Capture-seq and small RNA-seq to identify noncoding RNAs in the mouse ribosomal RNA gene repeat intergenic spacer

A thesis presented in partial fulfilment of the requirements for the degree of

Master of Science

In

Genetics

At Massey University (Albany), New Zealand

Jessica Leigh Fitch

2018

Abstract

Cancer is a leading cause of mortality in developed countries. Due to the genetic and epigenetic heterogeneity of this disease, we still don't have effective long-term therapies for many cancers. A characteristic of many cancer cells is an alteration in the structure of the nucleolus - the primary location of the ribosomal DNA (rDNA). The rDNA encodes ribosomal RNA, which is the major structural and catalytic component of ribosomes – the cellular machinery responsible for protein biosynthesis. Accordingly, the rDNA and its transcription is a key regulator of cell proliferation. Despite this critical role, the highly repetitive nature of the rDNA has made it difficult to study, thus it remains an attractive target for anti-cancer therapies. Indeed, the promising anti-cancer drug, CX-5461, developed by our collaborators, targets the rDNA through the inhibition of the rDNA dedicated RNA polymerase I (currently in clinical trials).

In preliminary experimentation, there is a dramatic change in expression of non-coding RNAs (ncRNAs) from the rDNA during the transition to malignancy. Although the function of rDNA ncRNAs is almost entirely unknown, ncRNAs from other regions of the genome have a multitude of regulatory functions, including involvement in cancer. We hypothesise that these transcripts play a role in malignancy and CX-5461 sensitivity.

Utilising a mouse B-lymphoma model (Eμ-myc), we first applied a high throughput hybridisation-based RNA-sequencing approach (capture-seq), to enrich for rDNA intergenic spacer (IGS) ncRNA transcripts within 11 cDNA sequencing libraries. Regions of transcription throughout the IGS were identified using several bioinformatic tools, and qPCR was performed to validate transcription status as well as assess for CX-5461-dependent transcriptional changes. We also utilised other bioinformatics tools, to predict small RNAs arising from the IGS and other regions of the Eμ-myc genome, and briefly assessed their response to CX-5461 treatment. miRNAs of interest were assessed for potential pathway targets using several bioinformatic targets. Lastly, we aimed to further characterise the Eμ-myc model. With this, we assessed efficacy of methods that could be used for downstream knockdown/over expression analysis.

Overall, using the capture-seq method we identified 8 major clusters of exons (known as exon cluster groups), that were consistently predicted between RNA library preparations. These were confirmed to be transcriptionally active by qPCR, with one of these clusters. Additionally, we identified several sites in the mouse rDNA IGS that may express small RNAs, with small RNA reads aligning to these sites with some consistency between library preparations. Some of these, due to presence and absence patterns in either CX-5461 treated or control libraries, may show some signs of treatment-dependent differential expression. We also identified miRNAs from other regions of the genome which show similar patterns. We assessed potential small RNAs for gene target enrichment. No pathways/cellular components appeared to be biologically significant . We assessed a method of viral-mediated gene knockdown in a number of cell lines, which did not show efficacy in the mouse lines we had available.

In conclusion, if these exons produce ncRNAs that contribute to malignancy, the ncRNAs will form attractive new targets for therapy, independently or in combination with CX-5461, and could be used as diagnostic and prognostic markers of cancer. The future trajectories of this project include selecting promising IGS transcripts, particularly those differentially expressed, to confirm their size by northern blot. Then, to assess their role in malignant cells, to perform knockdown/overexpression assays and assess cellular response. Further, we would target the rDNA ncRNAs in several cancer and non-cancer cell lines, to broaden our understanding of anti-cancer application.

Acknowledgements

I'd like to extend my sincerest gratitude to my supervisors, Austen and Sebastian, for their on-going support and guidance throughout my masters. I have always been one to ask a lot of questions, some better than others, so I have appreciated your patience. My confidence in my own ability has greatly improved over the last few years, and they (as well as all my colleagues and friends) are the reason behind it. I will always be immensely grateful.

I'd also like to thank the entire Ganley Lab for their help, by means of suggestions, critique and some laughs along the way; it has been a pleasure being part of the team. Particularly, thank you Daria Chudakova and Diksha Sharma for all your help and advice in the lab and around computers. A lot of the work completed was completely foreign (and in some cases extremely intimidating). I wouldn't have been able to finish this without you.

Thank you to my parents, family and friends for all your love and encouragement over the last few years. Your support (in all the many shapes and forms) will always be cherished.

Lastly, thank you to all our new friends at The University of Auckland. Special acknowledgements to Liam Williams and Kristine Boxen, who day and night work tirelessly sequencing and helping with genomics equipment. Your tips and tricks have been fantastic, and it has been great working with you.

Contents

Abstract ii
Acknowledgements iv
List of tables ix
List of Figures x
List of abbreviations 1
1.Introduction 2
1.1 A brief introduction to cancer and current cancer therapies 2
1.2 Targeting the ribosomal DNA and noncoding RNA for the treatment of cancers 4
1.2.1 Ribosomal DNA (rDNA) 4
1.2.2 Non-coding RNAs 11
2. Project aims 15
3. Materials and Methods 16
3.1 Culturing mammalian suspension cell lines (Eμ-myc) *in vitro* 16
3.2 Culturing adherent cell lines 17
3.3 Cytotoxicity assay 17
3.4 Treating cells with CX-5461 18
3.5 Preparing RNA for Capture-Seq 19
3.5.1 RNA extractions for total RNA 19
3.5.2 Ribodepletion of total RNA 19
3.5.3 Measuring RNA quality 19
3.5.4 DNA extraction 19
3.5.5 DNA sonication 20
3.6 Western blot analysis for analysing UBF knockdown efficiency 20
3.7 Quantitative polymerase chain reaction (qPCR) for shUBF knockdown efficiency 22
3. 8 cDNA synthesis, library preparation and hybridisation-based enrichment for lncRNAs and small RNAs. 23
3.9 cDNA synthesis and library preparation for <120 bp small RNAs 25
3.10 Bioinformatic analysis Capture long ncRNA RNA-Seq Multiplex libraries 25
3.10.1 Producing reference genome 25
3.10.2 Indexing reference genome 26
3.10.3 Aligning Capture-seq reads to reference genome, and sorting/cleaning alignment outputs using Samtools 26
3.10.4 Using Stringtie to assemble reads into Transcripts 27

3.10.5 Determining per base coverage levels of the rDNA IGS to assess for areas of high transcription 28
3.10.6 Comparing IGS aligned read numbers to reads aligning to the whole genome in the DNA-derived library to estimate theoretical enrichment potential 30
3.10.7 Normalising between libraries using ERCC spike ins 31
3.10.8 Normalising exons to the captured-DNA to reduce effect of capture bias 33
3.10.9 Producing a Repeatmasker GTF file for the mouse rDNA IGS 33
3.11 Designing and optimising qPCR primers to validate expression of IGS exons 34
3.11.1 Assessing efficiency/specificity of IGS exon qPCR primers and preliminary qPCR testing for validating exon transcription 34
3.11.2 Testing for presence of gDNA contamination 35
3.11.3 RNA extraction and final to validate transcription from predicted transcribed regions of the mouse rDNA IGS 35
3.11.4 Assessing rRNA transcription changes with CX-5461 treatment via qPCR 36
3.12 Bioinformatic analysis of small (>120bp) RNA-seq data 36
3.12.1 Quality assessment and trimming raw reads 36
3.12.2 Aligning and visualising small RNA reads to identify potential small RNAs 36
3.12.3 Finding miRNA from the rDNA IGS using Bowtie and mirDeep2 38
3.13 Lentiviral transfection and infection optimisation 40
3.13.1 Lentiviral transfection of HEK and MEF cells 40
3.13.2 Lentiviral transduction of 4242 cells 41
4. Results 42
Section 4.1 Identifying regions of lncRNA transcription in the mouse rDNA IGS 43
4.1.1 Capture-seq experimental design 43
4.1.2 Producing the captured libraries enriched for mouse rDNA IGS transcripts 44
4.1.2.1 First attempt cDNA library synthesis from high quality ribodepleted RNA 45
4.1.2.2 Second attempt cDNA library synthesis from high quality ribodepleted RNA 50
4.1.2.3 Measuring efficacy of CX-5461 treatment on 4242 and shUBF cells 53
4.1.2.4 Preparing and capturing pooled cDNA libraries 54
4.1.3 Bioinformatic analysis of capture-RNA-seq for IGS noncoding-transcripts 57
4.1.3.1 The theoretical efficiency of the Capture-seq method at enriching for noncoding RNA from the IGS 58
4.1.2.2 Using bedtools coverage to assess for areas of transcription within the mouse rDNA IGS 60
4.1.2.3 Finding rDNA IGS exons using a bioinformatic approach 65

4.1.2.4 ERCC analysis for normalising Stringtie FPKM outputs between cDNA library samples 66
4.1.3 Initial qPCR assessing IGS exon transcription from exon clusters 71
4.1.3.1 Assessing gDNA contamination in qPCR RNA samples 75
4.1.4 Final qPCR validation of mouse rDNA IGS transcription within exon clusters..... 76
Section 4.2 Identifying small RNAs derived from the mouse rDNA IGS and assessing small RNA response to CX-5461 81
4.2.1 Preparing small RNA for library preparation, and early sequencing output cleaning and manipulation 81
4.2.2.1 Small RNA analysis using standard alignment/visualisation, and downstream target and GO enrichment analysis 83
4.2.2.2 GO enrichment analysis of potential seed sequences as determined from STAR aligner 91
4.2.3 Small RNA analysis using miRDeep2 software 93
4.3 shRNA characterisation in the Eμ-myc model 97
4.3.1 shUBF knockdown confirmation analysis 97
..... 99
44.3.2 Lentiviral transduction 99
5. Discussion..... 105
5.1.1 Identification of noncoding exons within the Eμ-myc rDNA IGS 105
5.2 Small RNA-seq data reveals small RNAs with potential differential expression upon CX-5461 treatment 109
5.3 Transduction into mouse cell lines 110
5.4 Future trajectories 111
5.5 Final summary..... 114
6. Bibliography 115
7. Appendices..... 126
Appendix 1 :Primer sequences for qPCR and conventional PCR 126
Appendix 2 : Table of S1 library full Stringtie output example showing transcript coverage, exons contributing to the transcript and coverage of exons..... 130
Appendix 3: Table of IGS exons predicted by Stringtie in day two library data, normalised to DNA and ranked from highest to lowest in regards to coverage. Library name in bold. 132
Appendix 4: Full GO SLIM enrichment outputs (biological processes and cellular components) from STAR alignment predicted seed sequences 133
Appendix 5: Full GO SLIM enrichment biological processes (bio pro) and cellular components(cell comp) outputs from miRDeep2 predicted seed sequences 134

Appendix 6: Buffer table 135

Appendix 7: Table of ERCC input concentrations calculated for day one or day two libraries (in ERCC mix 1 or mix 2) 136

Appendix 7 continue: ERCC input concentrations calculated for day one or day two libraries (in ERCC mix 1 or mix 2) 137

List of tables

Table 1 Mammalian cell lines used in experiments 16
Table 2 First extraction RNA concentrations pre- ,post- ribodepletion and post lm-PCR 49
Table 3 Second extraction RNA concentrations pre- ,post- ribodepletion and post lm-PCR 52
Table 4 Outline of day one and day two libraries pooled for capture 55
Table 5 Raw read and STAR aligner outputs from rDNA IGS capture sequencing data 57
Table 6 Theoretical capture efficiency using DNA derived library comparing all mouse chromosomes (and rDNA coding region) to the IGS 59
Table 7 Cycle differences between RNA to cDNA input and RNase treated RNA input into qPCR with capture-seq noncoding exon primers 1-10 76
Table 8 Small RNA library raw read output numbers 82
Table 9 STAR aligner read outputs from small library sequencing for 4242 and shUBF lines either treated or untreated (DMSO) with CX-5461 85
Table 10 Small RNA reads fitting our set criteria that may reflect small RNA exons, their location in the IGS and seed sequence 87
Table 11 Comparing read numbers aligning to spacer promoter region between different libraries 88
Table 12 Seed sequences of our predicted small RNAs, their sequence and treatment scheme found in 91
Table 13 GO slim biological processes and cellular compartment outputs from Targetscan results of seed sequences (table 11) 92
Table 14 miRNAs predicted by miRDeep2 found in more than one library 94
Table 15 miRDeep2 predicted miRNAs GO slim pathway enrichment output from Targetscan-predicted targets 95

List of Figures

Figure 1 RNA polymerase I (RNA Pol I) complex at the rDNA promoter 6
Figure 2 Schmetic summarising RNA extraction, Ribodepletion and the steps within the SeqEZ RNA Enrichment System User guide 23
Figure 3 rDNA IGS regions targeted by capture probes 44
Figure 4 Figure 4 Examples of Bioanalyser outputs with high RNA quality 45
Figure 5 Assessing ribodepletion efficiency in first RNA extractions 46
Figure 6 Comparing Bioanalyser results from day one cDNA library preparations after ligation-mediated PCR (lm-PCR) to ideal capture-seq cDNA libraries 48
Figure 7 Assessing ribodepletion efficiency in second RNA extractions 51
Figure 8 Bioanalyser results of day two cDNA library preparations after ligation-mediated PCR (lm-PCR**)** 51
Figure 9 Example of Cytotoxicity assay results 54
Figure 10 Final pooled captured library output 56
Figure 11 Coverage graphed against rDNA base position (coding and IGS) 62
Figure 12 Coverage graphed against IGS base position 63
Figure 13 Coverage graphed against IGS base position after normalisation to DNA capture 64
Figure 14 Stringtie-predicted exon output visualised in IGV 66
Figure 15 ERCC plots used for step one of two-step IGS exon coverage normalisation 68
Figure 16 ERCC slopes against volume of input or read number output 69
Figure 17 Location of exon clusters within the mouse rDNA IGS 71
Figure 18 Examples of outputs during primer validation 72
Figure 19 First attempt of qPCR validating transcription from predicted IGS noncoding exons clusters 74
Figure 20 qPCR validation of bioinformatically identified mouse rDNA IGS ncRNA exon clusters 79
Figure 21 Measuring 47S pre-rRNA expression difference in CX-5461 treated and untreated samples 80
Figure 22 Pre- and post- trimming FastQC graphs of a raw small RNA sequencing output 83
Figure 23 Small library read outputs compared to numbers of uniquely mapped and unmapped 84
Figure 24 IGV visualisation of all small RNA reads aligning across the mouse rDNA unit 86
Figure 25 Self-dotplot of mouse rDNA unit produced by Geneious software 90
Figure 26 UBF knockdown analysis in shUBF Eμ-myc compared to control 99
Figure 27 Packaging and infection into HEK cells using lentiviral system 100
Figure 28 Lentiviral infection attempt into Eμ-myc from HEK packing cells 102
Figure 29 Lentiviral infection into MEF cells from HEK packaging cells 103

List of abbreviations

Unit abbreviations

µl/ µlg : micro-litre/-gram

G: gram

Hrs: hours

L: litre

Mins : minutes

ml/mg: mili-litre/-gram

nl/ng: nano-litre/-gram

Frequently used abbreviations

cDNA: complementary DNA

IGS: intergenic spacer

miRNA: micro-RNA

mRNA: messenger RNA

qPCR: quantitative PCR

rDNA: ribosomal DNA

RIN: RNA integrity number

RNA pol I : RNA polymerase I

RNA-seq: RNA- sequencing

shRNA: short hairpin RNA