# COVER PAGE

# TITLE PAGE

Possible titles:

* Investigation of long RNAs using a novel ligation-based approach
* Investigation of long RNAs using RNA templated DNA ligation
* Long RNAs: biology, technology, and perspective
* Examination of dynamic long RNAs

A Dissertation Presented

By

Christian Knauf Roy

Submitted to the Faculty of the

University of Massachusetts Graduate School of Biomedical Sciences, Worcester

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSPHY

(MONTH, DAY, YEAR)

BIOCHEMISTRY

# SIGNATURE PAGE

A Dissertation Presented

By

Christian Knauf Roy

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

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

(Signature) (Name Typed),

Thesis Advisor (Signature) (Name Typed),

Thesis Advisor (Signature) (Name Typed),

Member of Committee

(Signature) (Name Typed),

Member of Committee

(Signature) (Name Typed),

Member of Committee

(Signature) (Name Typed),

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

(Signature) (Name Typed),

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. (Signature)

Anthony Carruthers, Ph.D., Dean of the Graduate School of Biomedical Sciences Program (Typed)

Month, Day and Year (Typed)

# Front Matter

## Dedication

## Acknowledgements

1. Talk about going to Grad School
2. Advisors
3. MJM
4. PDZ
5. Committee members
6. Zhiping
7. Scot
8. Job
9. Lab Members
10. Aaron
11. Alper
12. Amrit
13. Eric and Erin
14. Collaborators
15. Dave Weaver
16. Muro
17. Graveley
18. Heinrich
19. Anna
20. Ogo
21. Family

Table of Contents

Front Matter iv

Dedication iv

Acknowledgements v

Table of Contents vi

Abstract vii

List of Tables viii

List of Figures ix

List of Symbols, Abbreviations or Nomenclature (optional) x

Preface xi

Body Matter xii

Introduction **Error! Bookmark not defined.**

Research Chapters 5

Back Matter 9

Appendices 9

Biblography 9

## Abstract

## List of Tables

## List of Figures

* RNA Sequencing technology figure
* SeqZip methodology
* piRNA expression in mice

## List of Symbols, Abbreviations or Nomenclature (optional)

### Abbreviations

|  |  |
| --- | --- |
| AS | Alternative Splicing |
| DNA | Deoxyribonucleic acid |
| RNA | Ribonucleic acid |
| 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 basepair of DNA |
|  |  |

### Symbols

|  |  |
| --- | --- |
| 5´ |  |
| 3´ |  |
| μ |  |
|  |  |

### Definitions

|  |  |
| --- | --- |
| RNA-Seq |  |
| A ‘Read’ |  |

## Preface

*Here is where you will put wording discussing previous publications*

*Here is example wording from the Gupta thesis*

The work reported in this dissertation has been published in the following articles.

Chapter II has been published previously as: Ankit Gupta, Xiangdong Meng, Lihua J. Zhu, Nathan D. Lawson, and Scot A. Wolfe (2011). Zinc finger protein-dependent and -independent contributions to the in vivo off-target activity of zinc finger nucleases. Nucleic acids research 39, 381-392

Chapter III has been published previously as: Christensen RG, Gupta A, Zuo Z, Schriefer LA, Wolfe SA, Stormo GD (2011)

A modified bacterial one-hybrid system yields improved quantitative models of transcription factor specificity. Nucleic acids research 39, e83

Contents of Chapter IV have been accepted for publication Ankit Gupta, Ryan G. Christensen, Amy L. Rayla, Abirami Lakshmanan,

Gary D. Stormo, Scot A. Wolfe (2012)

An optimized two‐finger archive for ZFN‐mediated gene targeting, Nat. Methods, (Manuscript accepted).

1. A

# Body Matter

## Introduction

#### Motivation

Type your text under subheadings like this in paragraph form.

1. Topics that need to be discussed
2. Coordination in splicing
   1. Fededa, Fagnani, Parra (Deep intron), Peng (exons)
   2. Long range RNA interactions
      1. Bindewald, Eckart, and Bruce a Shapiro. 2013. “Computational Detection of Abundant Long-range Nucleotide Covariation in Drosophila Genomes.” RNA (New York, N.Y.) (July 25): 1171–1182. doi:10.1261/rna.037630.112. http://www.ncbi.nlm.nih.gov/pubmed/23887147.
3. Introduction to piRNAs
4. Historical Review of the literature
   1. RNA Expression
   2. Splicing
   3. Alternative Splicing
   4. SR Proteins
   5. RNA-Processing
   6. mRNPs
5. piRNAs
   1. Small RNA
   2. Tissue specific RNAs
6. Technical limitations
   1. Sequencing history
   2. Deep sequencing history

Soon after it was realized that DNA is the source of genetic information in all living organisms [Ref – Watson and Crick], and the ‘pretty’ and ‘elegant’ arrangement of complementary, antiparrellel DNA strands was known [Ref-The Double Helix, Watson, James D], 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 XX years after the nature of DNA’s architecture for the ability to ‘sequence’ it to become a reality. In 19XX (& 19XX(?)), two very different methods reported by Sanger [Ref] and Maxim-Gilbert [Ref] were reported. These sequencing technologies, from then on referred eponymously as ‘Sanger’ or ‘Maxem-Gilbery’ sequencing, were used to determine the specific order of DNA bases in a small piece (XX length) of viral (?) DNA. Sanger sequencing soon dominated most sequencing reactions, likely due to the conceptually more intuitive nature of the technology, and for over XX years, DNA sequences were slowly cloned, sequenced, analyzed, and dutifully catalogued into databases. For the 70’s and 80’s sequences were typically communicate in important publications [Ref-Some cell paper with the sequence of a human gene – insulin?]. The birth of the internet in the 1990’s brought essential publically-funded repositories for sequence information, such as PubMed [REF], and Ensembl [Ref]. However, it was really the human genome project [Ref – some review], that provided the important activation energy that brought DNA sequencing from a hard-to-perform but necessary piece of analysis, to an organized, large-scale effort to assemble complete source genetic material for relatively simple (*e.Coli* [Ref]) to complex (human [Ref]) genomes. A often criticized, but undeniably disrupting force, in the human genome project was the competing efforts of the privately-owned company Celera [Ref, Venter’s book]. Taking a higher-throughput and centralized approach to determining the sequence of the human genome, Celera fundamentally changed to approach to a large-scale sequencing project [Ref]. Instead of assigning specific sections of the genome to be worked out by individual labs, Celera parallelized the effort, in one room, by collecting many of the first high-throughput Sanger-sequencing devices from Agilent [Ref for a product?]. Using an EST-based approach [Ref], combined with sequence scaffolds [Definition] made available by the publically-funded project, Celera was able to create high-quality genomic sequences very quickly. Arguable, this was the first deep sequencing [Definition] effort, and changed the landscape of molecular and biochemical research, coincident with the beginning of a new millennium.

Sequencing DNA using Sanger’s technology remains, to this day, 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 & clonally amplified, typically using bacteria to do the heavy lifting [Ref]. Given that the human genome (hg19) comprises 3,137,161,264 bases (on just one strand), and that each Sanger reaction will provide ~800 nt of quality sequence, simple arithmetic tells us that we need at least ~4 Million individual reactions to determine the sequence of the human genome, assuming that all of our reads [definition] are of sufficient quality, length, and do not overlap by even 1 nt. Even the best practical improvements to workflows could not bring the Sanger approach to DNA sequencing in-line with asperations of analyzing the genomes of many different species or individuals.

In the early 2000’s appear to first efforts to change to approach to DNA sequencing, first using MPSS [Ref], but perhaps more importantly, and disruptively from a technology perspective, by Pyrosequencing [Ref].

* 1. The Isoform problem
  2. lncRNAs
  3. Computaitonal tools

1. State of Current literature
   1. ENCODE
   2. Integration of different datasets for more complete transcirptonal pictures
      1. Mention lincRNAs
   3. Single Cell RNA-Seq
      1. Shapiro, Ehud, Tamir Biezuner, and Sten Linnarsson. 2013. “Single-cell Sequencing-based Technologies Will Revolutionize Whole-organism Science.” Nature Reviews. Genetics 14 (9) (July 30): 618–630. doi:10.1038/nrg3542. http://www.ncbi.nlm.nih.gov/pubmed/23897237.
      2. 1. Shalek AK, Satija R, Adiconis X, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*. 2013:1–5. doi:10.1038/nature12172.
      3. Wills, Quin F, Kenneth J Livak, Alex J Tipping, Tariq Enver, Andrew J Goldson, Darren W Sexton, and Chris Holmes. 2013. “Single-cell Gene Expression Analysis Reveals Genetic Associations Masked in Whole-tissue Experiments.” Nature Biotechnology 31 (8) (July 21): 748–752. doi:10.1038/nbt.2642. http://www.ncbi.nlm.nih.gov/pubmed/23873083.
2. Relationship between alternative splicing and protein domains
3. 1. Light, S. & Elofsson, A. The impact of splicing on protein domain architecture. *Current opinion in structural biology* 1–8 (2013).doi:10.1016/j.sbi.2013.02.013

## Research Chapters

### CHAPTER I: SeqZip methodology

1. Nature methods paper will do 90% of the work here
2. SeqZip Development
   1. Discovery of novel enzyme activity
3. Search for long and complicated RNAs
   1. CD45
   2. FN1
4. You should tie together the fact that Dscam ALSO has Fibronectin sections!
5. DSCAM

### CHAPTER II:HIV and piRNA precursors

1. piRNA precursors
2. HIV Transcript integrity

### CHAPTER III : piRNA Chapter & MolCel2013

## Perspective/Final Summary & Conclusions

Things to remember about this section:

1. Multiple people have stated that this can be highly speculative
2. This is my chance to show that I am a scientist
3. If I were to die tomorrow, what would be the experiments to conduct
4. Discuss long range RNA secondary structure and implecations of regulating alternative splicing (See Li, S., & Breaker, R. R. (2013). Eukaryotic TPP riboswitch regulation of alternative splicing involving long-distance base pairing. Nucleic acids research, 41(5), 3022–31. doi:10.1093/nar/gkt057) and Reg of AS by long RANGE SS folder in Mendeley
5. Full length analysis of mRNAs
6. Implications for discrination past one's DNA as it is the actual PRODUCT of the DNA and the actual biology (or at least closer to the functional biology) that is going on inside of every person
7. Catalogue of every possible mRNA isoform
8. Seperate the signal from the noise of transcription
9. For out there comments:
10. Most high cited pnas papers are methodology papers
11. www.pnas.org/reports/most-cited

# Back Matter

## Appendices

### Methods

#### SeqZip

#### Computational Scripts

#### Computational Approaches

##### Git

##### Perl

##### IGV

##### MySQL

#### ImageQuant

#### LOD of Radioactivity

### Oligo Database

### Important MySQL data tables

## Biblography