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

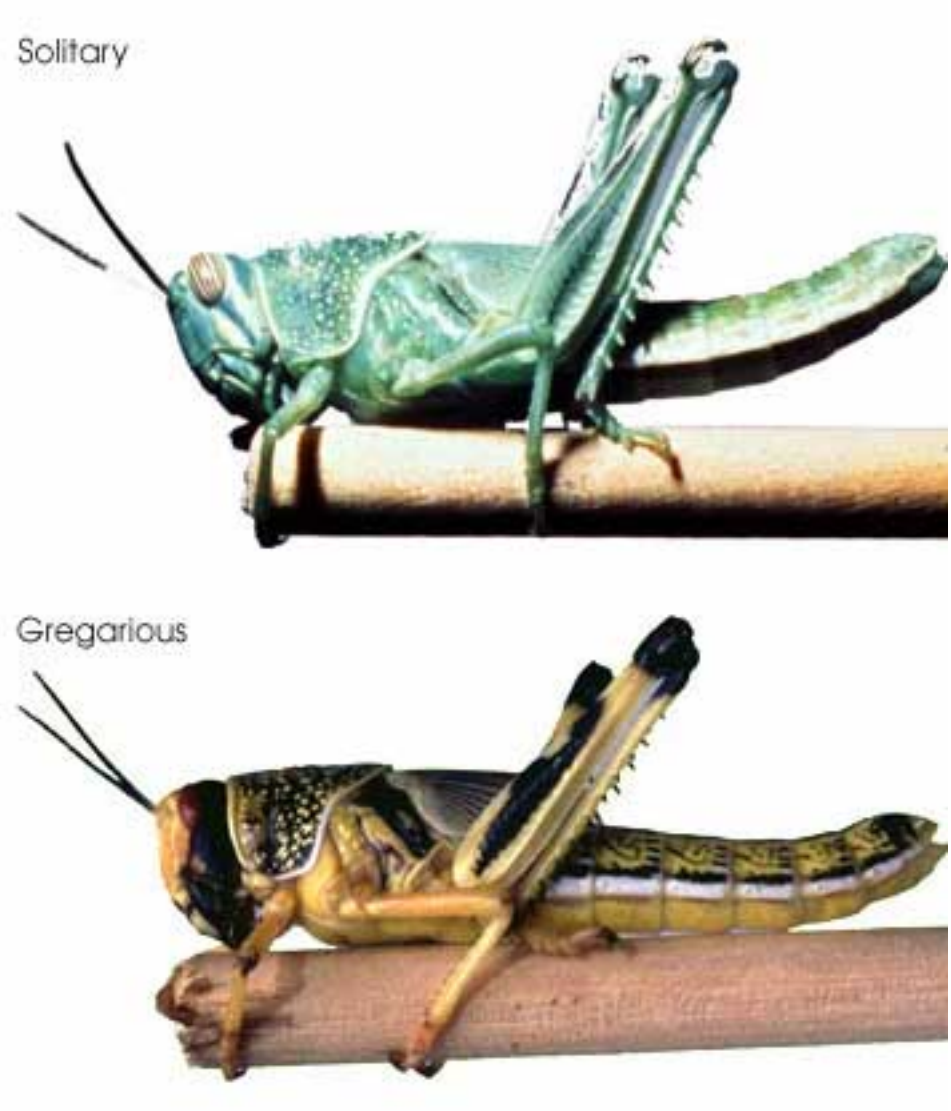

FIGURE 1.1: The Solitary and Gregarious forms of *Schistocerca gregaria*

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 [Shingleton et al., 2003, Purandare et al., 2014].

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.

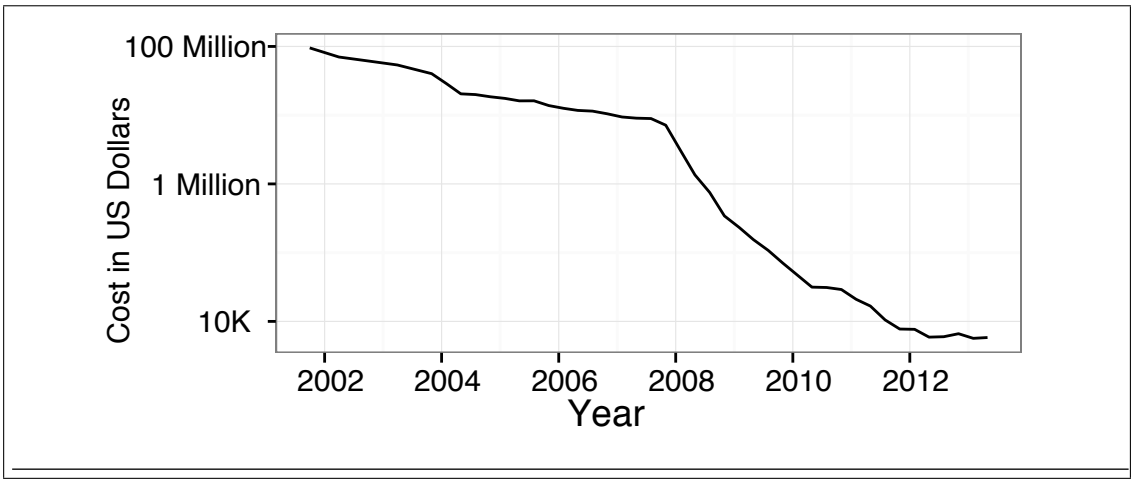

FIGURE 1.2: Cost of sequencing the human genome over time

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 [Shendure and Ji, 2008, Marioni et al., 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 [Mortazavi et al., 2008, Nagalakshmi et al., 2008, Lister 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).

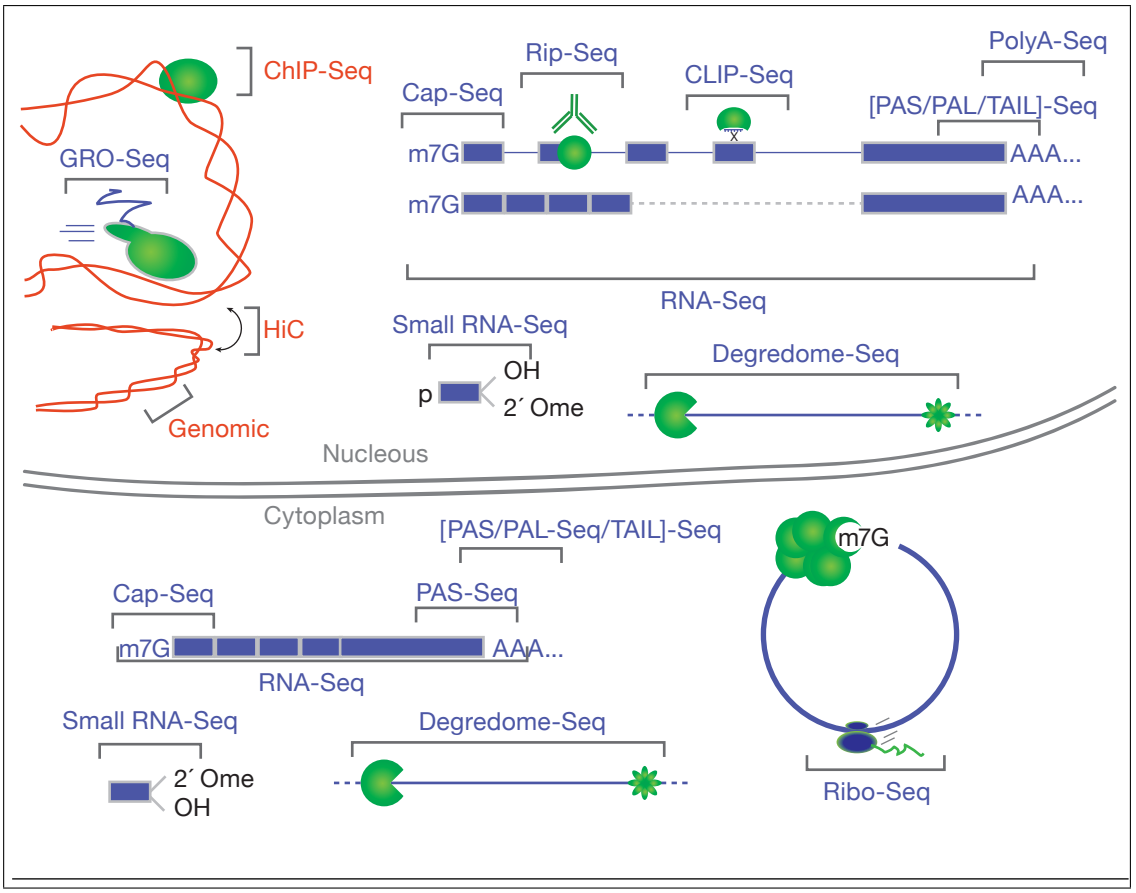

FIGURE 1.3: Methods for High-throughput sequencing of RNA

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.

Please refer to figure **??** to see if my Macro works!

Bibliography

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

Appendix A

Appendix Title Here

Write your Appendix content here.

Bibliography

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.