green, David L Gustafson, Nancy F Hansen, Kevin Harnish, Christian D Haudenschild, Narinder I Heyer, Matthew M Hims, Johnny T Ho, Adrian M Horgan, Katya Hoschler, Steve Hurwitz, Denis V Ivanov, Maria Q Johnson, Terena James, T a Huw Jones, Gyoung-Dong Kang, Tzvetana H Kerelska, Alan D Kersey, Irina Khrebtukova, Alex P Kindwall, Zoya Kingsbury, Paula I Kokko-Gonzales, Anil Kumar, Marc a Laurent, Cynthia T Lawley, Sarah E Lee, Xavier Lee, Arnold K Liao, Jennifer a Loch, Mitch Lok, Shujun Luo, Radhika M Mammen, John W Martin, Patrick G McCauley, Paul McNitt, Parul Mehta, Keith W Moon, Joe W Mullens, Taksina Newington, Zemin Ning, Bee Ling Ng, Sonia M Novo, Michael J O'Neill, Mark a Osborne, Andrew Osnowski, Omead Ostadan, Lambros L Paraschos, Lea Pickering, Andrew C Pike, Alger C Pike, D Chris Pinkard, Daniel P Pliskin, Joe Podhasky, Victor J Quijano, Come Raczy, Vicki H Rae, Stephen R Rawlings, Ana Chiva

Rodriguez, Phyllida M Roe, John Rogers, Maria C Rogert Bacigalupo, Nikolai Romanov, Anthony Romieu, Rithy K Roth, Natalie J Rourke, Silke T Ruediger, Eli Rusman, Raquel M Sanches-Kuiper, Martin R Schenker, Josefina M Seoane, Richard J Shaw, Mitch K Shiver, Steven W Short, Ning L Sizto, Johannes P Sluis, Melanie a Smith, Jean Ernest Sohna Sohna, Eric J Spence, Kim Stevens, Neil Sutton, Lukasz Szajkowski, Carolyn L Tregidgo, Gerardo Turcatti, Stephanie Vandevondele, Yuli Verhovsky, Selene M Virk, Suzanne Wakelin, Gregory C Walcott, Jingwen Wang, Graham J Worsley, Juying Yan, Ling Yau, Mike Zuerlein, Jane Rogers, James C Mullikin, Matthew E Hurles, Nick J McCooke, John S West, Frank L Oaks, Peter L Lundberg, David Klenerman, Richard Durbin, and Anthony J Smith. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–9, November 2008. ISSN 1476-4687. doi: 10.1038/nature07517. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2581791&tool=pmcentrez&rendertype=abstract.

Sanford I. Bernstein, Kaname Mogami, J. James Donady, and Charles P. Emerson. Drosophila muscle myosin heavy chain encoded by a single gene in a cluster of muscle mutations. *Nature*, 302(5907):393–397, March 1983. ISSN 0028-0836. doi: 10.1038/302393a0. URL http://www.nature.com/nature/journal/v302/n5907/pdf/302393a0.pdf.

P. Besmer, R. C. Miller Jr., M. H. Caruthers, A. Kumar, K. Minamoto, J.H. van de Sande, N. Sidarova, and H.G. Khorana. Studies on polynucleotides. CXVII. Hybridization of polydeoxynucleotides with tyrosine transfer RNA sequences to the r-strand of phi80psu + 3 DNA. *Journal of Molecular Biology*, 72:503–522, 1972.

Benjamin J Blencowe, Sidrah Ahmad, and Leo J Lee. Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes. *Genes & development*, 23(12):1379–86, June 2009. ISSN 1549-5477. doi: 10.1101/gad.1788009. URL http://www.ncbi.nlm.nih.gov/pubmed/19528315.

S Brenner, M Johnson, J Bridgham, G Golda, D H Lloyd, D Johnson, S Luo, S McCurdy, M Foy, M Ewan, R Roth, D George, S Eletr, G Albrecht, E Vermaas, S R Williams, K Moon, T Burcham, M Pallas, R B DuBridge, J Kirchner, K Fearon, J Mao, and K Corcoran. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature biotechnology*, 18(6):630–4, June 2000. ISSN 1087-0156. doi: 10.1038/76469. URL http://www.ncbi.nlm.nih.gov/pubmed/10835600.

James B. Brown, Nathan Boley, Robert Eisman, Gemma E. May, Marcus H. Stoiber, Michael O. Duff, Ben W. Booth, Jiayu Wen, Soo Park, Ana Maria Suzuki, Kenneth H. Wan, Charles Yu, Dayu Zhang, Joseph W. Carlson, Lucy Cherbas, Brian D. Eads, David Miller, Keithanne Mockaitis, Johnny Roberts,

Carrie a. Davis, Erwin Frise, Ann S. Hammonds, Sara Olson, Sol Shenker, David Sturgill, Anastasia a. Samsonova, Richard Weiszmann, Garret Robinson, Juan Hernandez, Justen Andrews, Peter J. Bickel, Piero Carninci, Peter Cherbas, Thomas R. Gingeras, Roger a. Hoskins, Thomas C. Kaufman, Eric C. Lai, Brian Oliver, Norbert Perrimon, Brenton R. Graveley, and Susan E. Celniker. Diversity and dynamics of the Drosophila transcriptome. *Nature*, pages 1–7, March 2014. ISSN 0028-0836. doi: 10.1038/nature12962. URL http://www.nature.com/doifinder/10.1038/nature12962.

DesmondR. Bullard and RichardP. Bowater. Direct comparison of nick-joining activity of the nucleic acid ligases from bacteriophage T4. *Biochemical Journal*, 398(Pt 1):135–144, August 2006. ISSN 0264-6021. doi: 10.1042/BJ20060313. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1525015/pdf/bj3980135.pdf.

John A Calarco, Arneet L Saltzman, Joanna Y Ip, and Benjamin J Blencowe. Technologies for the global discovery and analysis of alternative splicing. In *Advances in Experimental Medicine and Biology*, volume 623, pages 64–84. 2007a. URL http://www.ncbi.nlm.nih.gov/pubmed/18380341.

John A Calarco, Yi Xing, Mario Cáceres, Joseph P Calarco, Xinshu Xiao, Qun Pan, Christopher Lee, Todd M Preuss, and Benjamin J Blencowe. Global analysis of alternative splicing differences between humans and chimpanzees. *Genes & Development*, 21(22):2963–75, November 2007b. ISSN 0890-9369. doi: 10.1101/gad.1606907. URL http://www.ncbi.nlm.nih.gov/pubmed/17978102.

Mathieu Chauleau and Stewart Shuman. Kinetic mechanism of nick sealing by T4 RNA ligase 2 and effects of 3'-OH base mispairs and damaged base lesions. *RNA (New York, N.Y.)*, 19(12):1840–1847, October 2013. ISSN 1469-9001. doi: 10.1261/rna.041731.113. URL http://www.ncbi.nlm.nih.gov/pubmed/24158792.

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 50(2):162–8, February 2004. ISSN 0039-9450. URL http://www.ncbi.nlm.nih.gov/pubmed/15704464http://www.nature.com/nature/journal/v431/n7011/abs/nature03001.html.

Tim Conze, Alysha Shetye, Yuki Tanaka, Jijuan Gu, Chatarina Larsson, Jenny Göransson, Gholamreza Tavoosidana, Ola Söderberg, Mats Nilsson, and Ulf Landegren. Analysis of genes, transcripts, and proteins via DNA ligation. *Annual review of analytical chemistry (Palo Alto, Calif.)*, 2:215–39, January 2009. ISSN 1936-1335. doi: 10.1146/annurev-anchem-060908-155239. URL http://www.ncbi.nlm.nih.gov/pubmed/20636060.

Tim Conze, Jenny Goransson, Hamid Reza Razzaghian, Olle Ericsson, Daniel Oberg, Goran Akusjarvi, Ulf Landegren, and Mats Nilsson. Single molecule analysis of combinatorial splicing. *Nucl. Acids Res.*, page gkq581, June 2010. doi: 10.1093/nar/gkq581. URL http://nar.oxfordjournals.org/cgi/content/abstract/gkq581v1http://nar.oxfordjournals.org/content/38/16/e163.full.pdf.

Leighton J Core, Joshua J Waterfall, and John T Lis. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, 322:1845–1848, 2008.

Sarah Djebali, Philipp Kapranov, Sylvain Foissac, Julien Lagarde, Alexandre Reymond, Catherine Ucla, Carine Wyss, Jorg Drenkow, Erica Dumais, Ryan R Murray, Chenwei Lin, David Szeto, France Denoeud, Miquel Calvo, Adam Frankish, Jennifer Harrow, Periklis Makrythanasis, Marc Vidal, Kourosh Salehi-Ashtiani, Stylianos E Antonarakis, Thomas R Gingeras, and Roderic Guigo. Efficient targeted transcript discovery via array-based normalization of RACE libraries. *Nat Meth*, 5(7):629–635, July 2008. ISSN 1548-7091. doi: 10.1038/nmeth.1216. URL http://dx.doi.org/10.1038/nmeth.1216http://www.nature.com/nmeth/journal/v5/n7/pdf/nmeth.1216.pdf.

Sarah Djebali, Carrie a Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, Chenghai Xue, Georgi K Marinov, Jainab Khatun, Brian a Williams, Chris Zaleski, Joel Rozowsky, Maik Röder, Felix Kokocinski, Rehab F Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T Baer, Nadav S Bar, Philippe Batut, Kimberly Bell, Ian Bell, Sudipto Chakrabortty, Xian Chen, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jacqueline Dumais, Radha Duttagupta, Emilie Falconnet, Meagan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Harsha Gunawardena, Cedric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Oscar J Luo, Eddie Park, Kimberly Persaud, Jonathan B Preall, Paolo Ribeca, Brian Risk, Daniel Robyr, Michael Sammeth, Lorian Schaffer, Lei-Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters, Huaien Wang, John Wrobel, Yanbao Yu, Xiaoan Ruan, Yoshihide Hayashizaki, Jennifer Harrow, Mark Gerstein, Tim Hubbard, Alexandre Reymond, Stylianos E Antonarakis, Gregory Hannon, Morgan C Giddings, Yijun Ruan, Barbara Wold, Piero Carninci, Roderic Guigó, and Thomas R Gingeras. Landscape of transcription in human cells. *Nature*, 489 (7414):101–8, September 2012. ISSN 1476-4687. doi: 10.1038/nature11233. URL http://www.ncbi.nlm.nih.gov/pubmed/22955620.

Mark Emerick, Giovanni Parmigiani, and William Agnew. Multivariate Analysis and Visualization of Splicing Correlations in Single-Gene Transcriptomes. *BMC Bioinformatics*, 8(1):16, 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-16. URL http://www.biomedcentral.com/1471-2105/8/16http://www.biomedcentral.com/content/pdf/1471-2105-8-16.pdf.

Matthew Fagnani, Yoseph Barash, Joanna Y Ip, Christine Misquitta, Qun Pan, Arneet L Saltzman, Ofer Shai, Leo Lee, Aviad Rozenhek, Naveed Mohammad, Sandrine Willaime-Morawek, Tomas Babak, Wen Zhang, Timothy R Hughes, Derek van der Kooy, Brendan J Frey, and Benjamin J Blencowe. Functional coordination of alternative splicing in the mammalian central nervous system. *Genome Biology*, 8(6):R108, 2007. ISSN 14656906. doi: 10.1186/gb-2007-8-6-r108. URL http://genomebiology.com/content/8/6/R108.

George C Fareed, Elaine M Wilt, and Charles C Richardson. Enzymatic Breakage and Joining of Deoxyribonucleic Acid. *Journal of Biological Chemistry*, 246:925–932, 1971.

Juan P. Fededa, Ezequiel Petrillo, Mikhail S. Gelfand, Alexei D. Neverov, Sebastián Kadener, Guadalupe Nogués, Federico Pelisch, Francisco E. Baralle, Andrés F. Muro, and Alberto R. Kornblihtt. A Polar Mechanism Coordinates Different Regions of Alternative Splicing within a Single Gene. *Molecular Cell*, 19(3):393–404, August 2005. ISSN 1097-2765. doi: 10.1016/j.molcel.2005.06.035. URL http://www.sciencedirect.com/science?_ob=GatewayURL&_origin=CELLPRESS&_urlversion=4&_method=citationSearch&_version=1&_src=FPDF&_piikey=S1097276505014425&md5=6792ea574f8eda1316388b24d0e2d655.

Anthony a Fodor and Richard W Aldrich. Convergent evolution of alternative splices at domain boundaries of the BK channel. *Annual review of physiology*, 71:19–36, January 2009. ISSN 1545-1585. doi: 10.1146/annurev.physiol.010908.163124. URL http://www.ncbi.nlm.nih.gov/pubmed/18694345.

Megha Ghildiyal, Hervé Seitz, Michael D Horwich, Chengjian Li, Tingting Du, Soohyun Lee, Jia Xu, Ellen L. W Kittler, Maria L Zapp, Zhiping Weng, and Phillip D Zamore. Endogenous siRNAs Derived from Transposons and mRNAs in Drosophila Somatic Cells. *Science*, 320(5879):1077–1081, May 2008. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1157396. URL http://www.sciencemag.org/content/320/5879/1077http://www.sciencemag.org/content/320/5879/1077.full.pdf.

Shubhendu Ghosh and Allan Jacobson. RNA decay modulates gene expression and controls its fidelity. *WIREs RNA*, 1:351–361, 2010. ISSN 17577012. doi: 10.1002/wrna.25.

W Gilbert. Why genes in pieces? *Nature*, 271(5645):501, February 1978. ISSN 0028-0836. URL http://www.ncbi.nlm.nih.gov/pubmed/622185.

D. A. Glauser, B. E. Johnson, R. W. Aldrich, and M. B. Goodman. Intragenic alternative splicing coordination is essential for Caenorhabditis elegans slo-1 gene function. *Proceedings of the National Academy of Sciences*, 108 (51):20790–20795, November 2011. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1116712108. URL https://ssl.umassmed.edu/content/108/51/,DanaInfo=www.pnas.org+20790.short.

B R Graveley. Sorting out the complexity of SR protein functions. *RNA (New York, N.Y.)*, 6(9):1197–211, September 2000. ISSN 1355-8382. URL http://www.ncbi.nlm.nih.gov/pubmed/10999598.

Markus Hafner, Pablo Landgraf, Janos Ludwig, Amanda Rice, Tolulope Ojo, Carolina Lin, Daniel Holoch, Cindy Lim, and Thomas Tuschl. Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods (San Diego, Calif.)*, 44(1): 3–12, January 2008. ISSN 1046-2023. doi: 10.1016/j.ymeth.2007.09.009. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2847350&tool=pmcentrez&rendertype=abstract.

Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano Jr., Anna-Carina Jungkamp, Mathias Munschauer, Alexander Ulrich, Greg S. Wardle, Scott Dewell, Mihaela Zavolan, and Thomas Tuschl. Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell*, 141(1):129–141, April 2010. ISSN 0092-8674. doi: 10.1016/j.cell.2010.03.009. URL http://www.sciencedirect.com/science/article/B6WSN-4YRY58X-H/2/506aed3f8c6aebed3d3b0918ed7e8ec1http://www.sciencedirect.com/science?_ob=MImg&_imagekey=B6WSN-4YRY58X-H-1W&_cdi=7051&_user=115184&_pii=S009286741000245X&_origin=gateway&_coverDate=04%2F02%2F2010&_sk=998589998&view=c&wchp=dGLzVzz-zSkWb&md5=831980768d084e9c9b52517b7dcc371f&ie=/sdarticle.pdf.

C K Ho, J L Van Etten, and S Shuman. Characterization of an ATP-dependent DNA ligase encoded by Chlorella virus PBCV-1. *Journal of virology*, 71(3), 1997.

C Kiong Ho and Stewart Shuman. Bacteriophage T4 RNA ligase 2 (gp24.1) exemplifies a family of RNA ligases found in all phylogenetic domains. *Proceedings of the National Academy of Sciences*, 99(20):12709–12714, 2002. URL http://www.pnas.org/content/99/20/12709.long.

C Kiong Ho, Li Kai Wang, Christopher D Lima, and Stewart Shuman. Structure and mechanism of RNA ligase. *Structure (London, England: 1993)*, 12(2): 327–339, February 2004. ISSN 0969-2126. doi: 10.1016/j.str.2004.01.011. URL http://www.ncbi.nlm.nih.gov/pubmed/14962393http://pdn.sciencedirect.com/science?_ob=MiamiImageURL&_cid=272027&_user=

115184&_pii=S0969212604000231&_check=y&_origin=article&_zone=toolbar&_coverDate=29-Feb-2004&view=c&originContentFamily=serial&wchp=dGLbVlt-zSkWA&md5=018e26b1b114f6a73bef97ac8a7ccefd/1-s2.0-S0969212604000231-main.pdf.

Amy E House and Kristen W Lynch. Regulation of alternative splicing: more than just the ABCs. *The Journal of Biological Chemistry*, 283(3):1217–21, January 2008. ISSN 0021-9258. doi: 10.1074/jbc.R700031200. URL http://www.ncbi.nlm.nih.gov/pubmed/18024429.

Nicholas T Ingolia, Sina Ghaemmaghami, John R S Newman, and Jonathan S Weissman. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324:218–223, 2009.

Nicholas T Ingolia, Liana F Lareau, and Jonathan S Weissman. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*, 147(4):789–802, November 2011. ISSN 1097-4172. doi: 10.1016/j.cell.2011.10.002. URL http://www.sciencedirect.com/science/article/pii/S0092867411011925.

Yui Jin, Hitoshi Suzuki, Shingo Maegawa, Hitoshi Endo, Sumio Sugano, Katsuyuki Hashimoto, Kunio Yasuda, and Kunio Inoue. A vertebrate RNA-binding protein Fox-1 regulates tissue-specific splicing via the pentanucleotide GCAUG. *The EMBO Journal*, 22(4):905–12, February 2003. ISSN 0261-4189. doi: 10.1093/emboj/cdg089. URL http://www.ncbi.nlm.nih.gov/pubmed/12574126.

Brandon E Johnson, Dominique A Glauser, Elise S Dan-Glauser, D. Brent Halling, Richard W Aldrich, and Miriam B Goodman. Alternatively Spliced Domains Interact to Regulate BK Potassium Channel Gating. *Proceedings of the National Academy of Sciences*, 108(51):20784–20789, December 2011. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1116795108. URL http://www.pnas.org/content/108/51/20784.

Jason M Johnson, John Castle, Philip Garrett-Engele, Zhengyan Kan, Patrick M Loerch, Christopher D Armour, Ralph Santos, Eric E Schadt, Roland Stoughton, and Daniel D Shoemaker. Genome-Wide Survey of Human Alternative Pre-mRNA Splicing with Exon Junction Microarrays. *Science*, 302 (5653):2141–2144, 2003. URL http://www.sciencemag.org/cgi/content/abstract/302/5653/2141.

E. H. Kennard. Zur Quantenmechanik einfacher Bewegungstypen. *Zeitschrift fr Physik*, 44(4-5):326–352, April 1927. ISSN 1434-6001. doi: 10.1007/BF01391200. URL http://link.springer.com/10.1007/BF01391200.

K Kleppe, J H Van de Sande, and H G Khorana. Polynucleotide ligase-catalyzed joining of deoxyribo-oligonucleotides on ribopolynucleotide templates and of ribo-oligonucleotides on deoxyribopolynucleotide templates. *Proceedings of the National Academy of Sciences of the*

United States of America, 67(1):68–73, September 1970. ISSN 0027-8424. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=283168&tool=pmcentrez&rendertype=abstract.

Andrea N Ladd and Thomas A Cooper. Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biology*, 3(11):reviews0008, October 2002. ISSN 1465-6914. URL http://www.ncbi.nlm.nih.gov/pubmed/12429065.

Eric S Lander. Initial impact of the sequencing of the human genome. *Nature*, 470(7333):187–97, February 2011. ISSN 1476-4687. doi: 10.1038/nature09792. URL http://www.ncbi.nlm.nih.gov/pubmed/21307931.

Tina Lenasi, B. Matija Peterlin, and Peter Dovc. Distal regulation of alternative splicing by splicing enhancer in equine β-casein intron 1. *RNA*, 12(3):498 –507, March 2006. doi: 10.1261/rna.7261206. URL http://rnajournal.cshlp.org/content/12/3/498.abstracthttp://rnajournal.cshlp.org/content/12/3/498.full.pdf.

Hairi Li, Jinsong Qiu, and Xiang-Dong Fu. RASL-seq for massively parallel and quantitative analysis of gene expression. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]*, Chapter 4(April):Unit 4.13.1–9, April 2012. ISSN 1934-3647. doi: 10.1002/0471142727.mb0413s98. URL http://www.ncbi.nlm.nih.gov/pubmed/22470064.

Mingyao Li, Isabel X Wang, Yun Li, Alan Bruzel, Allison L Richards, Jonathan M Toung, and Vivian G Cheung. Widespread RNA and DNA sequence differences in the human transcriptome. *Science (New York, N.Y.)*, 333(6038):53–8, July 2011. ISSN 1095-9203. doi: 10.1126/science.1207018. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3204392&tool=pmcentrez&rendertype=abstract.

Donny D. Licatalosi, Aldo Mele, John J. Fak, Jernej Ule, Melis Kayikci, Sung Wook Chi, Tyson A. Clark, Anthony C. Schweitzer, John E. Blume, Xuning Wang, Jennifer C. Darnell, and Robert B. Darnell. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456(7221):464–469, November 2008. ISSN 0028-0836. doi: 10.1038/nature07488. URL http://dx.doi.org/10.1038/nature07488.

Ryan Lister, Ronan C O'Malley, Julian Tonti-Filippini, Brian D Gregory, Charles C Berry, A Harvey Millar, and Joseph R Ecker. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell*, 133(3):523–36, May 2008. ISSN 1097-4172. doi: 10.1016/j.cell.2008.03.029. URL http://www.ncbi.nlm.nih.gov/pubmed/18423832http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2723732&tool=pmcentrez&rendertype=abstract.

Gregory J S Lohman, Yinhua Zhang, Alexander M Zhelkovsky, Eric J Cantor, Thomas C Evans, and Thomas C Evans Jr. Efficient DNA

ligation in DNA – RNA hybrid helices by Chlorella virus DNA ligase. *Nucleic acids research*, 42(36):1–14, November 2013. ISSN 1362-4962. doi: 10.1093/nar/gkt1032. URL http://www.ncbi.nlm.nih.gov/pubmed/24203707http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3919565&tool=pmcentrez&rendertype=abstract.

Jennifer C Long and Javier F Caceres. The SR protein family of splicing factors: master regulators of gene expression. *The Biochemical journal*, 417(1):15–27, 2009. URL http://www.ncbi.nlm.nih.gov/pubmed/19061484.

Kristen W. Lynch. Consequences of regulated pre-mRNA splicing in the immune system. *Nat Rev Immunol*, 4(12):931–940, December 2004. ISSN 1474-1733. doi: 10.1038/nri1497. URL http://dx.doi.org/10.1038/nri1497.

John C Marioni, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–17, September 2008. ISSN 1088-9051. doi: 10.1101/gr.079558.108. URL http://www.ncbi.nlm.nih.gov/pubmed/18550803http://genome.cshlp.org/content/18/9/1509.full.pdf.

Rebeca Martinez-Contreras, Philippe Cloutier, Lulzim Shkreta, Jean-François Fisette, Timothée Revil, and Benoit Chabot. hnRNP proteins and splicing control. *Advances in experimental medicine and biology*, 623:123–147, 2007.

A M Maxam and W Gilbert. A new method for sequencing DNA. 1977. *Biotechnology (Reading, Mass.)*, 24(2):99–103, January 1992. ISSN 0740-7378. URL http://www.ncbi.nlm.nih.gov/pubmed/1422074.

Barmak Modrek and Christopher Lee. A genomic view of alternative splicing. *Nature Genetics*, 30(1):13–9, January 2002. ISSN 1061-4036. doi: 10.1038/ng0102-13. URL http://www.ncbi.nlm.nih.gov/pubmed/11753382.

Paul Modrich, Yasuhiro Anraku, and I. R. Lehman. Deoxyribonucleic Acid Ligase ISOLATION AND PHYSICAL CHARACTERIZATION OF THE HOMOGENEOUS ENZYME FROM ESCHERICHIA COLI. *Journal of Biological Chemistry*, 248(21):7495–7501, November 1973. ISSN 0021-9258, 1083-351X. URL http://www.jbc.org/content/248/21/7495http://www.jbc.org/content/248/21/7495.full.pdf.

Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Meth*, 5(7):621–628, July 2008. ISSN 1548-7091. doi: 10.1038/NMETH.1226. URL http://www.ncbi.nlm.nih.gov/pubmed/18516045http://dx.doi.org/10.1038/nmeth.1226.

Ugrappa Nagalakshmi, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science (New York, N.Y.)*, 320(5881):1344–9, June 2008. ISSN 1095-9203. doi:

10.1126/science.1158441. URL http://www.ncbi.nlm.nih.gov/pubmed/18451266http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2951732&tool=pmcentrez&rendertype=abstract.

Michihiko Nakano, Jun Komatsu, Shun-ichi Matsuura, Kazunori Takashima, Shinji Katsura, and Akira Mizuno. Single-molecule PCR using water-in-oil emulsion. *Journal of Biotechnology*, 102(2):117–124, April 2003. ISSN 0168-1656. doi: http://dx.doi.org/10.1016/S0168-1656(03)00023-3. URL http://www.sciencedirect.com/science/article/pii/S0168165603000233.

Jayakrishnan Nandakumar, C Kiong Ho, Christopher D Lima, and Stewart Shuman. RNA substrate specificity and structure-guided mutational analysis of bacteriophage T4 RNA ligase 2. *The Journal of biological chemistry*, 279(30): 31337–47, July 2004. ISSN 0021-9258. doi: 10.1074/jbc.M402394200. URL http://www.ncbi.nlm.nih.gov/pubmed/15084599.

Jayakrishnan Nandakumar, Stewart Shuman, and Christopher D Lima. RNA ligase structures reveal the basis for RNA specificity and conformational changes that drive ligation forward. *Cell*, 127(1):71–84, October 2006. ISSN 0092-8674. doi: 10.1016/j.cell.2006.08.038. URL http://www.ncbi.nlm.nih.gov/pubmed/17018278http://pdn.sciencedirect.com/science?_ob=MiamiImageURL&_cid=272196&_user=115184&_pii=S0092867406011603&_check=y&_origin=article&_zone=toolbar&_coverDate=06-Oct-2006&view=c&originContentFamily=serial&wchp=dGLbVBA-zSkWz&md5=73f1e929c58ce0a0d8812ec3873211f5/1-s2.0-S0092867406011603-main.pdf.

Timothy W. Nilsen and Brenton R. Graveley. Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463(7280):457–463, January 2010. ISSN 0028-0836. doi: 10.1038/nature08909. URL http://www.nature.com/nature/journal/v463/n7280/full/nature08909.htmlhttp://www.nature.com/nature/journal/v463/n7280/pdf/nature08909.pdf.

Mats Nilsson, G Barbany, D O Antson, K Gertow, and U Landegren. Enhanced detection and distinction of RNA by enzymatic probe ligation. *Nature biotechnology*, 18(7):791–3, July 2000. ISSN 1087-0156. doi: 10.1038/77367. URL http://www.ncbi.nlm.nih.gov/pubmed/10888852.

Mats Nilsson, Dan-Oscar Antson, Gisela Barbany, and Ulf Landegren. RNA-templated DNA ligation for transcript analysis. *Nucleic Acids Research*, 29 (2):578–581, January 2001. ISSN 0305-1048. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC29667/pdf/GKE133.pdf.

BM Olivera and IR Lehman. Linkage of polynucleotides through phosphodiester bonds by an enzyme from Escherichia coli. *Proceedings of the National Academy of . . .*, pages 1426–1433, 1967. URL http://www.ncbi.nlm.nih.gov/pmc/articles/pmc224490/.

Qun Pan, Ofer Shai, Leo J Lee, Brendan J Frey, and Benjamin J

Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40 (12):1413–5, December 2008. ISSN 1546-1718. doi: 10.1038/ng.259. URL http://www.ncbi.nlm.nih.gov/pubmed/18978789http://www.nature.com/ng/journal/v40/n12/pdf/ng.259.pdf.

Jung Woo Park and Brenton R Graveley. Complex alternative splicing. *Advances in Experimental Medicine and Biology*, 623:50–63, 2007. ISSN 0065-2598. URL http://www.ncbi.nlm.nih.gov/pubmed/18380340.

Marilyn K Parra, Jeff S Tan, Narla Mohandas, and John G Conboy. Intrasplicing coordinates alternative first exons with alternative splicing in the protein 4.1R gene. *EMBO J*, 27(1):122–131, January 2008. ISSN 0261-4189. doi: 10.1038/sj.emboj.7601957. URL http://dx.doi.org/10.1038/sj.emboj.7601957http://www.nature.com/emboj/journal/v27/n1/pdf/7601957a.pdf.

Marilyn K Parra, Thomas L Gallagher, Sharon L Amacher, Narla Mohandas, and John G Conboy. Deep Intron Elements Mediate Nested Splicing Events at Consecutive AG Dinucleotides To Regulate Alternative 3 Splice Site Choice in Vertebrate 4.1 Genes. *Molecular and Cellular Biology*, 32(11): 2044–2053, June 2012. ISSN 0270-7306, 1098-5549. doi: 10.1128/MCB.05716-11. URL http://mcb.asm.org/content/32/11/2044http://mcb.asm.org/content/32/11/2044.full.pdf.

Tao Peng, Chenghai Xue, Jianning Bi, Tingting Li, Xiaowo Wang, Xuegong Zhang, and Yanda Li. Functional importance of different patterns of correlation between adjacent cassette exons in human and mouse. *BMC Genomics*, 9(1):191, 2008. ISSN 1471-2164. doi: 10.1186/1471-2164-9-191. URL http://www.biomedcentral.com/1471-2164/9/191http://www.biomedcentral.com/content/pdf/1471-2164-9-191.pdf.

Helmut Ponta, Larry Sherman, and Peter a Herrlich. CD44: from adhesion molecules to signalling regulators. *Nature reviews. Molecular cell biology*, 4(1):33–45, January 2003. ISSN 1471-0072. doi: 10.1038/nrm1004. URL http://www.ncbi.nlm.nih.gov/pubmed/12511867.

Swapna R. Purandare, Brigitte Tenhumberg, and Jennifer a. Brisson. Comparison of the wing polyphenic response of pea aphids (Acyrthosiphon pisum) to crowding and predator cues. *Ecological Entomology*, 39(2):263–266, April 2014. ISSN 03076946. doi: 10.1111/een.12080. URL http://doi.wiley.com/10.1111/een.12080.

CC Richardson. Phosphorylation of Nucleic Acid by and Enzyme from T4 Bacteriophate-infected Escherichia Coli. *Proceedings of the National Academy of ...*, 1372(1961):158–165, 1965. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC285814/.

Kasandra J Riley and Joan A Steitz. Minireview The " Observer Effect " in

Genome-wide Surveys of Protein-RNA Interactions Minireview. pages 2011–2014, 2013.

JC Jared C Roach, Cecilie Boysen, K Wang, Leroy Hood, and I K A I Wang. Pairwise End Sequencing: A unified approach to genomic mapping and sequencing. *Genomics*, 353:345–353, 1995. URL http://www.sciencedirect.com/science/article/pii/088875439580219C.

M Ronaghi, M Uhlén, and P Nyrén. A sequencing method based on real-time pyrophosphate. *Science*, 281:363, 365, 1998.

Kourosh Salehi-Ashtiani, Xinping Yang, Adnan Derti, Weidong Tian, Tong Hao, Chenwei Lin, Kathryn Makowski, Lei Shen, Ryan R Murray, David Szeto, Nadeem Tusneem, Douglas R Smith, Michael E Cusick, David E Hill, Frederick P Roth, and Marc Vidal. Isoform discovery by targeted cloning, 'deep-well' pooling and parallel sequencing. *Nature Methods*, 5(7):597–600, July 2008. ISSN 1548-7105. doi: 10.1038/nmeth.1224. URL http://www.ncbi.nlm.nih.gov/pubmed/18552854.

F Sanger and A R Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*, 94 (3):441–8, May 1975. ISSN 0022-2836. URL http://www.ncbi.nlm.nih.gov/pubmed/1100841.

F Sanger, S Nicklen, and A R Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–7, December 1977. ISSN 0027-8424. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=431765&tool=pmcentrez&rendertype=abstract.

F Sanger, A R Coulson, T Friedmann, G M Air, B G Barrell, N L Brown, J C Fiddes, C A Hutchison, P M Slocombe, and M Smith. The nucleotide sequence of bacteriophage phiX174. *Journal of molecular biology*, 125(2):225–46, October 1978. ISSN 0022-2836. URL http://www.ncbi.nlm.nih.gov/pubmed/731693.

M Schena, D Shalon, R W Davis, and P O Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y.)*, 270(5235):467–70, October 1995. ISSN 0036-8075. URL http://www.ncbi.nlm.nih.gov/pubmed/7569999.

Dietmar Schmucker, James C Clemens, Huidy Shu, Carolyn A Worby, Jian Xiao, Marco Muda, Jack E Dixon, and S.Lawrence Zipursky. Drosophila Dscam Is an Axon Guidance Receptor Exhibiting Extraordinary Molecular Diversity. *Cell*, 101(6):671–684, June 2000. ISSN 00928674. doi: 10.1016/S0092-8674(00)80878-8. URL http://www.sciencedirect.com/science/article/pii/S0092867400808788http://pdn.sciencedirect.com/science?_ob=MiamiImageURL&_cid=272196&_user=115184&_pii=S0092867400808788&_check=y&_origin=article&

_zone=toolbar&_coverDate=09-Jun-2000&view=c&originContentFamily=serial&wchp=dGLbVlV-zSkzS&md5=76cf30668637005c4da6fc6dc8de1873&pid=1-s2.0-S0092867400808788-main.pdfhttp://linkinghub.elsevier.com/retrieve/pii/S0092867400808788.

J E Schwarzbauer, J W Tamkun, I R Lemischka, and R O Hynes. Three different fibronectin mRNAs arise by alternative splicing within the coding region. *Cell*, 35(2 Pt 1):421–31, December 1983. ISSN 0092-8674. URL http://www.ncbi.nlm.nih.gov/pubmed/6317187.

Alex K Shalek, Rahul Satija, Xian Adiconis, Rona S Gertner, Jellert T Gaublomme, Raktima Raychowdhury, Schragi Schwartz, Nir Yosef, Christine Malboeuf, Diana Lu, John T Trombetta, Dave Gennert, Andreas Gnirke, Alon Goren, Nir Hacohen, Joshua Z Levin, Hongkun Park, and Aviv Regev. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, pages 1–5, May 2013. ISSN 1476-4687. doi: 10.1038/nature12172. URL http://www.ncbi.nlm.nih.gov/pubmed/23685454.

Ehud Shapiro, Tamir Biezuner, and Sten Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature reviews. Genetics*, 14(9):618–630, July 2013. ISSN 1471-0064. doi: 10.1038/nrg3542. URL http://www.ncbi.nlm.nih.gov/pubmed/23897237.

Phillip A. Sharp. Nobel Lectures in Physiology or Medicine 1991-1995, 2014. URL http://www.nobelprize.org/nobel_organizations/nobelfoundation/publications/lectures/WSC/physio-91-95.htm.

Jay Shendure and Erez Lieberman Aiden. The expanding scope of DNA sequencing. *Nature Biotechnology*, 30(11):1084–1094, 2012. ISSN 1087-0156. doi: 10.1038/nbt.2421. URL http://www.nature.com/nbt/journal/v30/n11/full/nbt.2421.htmlhttp://www.nature.com/nbt/journal/v30/n11/pdf/nbt.2421.pdf.

Jay Shendure and Hanlee Ji. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145, 2008. URL http://www.ncbi.nlm.nih.gov/pubmed/18846087.

Jay Shendure, Robi D Mitra, Chris Varma, and George M Church. Advanced sequencing technologies: methods and goals. *Nature Reviews. Genetics*, 5 (5):335–44, May 2004. ISSN 1471-0056. doi: 10.1038/nrg1325. URL http://www.ncbi.nlm.nih.gov/pubmed/15143316.

Jay Shendure, Gregory J. Porreca, Nikos B. Reppas, Xiaoxia Lin, John P. McCutcheon, Abraham M. Rosenbaum, Michael D. Wang, Kun Zhang, Robi D. Mitra, and George M. Church. Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. *Science*, 309(5741):1728–1732, September 2005. doi: 10.1126/science.1117389. URL http://www.sciencemag.org.ezproxy.umassmed.edu/cgi/content/abstract/309/5741/1728.

AW Shingleton, GC Sisk, and DL Stern. Diapause in the pea aphid

(Acyrthosiphon pisum) is a slowing but not a cessation of development. *BMC developmental biology*, 12:1–12, 2003. URL http://www.biomedcentral.com/1471-213X/3/7.

E M Southern. DNA microarrays. History and overview. *Methods in Molecular Biology (Clifton, N.J.)*, 170:1–15, 2001. ISSN 1064-3745. doi: 10.1385/1-59259-234-1:1. URL http://www.ncbi.nlm.nih.gov/pubmed/11357674.

V Sriskanda and S Shuman. Specificity and fidelity of strand joining by Chlorella virus DNA ligase. *Nucleic acids research*, 26(15):3536–41, August 1998. ISSN 0305-1048. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=147728&tool=pmcentrez&rendertype=abstract.

R Staden. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research*, 6(7):2601–2610, 1979. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC327874/http://nar.oxfordjournals.org/content/6/7/2601.short.

Marc Sultan, Marcel H. Schulz, Hugues Richard, Alon Magen, Andreas Klingenhoff, Matthias Scherf, Martin Seifert, Tatjana Borodina, Aleksey Soldatov, Dmitri Parkhomchuk, Dominic Schmidt, Sean O'Keeffe, Stefan Haas, Martin Vingron, Hans Lehrach, and Marie-Laure Yaspo. A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome. *Science*, 321(5891):956–960, August 2008. doi: 10.1126/science.1160342. URL http://www.sciencemag.org/cgi/content/abstract/321/5891/956.

WC Summers and RB Siegel. Transcription of Late Phage RNA by T7 RNA Polymerase. *Nature*, 228:1160–1162, 1970. URL http://www.nature.com/nature/journal/v228/n5277/abs/2281160a0.html.

Stanley Tabor. DNA Ligases. *Current Protocols in Molecular Biology*, pages 3.14.1–3.14.4, 1987. URL http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:DNA+Ligases#8.

Jamal Tazi, Nadia Bakkour, and Stefan Stamm. Alternative splicing and disease. *Biochimica et biophysica acta*, 1792(1):14–26, January 2009. ISSN 0006-3002. doi: 10.1016/j.bbadis.2008.09.017. URL http://www.ncbi.nlm.nih.gov/pubmed/18992329.

Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotech*, 28(5):511–515, May 2010. ISSN 1087-0156. doi: 10.1038/nbt.1621. URL http://dx.doi.org/10.1038/nbt.1621http://www.nature.com/nbt/journal/v28/n5/pdf/nbt.1621.pdf.

Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim,

David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, 7(3):562–578, March 2012. ISSN 1754-2189. doi: 10.1038/nprot.2012.016. URL http://www.nature.com/nprot/journal/v7/n3/full/nprot.2012.016.htmlhttp://www.nature.com/nprot/journal/v7/n3/pdf/nprot.2012.016.pdf.

Jernej Ule, Kirk Jensen, Aldo Mele, and Robert B. Darnell. CLIP: A method for identifying protein-RNA interaction sites in living cells. *Methods*, 37(4):376–386, December 2005. ISSN 1046-2023. doi: 10.1016/j.ymeth.2005.07.018. URL http://www.sciencedirect.com/science/article/B6WN5-4HN9Y57-D/2/1718faa3c3c5bd5bcc6f6959e0316231.

V E Velculescu, L Zhang, B Vogelstein, and K W Kinzler. Serial analysis of gene expression. *Science (New York, N.Y.)*, 270(5235):484–7, October 1995. ISSN 0036-8075. URL http://www.ncbi.nlm.nih.gov/pubmed/7570003.

J C Venter, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, M Yandell, C a Evans, R a Holt, J D Gocayne, P Amanatides, R M Ballew, D H Huson, J R Wortman, Q Zhang, C D Kodira, X H Zheng, L Chen, M Skupski, G Subramanian, P D Thomas, J Zhang, G L Gabor Miklos, C Nelson, S Broder, a G Clark, J Nadeau, V a McKusick, N Zinder, a J Levine, R J Roberts, M Simon, C Slayman, M Hunkapiller, R Bolanos, A Delcher, I Dew, D Fasulo, M Flanigan, L Florea, A Halpern, S Hannenhalli, S Kravitz, S Levy, C Mobarry, K Reinert, K Remington, J Abu-Threideh, E Beasley, K Biddick, V Bonazzi, R Brandon, M Cargill, I Chandramouliswaran, R Charlab, K Chaturvedi, Z Deng, V Di Francesco, P Dunn, K Eilbeck, C Evangelista, a E Gabrielian, W Gan, W Ge, F Gong, Z Gu, P Guan, T J Heiman, M E Higgins, R R Ji, Z Ke, K a Ketchum, Z Lai, Y Lei, Z Li, J Li, Y Liang, X Lin, F Lu, G V Merkulov, N Milshina, H M Moore, a K Naik, V a Narayan, B Neelam, D Nusskern, D B Rusch, S Salzberg, W Shao, B Shue, J Sun, Z Wang, A Wang, X Wang, J Wang, M Wei, R Wides, C Xiao, C Yan, A Yao, J Ye, M Zhan, W Zhang, H Zhang, Q Zhao, L Zheng, F Zhong, W Zhong, S Zhu, S Zhao, D Gilbert, S Baumhueter, G Spier, C Carter, A Cravchik, T Woodage, F Ali, H An, A Awe, D Baldwin, H Baden, M Barnstead, I Barrow, K Beeson, D Busam, A Carver, A Center, M L Cheng, L Curry, S Danaher, L Davenport, R Desilets, S Dietz, K Dodson, L Doup, S Ferriera, N Garg, A Gluecksmann, B Hart, J Haynes, C Haynes, C Heiner, S Hladun, D Hostin, J Houck, T Howland, C Ibegwam, J Johnson, F Kalush, L Kline, S Koduru, A Love, F Mann, D May, S McCawley, T McIntosh, I McMullen, M Moy, L Moy, B Murphy, K Nelson, C Pfannkoch, E Pratts, V Puri, H Qureshi, M Reardon, R Rodriguez, Y H Rogers, D Romblad, B Ruhfel, R Scott, C Sitter, M Smallwood, E Stewart, R Strong, E Suh, R Thomas, N N

Tint, S Tse, C Vech, G Wang, J Wetter, S Williams, M Williams, S Windsor, E Winn-Deen, K Wolfe, J Zaveri, K Zaveri, J F Abril, R Guigó, M J Campbell, K V Sjolander, B Karlak, A Kejariwal, H Mi, B Lazareva, T Hatton, A Narechania, K Diemer, A Muruganujan, N Guo, S Sato, V Bafna, S Istrail, R Lippert, R Schwartz, B Walenz, S Yooseph, D Allen, A Basu, J Baxendale, L Blick, M Caminha, J Carnes-Stine, P Caulk, Y H Chiang, M Coyne, C Dahlke, A Mays, M Dombroski, M Donnelly, D Ely, S Esparham, C Fosler, H Gire, S Glanowski, K Glasser, A Glodek, M Gorokhov, K Graham, B Gropman, M Harris, J Heil, S Henderson, J Hoover, D Jennings, C Jordan, J Jordan, J Kasha, L Kagan, C Kraft, A Levitsky, M Lewis, X Liu, J Lopez, D Ma, W Majoros, J McDaniel, S Murphy, M Newman, T Nguyen, N Nguyen, M Nodell, S Pan, J Peck, M Peterson, W Rowe, R Sanders, J Scott, M Simpson, T Smith, A Sprague, T Stockwell, R Turner, E Venter, M Wang, M Wen, D Wu, M Wu, A Xia, A Zandieh, and X Zhu. The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507):1304–51, February 2001. ISSN 0036-8075. doi: 10.1126/science.1058040. URL http://www.ncbi.nlm.nih.gov/pubmed/11181995.

JC Venter. *A life decoded: my genome, my life*. Penguin (Non-Classics), 2007. ISBN 0143114182. URL http://www.amazon.com/Life-Decoded-My-Genome/dp/B002HREL9Khttp://books.google.com/books?hl=en&lr=&id=jx9JsHry1PgC&oi=fnd&pg=PA1&dq=A+Life+Decoded:+My+Genome:+My+Life&ots=Hqq64xIjzP&sig=VyT45kqGEI0bLqNBO8p_J86sbT8.

Sebastien Viollet, Ryan T Fuchs, Daniela B Munafo, Fanglei Zhuang, and Gregory B Robb. T4 RNA Ligase 2 truncated active site mutants: improved tools for RNA analysis. *BMC Biotechnology*, 11(1): 72, 2011. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3149579&tool=pmcentrez&rendertype=abstract.

Eric T Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F Kingsmore, Gary P Schroth, and Christopher B Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–6, November 2008. ISSN 1476-4687. doi: 10.1038/nature07509. URL http://www.nature.com/nature/journal/v456/n7221/pdf/nature07509.pdf.

Li Kai Wang, Christopher D Lima, and Stewart Shuman. Structure and mechanism of T4 polynucleotide kinase: an RNA repair enzyme. *The EMBO journal*, 21(14):3873–80, July 2002. ISSN 0261-4189. doi: 10.1093/emboj/cdf397. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=126130&tool=pmcentrez&rendertype=abstract.

Fiona L Watson, Roland Püttmann-Holgado, Franziska Thomas, David L Lamar, Michael Hughes, Masahiro Kondo, Vivienne I Rebel, and Dietmar Schmucker.

Extensive diversity of Ig-superfamily proteins in the immune system of insects. *Science (New York, N.Y.)*, 309(5742):1874–8, September 2005. ISSN 1095-9203. doi: 10.1126/science.1116887. URL http://www.ncbi.nlm.nih.gov/pubmed/16109846.

James D. Watson, Alexander Gann, and Jan Witkowski. *The Annotated and Illustrated Double Helix*. Simon & Schuster, 2012. ISBN 1476715491. URL http://www.amazon.com/The-Annotated-Illustrated-Double-Helix/dp/1476715491.

James D. JD Watson and FHC Crick. Molecular structure of nucleic acids. *Nature*, 4356(171):737–738, 1953. URL http://www.nature.com/physics/looking-back/crick/www.nature.com+171737a0.pdf.

B Weiss and CC Richardson. Enzymatic breakage and joining of deoxyribonucleic acid, I. Repair of single-strand breaks in DNA by an enzyme system from Escherichia coli infected with T4 bacteriophage. *Proceedings of the National Academy of . . .*, 1967. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC224649/?rendertype=abstract.

Eric S White and Andrés F Muro. Fibronectin splice variants: understanding their multiple roles in health and disease using engineered mouse models. *IUBMB life*, 63(7):538–46, July 2011. ISSN 1521-6551. doi: 10.1002/iub.493. URL http://www.ncbi.nlm.nih.gov/pubmed/21698758.

Quin F Wills, Kenneth J Livak, Alex J Tipping, Tariq Enver, Andrew J Goldson, Darren W Sexton, and Chris Holmes. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nature biotechnology*, 31(8):748–752, July 2013. ISSN 1546-1696. doi: 10.1038/nbt.2642. URL http://www.ncbi.nlm.nih.gov/pubmed/23873083.

K Yamakawa, Y K Huot, M a Haendelt, R Hubert, X N Chen, G E Lyons, and J R Korenberg. DSCAM: a novel member of the immunoglobulin superfamily maps in a Down syndrome region and is involved in the development of the nervous system. *Human molecular genetics*, 7(2):227–37, February 1998. ISSN 0964-6906. URL http://www.ncbi.nlm.nih.gov/pubmed/9426258.

Joanne M. Yeakley, Jian-Bing Fan, Dennis Doucet, Lin Luo, Eliza Wickham, Zhen Ye, Mark S. Chee, and Xiang-Dong Fu. Profiling alternative splicing on fiber-optic arrays. *Nature Biotechnology*, 20(4):353–358, April 2002. ISSN 1087-0156. doi: 10.1038/nbt0402-353. URL http://www.nature.com/nbt/journal/v20/n4/full/nbt0402-353.htmlhttp://www.nature.com/nbt/journal/v20/n4/pdf/nbt0402-353.pdf.

Shenmin Yin, C Kiong Ho, and Stewart Shuman. Structure-function analysis of T4 RNA ligase 2. *The Journal of biological chemistry*, 278(20):17601–8, May 2003. ISSN 0021-9258. doi: 10.1074/jbc.M300817200. URL http://www.ncbi.nlm.nih.gov/pubmed/12611899.

Xiao-Li Zhan, James C. Clemens, Guilherme Neves, Daisuke Hattori,

John J. Flanagan, Thomas Hummel, M. Luisa Vasconcelos, Andrew Chess, and S. Lawrence Zipursky. Analysis of Dscam Diversity in Regulating Axon Guidance in Drosophila Mushroom Bodies. *Neuron*, 43(5):673–686, September 2004. ISSN 0896-6273. doi: 10.1016/j.neuron.2004.07.020. URL http://www.cell.com/neuron/abstract/S0896-6273(04)00431-3http://www.sciencedirect.com/science?_ob=GatewayURL&_origin=CELLPRESS&_urlversion=4&_method=citationSearch&_version=1&_src=FPDF&_piikey=S0896627304004313&md5=dc4c49dba613e4c5b586cf0971dee1f2http://dx.doi.org/10.1016/j.neuron.2004.07.020.

Jun Zhu, Jay Shendure, Robi D Mitra, and George M Church. Single molecule profiling of alternative pre-mRNA splicing. *Science (New York, N.Y.)*, 301 (5634):836–8, August 2003. ISSN 1095-9203. doi: 10.1126/science.1085792. URL http://www.ncbi.nlm.nih.gov/pubmed/12907803.

Julie Zikherman and Arthur Weiss. Alternative Splicing of CD45: The Tip of the Iceberg. *Immunity*, 29(6):839–841, December 2008. ISSN 1074-7613. doi: 10.1016/j.immuni.2008.12.005. URL http://www.sciencedirect.com/science/article/B6WSP-4V5PHVG-2/2/85564fb6cf8edff69e39ac25ef6c8484http://www.sciencedirect.com/science?_ob=MImg&_imagekey=B6WSP-4V5PHVG-2-1&_cdi=7052&_user=115184&_pii=S1074761308005153&_origin=&_coverDate=12/19/2008&_sk=999709993&view=c&wchp=dGLbVzz-zSkWb&md5=6395575253d87597c6465b78f94f7cb6&ie=/sdarticle.pdf.