

Investigating the microRNA-mediated regulation of protein-coding transcripts in animals using RNA-Seq data

Thomas Bradley

A thesis submitted for the degree of Doctor of Philosophy (PhD)

School of Biological Sciences,
University of East Anglia,
Norwich,
United Kingdom

Earlham Institute,
Norwich,
United Kingdom

July 2020

This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the author and that use of any information derived therefrom must be in accordance with current UK Copyright Law. In addition, any quotation or extract must include full attribution.

Abstract

MicroRNAs (miRNA) are small non-coding RNAs, of approximately 22 nucleotides in length, which play an important role in the post-transcriptional regulation of gene expression. Post-transcriptional regulation by miRNAs is achieved by direct translational inhibition or decay of other RNA molecules, or a combination of both of these mechanisms. miRNAs are implicated in a large number of developmental processes across the animal kingdom, underscoring their importance to biological research, and in particular, research relating to diseases states as a result of aberrant development. Investigation of the precise role of individual miRNAs or groups of miRNAs within cells requires the accurate identification of miRNA targets. Limitations in experimental methods for the identification of miRNA targets, necessitates the use of computational algorithms for this purpose. However, accurate computational identification of miRNA targets can be difficult, due to the short six or seven nucleotide seed sequence of the miRNA which is used for the recognition of targets, leading to a large number of false positive predictions being made. In this thesis, I demonstrate how data from transcriptome-wide bulk RNA sequencing experiments can be used to increase the accuracy of miRNA target prediction workflows. Firstly, I show how data of this type can be used to generate 3'UTR annotations specific to the biological context in which sequencing occurred, and secondly, how it can be used to remove lowly expressed mRNA transcripts from the target prediction process. Implementation of both of these steps in miRNA predictions workflows is shown in this thesis to increase prediction accuracy. In addition, I explore how data from bulk RNA Sequencing can be used in combination with data generated from

small RNA sequencing experiments in order to infer the regulatory activity of individual miRNAs during given developmental processes.

Access Condition and Agreement

Each deposit in UEA Digital Repository is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the Data Collections is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form. You must obtain permission from the copyright holder, usually the author, for any other use. Exceptions only apply where a deposit may be explicitly provided under a stated licence, such as a Creative Commons licence or Open Government licence.

Electronic or print copies may not be offered, whether for sale or otherwise to anyone, unless explicitly stated under a Creative Commons or Open Government license. Unauthorised reproduction, editing or reformatting for resale purposes is explicitly prohibited (except where approved by the copyright holder themselves) and UEA reserves the right to take immediate 'take down' action on behalf of the copyright and/or rights holder if this Access condition of the UEA Digital Repository is breached. Any material in this database has been supplied on the understanding that it is copyright material and that no quotation from the material may be published without proper acknowledgement.

Table of Contents

<i>Abstract</i>	2
<i>List of Tables</i>	9
<i>List of Figures</i>	11
<i>Acknowledgements</i>	17
<i>Declaration</i>	21
<i>Outcomes</i>	23
Publications	23
Software Developed	23
Chapter 1: Introduction	25
Chapter 2: Background	28
2.1 Overview	28
2.2 Evolution of miRNAs	30
2.3 Biogenesis and crosstalk with RNAi pathways	32
2.4 The relationship between miRNAs and other classes of metazoan sRNAs ...	39
2.5 miRNA annotation and database resources	41
2.6 Roles of miRNAs in the cell	45
2.7 Molecular mechanisms of miRNA-mediated target repression	48
2.8 Non-coding RNA targets of microRNAs	51
2.9 Experimental validation of miRNA targets	53
2.9.1 3'UTR reporter assays.....	53
2.9.2 miRNA perturbation and sequencing	54
2.9.3 Cross-linking and immunoprecipitation	57
2.9.4 Databases of validated miRNA interactions	60
2.10 Principles of miRNA targeting	60
2.10.1 Canonical models of miRNA targeting	61
2.10.2 Noncanonical models of miRNA target prediction	72

2.11 The accuracy of miRNA target prediction models	81
2.12 RNA-Seq and differential expression analysis.....	84
2.12.1 RNA Sequencing and transcript quantification	84
2.12.2 Transcript quantification tools	90
2.12.3 Differential expression analysis	93
2.13 Alternative cleavage and polyadenylation	97
2.14 Combined target prediction and expression data tools and analyses	100
<i>Chapter 3: FilTar and FilTarDB design and development.....</i>	<i>104</i>
3.1 Contributions	104
3.2 Introduction	104
3.3 Motivation	105
3.4 Aims statement.....	106
3.4.1 Target user base	107
3.4.2 MicroRNA Target Type	108
3.5 Community needs addressed	111
3.5.1 Current issues with existing software.....	112
3.6 General Design & Implementation.....	112
3.6.1 Workflow Management.....	112
3.6.2 General Schema	114
3.6.3 Modular Design	117
3.6.4 Module Configuration	120
3.6.5 Modules Schema.....	121
3.6.6 FilTar Modules and configuration.....	122
3.6.7 Dependency Management.....	139
3.6.8 Automated Testing, Automated Building & Continuous Integration	143
3.7 Command-Line Application	145
3.7.1 User Interface.....	145
3.7.2 Project Deployment, licensing & maintenance	146
3.7.3 Documentation.....	148
3.7.4 Performance.....	148
3.8 Web Application and Database	152
3.8.1 System Architecture	152

3.8.2 Additional Backend Modules	154
3.8.3 The FilTar Database.....	155
3.8.4 The FilTarDB web application	158
3.8.5 Performance	169
3.9 Conclusion.....	170
<i>Chapter 4: Validation of the FilTar approach.....</i>	<i>172</i>
4.1 Contributions	172
4.2 Introduction	172
4.3 Methods.....	173
4.3.1 Data selection	173
4.3.2 Quality control and statistics	175
4.3.3 Differential expression analysis.....	176
4.3.4 Data Visualisation.....	178
4.3.5 FilTar Implementation.....	179
4.4 Results	180
4.4.1 Expression filtering.....	180
4.4.2 3' UTR extension.....	185
4.4.3 3' UTR truncation.....	191
4.4.4 Cumulative effect of filtering and reannotation	200
4.5 Discussion.....	209
4.6 Conclusion.....	213
<i>Chapter 5: The regulation of the post-mating response in Drosophila melanogaster by miRNAs.....</i>	<i>215</i>
5.1 Contributions	215
5.2 Introduction	216
5.3 Background.....	216
5.4 Methodology.....	220
5.4.1 Experimental Design.....	220
5.4.2 Sample preparation	222
5.4.3 RNA Extraction.....	223
5.4.4 Library construction and sequencing	223
5.4.5 Sequence analysis and differential expression analysis.....	223

5.4.6 miRNA target prediction	224
5.4.7 Data Pre-processing and Normalisation.....	229
5.4.8 Integrated Analysis.....	230
5.5 Results	232
5.5.1 QC.....	232
5.5.2 Differential Expression Analysis.....	233
5.5.3 miRNA target prediction	237
5.5.4 Integrated Analysis.....	259
5.6 Discussion	264
<i>Chapter 6: The regulation of sex transition in Lates calcarifer (Asian seabass) by miRNAs.....</i>	<i>270</i>
6.1 Contributions	270
6.2 Introduction	270
6.3 Background	272
6.4 Experimental Design	274
6.5 Methodology.....	275
6.5.1 Sample preparation.....	275
6.5.2 RNA extraction	275
6.5.3 Library construction and sequencing.....	276
6.5.4 miRNA sequence analysis, annotation and quantification.....	276
6.5.5 Differential expression analysis	277
6.5.6 miRNA target analysis	278
6.5.7 Clustering and data visualisation.....	279
6.5.8 GO Term enrichment analysis	279
6.6 Results	280
6.6.1 Differential expression analysis	280
6.6.2 miRNA targeting analysis.....	288
6.6.3 GO term enrichment Analysis.....	291
6.7 Discussion	293
6.8 Conclusion	296
<i>Chapter 7: Future Work and Conclusion</i>	<i>297</i>
7.1 Future Work.....	297

7.2 Conclusion.....	302
<i>Definitions</i>	<i>304</i>
<i>Glossary</i>	<i>305</i>
<i>Bibliography</i>	<i>312</i>
<i>Index.....</i>	<i>330</i>
<i>Appendix A</i>	<i>331</i>
<i>Appendix B</i>	<i>352</i>
<i>Appendix C</i>	<i>369</i>

List of Tables

Table 2.1 - A summary of the most common computational methods used for miRNA target prediction.	80
Table 3.1 - A list of all dependencies needed to use the FilTar command line application.	141
Table 4.1 - FilTar 3'UTR reannotation summary statistics for cell line and tissue data used in this study.	202
Table 4.2 - The total number of miRNA seed sites lost through expression filtering or 3'UTR reannotation of transcripts.	204
Table 4.3 - Summary statistics of the effects of filtering protein-coding transcripts at an expression threshold of 0.1 TPM.	206
Table 4.4 – Combined statistics relating to 3'UTR reannotation and expression filtering	208
Table 5.1 - The experimental conditions examined in this study.	221
Table 5.2 - The results of the differential expression analysis of miRNA and protein coding genes for comparisons relating to both sex and body part. ..	235
Table 5.3 - A table providing information relating to differentially expressed genes co-targeted by miR-927-3p and miR-927-5p.	240
Table 5.4 - A statistics table for the integrated analysis of D. melanogaster sequencing data for single mature miRNAs.	260
Table 5.5 - A table of p and adjusted p values testing for the combinatorial effect of multiple miRNAs differentially expressed in the same direction targeting the same set of targets.	262
Table 6.1 - A summary of the results of the miRNA differential expression analysis with demarcations between the number of downregulated, upregulated and differentially expressed miRNA in each comparison.	283
Table 6.2 - A summary of the results of the RNA seq differential expression analysis.	285
Table 6.3 - A summary of the results of the analysis of the predicted targets of differentially expressed miRNAs.	290
Table 6.4 - GO term enrichment analysis.	292

Table A.1 - A table of summary and quality control statistics for all sequencing runs used in this analysis.	336
Table A.2 - A summary of RNA-sequencing datasets analysed for chapter 4 of this thesis.	348
Table A.3 - A summary of data considered during preliminary analysis, but were not used for further analysis.	349
Table A.4 - An assessment of the signal-noise ratio in each miRNA mimic transfection experiment.	351
Table B.1 - mRNA sequencing depth metrics and values, along with library metadata.	354
Table B.2 - sRNA sequencing depth metrics and values, along with library metadata.	355
Table B.3 - A table of oppositely differentially expressed targets of differentially expression miRNAs.	368
Table C.1 - Identifiers of novel Asian seabass (<i>Lates calcarifer</i>) miRNAs discovered during the course of the research described in chapter 6.	372

List of Figures

Figure 2.1 - A comparison of structural (blue text), biogenesis (green text) and functional (red text) differences and similarities between three different classes of small RNA involved in cellular gene silencing mechanisms: miRNAs, piRNAs and siRNAs.	29
Figure 2.2 - A model for the evolution of miRNA targeting mechanisms in different eukaryotic lineages as proposed in Moran et al. 2017 (Moran, et al., 2017).	31
Figure 2.3 - summary of miRNA biogenesis in animals.	38
Figure 2.4 - A summary of miRNA-mediated destabilisation and translation repression of mRNA molecules.	50
Figure 2.5 – The domains and crystal structure of human argonaute-2.	65
Figure 2.6 – A summary of miRNA target site types.	67
Figure 2.7 – A summary of some of the duplex, local contextual and global contextual sequence features used by miRSVR to score putative targets, which are representative of features used by other target prediction algorithms. ...	71
Figure 2.8 – A standard RNA-Seq library preparation and data analysis workflow.	89
Figure 2.9 – The different forms of alternative polyadenylation and cleavage.	99
Figure 3.1 – A high-level overview of the FilTar workflow.	114
Figure 3.2 – Basic FilTar schema.	115
Figure 3.3 – The recursive relationship of snakemake modules and subsidiary modules exploited by FilTar for the purposes of efficient workflow management.	119
Figure 3.4 – The configurability of the FilTar pipeline architecture.	121
Figure 3.5 – Schema of FilTar modules.	122
Figure 3.6 – The estimated distribution of the context++ scores of the TargetScan algorithm.	135
Figure 3.7 – The predicted probability mass function of miRanda alignment scores.	138

Figure 3.8 - Core installation duration for the FilTar command line tool..	149
Figure 3.9 – The effect of library number and total library size on FilTar run time.....	151
Figure 3.10 – The systems architecture of the FilTar web tool.....	154
Figure 3.11 – FilTarDB database design.....	157
Figure 3.12 – The basic design of the FilTarDB web application.	160
Figure 3.13 – The home page of the FilTarDB website.	162
Figure 3.14 – Forms to be completed by the user exhibit field chaining and auto-complete functionality.....	163
Figure 3.15 – Exception handling mechanisms prevents the user from submitting invalid queries.	165
Figure 3.16 – An example of a results page from the FilTar website. Relevant metadata is displayed above the results table.....	165
Figure 3.17 – The FilTar results data table can be searched using a search bar.	167
Figure 3.18 – The ordering of columns of the results table.....	168
Figure 3.19 - Test of FITarDB query latency.....	170
Figure 4.1 - Implementing an expression threshold on predicted miRNA targets improves miRNA target prediction accuracy.....	181
Figure 4.2 – Differential expression of lowly expressed predicted miRNA targets upon miRNA mimic transfection.....	182
Figure 4.3 - The effect of expression filtering on retained protein-coding transcripts using multiple expression thresholds.....	183
Figure 4.4 - The effect of expression filtering on removed protein-coding transcripts using multiple expression thresholds.....	184
Figure 4.5 - 3'UTR elongation by FilTar leads to the identification of additional valid miRNA targets. mRNA transcripts contained in each distribution.....	186
Figure 4.6 - Greater sequencing depth leads to greater 3'UTR elongation up to a point of saturation.....	187
Figure 4.7 - The relationship between the number of mapped reads and the extent of 3'UTR elongation observed when using FilTar.....	189

Figure 4.8 - A scatter plot of the percentage gain in total miRNA target site predictions vs. percentage gain in 3'UTR bases for a number of cell lines and tissue datasets analysed (black dots).	190
Figure 4.9 - 3'UTR truncation by FilTar leads to the removal of false positive miRNA target predictions.	192
Figure 4.10 - As in figure 4.9, with the exception that no expression threshold has been implemented to filter data points contained with the removed seed site distribution.	193
Figure 4.11 – An example of erroneous 3'UTR model predictions for lowly expressed genes.	194
Figure 4.12 - Predicted targets removed by FilTar exhibit weaker repression in response to miRNAs than 6mer targets.	196
Figure 4.13 - Greater sequencing depth leads to greater 3'UTR truncation up to a point of saturation.	197
Figure 4.14 - The relationship between the number of mapped reads and the extent of 3'UTR truncation observed when using FilTar.	198
Figure 4.15 - A scatter plot of the percentage loss in total miRNA target predictions vs. percentage loss in total 3'UTR bases.	199
Figure 4.16 - Total miRNA target site gain and loss when applying FilTar to multiple sample types.	201
Figure 4.17 – The effect of expression filtering on multiple cell lines.	205
Figure 5.1 – The relationship between context++ scores and fold change between two conditions for the predicted targets of a differentially expressed miRNA.	228
Figure 5.2 – Principle components analysis of expression data for both A) mRNA and b) miRNA in this study.	233
Figure 5.3 - Network visualisation of predicted miRNA interactions in the male abdomen: Nodes with thick borders denote miRNAs, whilst nodes without thick borders represent coding genes.	238
Figure 5.4 - The number of predicted miRNA seed target sites categorised by site type, after running the TargetScanS (Lewis, et al., 2005) algorithm with D. melanogaster miRNAs and 3'UTRs.	242
Figure 5.5 - The number of predicted seed miRNA target sites with respect to the gene in which those sites are found.	243

Figure 5.6 - The frequency of predicted seed miRNA target sites with respect to the targeting miRNA.	244
Figure 5.7 – The distribution of \log_{10} 3'UTR sequence lengths for <i>D. melanogaster</i> mRNA transcripts obtained from release 89 of Ensembl (Zerbino, et al., 2018).	246
Figure 5.8 – A histogram of length normalised positions of predicted miRNA target sites along <i>D. melanogaster</i> 3'UTRs.	248
Figure 5.9 – The mean percentage GC content of <i>D. melanogaster</i> 3'UTRs along the normalised length of the 3'UTR.	249
Figure 5.10 – An examination of the cumulative distributions of 3'UTR sequence length for both male and female fruit fly, grouped according to body type.	250
Figure 5.11 - Empirical cumulative distributions of coding genes which are either upregulated, downregulated or not differentially expressed, with respect to predicted miRNA target site frequency.	252
Figure 5.12 - Comparison: Mated female head/thorax vs. virgin female head/thorax – otherwise, as in figure 5.11.	253
Figure 5.13 - Empirical cumulative distributions of coding genes which are either upregulated, downregulated or not differentially expressed, with respect to 3'UTR length.	254
Figure 5.14 - Comparison: Mated female head/thorax vs. virgin female head/thorax.	255
Figure 5.15 - Empirical cumulative distributions of coding genes which are either upregulated, downregulated or not differentially expressed, with respect to predicted target site frequency of randomly generated miRNA seed sequences.	256
Figure 5.16 - Comparison: Mated female head/thorax vs. virgin female head/thorax.	257
Figure 5.17 - Empirical cumulative distributions of the predicted target and non-targets of dme-miR-14-3p with respect to the \log_2 mRNA fold change. dme-miR-14-3p was chosen as a miRNA exhibiting typical behaviour of a differentially expressed miRNA in this comparison.	263
Figure 6.1 - Dendrogram shows hierarchical clustering of RNA-seq derived gene abundance data using the Jensen-Shannon distance.	280

Figure 6.2 - Principal components analysis of the normalised miRNA read counts derived from sRNA sequencing of <i>L. calcarifer</i> gonadal tissue.	281
Figure 6.3 - miRNA transcript abundances recorded for (A) testis to T1/T2 comparison (B) T1/T2 to T3/T4 comparison (C) T3/T4 to ovary comparison (D) testis to T3/T4 (E) T1/T2 to ovary (F) testis to ovary comparison.....	286
Figure 6.4 - mRNA transcript abundances recorded for (A) testis to T1/T2 comparison (B) T1/T2 to T3/T4 comparison (C) T3/T4 to ovary comparison (D) testis to T3/T4 (E) T1/T2 to ovary (F) testis to ovary comparison.....	287
Figure 6.5 - A cumulative plot of the differential expression of predicted targets and predicted non- targets of miR-19a-3p when considering the testis to T1/T2 comparison.....	289
Figure A.1 – The total number of reads, as well as the number of aligned and pseudoaligned reads for all datasets analysed as part of chapter 4 of this thesis.	331
Figure A.2 – Preliminary analyses conducted on datasets, which were judged unsuitable for further analysis, due to extensive similarity between predicted target and non-target distributions, indicating a potential failure in transfection experiments used to generate the data.	332
Figure A.3 – Volcano plots for all RNA-Seq transfection datasets analysed as part of chapter 4 of this thesis.	339
Figure A.4 - As in figure 4.1, though with a greater number of datasets analysed.....	342
Figure A.5 – As in figure 4.5, though with a greater number of datasets analysed.....	343
Figure A.6 – As in figure 4.9, though with more datasets analysed.....	346
Figure B.1 – mRNA sequencing depth:.....	352
Figure B.2 – sRNA sequencing depth:	353
Figure B.3 – A volcano plot for protein-coding genes for the differential expression analysis presented in chapter 5 of this thesis.....	356
Figure B.4 – A volcano plot for miRNA for the differential expression analysis presented in chapter 5 of this thesis.....	357
Figure B.5 – A decomposition of the variance observed between biological replicates for protein-coding genes for the analysis presented in chapter 5 of this thesis.....	358

Figure B.6 – An MA plot for the <i>D. melanogaster</i> protein-coding genes. On the x-axis – the mean relative gene abundance across all replicates.	359
Figure B.8 – A comparison of the differential expression effect size with the uncertainty of the effect size estimate.	360
Figure B.9 – An MA plot for miRNA differential expression for the study presented in chapter 5.	361
Figure B.10 – The relationship between the log₂ fold change and its standard error for the differential expression analysis study presented in chapter 5 of this thesis.	362
Figure B.11 – A comparison of the cumulative target site frequency distribution of downregulated, non-differentially expressed, and upregulated genes in the female abdomen when comparing mated and virgin fruit flies.	363
Figure C.1 – Volcano plot for miRNA for the differential expression analysis presented in chapter 6 of this thesis.	373
Figure C.2 – volcano plot for protein-coding genes for the differential expression analysis presented in chapter 6 of this thesis.	374

Acknowledgements

Firstly, I would like to thank Simon Moxon for giving me the opportunity to study for a PhD at the Norwich Research Park, for being my primary supervisor for four years, and for always being willing to answer any questions and queries I may have had throughout this process as well as provide helpful feedback on submitted work.

I would also like to thank all other members of my supervisory team: I would like to thank Tamas Dalmay for allowing me to attend his group's lab meetings for four years, and through the presentations of lab members and subsequent discussions, gain a deeper understanding of molecular biology. I would like to thank Rob Davey for providing helpful comments during interim and annual review meetings, and for his group's management of CyVerse UK, the infrastructure allowing me to host one of the applications I have developed during this PhD. I would like to thank Daniel Mapleson for providing helpful comments and suggestions during the earlier stages of my PhD. I would also like to thank Andrea Münsterberg for chairing my probationary review meeting at the end of the first year of my PhD. I am also grateful that both Wilfried Haerty and Anton Enright agreed to examine my thesis. The quality of this thesis has been greatly improved as a result of their feedback.

I am also grateful for the opportunity I have had to learn from other students and researchers during the course of my studies. In particular, I would like to thank my brother David Bradley, as well as James Walker, Joshua Thody, Rocky Payet, Martina Billmeier, Claudia Paicu,

Daniele Braga, Leighton Folkes and Luca Penso-Dolfi, from whose work and discussions I have developed a greater understanding of either small RNA biology, computational biology or molecular biology as a whole. In addition, I would also like to thank Dagnė Daškevičiūtė, an undergraduate project student that I supported, who, through her enthusiasm to learn, allowed me to develop my mentoring and supervision skills.

I would also like to thank all of my collaborators that I have worked with, on whose work I have built upon during the course of my PhD. Without it, a lot of the work presented in this thesis would not have been possible.

I have also had the opportunity during my PhD to take part in projects and activities outside of my core studies and research: I would like to thank Lisa Crossman, who supervised me during an internship at SequenceAnalysis.co.uk during the course of my PhD. During this internship, I developed computational skills which would aid me through the remainder of my PhD. I would like to thank Earlham Institute's scientific outreach and communications team for allowing me to participate in public engagement opportunities during the earlier stages of my PhD. I would also like to thank Ben Miller, Ping Xu, Earlham Institute students, and members of the Münsterberg and Wheeler labs who I learned from whilst participating in journal clubs during the course of my PhD. I would also like to thank all those I worked with, and learned from whilst working as part of the Earlham Institute Student Committee.

I have benefitted from a large degree of institutional support through the course of my PhD which I am grateful for. Primarily, I would like to thank my funders, the British Biological Sciences Research Council (BBSRC), for giving me this opportunity and supporting me through the entirety of this PhD. I would like to thank the Earlham Institute for giving me the opportunity to study in Norwich, and hosting me during the earlier stages of my PhD. I would like to thank the University of East Anglia, in particular, the School of Biological Sciences for hosting me for most of my time as a PhD student. I would like to thank the Graduate Research Offices based at the NBI and UEA for providing administrative support for my studies throughout the course of my PhD. The high-performance computing team at UEA, the computing infrastructure support (CiS) group at NBI, the scientific computing group at EI, and the NBI computing group have all been essential for providing me with the necessary cyberinfrastructure and computing support to allow me to complete my studies. In addition, I would also like to thank Anthony Etuk and particularly Alice Minotto for infrastructure support in relation to setting up the FilTarDB web application. Beyond this, I am conscious of a network of administrative, managerial, catering, facilities and estates staff working at UEA and the Earlham Institute, which make research within these institutes possible, and for which I am grateful.

I would like to thank all my past teachers, lecturers, supervisors, mentors and educators who I learned from even before beginning the PhD process. Without them, even beginning this PhD may not have been possible. There are too many to name individually.

Finally, I would like to thank my family for their support during these four years. In particular, I would like to thank my mother, Jacqueline, and my brother, David, for their continued help and support.

Declaration

I declare that this thesis, in part or full, has not been submitted in any application for another degree or qualification at this university or any other institute of learning.

Some portions of this thesis form the part of the following jointly-authored publications:

1. Thomas Bradley, Simon Moxon, FilTar: using RNA-Seq data to improve microRNA target prediction accuracy in animals, *Bioinformatics*, Volume 36, Issue 8, 15 April 2020, Pages 2410–2416, <https://doi.org/10.1093/bioinformatics/btaa007>
2. Fowler, E.K., Bradley, T., Moxon, S. *et al.* Divergence in Transcriptional and Regulatory Responses to Mating in Male and Female Fruitflies. *Sci Rep* **9**, 16100 (2019). <https://doi.org/10.1038/s41598-019-51141-9>

The first publication relates to research presented in chapters 3 and 4 of this thesis. For this publication, I was responsible for software development, data analysis and some aspects of project design. Dr. Simon Moxon was also responsible for some aspects of project design for this publication.

The second publication relates to research presented in chapter 5 of this thesis. For this publication, I was responsible for an integrated analysis

of the mRNA and miRNA expression, and differential expression results, the subsequent miRNA targeting and network analysis as well as helping to write and edit the final manuscript.

I certify that this thesis, and the research to which it refers, are the product of my own work, and that the work of other people included in this thesis, published or otherwise, is fully acknowledged.

Outcomes

Publications

Bradley, T., & Moxon, S. (2017). An assessment of the next generation of animal miRNA target prediction algorithms. In *MicroRNA detection and target identification* (pp. 175-191). Humana Press, New York, NY.

Fowler, E.K., Bradley, T., Moxon, S. *et al.* Divergence in Transcriptional and Regulatory Responses to Mating in Male and Female Fruitflies. *Sci Rep* **9**, 16100 (2019). <https://doi.org/10.1038/s41598-019-51141-9>

Thomas Bradley, Simon Moxon, FilTar: using RNA-Seq data to improve microRNA target prediction accuracy in animals, *Bioinformatics*, Volume 36, Issue 8, 15 April 2020, Pages 2410–2416, <https://doi.org/10.1093/bioinformatics/btaa007>

Software Developed

FilTar (command-line application): <https://github.com/TBradley27/FilTar>

Documentation: <https://tbradley27.github.io/FilTar/>

FilTarDB (database and web application): filtar.db.earlham.ac.uk

Source code: <http://filtar.db.earlham.ac.uk>

filtrar_R (library used by FilTar): https://github.com/TBradley27/filtrar_R

Chapter 1: Introduction

As miRNAs play a key role in the regulation of a wide range of developmental processes in animal species, it is important to understand how miRNAs act within cells in order to enact developmental change. To achieve this goal, we first must come to an understanding of how miRNAs target mRNA molecules in the cell, and next how miRNAs effect developmental change by acting concurrently on an ensemble of targeted, cellular mRNA transcripts. In order to examine miRNA regulatory activity, it is helpful to gauge relative mRNA abundance levels within a cell for different experimental conditions or developmental time points. Bulk RNA sequencing technologies provide a method of determining mRNA expression levels across the combined transcriptomes of a large number of cells. This enables researchers to examine the effect of miRNA perturbation on a potentially large number of transcripts, making it a particularly useful tool for achieving the aims described above. In this thesis, I demonstrate how data from RNA-Seq experiments can be used to increase the accuracy of miRNA target prediction, and also how this data can be used in combination with small RNA sequencing data in order to determine the key regulators of developmental processes amongst a list of differentially expressed miRNAs.

In the **second chapter**, background information is provided which will aid understanding of the remaining contents of this thesis. I examine the available literature in order to provide a more thorough description of the role of miRNAs within the cell, their evolution within diverse eukaryotic lineages, their biogenesis, as well as their relationship to other cellular small RNAs, allowing a greater understanding of miRNAs

within the context of general RNA and cellular biology. Important methodological research is also described including experimental and computational methods for predicting miRNA targets, as well as an assessment of the accuracy, utility and limitations of these published methods. In addition, processes for the evaluation of miRNA and mRNA expression levels using RNA sequencing data, as well as methods used to annotate these classes of RNA molecules, with a focus on the annotation of the 3'UTRs of mRNA molecules, are discussed.

In the **third chapter**, I detail the design, development, implementation and performance metrics of two different software applications relating to the identification of putative miRNA targets in animal species. With *FilTar*, I have released a dedicated command-line application which acts as a configurable animal miRNA target prediction workflow using previously released target prediction algorithms, as well as the implementation of additional pre-processing (*i.e.* 3'UTR reannotation) and post-processing (*i.e.* expression filtering) steps in order to improve prediction accuracy. To complement *FilTar*, I also have released a web application, *FilTarDB*, which provides the user with a graphical user interface in order to allow them to interrogate a database of results generated using the *FilTar* workflow.

In the **fourth chapter**, I assess the biological validity of using the *FilTar* approach for identifying putative miRNA targets. In particular, I evaluate the effects of both expression filtering and 3'UTR reannotation on target prediction accuracy by observing the effects of implementing these steps on analyses of data deriving from miRNA perturbation experiments. In these experiments, miRNA mimics are transfected into

cell cultures which subsequently undergo RNA sequencing. RNA sequencing is also performed on mock transfected cell cultures. By performing a differential expression analysis on data deriving from these two types of transfected cell cultures, the accuracy of different miRNA target prediction methods can be inferred.

In the **fifth chapter**, using the specific biological context of transcriptomic post-mating responses in *Drosophila melanogaster* (common fruit fly), I examine how RNA-Seq data can be used to not only to help infer direct miRNA-mRNA interactions, but also to help determine the regulatory activity and effectiveness of differentially expressed miRNAs for a given developmental process. This is achieved by examining the degree of repression of all predicted mRNA targets of a differentially expressed miRNA in comparison to that of predicted non-targets of the same miRNA.

In **chapter six**, I perform a similar, but distinct analysis, though this time in the context of the naturally occurring sex transition developmental process known to occur in *Lates calcarifer* (Asian seabass). In doing so, I highlight the utility of using combined mRNA-Seq and sRNA-Seq data to help infer key miRNA regulators for a diversity of developmental processes.

Finally, in **chapter seven**, I conclude with a summary of the contents and main findings reported in this thesis, and also discuss future work which would be hoped to both extend the software applications developed as part of this thesis, and also advance knowledge and understanding in this general research area.

Chapter 2: Background

2.1 Overview

miRNAs are short RNA molecules of approximately 22nt in length which guide the repression of other RNA molecules. They predominantly arise from a primary pol II-transcribed RNA transcript, which are processed to produce ~70nt stem-loop precursor miRNA structures. The miRNA precursor is then cleaved to produce the canonical ~22nt miRNA duplex, one strand of which is loaded into the argonaute effector protein (Bartel, 2018).

In bilaterian animals, nucleotides 2-7 of the miRNA from the 5' end determines the specificity of the miRNA-argonaute complex, and guides this complex to corresponding recognition elements on other RNA molecules, leading to the repression of these RNA targets (Bartel, 2018).

The primary function of miRNAs within the cell is to provide a post-transcriptional layer of gene expression regulation, including, in particular the regulation of protein-coding genes, and sometimes within the context of intricate interaction networks between coding and non-coding RNAs.

miRNA pathways can best be understood as one of three parallel arms, along with piwi-interacting RNAs (piRNAs) and short-interfering RNAs (siRNAs) of a general RNA interference (RNAi) pathway

(Almeida, et al., 2019). Therefore, in order to understand the specific biology of miRNAs within cells, it is appropriate to compare and contrast their biogenesis, structure and function with that of both piRNAs and siRNAs.

As piRNAs and endogenous siRNAs can be described as having broad roles in genomic defence (Billi, et al., 2012), the specific role of miRNAs within the milieu of RNA-mediated control and regulation can be described as the post-transcriptional regulation of the expression of canonical, non-invading (*i.e.* ‘self’) genes of the genome, which has broad roles in development and general cellular homeostasis. This is in contrast to the role of piRNAs in regulating transposon expression and the generally unclear and heterogeneous role of siRNAs in animals. Figure 2.1 summarises some of the known commonalities and differences between these three classes of small RNA (Bartel, 2018; Okamura and Lai, 2008; Weick and Miska, 2014):

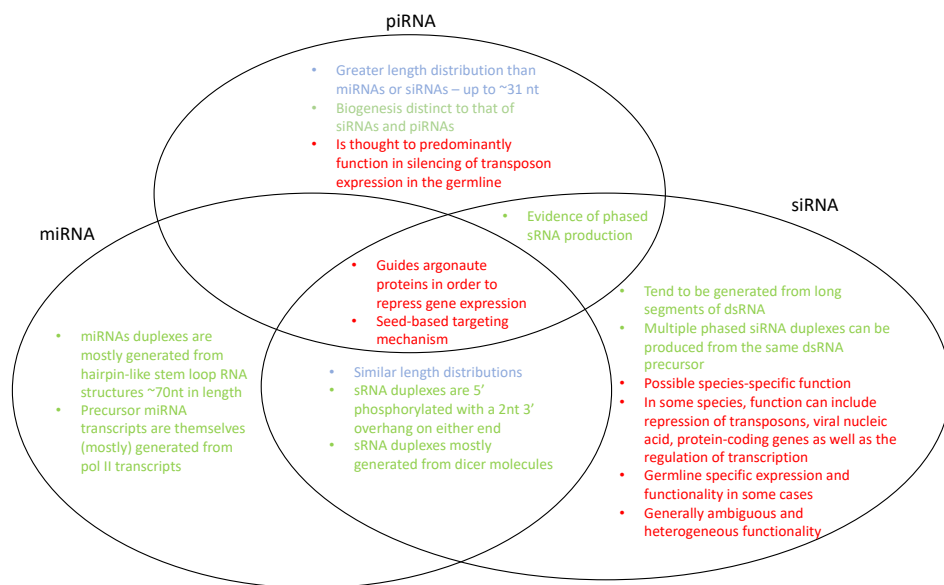


Figure 2.1 - A comparison of structural (blue text), biogenesis (green text) and functional (red text) differences and similarities between three different classes

of small RNA involved in cellular gene silencing mechanisms: miRNAs, piRNAs and siRNAs. All three classes of sRNA can be distinguished in terms of their unique biogenesis. There are many structural similarities between these sRNA classes, and whilst miRNAs and piRNAs are functionally distinct – the function of endogenous siRNAs seem to be heterogeneous and to a degree, unknown.

2.2 Evolution of miRNAs

miRNAs are found in a large and diverse number of basal eukaryotic lineages (Moran, et al., 2017). The origin of miRNAs cannot be traced to a single monophyletic group within the domain *Eukarya*, and basal lineages within this domain which possess miRNAs, such as the plant and animal kingdoms, lack miRNA sequence homology with each other (Bartel, 2004; Kozomara and Griffiths-Jones, 2010) suggesting the independent and thus convergent evolution of miRNAs on multiple occasions through the course of eukaryote evolution (Axtell, et al., 2011; Jones-Rhoades, et al., 2006; Tarver, et al., 2012). A relatively rapid gain and loss of miRNA gene families and sequences within plants and non-bilaterian animals (Moran, et al., 2017) could however suggest divergent evolution with subsequent rapid change in miRNA gene functionality as an alternative explanation for these findings (figure 2.2). The existence of orthologues to the common miRNA protein machinery found in species ancestral to plants and animals supports this hypothesis (Moran, et al., 2017). Despite opposing views, there is a consensus that high quality miRNA annotations for a diverse number of species sampled under diverse conditions are necessary to resolve issues of contention; either by revealing some previously undiscovered miRNA gene homology between species (Moran, et al., 2017), or by identifying *false positive* designations of miRNA gene loss (Tarver, et al., 2018).

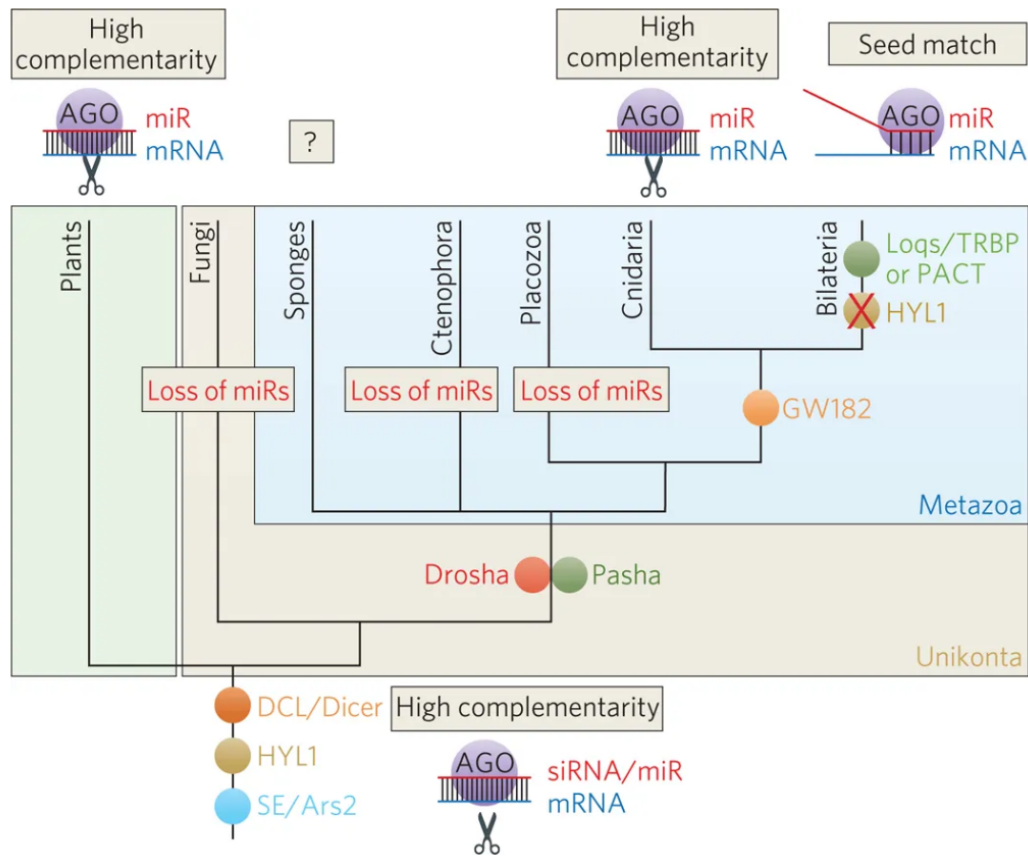


Figure 2.2 - A model for the evolution of miRNA targeting mechanisms in different eukaryotic lineages as proposed in Moran et al. 2017 (Moran, et al., 2017). In this model, all extant miRNA pathways divergently evolved from basal RNA-interference or ‘miRNA-like’ pathways, with subsequent loss of miRNA functionality in some clades. In addition, the seed match miRNA targeting mechanism is proposed to have evolved in bilateria.

Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Nature Ecology & Evolution, The evolutionary origin of plant and animal miRNAs, Moran, Y., Agron, M., Praher, D., & Technau, U., Copyright 2017 (<https://www.nature.com/articles/s41559-016-0027>).

At the molecular level, there are a number of mechanisms by which new miRNAs can evolve. Most simply, paralogous miRNAs can be formed from the duplication of existing miRNAs with subsequent nucleotide

substitution (Nozawa, et al., 2010). New miRNA genes can also be formed from the inverted duplication of existing genetic elements, such as protein-coding genes and transposable elements (Smalheiser and Torvik, 2005). Other known mechanism of miRNA formation are from introns of protein-coding genes, or from the formation of miRNAs from random sequences with a substantial degree of self-complementarity (De Felippes, et al., 2008). As regulatory elements, miRNAs only exist in relation to their targets, and perhaps unsurprisingly, a large degree of coevolution between miRNAs and their targets have been observed: Not only do predicted target sites generally exhibit greater conservation than surrounding regions (Lewis, et al., 2005), but the predicted targets of conserved miRNA seed families exhibit greater conservation than those of species-specific miRNA seed families (Penso-Dolfin, et al., 2018). In addition, another study found that in general, once evolved, that miRNA target site loss is generally disfavoured, and even mutations that alter the strength of existing target sites are also disfavoured, indicating the functional importance of not only the existence of miRNA target sites but the precise magnitude of their effect on the transcriptome (Simkin, et al., 2019).

2.3 Biogenesis and crosstalk with RNAi pathways

The biogenesis of miRNAs exists as one entry point, amongst others, of the broader regulatory pathway termed *RNA interference* (RNAi) (Fire, et al., 1998; Hannon, 2002; Sharp, 2001). RNAi involves the loading of double-stranded RNA (dsRNA) into the RNA-induced silencing complex (RISC), followed by a selective degradation of one of the dsRNA strands (termed the *passenger strand*) and retention of the

remaining strand (*i.e. the guide strand*), which is subsequently used to guide RISC to RNA molecules with complementarity to the guide RNA (gRNA) for subsequent repression (Hannon, 2002).

Before the initial discovery in 1998 by the labs of Andrew Fire and Craig Mello of the specific potency of dsRNA in directing RNA interference as opposed to the relatively weak effects of single-stranded RNA (ssRNA) (Fire, et al., 1998), it had long been known that both sense (Guo and Kemphues, 1995) and antisense (Fire, et al., 1991; Guo and Kemphues, 1995; Izant and Weintraub, 1984) ssRNA could lead to the repression of cognate RNA molecules. The mechanism by which the dsRNA led to the repression of ssRNA molecules, after these initial discoveries was unclear. In parallel to work completed in animal model organisms such as *Drosophila melanogaster* (common fruit fly) and *Caenorhabditis elegans* (a species of nematode worm), was work relating to the silencing of viral genes and transgenes in plants species: Crucially, in the lab of David Baulcombe, it was discovered that ~25 nucleotide antisense RNA molecules accumulated in samples in which there had been post-transcriptional gene silencing of a transgene and viral RNA (Hamilton and Baulcombe, 1999). In the study, it was hypothesised that ‘...the 25-nucleotide antisense RNA is likely synthesized from an RNA template...’ which was later experimentally confirmed for some cases when it was discovered that a plant RNA-dependent RNA polymerase (RdRP) was necessary for transgene silencing but not for the silencing of RNA belonging to a virus (which encodes its own RNA polymerase) (Dalmay, et al., 2000). Transgene silencing was also known to occur in cases of inverted repeated transgenes (Stam, et al., 1997), and in the case of sense and antisense

transgene co-expression (Waterhouse, et al., 1998), which was hypothesised to lead to dsRNA formation by an RdRP-independent mechanism (Dalmay, et al., 2000). These discoveries together linked the processes of transcription, dsRNA formation, post-transcriptional gene silencing, and the accumulation of small RNA antisense to targeted transcripts.

Further work was conducted in order to elucidate a mechanism for RNAi. Gregory Hannon and colleagues discovered RISC when they had taken extracts from *Drosophila* cells transfected with dsRNA *in vivo*, and co-fractionated a nuclease associated with a ~25nt RNA which degraded antisense transcripts (Hammond, et al., 2000). Additional work in the Hannon lab led to the discovery that an RNase III enzyme named *dicer*, possessing both helicase and endoribonuclease domains, could produce short, approximately ~22nt guide RNA from dsRNA, which is subsequently associated with RISC (Bernstein, et al., 2001). In addition, it was discovered that it is an ATP-dependent unwinding of the dsRNA into ssRNA which is necessary for the formation of the active RISC complex (Nykänen, et al., 2001). Further work revealed that the argonaute-2 protein (AGO2) is an essential component of functional RISC complexes assembled as a response to dsRNA transfection in *D. melanogaster* (Hammond, et al., 2001), which was later discovered to derive from AGO2's role in the endonucleolytic cleavage of the target molecule (Liu, et al., 2004; Meister, et al., 2004).

As mentioned previously, miRNA biogenesis serves as a one entry point into the more general RNA interference pathway. Experiments described so far elucidating the RNAi pathway mostly involved the transfection of exogenous RNA, or otherwise the viral infection or

transformation of plants by foreign genetic material. The discovery of miRNAs revealed one particular pathway by which RNA interference is activated by endogenous RNA.

The discovery of miRNAs occurred in parallel to that of RNA interference, although the relationship between miRNA processing and function and RNA interference was initially unclear or not known. In the labs of Victor Ambros and Gary Ruvkun it was discovered that in *C. elegans*, the *lin-4* small RNA represses the LIN-14 protein, which was hypothesised to result from the antisense complementarity between *lin-4* and the 3'UTR of the *lin-14* mRNA (Lee, et al., 1993; Wightman, et al., 1993). The generality of this proposed targeting mechanism for multiple *C. elegans* genes, and also across many species however was not initially realised. Many years later it was discovered that a second small RNA, *let-7*, with complementarity to the *lin-14* 3'UTR and many other *C. elegans* genes, regulated LIN-14 expression levels (Reinhart, et al., 2000). Crucially, unlike *lin-4*, *let-7* was found to be conserved in a large number of bilaterian species (Pasquinelli, et al., 2000), suggesting the evolution of a highly conserved mechanism for the post-transcriptional regulation of gene expression by endogenously produced small RNA molecules. The term 'heterochronic' was initially used to describe these RNA species, and later 'small temporal RNAs' (Ambros, 2001) to reflect observations that these RNA molecules are expressed at different stages of development, and regulate the transition between developmental states. The term 'microRNA' (Lagos-Quintana, et al., 2001) came to be coined and extensively used for what was now a large number of discovered ~21-24nt RNA molecules which were predicted to derive from larger RNA molecules with hairpin like structures (Lagos-Quintana, et al., 2001; Lau, et al., 2001; Lee and Ambros,

2001), most of which were transcribed from genomic clusters of 3-6 miRNA genes (Lagos-Quintana, et al., 2001).

Studies provided evidence that miRNAs were utilising components of the RNA interference protein machinery in order to regulate developmental timing: Inactivation of dicer and argonaute orthologues in *C. elegans* led to similar phenotypes as observed in *lin-4* and *let-7* mutants (Grishok, et al., 2001), whereas in *D. melanogaster* it was discovered a precursor RNA was cleaved in a dicer-like mechanism to produce *let-7*, and that inactivation of dicer mRNA led to accumulation of the *let-7* precursor in humans (Hutvagner, et al., 2001).

Although the previously described short-hairpin miRNA structures were known to be produced endogenously, at the beginning of the century, their exact biogenesis was still unclear. By 2004, it was discovered that miRNA genes are transcribed by RNA polymerase II (Pol II), producing transcripts which are both 5' capped and polyadenylated at the 3' terminus (Cai, et al., 2004; Lee, et al., 2004), which came to be referred to as the *primary miRNA* (pri-miRNA). Within the pri-miRNA are one or several miRNA hairpin-like structures corresponding to the number of miRNA sequences controlled by a single promoter at the miRNA gene or miRNA gene cluster locus. In animals, the hairpin miRNA structures, termed *precursor miRNAs* (pre-miRNAs) are excised from the primary transcript through the action of the *microprocessor complex*, composed of a dsRNA binding protein called *DGCR8*, and a RNase III enzyme, *droscha* (Denli, et al., 2004; Gregory, et al., 2004), which cleaves the stem-loop structures within the pri-miRNA, generating isolated precursor RNA molecules with a 2nt overhang at

the 3' end. Alternatively, pre-miRNAs can be generated independently of the microprocessor complex if the miRNA hairpin structures are found in intronic regions of a precursor messenger RNA, and lack the basal region of a long stem, which is used for recognition by the microprocessor complex. Termed 'mirtrons', these miRNA sequences are removed during splicing of the mRNA, forming a lariat-like structure which is subsequently debranched, generating a pre-miRNA molecule (Berezikov, et al., 2007; Okamura, et al., 2007; Ruby, et al., 2007). An exportin protein (XPO5) mediates the export of the pre-miRNA from the nucleus to the cytosol (Bohnsack, et al., 2004; Yi, et al., 2003), where it associates with dicer, and hence at this point the miRNA processing pathway merges with the canonical RNA interference pathway. A visual summary of these described processes is given in figure 2.3.

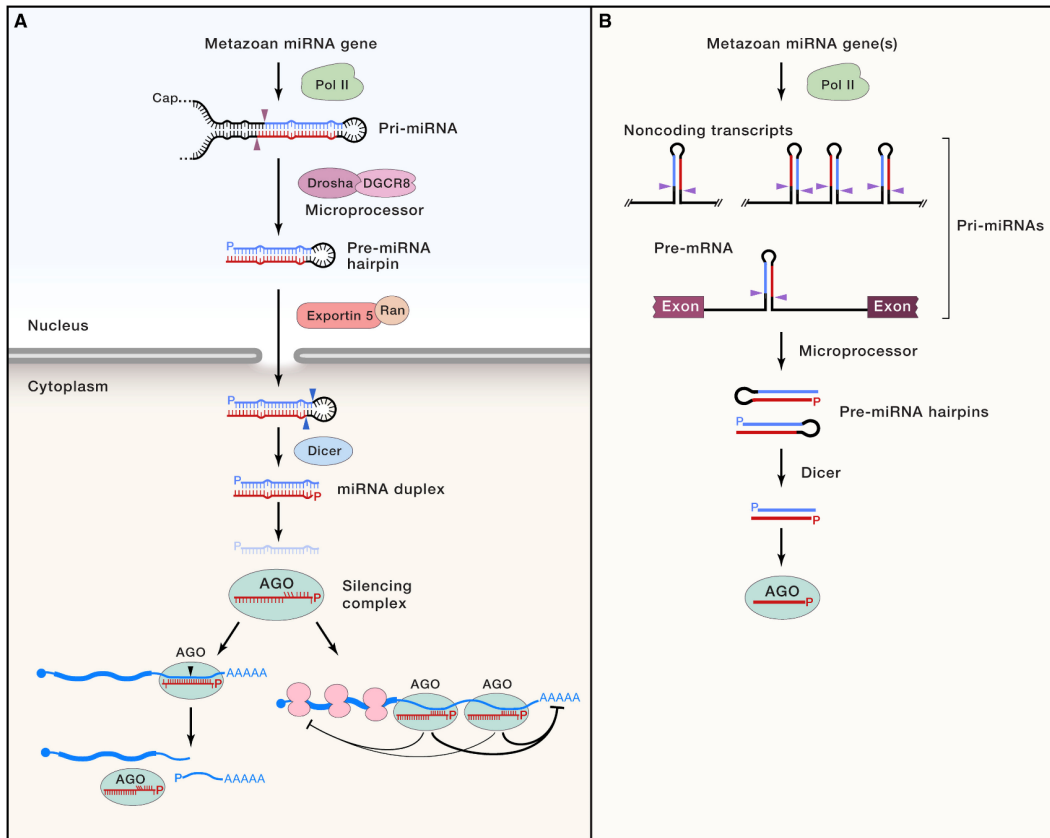


Figure 2.3 - summary of miRNA biogenesis in animals. A) The biogenesis of a typical miRNA: Metazoan miRNA genes are transcribed from RNA polymerase II, generating a primary miRNA transcript (pri-miRNA). The microprocessor complex, which includes drosha and DGCR8, endonucleolytically cleaves both strands of the pri-miRNA, generating an RNA hairpin with an approximately two-nucleotide overhang at the 3' end referred to as the precursor miRNA (pre-miRNA). Exportin 5, in complex with the Ran cofactor is used to facilitate export of the pre-miRNA into the cytoplasm where the 'loop' structure of the RNA hairpin is excised by dicer, generating a miRNA duplex with an ~2nt 3' overhang at either end. Each 5' end of the RNA duplex is phosphorylated. The passenger strand is degraded, whilst the guide strand is loaded into an AGO protein. B) Different transcriptomic sources of pri-miRNAs: pri-miRNAs can derive from either non-coding transcripts, or the introns of precursor mRNA (pre-mRNA) molecules.¹

¹ Reprinted from Cell, 173(1), Bartel DP., Metazoan miRNAs, 20-51., Copyright (2018), with permission from Elsevier.

2.4 The relationship between miRNAs and other classes of metazoan sRNAs

Knowledge of the miRNA biogenesis pathway, and its relationship to other sRNA pathways in the cell is important for distinguishing miRNAs from other classes of sRNA. Some of these other classes of sRNA at some point interact with one or more components of miRNA and RNAi pathways. Small interfering RNA (siRNA) is the name given to sRNAs derived from long segments of perfectly complementary double-stranded RNA (Bernstein, et al., 2001). This structure contrasts with that of the pre-miRNA, which contains a much shorter dsRNA region, and usually contain several symmetric or asymmetric bulges. Common, ancestral roles of siRNAs was as a defence for cells against nucleic acids external to (*e.g.* viral RNA), and within (*e.g.* transposable elements) the cell. In most bilaterian species, these functions have mostly been superseded by the interferon (Okamura and Lai, 2008) and piRNA pathways respectively. Although endogenous siRNA genes are still found in bilaterian species, their roles are generally unclear (Okamura and Lai, 2008). In contrast, piRNAs, found in animals, are between 21-30nt in length, are enriched in germline tissue, and have a clearly identified function in the defence of germline genomes (Weick and Miska, 2014). The name 'piRNA' which is an abbreviation of the term 'piwi-interacting RNA', refers to the piwi proteins which bind piRNAs, and are a subfamily of the argonaute protein family. The piRNA-piwi ribonucleoprotein is involved in the repression of the activity of transposable elements (Höck and Meister, 2008). Small RNAs can also be derived from a class of larger RNA molecules called Y RNAs, which are believed to have roles in the initiation of DNA replication (Christov, et

al., 2006), and the inhibition of Ro60, a protein believed to be involved in processes relating to the quality control of some RNA molecules (Hall, et al., 2013). Small RNAs are derived from Y RNAs during apoptosis, and some of these small RNAs have been shown via CLIP experiments to be in the miRNA size range (Rutjes, et al., 1999) and associated with argonaute proteins (Thomson, et al., 2014). However, some of these sRNAs have been shown to be produced in a dicer-independent manner, and are not associated with AGO2 (Nicolas, et al., 2012), the member of the argonaute protein sub-family which exhibits cleavage activity. These Y RNA derived sRNA also failed to exhibit repression of candidate targets during dual luciferase reporter assays (Meiri, et al., 2010; Thomson, et al., 2014), suggesting together that these sRNAs do not operate in miRNA or RNA interference type pathways. Similarly, small RNAs derived from vault RNAs and small nuclear RNAs whilst shown to be AGO-bound, have not so far exhibited repressive activity (Thomson, et al., 2014). Conversely, there is abundant evidence to suggest that tRNA-derived small RNA fragments (tRF) operate at least partially in miRNA/RNA interference-like pathways: PAR-CLIP data demonstrates complexing of tRFs and human argonaute proteins (Hafner, et al., 2010; Kumar, et al., 2014), and further evidence from CLASH (cross-linking and sequencing of hybrid) data demonstrates a proximity between AGO proteins, tRFs and messenger RNA (Helwak, et al.). Crucially, these small RNAs also repress mRNA targets with complementary sequences in their 3'UTRs, in an argonaute-dependent and dicer-independent manner, potentially in association with GW182 (Kuscu, et al., 2018), a protein of the miRNA-RISC complex necessary for the translation repression and decay of mRNA (Eulalio, et al., 2008). Similarly, there is evidence to suggest that sRNAs derived from small nucleolar RNAs (snoRNAs) and ribosomal RNAs (rRNAs) can act in a

miRNA-like manner at high abundances (Ender, et al., 2008; Thomson, et al., 2014).

2.5 miRNA annotation and database resources

The number of existing sRNA classes, as well as the existence of sRNA molecules with no identifiable functional role in the cell can make it difficult to discern genuine miRNAs from other sRNA molecules in sequencing datasets. This can be a particular problem considering the common use of high-throughput sequencing of small RNAs populations as a method of profiling miRNA expression levels in a given sample, and for annotating novel miRNAs. To resolve this problem, criteria for annotating miRNA have been established (Ambros, et al., 2003), including criteria relating to the expression of the mature miRNA, and relating to the identity of the predicted miRNA precursor molecule from which the mature miRNA molecule derives. Using such criteria or similar criteria, algorithms have been developed to identify miRNAs from deep sequencing datasets (Friedländer, et al., 2008; Friedländer, et al., 2011; Moxon, et al., 2008). As the knowledge of miRNA and their biogenesis has increased with continued research, methods of annotating, and systems of naming miRNAs continue to be proposed (Fromm, et al., 2015), and methods for detecting miRNAs continue to be developed (Mapleson, et al., 2013; Paicu, et al., 2017; Vitsios, et al., 2017), indicating the continued active research in this area.

Data associated with miRNA research is typically deposited in dedicated databases, which are publicly available and accessible by the re-

search community. For example, miRBase (Kozomara, et al., 2018) reports information relating to both mature and precursor miRNA sequences, their genomic co-ordinates, relevant gene or sequence identifiers, experimental evidence supporting miRNA annotations, as well as hyperlinks to online resources relating to published research supporting annotations, and crucially allows user to submit their own sequence data and annotations for possible entry in the database. The initial release of a precursor version of miRBase was as a database and web interface termed the ‘miRNA registry’ (Griffiths-Jones, 2004) hosted by the Wellcome Sanger institute, which predominantly acted as a system to assign names to putatively annotated miRNA sequences prior to publication, but later was developed to include more sequence information, as well as information relating to predicted miRNA targets (Griffiths-Jones, et al., 2006; Griffiths-Jones, et al., 2007). In later iterations of the database, there was an increased emphasis in incorporating miRNAs annotations and data relating to sRNA high-throughput sequencing experiments (Kozomara and Griffiths-Jones, 2010; Kozomara and Griffiths-Jones, 2014), as well as a greater emphasis on trying to report the function of miRNA entries in the database (Kozomara, et al., 2018). In addition, because of the problem of low quality submissions to miRBase, a set of ‘high-confidence’ miRNA annotations within miRBase have been defined (Kozomara and Griffiths-Jones, 2014), which relates to defined criteria based on the number of sequencing reads aligning to precursor hairpins, the thermodynamic stability of predicted hairpin structures, as well as the consistency of aligned reads to known product signatures of drosha/dicer processing.

As miRBase is a database which does not generally manually curate miRNA annotations deriving from next-generation sequencing datasets, the contamination of the database with erroneously annotated miRNAs is a known issue (Fromm, et al., 2015; Ludwig, et al., 2017). Sources of noise in miRBase are RNA degradation products misannotated as miRNAs (Ludwig, et al., 2017), poor discrimination between miRNAs and other sRNA classes (*e.g.* endogenous siRNAs), a failure to provide evidence of pre-miRNA processing by dicer, a failure to employ phylogenetic approaches when assessing the validity of candidate miRNA loci – as well as the incorrect and inconsistent naming of miRNA loci (Taylor, et al., 2017).

In an attempt to mitigate against this problem, miRBase have introduced a set of ‘high-confidence’ miRNA annotations within miRBase (Griffiths-Jones, et al., 2006), which relates to defined criteria based on the number of sequencing reads aligning to precursor hairpins and the pattern of alignment of reads to the hairpin. For example, the expectation would be, that a proportionately larger number of reads would align to the locus corresponding to the canonical miRNA, rather than the miRNA* (*i.e.* the miRNA found on the passenger strand of the miRNA precursor). The thermodynamic stability of predicted hairpin structures is also a factor, as well as the consistency of aligned reads to known product signatures of drosha/dicer processing, such as a 2nt 3’ overhang on the precursor miRNA as a result of processing by the microprocessor complex, as well as a 2nt 3’ overhang on either end of the mature miRNA duplex after dicer processing. An additional strategy to validate miRNA annotations found in miRBase, is to, where possible, download the original high-throughput sRNA sequencing data, and to run on this data high-quality miRNA prediction annotation algorithms (*e.g.*

(Friedländer, et al., 2011; Paicu, et al., 2017)), in order to determine whether reported miRNAs can be recapitulated using these algorithms, especially if executed using stringent parameters. Manual curation of the output of miRNA prediction algorithms using rule-based approaches is also one strategy to increase the stringency of miRNA annotations and to reduce noise (*e.g.* (Penso-Dolfin, et al., 2016)).

Complementing miRBase, is the mirGeneDB database, which emphasises strict evolutionarily informed miRNA naming and nomenclature systems, and also strict manual curation of putative miRNA sequences even if they have been designated as miRNAs in peer-reviewed publications (Fromm, et al., 2015; Fromm, et al., 2019). As well as dedicated miRNA databases, miRNA information can also be found in more general online data stores, such as RNA-specific databases and web interfaces such as rfam (Kalvari, et al., 2017; Kalvari, et al., 2018) and RNACentral (2018), allowing miRNA data to be accessed, viewed and understood within the broader context of RNA biology.

2.6 Roles of miRNAs in the cell

An understanding of the existence and function of different classes of non-coding small RNA within the cell is helpful for coming to an understanding of the particular role of miRNAs. Perhaps with the exception of tRFs, evidence suggests that none or only a relatively small number of small RNAs derived from other classes of ncRNA operate within miRNA or RNA interference pathways, which suggests marginally or non-overlapping molecular functions between these small ncRNA classes and miRNAs. Conversely, as discussed, there is pronounced overlap between the miRNA and RNA interference pathways. The ancestral, and in some eukaryotic lineages, extant function of siRNAs, is to protect the cell against foreign genetic material deriving from viruses or transposable elements. The suggestion within the literature is that miRNAs evolved as an exaptation of the pre-existing siRNA gene regulatory mechanism, in order to selectively regulate non-transposable element host genes, in effect creating a post-transcriptional layer of gene regulation (Moran, et al., 2017). In addition, miRNAs have been found to affect protein levels not only by mRNA destabilisation and decay, but also by a process of directly inhibiting translational initiation (Pillai, et al., 2005).

This raises the question of the function and necessity of post-transcriptional and translational regulation of gene expression levels given pre-existing transcriptional and post-translational gene regulatory mechanisms. Presumably, the most efficient method of regulation is at the level of transcription, to minimise the energetic cost of producing unwanted mRNA and proteins. However, regulating gene expression at

this level cannot alter pre-existing mRNA and protein. By clearing pre-existing mRNAs, the miRNA can influence or ‘reprogramme’ cell fate and identity (Guan, et al., 2013; Pauli, et al., 2011). There is abundant evidence for this role of miRNAs for many processes during embryogenesis and general organism development: miR-430 is instrumental in first arresting the translation of maternal transcripts (Bazzini, et al., 2012), before a process of mRNA degradation (Giraldez, et al., 2006) during the maternal-to-zygotic transition. In the transition of human embryonic stem cells to a state of pluripotency, miR-145 has been shown to target several pluripotency associated transcription factors, and to be sufficient in inhibiting the self-renewal of these stem cells (Xu, et al., 2009). miR-430 as well as performing functions in the zygote, also targets components of the nodal signalling pathway in order to promote the formation of endoderm and mesoderm during germ layer specification in early embryogenesis (Choi, et al., 2007). In addition, there is evidence that miR-21 represses tumour repressor genes such as *PTEN* and *PDCD4* in order to promote a mesenchymal cell fate with its associated motile and invasive cell properties (Asangani, et al., 2008; Frankel, et al., 2008; Pauli, et al., 2011).

Not only do miRNAs have an essential role in the transition between cell identities, but they also have an essential role in the maintenance of a particular cellular state. Once a set of mRNAs associated with a particular cell state has been cleared from the cytosol, it is necessary to maintain low abundances of these messengers, in order to maintain the existing cell state. This may be necessary because of the inherently noisy nature of eukaryotic gene expression (Blake, et al., 2003; Thattai and Van Oudenaarden, 2001). This perspective corroborates the work of Oudenaarden and colleagues demonstrating that miRNA can highly

repress genes which are not transcribed at rates above a threshold (Mukherji, et al., 2011), and reduce the variability in protein expression of lowly expressed genes (Schmiedel, et al., 2015). Conversely, once transcription rates exceed a given threshold, the function of the miRNA is proposed to convert from that of a ‘switch’ of gene expression to a ‘fine-tuner’ in which the miRNAs buffers the gradual increase of partially complementary mRNA target levels in the cytosol (Mukherji, et al., 2011). As a result of this type of activity, miRNAs have been theorised as being heavily involved in the canalisation of animal developmental processes *i.e.* ‘genetic buffering that has evolved under natural selection in order to stabilise the phenotype and decreases its variability’ (Hornstein and Shomron, 2006). Indeed, miRNAs not only act in relation to developmentally pre-programmed events, but also in relation to external stressors (Ambros, 2003), implicating miRNAs in more general homeostatic mechanisms for the regulation of gene expression.

Together this evidence supports the perspective that the role of miRNAs is to provide the post-transcriptional and translational components of regulatory networks which drive the transition between, and maintain different cell states (Chakraborty, et al., 2019). In favour of this view, many authors have identified a corresponding relationship between miRNA evolution, multicellularity and general organismal complexity (Bartel, 2004; Grimson, et al., 2008; Peterson, et al., 2009; Tarver, et al., 2015). Arguments against this view, are that miRNAs are found in some unicellular species, whilst some multicellular organisms do not possess miRNAs. However, this could rather demonstrate the necessity but not the sufficiency of miRNAs for the listed features, and secondly, multicellular organisms not containing miRNAs may substitute the

functionality of these molecules for miRNA-like classes of RNA such as endogenous siRNAs (Calcino, et al., 2018; Lee, et al., 2010).

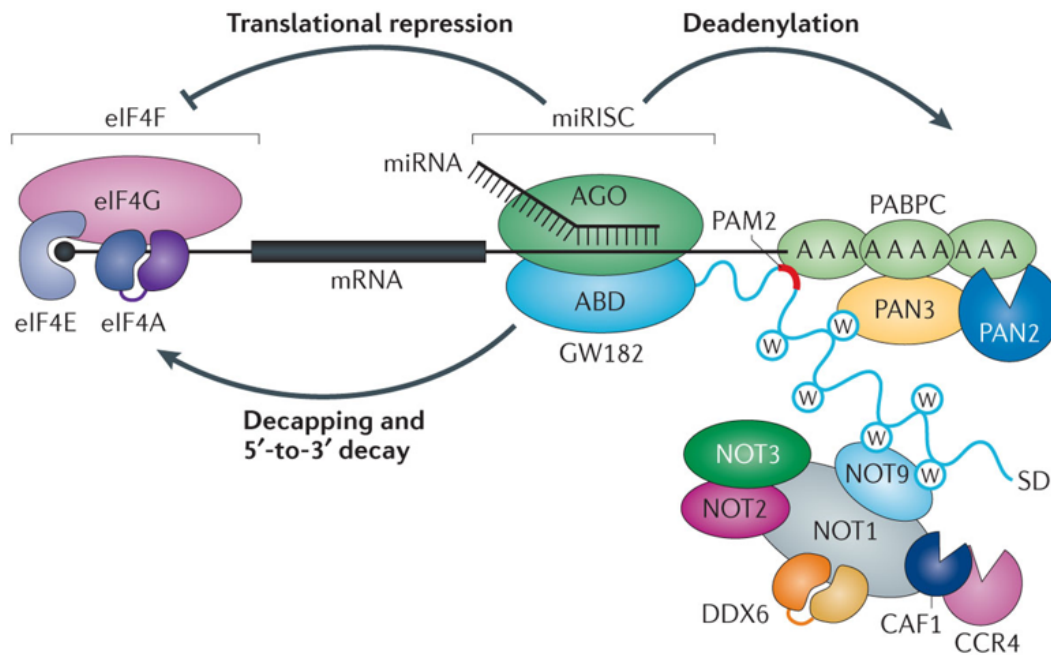
2.7 Molecular mechanisms of miRNA-mediated target repression

The principles of miRNA targeting in bilaterian species is distinct from that of miRNA targeting in plants and more basal (*i.e.* earlier branching) eukaryotic lineages, and also the targeting principles observed during RNA interference. In the latter cases, targeting of the siRNA bound RISC complex to single stranded RNA molecules, and subsequent cleavage of the target molecule requires perfect or near perfect complementarity between the guide and target RNA molecules (Allen, et al., 2005; Martinez, et al., 2002; Schwab, et al., 2005). In contrast, in animals, hybridisation between the 5' *seed* region, typically at nucleotides 2-7 of the miRNA, can be sufficient to induce degradation or direct translational inhibition of the miRNA target (Lewis, et al., 2003).

There is however overlap between siRNA and miRNA targeting principles in animals: RISC complexes bound with siRNA can repress partially complementary target mRNA molecules, accounting for some of the widely observed 'off-target' effects in RNA interference (Doench, et al., 2003; Jackson, et al., 2006). In addition, miRNAs can cause cleavage of targets when there is extended complementarity between the miRNA and its target, and the miRNA is loaded specifically into paralogues of argonaute with endoribonuclease functionality, such as AGO2 in humans (Meister, et al., 2004; Yekta, et al., 2004). This evidence suggests that once the miRNA or siRNA duplex is loaded into

argonaute, and the passenger strand degraded, then the miRNA-RISC complex is agnostic to the origins of the sRNA molecule which it is using as its guide.

Despite the endonuclease activity of some AGO2-bound miRNAs, the predominant method by which bilaterian miRNAs destabilise RNA targets, is through partial complementary targeting, leading to deadenylation, 5' decapping and exonucleolytic digestion of the target (Huntzinger and Izaurralde, 2011). These post-hybridisation events are mediated by the protein GW182, which interacts directly with argonaute (Eulalio, et al., 2008), and recruits downstream effector proteins (Braun, et al., 2011). Whilst the N-terminal domain of GW182 interacts with AGO, the C-terminal domain interacts with PABPC, a poly-A binding protein (Huntzinger, et al., 2010), and the PAN2-PAN3 and CCR4-NOT deadenylation complexes (Braun, et al., 2011). Deadenylation is coupled to 5' decapping through a DDX6 mediator (Chen, et al., 2014; Mathys, et al., 2014) triggering the DCP2 decapping protein (Rehwinkel, et al., 2005). The uncapped transcript is subsequently degraded in the 5'-3' direction by the XRN1 exoribonuclease (Huntzinger and Izaurralde, 2011). Many of the molecular components identified in this form of miRNA-mediated repression have been found to be localised in subcellular structures, phase-separated from the remaining cytosol, called *p-bodies* (Kulkarni, et al., 2010). However, miRNA-mediated gene silencing has been observed to occur in cells lacking detectable p-bodies, suggesting that these structures are not necessary for general miRNA activity, and may instead exist as a result of miRNA silencing activity (Eulalio, et al., 2007).



Nature Reviews | **Genetics**

Figure 2.4 - A summary of miRNA-mediated destabilisation and translation repression of mRNA molecules. The mRNA-bound miRISC complex associates with the GW182 protein via the AGO-binding domain (ABD) of GW182. The silencing domain (SD) of GW182 contains a PAM2 (poly-A binding protein interacting motif 2) motif and a tandem of tryptophan motifs which binds with PABPC (cytoplasmic poly-A binding protein), the PAN2 and PAN3 deadenylase complexes, and also the CCR4-NOT complex which also catalyses the deadenylation of the mRNA target. The PAN2-PAN3 complex is thought to catalyse the first stage of deadenylation, which is then continued by the action of the CCR4-NOT complex. 5' decapping is facilitated by DCP2 (decapping protein 2), a process stimulated by multiple DCP and EDC (enhancer of decapping) proteins as well as DEAD box protein (DDX6). After decapping, 5' to 3' degradation of that target is catalysed by an exonuclease (XRN1) (not shown). miRNA-mediated translation repression is thought to involve the eIF4F eukaryotic translation initiation protein complex. This complex contains proteins related to cap-binding (eIF4E), protein scaffolding (eIF4G) and RNA helicase (eIF4A) activity.

Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Nature Reviews Genetics, Towards a molecular understanding of

miRNA-mediated gene silencing, Jonas, S. Izaurralde, E., Copyright 2015 (<https://www.nature.com/articles/nrg3965>).

As well as destabilisation of the RNA molecule, miRNAs can mediate gene expression by direct translational repression of a transcript independent of RNA decay mechanisms resulting in a decrease in protein production, but stable mRNA abundance levels (Huntzinger and Izaurralde, 2011). Although the precise mechanism is still somewhat unclear, repression is thought to most likely occur at the initiation stage of translation, partially due to evidence that miRNAs do not repress mRNAs translated through use of internal ribosome entry sites (IRES) which bypasses the 5' cap altogether (Humphreys, et al., 2005; Pillai, et al., 2005), as well as evidence that miRNAs are unable to repress transcripts with artificial cap structures which are unable to recruit translation initiation factors (Mathonnet, et al., 2007). However, despite miRNAs role in the repression of translation, mRNA destabilisation is still thought to be the major contributor to the regulation of protein-coding transcripts by miRNA, and has been estimated to contribute '≥ 84%' of miRNA-related decreased protein production (Guo, et al., 2010). A summary of the mechanisms by which miRNAs repress targeted mRNA transcripts is given in figure 2.4.

2.8 Non-coding RNA targets of microRNAs

miRNAs are also known to target other non-coding RNA molecules, including long non-coding RNAs and (lncRNAs) and circular RNAs (circRNAs). The potential role of these molecules in this context, in concert potentially with other classes of RNA molecule in the cell, is to

act as molecular ‘sponges’ for the miRNA (Ebert, et al., 2007; Ebert and Sharp, 2010). It is proposed that this sponging effect would provide a form of competitive inhibition for miRNA activity, and thereby derepress the protein-coding targets of miRNA molecules, in what is known as the ‘competitive endogenous RNA’ or ‘ceRNA’ model of general miRNA-RNA interactions (Salmena, et al., 2011). Research has been conducted arguing against the validity of this model, with arguments made that the total target site abundance for a single miRNA is at such a high level, that unphysiological levels of ceRNA would be needed in order to mediate derepression of, for example, mRNA targets of the miRNA (Denzler, et al., 2014; Denzler, et al., 2016). In the Denzler *et al.* 2016 study, thresholds of additional sites required to mediate derepression of targets are stated as ‘~10%–40%’ of a single miRNA’s total 3’UTR target site abundance. In this study, it is argued because each transcript type (*i.e.* individual transcript identifier) typically possesses ‘< 5%’ of total 3’UTR target site abundance, then a single transcript type is very unlikely to mediate derepression of other miRNA targets via a ceRNA type sponging mechanism. However, it is not unlikely that multiple RNA molecules could act co-operatively to mediate a depressive ceRNA effect on miRNA-targeted mRNA molecules. Such competitive endogenous RNA effects have been experimentally observed, for example, in a mouse brain ncRNA network involving a lncRNA, a circRNA and two miRNAs (Kleaveland, et al., 2018), in which binding of miR-7 by the *Cyrano* lncRNA, prevents repression of *Cdr1as* circRNA. This sponging effect can be attenuated, as in this particular case, by the process of target-directed miRNA degradation (TDMD) (Ameres, et al., 2010; Baccarini, et al., 2011), which can occur in cases where there is extensive complementarity between the miRNA

and its target. In addition, multiple closely spaced miRNA target sites can mediate a stronger ceRNA depressive effect (Denzler, et al., 2016).

2.9 Experimental validation of miRNA targets

2.9.1 3'UTR reporter assays

There exist many experimental methods that can be used to identify miRNA targets. 3'UTR reporter assays are commonly used to perform low-throughput assays of potential miRNA targeting activity, in which a candidate 3'UTR of interest is typically cloned into an expression plasmid replacing the 3'UTR of a reporter gene. Once the cell or organism is transformed with the expression vector, the efficacy of the 3'UTR for transcript repression can be assayed using the reporter gene. Such a reporter system was used in the identification of the first *bona fide* miRNA-mRNA interaction in which for a *lacZ* reporter system, X-Gal staining was used to confirm the interaction between the *lin-14* mRNA and the *lin-4* miRNA, at a particular stage in *C. elegans* development (Wightman, et al., 1993). The luciferase reporter system has also been used for this purpose (Kertesz, et al., 2007), in which light emitted by the luminescent luciferase protein has been used as an indicator of target site effectiveness. A dual luciferase reporter system is commonly used, using both firefly and *Renilla* luciferase within the same reporter construct, one of which will contain a cloned 3'UTR of interest, and the other luciferase gene serving as a negative control for miRNA activity. GFP, YFP and mCherry reporter systems have also been successfully used to assay miRNA activity (McJunkin and Ambros, 2017; Mukherji, et al., 2011). In more recent years, massively

parallel reporter assays have been designed and implemented in order to assay miRNA activity using higher throughput methods. In these assays, large numbers of candidate 3'UTR sequences are cloned into expression vectors in parallel, and their activity assayed (Litterman, et al., 2019; Slutskin, et al., 2018).

Noise considerations must also be evaluated when considering 3'UTR reporter assays for the detection of miRNA targets. As in the case of miRNA mimic transfection experiments, unwanted variance is removed via normalisation against a negative control condition. In the negative control condition, a different reporter is co-transfected of a similar but distinct reporter construct (*e.g.* Renilla luciferase) in which the fused 3'UTR reporter does not contain a predicted miRNA target site. Alternatively, both reporter systems can be contained within the same cloned construct. In a massively parallel reporter assay, utilising a dual reporter system for each construct, it was shown when examining protein fold repression values between constructs containing and not containing a predicted miRNA response element, that the median relative standard deviation between constructs differing only by their barcodes was 10.5%, indicating a small amount of technical noise in this system (Slutskin, et al., 2018).

2.9.2 miRNA perturbation and sequencing

The activity of miRNAs can be investigated more indirectly by determining bulk cellular expression profiles upon perturbation of intracellular concentration of a single, or multiple miRNAs. Perturbation is

usually achieved by transfection of miRNA mimics into cells using liposome vectors (Agarwal, et al., 2015). Negative control transfections are also performed using plasmids or scrambled oligonucleotides. The abundance levels of RNA can be assayed using microarray and RNA-Seq technologies in order to compare treatment and negative control conditions, in order to gauge the effect of the perturbed miRNA on destabilising a given RNA, or set of RNAs. As this method does not test for a direct chemical interaction between the perturbed miRNA, and potential targets, it may detect RNA molecules only indirectly repressed or perturbed as a result of miRNA activity. As this method involves the assaying of RNA expression levels, rather than that of protein, as seen in the use of 3'UTR reporter systems, it is not sensitive to genes which are directly translationally repressed by the miRNA without associated changes in RNA stability levels, however, as previously discussed, direct translational inhibition is not thought to be a major contribution to miRNA activity in animals (Guo, et al., 2010). There are reports, as evidenced by northern blotting, that miRNA transfection leads to supraphysiological, ~100-fold increases in intracellular miRNA levels of the perturbed miRNA, and therefore should be used with caution (Jin, et al., 2015). However, it has also been noted that methods for assaying increases in miRNA abundance levels such as northern blotting and qPCR do not distinguish active AGO-bound miRNA, from inactive miRNA sequestered in vesicles, and hence the effect of miRNA transfection may not be as potent as once thought (Thomson, et al., 2013). Quantitative proteomics approaches have also been used to assay the effect of miRNA perturbation on intracellular protein levels (Baek, et al., 2008).

There are both biological and technical sources of noise in miRNA mimic transfection assays. Biological noise can arise from differences in gene expression upon the transfection of a nucleic acid which is unrelated to the investigated effect (*i.e.* miRNA targeting). For example, Agarwal *et al.* (Agarwal, et al., 2015) discovered a considerable association between transfection of a nucleic acid, and the dysregulation of transcripts according to their 3'UTR length, and also the AU content of their 3'UTRs. Technical sources of variation and noise for this assay include 'stochastic biochemistry during library preparation', the random sampling of cDNA fragments during sequencing, and non-deterministic processes during the computational analysis of reads.

A common method for accounting for noise in miRNA mimic transfection assays is to make a direct comparison between the log fold change cumulative distribution function (CDF) of predicted miRNA targets with the corresponding CDF of the predicted non-targets, and performing null-hypothesis significance testing using the Kolmogorov-Smirnov (KS) test. As the non-target distribution is expected to contain both biological and technical sources of noise, and the target distribution is expected to contain both this noise and the biological signal of interest (*i.e.* repression due to miRNA targeting), then comparing the target and non-target distributions using the KS tests acts as a form of normalisation in which biological signal is distinguished from noise.

It is helpful to quantify the proportion of the signal arising from this assay which is constituted by noise. One method of doing this is to conceptualise an idealised log fold change distribution for the condition of the mock transfection in which no noise at all is present within the system. For such an idealised distribution, all log fold change values are 0.

The deviation from the idealised ‘no-noise’ distribution can then be computed by taking the root of the average square fold change value for each distribution. The signal-noise ratio in the system can then be calculated as a ratio between the average deviation in the ‘signal’ target and ‘noise’ non-target log fold change distributions. The reciprocal of the signal-noise metric provides a measure of the proportion of the measured signal which is noise. This process can be formalised with the following equation:

$$SNR = \frac{\sqrt{\frac{\sum_i^{n_p} p_i^2}{n_p}}}{\sqrt{\frac{\sum_j^{n_q} q_j^2}{n_q}}}$$

Where p and q are arrays of target and non-target log fold change values respectively, and n_p and n_q give the number of elements (*i.e.* number of transcripts/genes) in n and q respectively. p_i and p_j give the i th and j th elements of p and q respectively, where $\{p_i, q_j\} \in \mathbb{R}$. SNR is the estimated signal-noise ratio.

2.9.3 Cross-linking and immunoprecipitation

A more recent set of approaches to identify miRNA targets is to use cross-linking of protein and bound RNA followed by immunoprecipitation chemistry in order to identify protein-RNA interactions (Niranjanakumari, et al., 2002), which in the context of miRNA biology is used specifically to identify AGO-RNA interactions. A commonly

used variant of this methodology is the cross-linking and immunoprecipitation (CLIP) approach (Ule, et al., 2003). In this methodology, ultraviolet light (UV) is typically used (although formaldehyde has been used in previous cross-linking protocols (Niranjanakumari, et al., 2002)) to create cross-links between RNA and proteins bound in ribonucleoprotein (RNP) complexes. Cell lysis then follows cross-linking. An RNase removes any overhanging unprotected RNA from the RNP. Antibodies are used to select for and pulldown a protein of interest, which due to crosslinking, will be pulled down as an RNP. RNP RNA is radiolabelled via phosphorylation, which is followed by gel electrophoresis and subsequent autoradiography. Relevant bands are excised. The protein is digested, leaving the previously crosslinked RNA free for adapter ligation, reverse transcription to cDNA, PCR amplification and subsequent high-throughput sequencing. Such protocols have been used to investigate chemical interactions between AGO and RNA *in vivo* (Chi, et al., 2009). A number of modifications and enhancements of CLIP approach have since been developed (Hafner, et al., 2010; Huppertz, et al., 2014; Van Nostrand, et al., 2016). Use of CLIP protocols *per se*, will identify segments of the guide or the target RNA bound to AGO, but will not identify both the guide RNA and target RNAs simultaneously, as a result, there is no direct inference of guide-target interactions from CLIP. To resolve this issue, protocols have been developed, labelled as CLIPL protocols (*i.e.* CLIP and ligation) (Wang, 2016), in which CLIP protocols are developed with an extra ligation step for the purposes of ligating the AGO-bound guide RNA and the target RNA, generating chimeric guide-target RNAs which can be sequenced (Grosswendt, et al., 2014; Helwak and Tollervey, 2014; Kudla, et al., 2011). In this way, the guide-target interaction can be unambigu-

ously identified. However, the functionality of many miRNA-target interactions identified using CLIP approaches has been contested (Agarwal, et al., 2015).

Like 3'UTR reporter assays and miRNA transfection experiments, CLIP assays contain many sources of noise (Darnell, 2010). Sources of noise include insufficient specificity of protein purification, an overabundance of competing low complexity RNA sequence, transient RNA-protein interactions, and PCR amplification artefacts. Such issues can be partially mitigated by ensuring stringent reagents and conditions for protein purification, filtering of unique reads, analysis of read clusters as opposed to individual reads, and filtering of reads for those matching a given RNA-protein binding motif. In addition, the use of negative control RNA binding proteins can be used to identify, and normalise against non-specific RNA-protein interactions (Darnell, 2010).

Algorithms have been developed for downstream analysis of read data from CLIP experiments in order to mitigate against this noise. In Omniclip (Drewe-Boss, et al., 2018), the strength of the protein-RNA interaction as well as the relative abundance of the mRNA is taken into account when calling CLIP peaks. Firstly, a series of generalised linear models (for peak and non-peak states) are created taking background gene expression from RNA-Seq data into account, in order to model the probability of peak and non-peak states given the coverage profile from RNA-seq and CLIP for different positions along the genome. Secondly, a multinomial Dirichlet mixture model is used to model nucleotide transitions which are artefacts of CLIP experiments (Kishore, et al., 2011). The coverage profile models and the diagnostic event models are used

together in order to parameterise a hidden Markov model from which the nucleotide-nucleotide ‘peak-state’ of the genome can be inferred.

2.9.4 Databases of validated miRNA interactions

Databases of validated miRNA interactions can help miRNA researchers easily identify high-confidence miRNA interactions, as well as the experiment type, and the specific publication from which the reported interaction derives. Both DIANA-TarBase (Karagkouni, et al., 2017) and miRTarBase (Chou, et al., 2017) are publicly available databases which are released for this purpose.

2.10 Principles of miRNA targeting

Despite these experimental advances, there is still no high-throughput method for identifying direct, functional targets of miRNAs, underlying the continued necessity of computational approaches for identifying putative targets. In addition, implementation of library preparation and sequencing protocols for these purposes is not always simple or cost-effective in terms of material resources, and skilled personnel required.

Primarily, target prediction is mostly performed using criteria relating to antisense complementarity between the miRNA and its target, physical and thermodynamics factors relating to the possibility of duplex formation, and the conservation of miRNA target sites between closely related species (Ritchie and Rasko, 2014).

2.10.1 Canonical models of miRNA targeting

If we examine criteria for the complementarity between miRNAs and putative targets further, we can broadly distinguish between two classes of miRNA target prediction algorithm; namely, *canonical* target prediction requiring full complementarity between the miRNA *seed region* (Lewis, et al., 2003), and conversely, *non-canonical* prediction methodologies which do not use this requirement. As described previously, the seed region constitutes nucleotides 2-8 of the miRNA, or less stringently nucleotides 2-7 of the miRNA. Unlike plant miRNAs, an enrichment of fully complementary targets of mature miRNAs for bilaterian species in their respective transcriptomes was not detected (Rhoades, et al., 2002). In addition, when developing one of the first seed-based algorithm, named *TargetScan* (Lewis, et al., 2003), it was noted that observations had been made that the conservation of the miRNA was greater at the 5' end (Lim, et al., 2003). Additionally, the validity of the identified seed region was tested by permuting the position of the designated seed region on the miRNA across its entire length, and subsequently examining the number of conserved miRNA targets detected. It was shown that the 2-8nt heptamer performed most successfully on this test, in addition, this particular heptamer location was shown to be the most conserved for all heptamer location permutations (*e.g.* 3-9nt, 4-10nt *etc.*) along the length of the miRNA (Lewis, et al., 2003). The seed-based approach to target prediction has been used for all further iterations of the TargetScan algorithm released from the Bartel lab (Agarwal, et al., 2015; Friedman, et al., 2009; Garcia, et al., 2011; Grimson, et al., 2007; Lewis, et al., 2005).

The original seed model was extended by dividing and categorising a set of different seed target sites types that could be used to aid target recognition (Lewis, et al., 2005). This divergence from the strict use of the original 2-8 nucleotide heptamer seed match, was the discovery that a conserved adenine base commonly occupied the 't1' position of the target immediately opposite the first nucleotide of the miRNA. This was irrespective of whether the first nitrogenous base of the miRNA was the Watson-Crick (WC) base complement of t1. It was also discovered that hexamer sequences position at nucleotide 2-7 were conserved above background across five mammalian genomes. This additional evidence led to defining seed types by using a combination of previously discovered features which were thought to be conducive to miRNA targeting, and also an evaluation of their relative strengths: The hexamer 2-7nt seed target (later commonly referred to in the literature as the *6mer*) exhibited the weakest sequence homology above background levels. A hexamer sequence with an A in the t1 position (*7mer-A1*), corresponding to the first nucleotide of the miRNA, considerably increased the signal-noise ratio from this core seed sequence. Indeed, the signal-noise ratio for *7mer-A1* target sites almost equalled that of target sites with 7 contiguous base pairs at nucleotides 2-8 of the miRNA, but without an adenine at t1 (*7mer-m8*). Perhaps unsurprisingly, the most conserved target site was the one that combined the features of the *7mer-A1* site and the *7mer-m8* site *i.e.* the *8mer* site. Though it should be noted that *7mer-A1* and *8mer* target sites only contain 6 and 7 Watson-Crick base pairs between the guide RNA and the target respectively (Lewis, et al., 2005). The specificity of this model has been validated in numerous experiments in which intracellular miRNA abundance levels have been perturbed *in vitro* via the transfection of miRNA mimics, or antagomiR oligonucleotides complementary to the miRNA for a

miRNA knock-down effect. These experiments also demonstrated the 7mer-m8 site tends to confer stronger repression than the 7mer-A1 site (Agarwal, et al., 2015; Friedman, et al., 2009; Garcia, et al., 2011; Grimson, et al., 2007).

Additional support for the seed model is provided from structural work relating to the argonaute effector, and more generally argonaute complexes containing the bound guide RNA. In 2012, Schirle and MacRae released a 2.3 angstrom (Å) resolution crystal structure of the human AGO2 protein (figure 2.5) (Schirle and MacRae, 2012). They discovered electron density in their structure attributable to an 8nt single-stranded RNA spanning the Mid and Piwi domains of AGO2. From this it was determined within the protein, that nucleotides 1-7 of the guide RNA are bound to the AGO2 molecule in a likely sequence independent manner, by a series of weak, non-specific and non-covalent interactions such as hydrogen bonds and van der Waals forces presumably to allow flexible association and dissociation between argonaute and the guide RNA. Crucially however, nucleotides 2-6 were found to be positioned, in an A-form conformation, so as to be exposed to the cytosol, aiding recognition by the seed region of different RNA elements. In subsequent work, the same authors crystallised the guide-RNA bound AGO2 structure, with only four nucleotides in the middle of the guide RNA unresolved (Schirle, et al., 2014). The first 5' nucleotide of the guide RNA was found to be anchored to the Mid domain of AGO2, potentially explaining the lack of use of the first gRNA nucleotide for Watson-Crick base-pairing to the target. Nucleotides positioned at the 3' end of the guide RNA were found to be locked in to the N-Paz channel of the protein, and facing away from the cytosol. Further crystal structures with bound target RNA to the 5' end of the guide RNA determined

that base-pairing between the guide and its target within this region causes a conformational change of helix-7 of the AGO-structure away from the RNA duplex. This is thought to enable base-pairing of guide RNA bases 6 and 7, which could cause disruption to the AGO-RNA complex, and potential dissociation of the bound target if there are mismatches or mispairing in this region, allowing the miRNA-AGO complex to rapidly accept or dismiss putative targets depending on their complementarity to the guide in this region (Klum, et al., 2018). This method of dynamically searching for targets by staged probing of target complementarity is thought to enable, coupled with lateral diffusion of the AGO-RNA complex along the target molecules, a rapid traversal of the cytosolic RNA search space (Chandradoss, et al., 2015). This conformational change of argonaute helix 7 is extended to nucleotides 11-16 of the guide RNA, enabling previously documented (Bartel, 2009) supplementary and compensatory binding of target nucleotides to this region of the guide RNA. Crucially, this additional pairing does not preclude the validity of the seed model, as pairing in the seed region is necessary to trigger the conformational changes in the argonaute protein which are required for 3' base pairing. In addition, further work from this lab demonstrated that the t1 adenine nucleotide is anchored to a pocket on the surface of the argonaute molecule, through a network of hydrogen-bonding water molecules, the effect of which is to increase the dwell time of argonaute on the target molecule and increase the probability that target repression occurs (Schirle, et al., 2015).

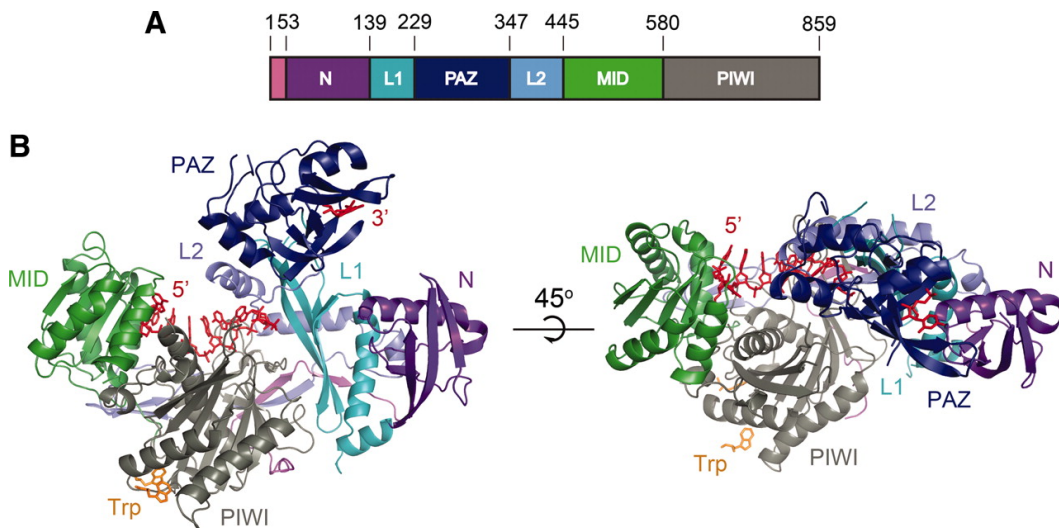


Figure 2.5 – The domains and crystal structure of human argonaute-2. The 5' nucleotides of bound RNA are anchored by interactions with the Mid and Piwi domains of this protein. The Paz domain weakly binds the 3' end of the miRNA.²

As well as structural evidence, there is also functional and computational evidence for miRNA 3' base pairing to its target. Work in the Bartel lab showed that 7mer sites with contiguous base pairing of 4bp or more in the 3' region, preferably starting at nucleotide 13 of the miRNA, enhanced targeting efficiency for this particular seed-target pairing. In fact, the 7mer-m8 match with good supplementary 3' pairing was found to be almost effective as 8mer target matches (Grimson, et al., 2007). Presumably the purpose of the supplementary pairing is to favour increased occupancy time of the AGO-RNA complex on the target molecule and disfavour dissociation. Increased occupancy of the argonaute on the target, will likely increase the probability of recruiting the scaffold protein GW182, and additional deadenylases and decapping complexes. Chandradoss and colleagues (Chandradoss, et al.,

² From Schirle, N. T., & MacRae, I. J. (2012). The crystal structure of human Argonaute2. *Science*, 336(6084), 1037-1040 (<https://science.sciencemag.org/content/336/6084/1037.full>). Reprinted with permission from AAAS.

2015), using FRET experiments, show a positive relationship between AGO dwell time and the number of base-pairing guide RNA nucleotides when they test this for a limited range of nucleotides (*i.e.* N=5 and N=6). A similar relationship would presumably exist for greater values of N. A summary of the core seed-based model for miRNA target prediction, including supplementary and compensatory base pairing is given in a recent review from the Bartel lab (figure 2.6).

miRNA, typically between nucleotides 13-16, which increases the efficacy of target repression. Noncanonical target sites are target sites which contain a mismatch or non-standard base pairing in the seed region of the miRNA. Base pairing in the 3' region as described previously can be used to compensate for these mismatches.³

There have been other canonical target prediction models produced outside of the Bartel lab. The EIMMo model for example requires strict seed pairing for example, but additionally uses a Bayesian phylogenetic method to infer the functionality of putative targets site, by analysing the pattern of conservation of a given site in relation to the conservation and selection patterns of other putative target sites of the same miRNA (Gaidatzis, et al., 2007).

A number of additional features contextual to the existence of the seed match, have been identified, which are believed to aid recognition and repression of predicted miRNA targets, and which may explain why the existence of a seed match is not always sufficient for the repression of a target (Grimson, et al., 2007). One discovery is that adjacent or proximal seed targets sites (within 50nt of each other) act co-operatively and synergistically, and are not linear sums of the predicted effects of the constituent sites (Grimson, et al., 2007; Sætrom, et al., 2007). Chandross and colleagues proposed a model, which can be interpreted as a mechanism for this observed co-operative effect, in which the argonaute RNP complex laterally diffuses and shuttles between adjacent target sites to increase the overall dwell time of argonaute on a narrow

³ Reprinted from Cell, 173(1), Bartel DP., Metazoan miRNAs, 20-51., Copyright (2018), with permission from Elsevier.

section of the RNA molecule (presumably to enable recruitment of scaffold and effector proteins), increasing the probability of a repressive effect occurring (Chandradoss, et al., 2015). Another factor identified as contributing to repression were AU-rich content flanking putative miRNA target sites (Agarwal, et al., 2015; Grimson, et al., 2007). Although the precise mechanism by which flanking AU-content influences miRNA targeting is unknown, local 3'UTR structural accessibility is thought to be a factor, although secondary-structure prediction was shown to be less informative than consideration of AU content suggesting the structure-independent mechanisms could contribute to the regulatory effects of AU rich regions, with evidence that AU-rich elements can act as general determinants of RNA stability independent of miRNA action (Chen and Shyu, 1995). A strong restriction implemented in all iterations of the TargetScan tool is that any putative target sites must reside within the 3'UTR of the mRNA molecule. Justification for this approach is that 5'UTR predicted targets were not found to be conserved above background levels, and whilst some open reading frame (ORF) miRNA targets were conserved above background levels (*i.e.* over and above the general conservation of codons in ORFs), the majority of conserved sites were still found in the 3'UTR (Lewis, et al., 2005). In addition, ORF and 5'UTR targets were collectively only marginally repressed or not repressed at all in analyses derived from miRNA mimic transfection experiments (Grimson, et al., 2007), although to reflect this, the number of ORF 8mer targets is used as a feature in the latest version of the TargetScan algorithm (Agarwal, et al., 2015). Different locations within the 3'UTR have also been assayed for responsiveness to miRNAs. An extended ORF luciferase reporter assay and conservation analyses were used to show that target sites < 25nt

away from stop codons are selected against, and if present are non-functional, presumably due to occlusive and steric hinderance from the ribosomal translational machinery. Somewhat counterintuitively, beyond this strict requirement, functional and conservation analyses indicated that for humans, sites positioned away from the centre of the 3'UTR and towards the stop codon and the poly-A start site were preferred (Agarwal, et al., 2015; Gaidatzis, et al., 2007; Grimson, et al., 2007; Majoros and Ohler, 2007). The offset 6mer in which the location of the miRNA seed region is offset by one nucleotide (to nucleotides 3-8 of the miRNA) was also found to confer a marginal repressive effect (Friedman, et al., 2009), and the number of offset 6mer sites in the 3'UTR is used as a feature in the latest version of the TargetScan algorithm (Agarwal, et al., 2015). Interestingly, in a relatively early publication associated with a release of a version of the TargetScan algorithm, high target site abundance in 3'UTRs is identified as a feature which minimises the repressive effect of a particular miRNA (Garcia, et al., 2011), and is also used as a major feature in the latest version of the TargetScan algorithm, indicating the contribution of a ceRNA effect for analyses using the TargetScan algorithm. The thermodynamic stability of the seed-target interaction has also been shown to contribute to the repressive effect of the miRNA, presumably by increasing the dwell time of argonaute on the RNA, and that this effect can be distinguished from the potentially confounding effect of target site abundance (Garcia, et al., 2011). In addition, in corroborating previously reported research (Hausser, et al., 2009), ORF and 3'UTR length seems to be inversely correlated with target site efficacy (Agarwal, et al., 2015), potentially indicating the formation of occlusive secondary structures in regions of the transcript distal to the stop codon and the poly-A tail. An alternative explanation is that there is increased difficulty in recruiting

deadenylase and decapping complexes in relatively distal and remote regions of the transcript. A summary of the different sequence-based and contextual features used in miRNA target prediction algorithms is given in figure 2.7.

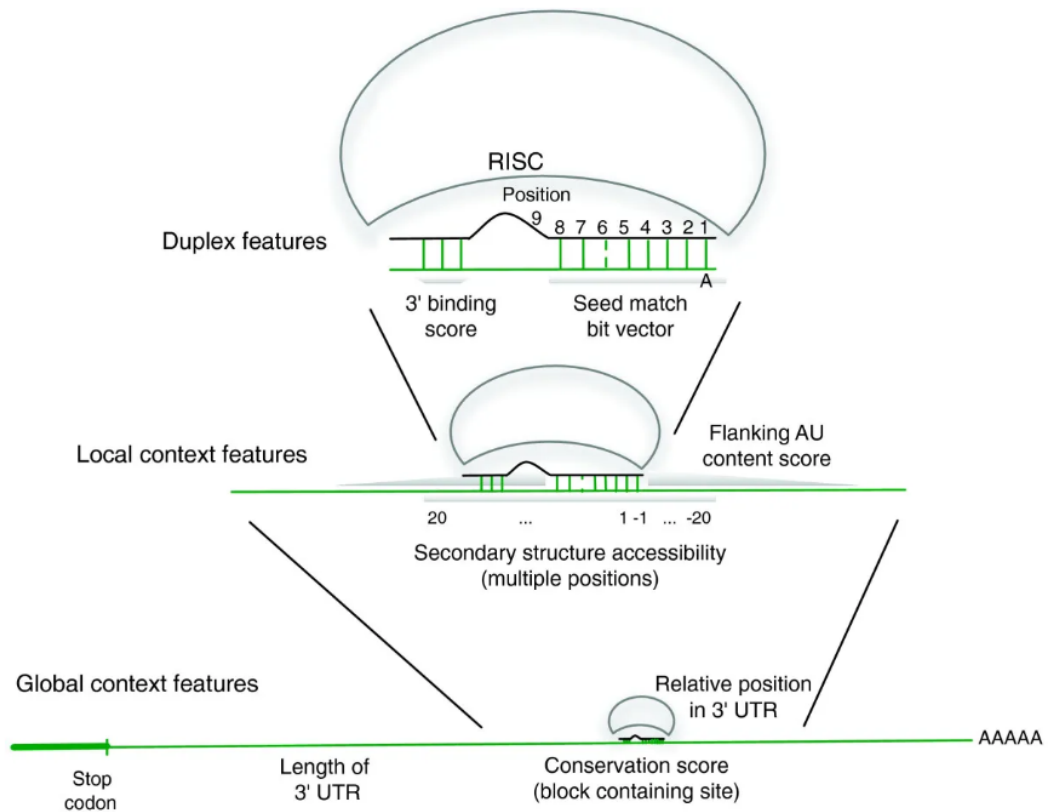


Figure 2.7 – A summary of some of the duplex, local contextual and global contextual sequence features used by miRSVR to score putative targets, which are representative of features used by other target prediction algorithms. As discussed previously, at the level of the RNA duplex, WC and non-standard base pairing features are used to score targets. Flanking AU content and the structural accessibility of secondary structures on the target are frequent local context features which are considered. Global contextual features considered include the length of

the 3'UTR, the conservation of the target site, and the relative position of the target site within the 3'UTR.⁴

2.10.2 Noncanonical models of miRNA target prediction

In parallel with the development of seed-based methods for predicting animal miRNA targets, algorithms have been developed which do not require perfect, contiguous Watson-Crick base pairing between every nucleotide of the miRNA seed region and the corresponding target *i.e.* non-canonical approaches to miRNA target prediction have been developed.

Experiments have been conducted in order to functionally validate some non-canonical miRNA target sites. The first type of non-canonical site to be discovered contains a GU-wobble base pair, which appears in both *let-7* target sites of *lin-41* (Reinhart, et al., 2000), and a *hid* target of the *bantam* miRNA in *D. melanogaster* (Brennecke, et al., 2003), all of which contain at least 6 nucleotides of base-pairing downstream of the seed, which likely compensate for the observed wobbles in the seed region. Additionally, some years later, centred sites in which there is at least 11 nucleotides of contiguous base pairing from nucleotides 4 or 5 of the miRNA onwards, without additional 3' or 5' base pairing has been shown to induce mRNA repression upon transfection of the miRNA (Shin, et al., 2010), however, because there is some requisite

⁴ Reproduced with permission, from Betel, D., Koppal, A., Agius, P., Sander, C., & Leslie, C. (2010). Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome biology*, 11(8), R90. doi:10.1186/gb-2010-11-8-r90

base-pairing in the seed region, this site type may be more simply classified as a form of compensatory target site

Stark *et al.* (Stark, *et al.*, 2003) and later Enright *et al.* (Enright, *et al.*, 2003), and Rajewsky and Socci (Rajewsky and Socci, 2004) predicted noncanonical miRNA target sites in the common fruit fly and later human (John, *et al.*, 2004). Commonalities between all of these approaches, is that they all preferentially weight complementarity at the 5' end of the miRNA (with a toleration of GU wobbles), try to determine the thermodynamic stability of putative RNA-RNA duplexes, and also include an assessment of the evolutionary conservation of target sites between closely related species, when evaluating putative targets. Despite requiring strict seed pairing, the previously described method reported by Lewis and colleagues (Lewis, *et al.*, 2003) does not differ considerably from these approaches. Indeed, the miRanda algorithm released by Enright and colleagues allows users to pass a '*strict*' flag when using the tool, allowing the tool to be optionally used as a canonical miRNA target prediction algorithm.

Both confirmatory and novel findings were reported in a study released from the Hatzigeorgiou lab not long afterwards (Kiriakidou, *et al.*, 2004). In this study, a mutagenesis approach was used to identify features of the miRNA-target interaction using a dual luciferase reporter assay performed on human and mouse cell lines. As in previous studies, the importance of the pairing of the 5' seed region was identified, as was the supplementary effects of 3' base pairing. However, Kiriakidou *et al.* also discovered through their mutagenesis screen that disruption of a symmetrical bulge or asymmetrical bulge (*i.e.* on either the miRNA

or the target) at the centre of the RNA duplex, caused derepression of the target for multiple miRNAs. As a result, in a particularly conservative approach, the DIANA-microT algorithm was released which required 5' seed pairing (GU wobbles and bulges are tolerated to an extent), 3' supplementary pairing, and the existence of a central bulge region of a specified number of nucleotides, yielding on average a relatively small 9.4 predicted targets per miRNA (Kiriakidou, et al., 2004).

Other non-canonical miRNA target prediction algorithms employ a different approach for identifying miRNA targets: RNA 22 is unique in that it searches for enriched patterns or motifs in previously annotated miRNAs from a number of species, searches for regions of the transcriptome containing reverse complements to enriched patterns, and identifies microRNAs which could associate with these 'target islands' (Miranda, et al., 2006). The scoring binding matrix (SBM), in a relatively unbiased approach, does not use a pattern matching or dynamic programming approach, but rather constructs a sequence matrix of the reverse complemented miRNA and all validated targets of that miRNA, and scores putative targets on the basis of their similarity to existing targets. Dinucleotides are used to incorporate information about RNA stacking energies (Moxon, et al., 2008). PITA focusses on the structural accessibility of putative target sites with perfect or near-perfect pairing in the seed region, and makes an explicit comparison when scoring targets between the free energy required to unpair existing secondary structures, and 'the free energy gained from the formation of the microRNA-target duplex' (Kertesz, et al., 2007). An improvement to the initial PicTar algorithm allowed dynamic calculation of the probability that an identified target site is functional, by an assessment of the conservation patterns of all putative targets of that miRNA sequence (Lall,

et al., 2006). TargetRank scores putative targets on a number of previously identified criteria such as seed type, flanking AU content – with a particular focus in this algorithm of being inclusive of species-specific miRNA targets, and potential siRNA off-targets (Nielsen, et al., 2007). In SVMicrO (Liu, et al., 2010) a seed filter for putative interactions is followed by two successive support-vector machine (SVM) classifiers, one for the seed pairing, and the second for the general 3'UTR context, trained on data stored on the miRecords database of experimentally validated targets (Xiao, et al., 2008), in order to classify putative miRNA targets.

The miRSVR model (Betel, et al., 2010) is somewhat similar to the latest TargetScan algorithm in the sense that it classes site conservation as a feature rather than a filter for target prediction, is trained on transfection data, and implements a regression model. However, a support vector regression model is used in this case, including features such as 3'UTR length, flanking AU content, and features relating to structural accessibility. A key difference however, is that miRSVR uses the seed-target duplex deriving from the miRanda algorithm as a set of binary features, and thus whilst perfect complementarity is preferred, it is not a requirement of this algorithm.

The advent of CLIP and CLIP ligation methods has led to the development of what has been termed a 'next-generation of animal miRNA target prediction algorithm' (Bradley and Moxon, 2017), which are trained on this data type.

The Hatzigeorgiou lab use PAR-CLIP data (Hafner, et al., 2010) to build a new target prediction model which includes putative target sites in the CDS region of mRNA transcripts (Reczko, et al., 2012). PAR-CLIP data is used to classify alignments of highly expressed miRNAs to 3'UTRs and CDS regions as true or false positives. Logistic regression is then used to identify potential target site features of interest, a subset of which is selected using the Akaike information criterion (AIC). A general linear model is then trained using this set of features, with separate models built for 3'UTR and CDS target sites. To assess the effect of multiple target sites, on a single transcript, a second general linear model is built using microarray transfection data. The model is subsequently tested using proteomics datasets (Selbach, et al., 2008).

In the chimiRic model trained using AGO CLIP and CLASH data, a dual SVM approach is used in which an SVM classifier first predicts miRNA-mRNA duplexes, whilst a second SVM implements an AGO binding model using contextual features relating to the 3'UTR sequence, in order to predict AGO binding (Lu and Leslie, 2016).

Another set of target prediction algorithms trained on CLIP or CLIP data are the MIRZA and MIRZA-G algorithms. In MIRZA (Khorshid, et al., 2013), a probabilistic model is trained from AGO-CLIP data in order to score a set of parameters relating to specific miRNA nucleotide base pairing, as well as parameters relating to the formation of bulges, loops, and specific base-pairing patterns (*e.g.* GU wobbles). Reassuringly, the model recapitulates a large number of previous findings including the importance of the seed sequence, a lack of base-pairing at the first nucleotide of the miRNA, and preferential base pairing at the

3' end of the miRNA. Interestingly, Khorshid *et al.* recapitulate the finding by Kiriakidou *et al.* (Kiriakidou, et al., 2004) that a single loop is favoured at the centre of the miRNA-mRNA duplex, which is likely a result of a highly disfavoured hybridisation to nucleotide 9 of the miRNA discovered by Khorshid *et al.* The interpretation provided is that this is a biophysically unfavourable interaction – however, an alternative explanation is that there is a specific selection pressure against full complementarity in the animal miRNA-mRNA duplex in order to prevent cleavage. miRNA-mediated cleavage can and does occur in animal species, which requires pairing in the central region, and so any potential energetic unfavourability of hybridisation in the central region cannot be sufficiently unfavourable as to prevent this type of interaction occurring altogether. In MIRZA-G, the MIRZA target quality score was used as a feature, along with contextual sequence information of the target transcript in order to train a suite of general linear models which differ on the basis of whether they predict miRNA targets canonically, and whether or not they consider target site conservation information (Gumienny and Zavolan, 2015).

The most recent version of the miRDB database (Wong and Wang, 2014) contains predictions from a recent method trained exclusively on data derived from CLIP-ligation protocols in which chimeric reads can be used to unambiguously link miRNAs and targets (Wang, 2016). The CLIP-ligation data was used to generate target and non-target sets in order to identify relevant features such as patterns of nucleotide and dinucleotide usage in the target site, the structural accessibility of the target site, seed site conservation, and the location of the target site within the 3'UTR, which were used to train an SVM model.

A list of the names of different miRNA target prediction algorithms is given in table 2.1, as well as how they relate to common features of miRNA target prediction algorithms (*e.g.* sequence conservation metrics):

<u>Algorithm Name</u>	<u>Training method</u>	<u>Testing data/method</u>	<u>Predominant Method</u>	<u>Seed-based</u>	<u>Conservation?</u>	<u>Contextual Features?</u>
Context++ (targetscan)	Transfection data	Transfection data	Multi-linear regression	strict	True	True
Context+ (targetscan)	Transfection experiments	Transfection experiments	Multi-linear regression	strict	True	True
Context (targetscan)	Transfection experiments	Transfection experiments	Multi-linear regression	strict	True	True
TargetScanS	Conservation analyses	SNR analyses using random sequences	Rule-based approach	strict	True	False
TargetScan	Validated interactions	SNR analyses using random sequences	Rule-based approach	strict	True	False
miRanda	Validated interactions	Validated Interactions	Dynamic Programming	5' bias	False	False
miRSVR	Transfection data	Transfection data	Support-vector regression	5' bias	True	True
MIRZA	CLIP data	Transfection data	Maximum likelihood estimation	5' bias	False	False
MIRZA-GC	Transfection data	Transfection data	Generalised linear model	5' bias	True	True
chimirc	CLIP+CLASH	CLIP+CLASH	Support vector machine	5' bias	False	True
miRTarget	miRNA transfection + CLIP data	NA	Support vector machine	5' bias	True	True
RNA22	miRNA database	Validated interactions	Reverse complement pattern matching	5' bias	False	False
PicTar	Validated interactions	Experimental validation + comparison against random miRNA predictions	HMM maximum likelihood estimation	5' bias	True	False

PITA	Luciferase experiments	Validated interactions	Thermodynamic model	5' bias	False	True
eIMMO	Conserved miRNA target sites	Validated interactions	Bayesian phylogenetic model	5' bias	True	False

Table 2.1 - A summary of the most common computational methods used for miRNA target prediction. Different methods are annotated in this table on the basis of the data used to train/design the algorithm, the data used to test the algorithm, the computational method underscoring the algorithm, whether the algorithm accounts for sequence homology and also whether the algorithm accounts for contextual features of the putative miRNA binding site.

2.11 The accuracy of miRNA target prediction models

Published articles describing newly released methods for miRNA target prediction normally include some form of assessment of the prediction accuracy of that method in comparison to previously released, and in particular, commonly used methods. As a large number of methods have been published, and these typically report contradictory findings between each other, it can be helpful to consult review articles for an evaluation and a comparison of currently released methods, with the caveat that any review quickly becomes outdated with the rapid publication of new research.

In a relatively early report in 2006 (Rajewsky, 2006), a reported problem of published target prediction algorithms was that some algorithms produce radically different sets of target predictions from each other, suggesting a lack of convergence at this early stage of miRNA targeting research about relevant criteria for target prediction. However, despite the poor overlap of predictions between some algorithms, taking the union of results between algorithms has been shown to increase sensitivity beyond the best individual performing algorithm, by approximately 25% using a benchmarking dataset containing 84 miRNA-target interactions from TarBase (Sethupathy, et al., 2006) for which a ‘direct miRNA effect’ had been detected. However in a review paper by Nikolaus Rajewsky (Rajewsky, 2006), it was also noted that the approach of identifying miRNA targets using a reporter assay in combination with miRNA expression may not yield genuine targets which are regulated by miRNAs under endogenous conditions. The same problem exists for

experiments in which the intracellular miRNA expression is perturbed, with follow-up sequencing of transfected cells. In principal, it is likely that these approaches are useful for identifying the sequence-based features of miRNA targets. However, the utility of these approaches for identifying miRNA targets regulated under endogenous conditions is uncertain. It is possible that whilst the general important features identified by these approaches are valid, specific parametrisations may not be relevant for endogenous conditions due to the nature of *in vitro* methods used for data collection and subsequent model generation.

As discussed previously, the advent of high-throughput CLIP and CLIP-ligation protocols has provided an *in vivo* method for assaying argonaute binding activity. However, the problem being that argonaute binding *per se*, is not necessarily evidence of a repressive relationship between argonaute and the bound molecule. For the RISC complex to function, the argonaute protein has to search at least a 3'UTR search space, and to some extent, transcript coding sequence. Some form of close, physical interaction is presumably required between argonaute and putative targets for this search to occur. As discussed earlier, a model has been proposed in which the bound argonaute laterally diffuses across RNA molecules in search of seed targets (Chandradoss, et al., 2015). Although the dwell time of argonaute at specific point on the RNA molecule will increase with increased base pairing, from a probabilistic perspective, it may be possible for argonaute to be cross-linked to RNA molecules even when repression of that RNA molecule does not occur, by the possibility of crosslinking occurring while the argonaute complex is still searching for a suitable target. In addition, the number of CLIP-derived reads will likely be biased to highly expressed transcripts, which may in fact possess low occupancy by the argonaute

protein (Agarwal, et al., 2015; Friedersdorf and Keene, 2014). Indeed, a large number of non-canonical targets of miRNAs, derived from data from CLIP and CLIPL experiments have been declared as being non-functional due to a lack of observed repression of these transcripts after miRNA perturbation (Agarwal, et al., 2015). This problem can easily be mitigated by filtering individual CLIP hits, for those containing a seed sequence to a known miRNA, however, this would not overcome the poor sensitivity of CLIP-based analyses for identified valid miRNA targets, which has been attributed to variable cross-linking efficiencies of different RNA-protein interactions, and similarly variable ligation efficiencies for CLIPL approaches (Agarwal, et al., 2015).

As identified in a previous review, when benchmarking is conducted there is the possibility of bias in the selection of benchmarking datasets, which favour a particular subset of prediction methods being tested (Rajewsky, 2006). One approach recently employed in order to counter these concerns was to compare and aggregate the ranking of prediction methods from multiple different benchmarking analyses (Bradley and Moxon, 2017). The hope being that individual biases present in different studies would be mitigated or ‘cancelled out’ upon aggregation, leaving a somewhat unbiased estimator of target prediction performance. From this analysis, it was cautiously and tentatively concluded that the latest version of the TargetScan algorithm (Agarwal, et al., 2015) was the best performing of the current set of animal miRNA target prediction algorithms. However, this model, as it is trained on miRNA perturbation data, is associated with previously discussed concerns relating to this method of experimentation. In addition, Agarwal *et al.* found that their own regression model, for 7mer and 8mer seed matches, explained at most 15% of fold change variability of mRNA

expression changes upon miRNA transfection (Agarwal, et al., 2015). This relatively low number could potentially be explained by peculiarities introduced by the miRNA transfection protocol or the experimental and computational methods contributing to the gauging of mRNA fold change values. Nevertheless, it would seem that there is a large degree of miRNA targeting activity which remains to be explained.

2.12 RNA-Seq and differential expression analysis

2.12.1 RNA Sequencing and transcript quantification

In order to gauge the effects of miRNAs on the transcriptome, methods must be used to quantify transcript abundance levels during different conditions – for example, a condition in which a cell culture has been transfected with miRNA mimics, in comparison to a control condition. For many decades assaying of RNA abundance levels has been achieved through a process of Northern blotting (Alwine, et al., 1977) and quantitative PCR (qPCR) (Heid, et al., 1996) which are low-throughput methods for assaying transcript abundance from biological samples. For Northern blotting, a radioactively or chemically labelled RNA probe sequence is used to indirectly report transcript abundance through a process of hybridisation of the probe to a target sequence, from an RNA extract size-separated by electrophoresis. The signal arising from the probe is then detected through a process of autoradiography. The relative abundance of a particular transcript is then typically gauged by the normalisation of the signal arising from that transcript in comparison to constitutively expressed transcripts such as ribosomal

RNAs. In qPCR however, a fluorescent reporter of some kind is added to the PCR solution, and is used to quantitatively report nucleic acid abundance during a PCR reaction. The number of PCR cycles needed for the PCR fluorescence signal to reach a given threshold is related to the starting RNA material, and can be used indirectly to assay transcript abundance.

More high-throughput methods for assaying transcript expression levels include the use of cDNA microarrays (Schena, et al., 1995). This technology shares similarities with Northern blotting in the sense that transcript abundance is assayed by the hybridisation of sequence-specific probes to a target sequence. In the particular case of DNA microarrays however, the probe sequences are fixed on a solid surface in a location-specific manner, and it is the sample/target sequences (arising from fragmented cDNA sequence) which are chemically labelled or radiolabelled rather than the probe sequence. The expression of each target sequence in the original sample can therefore be inferred by quantifying the strength of the signal originating from each 'spot' on the microarray. However, the number of target sequences assayed using this approach is limited by the number of distinct probe sequences on the microarray.

Use of microarrays however can be limiting in the sense that, because DNA probes must be designed before experimentation, a certain amount of *a priori* knowledge of the sequences of transcripts to be assayed is required. In addition, use of hybridisation to assay expression levels leads to inherent difficulties such as background hybridisation

obscuring the expression of low expression transcripts, sequence-biased hybridisation properties and a failure to distinguish between splice isoforms, and genetic variants of the same gene (Zhao, et al., 2014).

The use of sequencing technologies to quantify gene expression mitigates the inherent difficulties in using hybridisation-based approaches. The basic principle behind such approaches is that the number of sequences of a given type returned from a sequencing experiment can be counted in order to ascertain the expression of a transcript in a given sample. A commonly used method for this purpose is bulk *RNA-seq* (Mortazavi, et al., 2008) in which a cDNA library is generated from sampled RNA, and subsequently sequenced (figure 2.8). There are a number of factors present in this method which can potentially complicate downstream analyses (Conesa, et al., 2016). Firstly, sequenced reads are typically short (~50-150nt) requiring fragmentation of the full-length RNA molecule, so that fragments can be sequenced separately, generating sequence coverage across the entire length of the transcript. For most protocols, sequencing is performed on DNA molecules, which requires the reverse transcription of RNA molecules into cDNA. In order to generate enough nucleic acid for sequencing, cDNA typically is PCR amplified from adapters ligated to the 3' and 5' ends of the cDNA molecules. Sequenced reads are typically then mapped to a reference genome or transcriptome in order to infer the transcript from which the sequenced read likely originated from. In addition, in order to study protein-coding transcripts specifically, researchers can enrich the RNA sequenced for mRNA. This can be achieved either by a process of rRNA depletion of the sampled RNA, or the addition of a 'poly-A selection' step in the library preparation protocol in which mRNA is

isolated through hybridisation of their poly-A tails to beads coated with poly-T oligomers.

Library preparation and sequencing protocols for RNA sequencing informs methods for downstream analysis of data of this type (Conesa, et al., 2016). Typically, once reads are aligned to a reference, count-based metrics are used to infer the relative abundance of the transcript for which reads have been aligned. Transcript counts are normalised for transcript length and also sequencing depth supporting within-sample and between-sample comparison of transcript counts respectively. Sequence bias correction methods (Bray, et al., 2015; Patro, et al., 2017) can also be implemented in order to correct for biases associated with fragmentation, reverse transcription, ligation and PCR amplification steps included in library preparation protocols.

As discussed at a later point in this thesis, data from RNA sequencing experiments can be combined with data from sRNA sequencing experiments in order to help infer the efficacy of given miRNAs for regulating transcriptional activity. Small RNA sequencing is a form of RNA sequencing, in which the RNA from which cDNA libraries are generated are enriched in small RNAs. This is usually achieved by fractionating and excising a band of the relevant size after gel electrophoresis of total RNA input (Pfeffer, et al., 2005). From this point, cDNA libraries are generally prepared as previously described for total RNA or mRNA sequencing, with individual steps for 5' and 3' adapter ligation, reverse transcription and PCR amplification. One particular problem with this method of RNA sequencing, is that when using adapters, sequence-specific biases can exist in the step of the ligation of sRNAs to

5' or 3' adapters, which will bias downstream analyses. To mitigate against this problem, library preparation protocols in which adapters have been designed with multiple degenerate nucleotides at each ligating end, referred to as *high definition adapters* in some protocols, have been developed (Billmeier and Xu, 2017; Sorefan, et al., 2012; Xu, et al., 2015). In addition, this method has recently been improved through the use of 'blocking oligonucleotides', which are used to deplete the sequenced sRNA pool of abundant transcripts which are not of biological interest, such as rRNA and rRNA fragments (Fowler, et al., 2018). During analysis of the resultant sequencing data, adapter trimming of reads is important, as adapter sequences are more likely to appear in reads due to the small size of the inserts generated during cDNA library construction (Nobuta, et al., 2010). Suites of tools exist for the specialised purpose of downstream processing of data of this type, including quality control, miRNA normalisation and quantification and differential expression analysis steps (Beckers, et al., 2017; Mohorianu, et al., 2017; Moxon, et al., 2008; Stocks, et al., 2018; Stocks, et al., 2012).

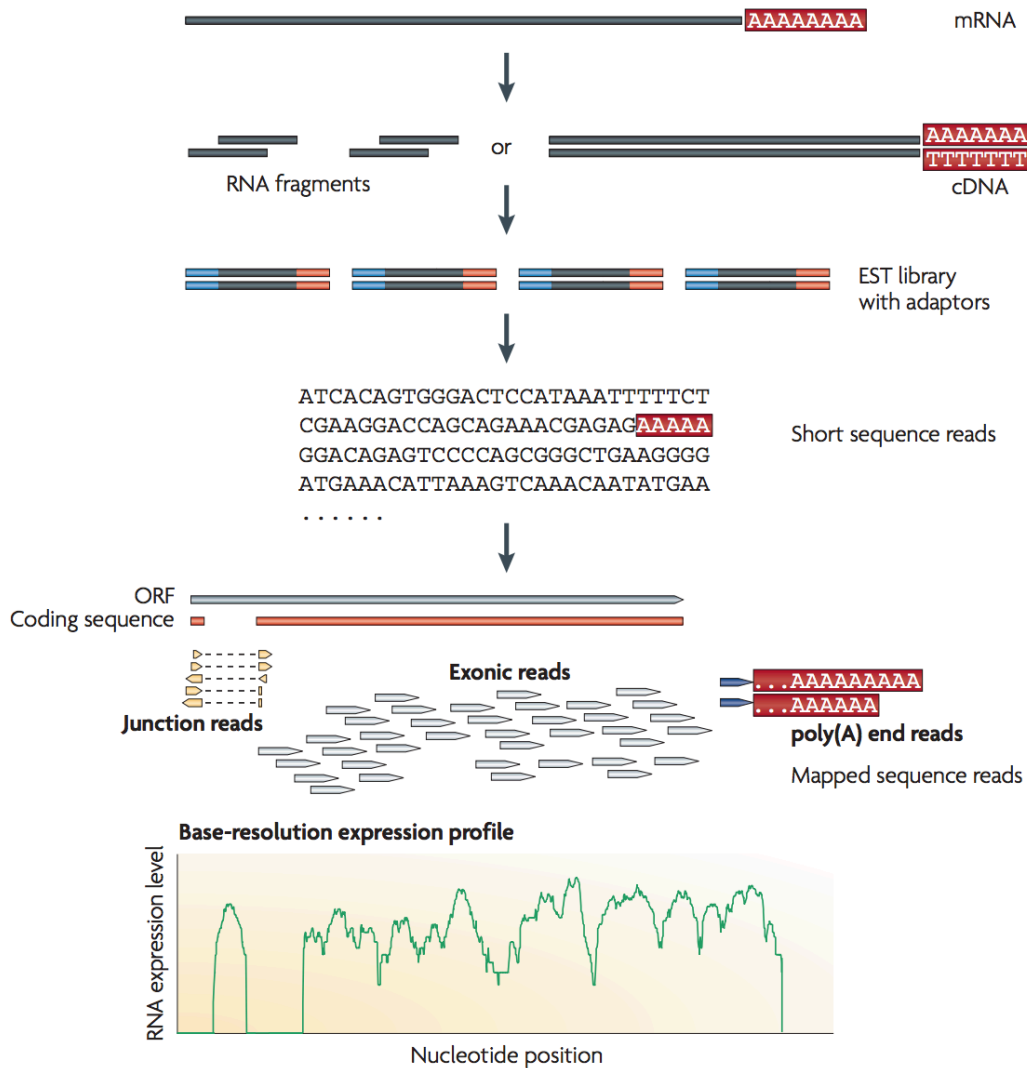


Figure 2.8 – A standard RNA-Seq library preparation and data analysis workflow. Isolated RNA is enriched for mRNA. This step is followed by reverse transcription and fragmentation of the nucleic acid, although the order of these two steps can be reversed. These cDNA fragments are PCR amplified, with subsequent ligation of adapters on either end of the cDNA molecule. Sequencing is performed, and sequenced reads are then mapped to a genomic reference, in order to infer RNA expression levels.

Reprinted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Nature Reviews Genetics, RNA-Seq: A revolutionary tool for transcriptomics, Wang, Z., Gerstein, M., & Snyder, M., Copyright 2009 (<https://www.nature.com/articles/nrg2484>).

2.12.2 Transcript quantification tools

For the FilTar tool which I present later in this thesis, both the Kallisto (Bray, et al., 2016) and Salmon (Patro, et al., 2017) tools are provided as options for use for transcript quantification from RNA-Seq data for users.

Salmon and Kallisto were developed with the intention of overcoming large computational bottlenecks in canonical RNA-Seq pipelines predominantly arising from the alignment of sequenced reads to a reference genome. The necessity of this approach came from, at the time, the increasing depth at which cDNA libraries from RNA-Seq experiments were being sequenced, and also the increasing number of samples and libraries being processed for studies utilising RNA-Seq.

To resolve this issue, the concept of lightweight algorithms for RNA-Seq data processing was proposed, in which, typically, strict alignment of reads to the transcriptome or the genome, is altogether avoided as a strategy for transcript quantification. Sailfish (Patro, et al., 2014), one of the first tools to use a lightweight approach for this problem, uses a k-mer based approach in which the transcriptome and the entire read set is indexed into a series of k-mers. The abundance of each transcript is then estimated by taking the average number of read k-mers mapping to each indexed k-mer within a given transcript. Relative transcript abundance estimates are then refined using an expectation maximisation approach.

Kallisto (Bray, et al., 2016) also adopts a k-mer based approach to RNA-Seq quantification, however, in this approach, k-mers are not individually mapped to a given set of matching transcripts. With Kallisto, k-mers within reads are collectively mapped to nodes on a transcriptome-De Bruijn Graph, in which each node represents a k-mer in the reference transcriptome. Each transcript in the reference transcriptome is represented as a path within the transcriptome-De Bruijn Graph (T-DBG). The mapping of reads to a subset of transcripts can then be deduced by taking the intersection of compatible transcripts for each constituent node for that read on the T-DBG. This process may return multiple compatible transcripts for a given read. An expectation-maximisation approach is again employed in order to infer transcript quantities, this time using sets of ‘equivalent’ transcripts for a read as inputs. By using expectation maximisation in this way, information from the entire pool of sequenced reads from a given library, and their respective mapped transcripts, can be used to predict relative transcript abundance values.

Salmon (Patro, et al., 2017) adopts an approach similar to Kallisto in the sense that a mapping between reads and transcripts are produced without generating complete base-to-base alignments, however, each mapping contains information regarding the sense-antisense orientation of the read, and the approximate location of the sourced read from the transcript. Salmon also differs from Kallisto in the sense that a suffix-array rather than a De Bruijn graph is used as the data structure in which the reference transcriptome is indexed. Such information can be extracted from Kallisto pseudo-alignments but requires additional computational steps after pseudo-alignment has been performed, and is therefore less easily accessible to the user.

In order to estimate relative transcript abundance values, both Kallisto and Salmon estimate the fragment length distributions from each library. For each transcript, estimated fragment length distributions are used to estimate what is termed the ‘effective length’ of the transcript *i.e.* which refers to ‘the number of start sites in a transcript which could have generated a fragment of a particular length’. The effective length of the transcript is then used when normalising read counts in order to give an estimate of relative transcript abundance. As the effective length metric is used to normalise pseudo-aligned read counts, estimated fragment length distribution means and standard deviations are parameters which are ultimately used in the calculation of TPM values within a given library.

The fragment length distribution can easily be inferred from paired-end sequencing data in which the insert size for relevant fragments can be deduced from taking the distance between matching paired-end reads. It is impossible to infer this information from the alignment of single-end read sequencing data however. When using single-end libraries, both Salmon and Kallisto contains default values for these parameters but expects users to input correct values for these for each of their libraries. These parameters can be deduced from the use of a BioAnalyzer on cDNA libraries generated for use in RNA-Sequencing experiments.

The inference of read library type is also an important consideration for both the Kallisto and Salmon tools during transcript quantification. ‘library type’ in this context predominantly refers to the strandedness of the reads deriving from cDNA fragments. Sequenced reads can match either exclusively the forward strand or the reverse strand of the cDNA

duplex (stranded protocol), or both (unstranded) depending on the cDNA library preparation and sequencing strategies. For paired-end read libraries, the reads may derive from both strands of the cDNA fragment, however, the order in which reads are derived from either respective strand can be used to designate the strandedness of the library. In addition, the relative orientation of pair-end reads with respect to each other (i.e. inward facing, outward facing, or matching) is another parameter which can be used to specify the library type.

2.12.3 Differential expression analysis

In order to compare gene expression values between a miRNA perturbation and a control condition, some form of differential expression analysis is needed. In qualitative or semi-quantitative methods of assaying gene expression such as Northern blotting, a rudimentary form of differential expression analysis can be performed by simply comparing the strength of signal emanating from corresponding bands in different Northern blot runs. More quantitative methods of differential expression analysis, for data deriving from microarray or RNA-Seq experiments requires more in-depth analyses.

When considering the difference between microarrays and next-generation sequencing (NGS) data, it is first important to consider the difference in the type of data generated by the two different technologies. The output of a microarray experiment is the fluorescence of sample molecules hybridised to their respective probes for each gene/feature of interest, which are continuous values. In contrast, NGS sequencing ex-

periment returns sequenced reads, which are aligned to transcripts/genes, providing discrete count data which can be used for downstream analyses.

Whilst there are a large number of differences between the two data types, differential expression analysis for both cases can be modelled using generalised linear models (GLMs) (Nelder and Wedderburn, 1972). However, because of the two different data types, the type of error probability distribution which can appropriately be used for GLMs for microarray and high-throughput sequencing experiments vary: Raw microarray fluorescence values can be modelled using a log-normal distribution (Hoyle, et al., 2002), whilst NGS sequencing data, can be fairly well approximated using the negative binomial distribution (Anders and Huber, 2010; Lu, et al., 2005; Robinson and Smyth, 2007):

Because RNA-sequencing involves the random sampling (with low probability) of reads for a given gene from a large set of reads for the total experiment, per gene read counts could potentially be modelled using a Poisson distribution – in which the expected value of the read counts and the variance of the read counts for that specific gene/transcript would be equal. And this is generally what is observed when examining the distribution of read count data for technical replicates (Marioni, et al., 2008). However, read count data from the biological replicates of next-generation sequencing datasets are generally overdispersed (Robinson and Smyth, 2007), meaning that the variance of the read counts exceeds the arithmetic mean read counts for the same gene/transcript. As a result, the observed mean-variance relationship

can be modelled using a distribution closely related to the Poisson, namely the negative binomial distribution. The negative binomial can also be interpreted as a Poisson-gamma mixture model in which the lambda rate parameter of the Poisson is a continuous random variable which is gamma distributed – producing a larger variance than what would be expected with a standard Poisson (Lipp, 2016). It has been claimed that the reason for the additional variance observed above that which would be expected with a single, non-compounded model is that different biological replicates being variable with respect to each other, cause slight changes in the parameterisation of the fundamentally stochastic Poisson sampling process (Lipp, 2016).

A commonly used method for differential expression analysis, and one that will be used for analyses discussed later in the thesis is the DESeq2 method (Love, et al., 2014). DESeq2 models read counts (*i.e.* the number of reads aligning to a given gene) using a generalised linear model, in which the distribution of reads counts for a given gene in a given sample is modelled by the negative binomial distribution, parameterised by the strength of expression of that gene and the variability of the expression of that gene. Gene expression strength in a given sample is estimated both using a normalisation constant relating to the size of a library in a given sample, and also a linear model of covariates thought to influence gene expression values (*e.g.* treatment conditions, batch effects *etc.*). From this model, using a matrix of raw gene feature counts (*e.g.* genes, transcripts, exons) as input, fold change parameters for designated covariates is estimated for each gene feature (*e.g.* the fold change of a gene between control and treatment conditions). In addition, null hypothesis significance testing is performed using the Wald test, to test for differential expression for a given covariate, by testing

whether the coefficient for the covariate (*i.e.* the logarithmic fold change parameter) differs significantly from zero – this in effect, is a test of whether a particular gene feature is differentially expressed or not.

DESeq2 uses empirical Bayes shrinkage methods in order to estimate dispersion and log-fold change parameters (*i.e.* β covariates) when fitting GLMs to each gene individually. The dispersion parameter is used to model read count variance for each gene of each sample, and is calculated by pooling information across samples and across genes (Love, et al., 2014):

$$\text{Var}(K_{ij}) = \mu_{ij} + \alpha_i \mu_{ij}^2$$

For K read counts calculated for i th genes in j th samples. $\text{Var}(K_{ij})$ represents gene and sample-specific read-count variance, μ_{ij} represents the gene and the sample-specific mean read count and α_i is the gene specific (but crucially not sample-specific) dispersion parameter.

For estimation of the dispersion parameter, an initial dispersion estimate is generated for each gene using replicate information from just that gene in order to generate maximum likelihood estimates (MLE). For all of these MLEs, a curve is fitted between gene dispersion and mean gene abundance, capturing the dependence of dispersion on mean gene expression. The fit is used as a prior for subsequent maximum *a posteriori* (MAP) estimates of gene-specific dispersion for each gene. In this way, genes of a similar mean abundance across replicates are assumed to possess similar levels of variability; information is shared

across genes, potentially overcoming the problem of uncertain dispersion estimates associated with estimating gene dispersion when experiment samples sizes are low. A similar approach is used to shrink log fold change estimates in cases in which there is a large degree of uncertainty when generating estimates (*e.g.* high gene dispersion or low sample number), reducing the probability of making spurious estimates of large log fold changes in gene expression. Such careful and methodical estimates of gene dispersion and log fold changes estimates are essential for the accurate investigation of the effects of miRNA perturbation on the regulation of gene expression.

Shrinkage of log fold change estimates helps overcome the inherent heteroskedasticity of log fold change data (*i.e.* the dependence of the variability of log fold change data on mean expression). This heteroskedasticity is a consequence of taking the ratios of count data (which occurs when calculating fold change), which produces largely variable and noisy results when counts are low (Love, et al., 2014).

2.13 Alternative cleavage and polyadenylation

A previously discussed benefit of using RNA-sequencing over other methods of RNA quantification is that it enables the reliable detection of transcript splice isoforms. Splice variants can impact the miRNA target predictions process, as differences in the primary sequence of transcripts can potentially lead to the gain and loss of predicted miRNA target sites. Another source of variation of the primary sequence of transcripts arising from a single gene derives from the phenomenon of alternative polyadenylation and cleavage (APA). The terminal point of

any 3'UTR sequence is determined by a co-transcriptional process in which the distal end of the 3'UTR of a nascent pre-mRNA is cleaved and subsequently polyadenylated. Variation in sites located on the nascent transcript for cleavage by APA machinery leads to the formation of transcript sequence isoforms.

Some elements of polyadenylation and cleavage, like splicing, occur co-transcriptionally. Transcription passes through the polyadenylation signal (typically 'AAUAAA') located on the 3'UTR, through to the transcription termination signal located on the DNA (Neugebauer, 2002; Proudfoot, et al., 2002). The polyadenylation signal is bound by CPSF (cleavage and polyadenylation factor), whilst a G/U rich region downstream of the eventual cleavage site is bound by CStF (cleavage stimulatory factor) (Neugebauer, 2002). The cleavage site is approximately 21 nt downstream of the polyadenylation signal and immediately upstream of the GU-rich region (Gruber and Zavolan, 2019).

Many different patterns of APA can take place (figure 2.9) (Gruber, et al., 2014). The simplest form of APA arises in cases in which different polyadenylation signals exist on the same terminal exon, leading to the formation of transcripts with the same patterns of exon usage, though with different 3'UTR lengths. Alternatively, alternative splicing events can lead to the selection of alternative terminal exons for the transcripts, leading to cleavage at an alternative polyadenylation signal. In addition, APA can also occur in introns in cases in which the activity of the APA machinery supersedes that of the splicing machinery, and also for polyadenylation signals located within constitutively expressed exons.

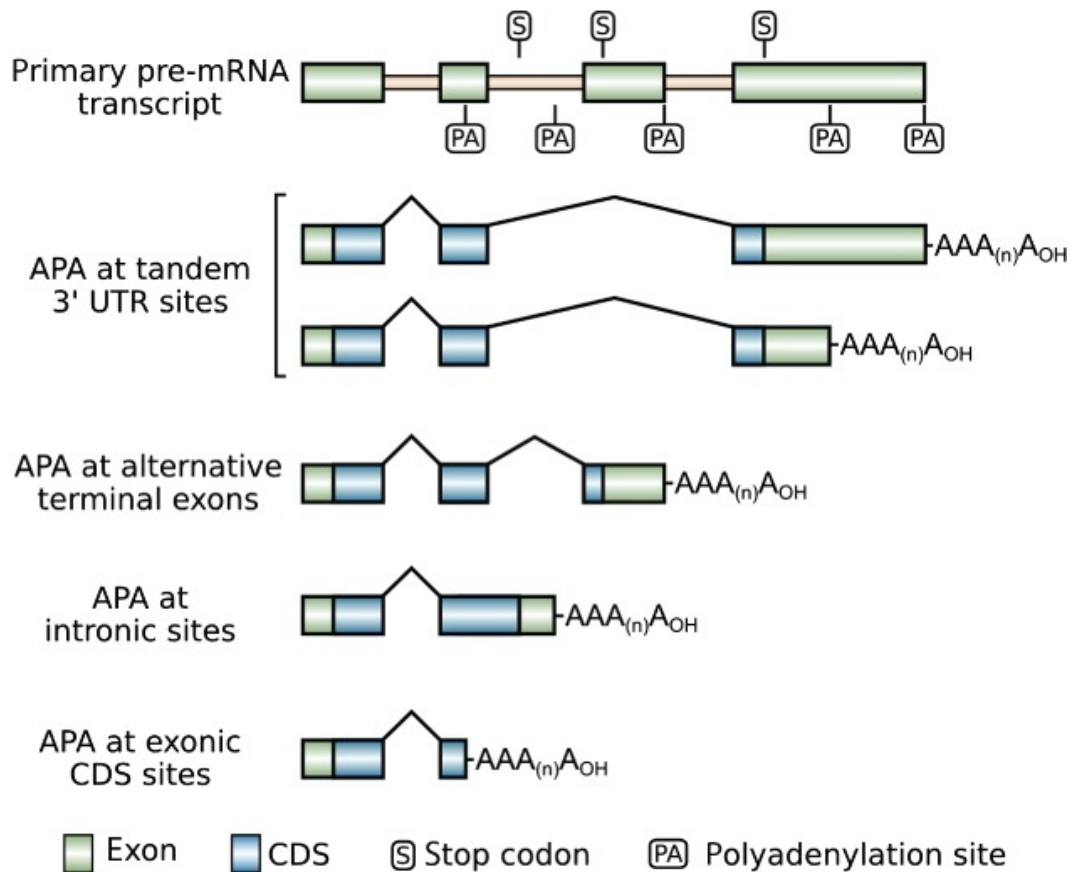


Figure 2.9 – The different forms of alternative polyadenylation and cleavage.

The most prevalent form of APA is in which alternative polyadenylation sites are used on the same terminal exons (i.e. tandem polyadenylation sites). In other cases, APA can occur at alternative terminal exons, some of which may contain coding sequence. In other cases, APA will occur at intronic sites – changing the primary sequence of the polypeptide produced from this mRNA transcript.⁵

APA also has implications for the targeting of mRNA transcripts by miRNA. Differential usage of polyadenylation sites in different biological contexts, can result in distinct 3'UTR isoform abundance profiles existing between different cell types (Nam, et al., 2014). As a result, some portion of the 3'UTR, and as a result some miRNA binding sites

⁵ Reproduced with permissions from Gruber, A. R., Martin, G., Keller, W., & Zavolan, M. (2014). Means to an end: mechanisms of alternative polyadenylation of messenger RNA precursors. *Wiley Interdisciplinary Reviews: RNA*, 5(2), 183-196.

may exist in some cellular contexts but not in others. Therefore, assuming identical 3'UTR profiles across biological context will likely lead to an inflation of both false negative and false positive miRNA target prediction results.

2.14 Combined target prediction and expression data tools and analyses

Throughout the course of this thesis, we will discuss analyses involving the combined use of both *in silico* miRNA target prediction and target expression data, however, there has been previous research conducted in this area with approaches proposed or tools deployed addressing the problem of how data of these two types can be combined in order to better understand miRNA activity for specific biological contexts.

There are a family of tools which address this problem using enrichment analyses. This approach can be useful in instances in which the user has prior knowledge of the miRNA or sRNA which is differentially expressed between two conditions. Sylamer is one example of this approach (Van Dongen, et al., 2008). Sylamer takes a list of genes ranked according to the magnitude of differential expression of those genes as a result of some form of miRNA perturbation (*e.g.* miRNA transfection or gene knockout), and then generates an enrichment (or depletion) profile for the k-mer complementary to the miRNA seed sequence across all genes, whilst correcting for 3'UTR length and compositional biases. For Sylamer, compositional biases are corrected using higher order Markov models in which the expectation of observing a given k-mer for a given bin is conditional on the identity of smaller sized k-mers

contained within that string (*e.g.* ‘AAAGT’ within ‘AAAAGTC’), and is parameterised as the background occurrence of the sub-word within that bin. This Markov model chain ensures that the enrichment of sub-words (which are not of biological interest) are not confounded with that of k-mers of specific length (*e.g.* miRNA seed sequences of a given length) which are of interest. It was shown that, using this approach, a transcriptomic signal for the knockout of a miRNA could be detected as a derepression (*i.e.* upregulation) of transcripts containing k-mers complementary to the seed of that miRNA in their 3’UTRs (Van Dongen, et al., 2008). This is further evidence that the regulatory effects of single miRNAs are not confined to a small number of targets, but can produce signals strong enough to be detected in transcriptome-wide analyses such as this. Other tools which also use enrichment analyses in order to link target predictions with expression data include miR-vestigator (Plaisier, et al., 2011) which utilises hidden Markov models for this purpose; miRonTop (Le Brigand, et al., 2010) which additionally allows the user to use multiple target prediction methods for the enrichment analysis and also interrogate the precise location of target sites on 3’UTRs, and also miRTrail which utilises a chi-squared test for overrepresentation analyses with subsequent pathway and network analysis (Laczny, et al., 2012).

Expression data cannot only be used to infer the regulatory potency of a given miRNA, but can also be used to infer individual sRNA-mRNA interactions. An example method of this type is FiRePat (Mohorianu, et al., 2012). FiRePat calculates the Pearson correlation coefficient between expression values of genes and sRNAs for a number of different sequencing runs, and then uses an unsupervised clustering method on

highly correlated gene-sRNA pairs in order to identify ‘putative networks of sRNA–gene interactions’ (Mohorianu, et al., 2012). In summary, for this approach, patterns of expression covariation are used to infer interactions between sRNA and mRNA gene pairs. However, for tools based on correlation analyses like FiRePat, whilst useful, such tools are only applicable when the users have data from multiple samples from different biological contexts. In addition, this approach assumes that regulatory relationship between sRNA-mRNA gene pairs are the same for different biological contexts, which may not always be true. Other tools perform a similar function to FiRePat, including MAGIA² (Bisognin, et al., 2012) which provides a choice of parametric and non-parametric measures to associate miRNA and mRNA expression datasets, as well as multiple algorithms for downstream miRNA target prediction, with subsequent inference of gene regulatory networks (GRN). In addition, MMIA (Nam, et al., 2009) also uses a covariation analysis to associate miRNA and gene expression data, but also includes information relating to disease states which have been associated with specific dysregulated miRNA levels. Other researchers have used expression data when defining one feature amongst many in a machine learning algorithm to predict miRNA targets, such as in TargetExpress (Ovando-Vázquez, et al., 2016), which predicts that expression of a gene in a given biological context is partially predictive of the ability of a miRNA to target that gene. With this approach, gene expression information is used to inform a larger, more extensive prediction model, but is not used as a filter for target prediction.

As well as applications to allow users to predict their own miRNA targets with the help of expression information, databases have been developed and released which provides users with pre-computed targets

of this type. CSmiRTar (condition-specific miRNA targets) (Wu, et al., 2017) is a database and web application which allows users to select miRNA targets (from a choice of four core miRNA target algorithms) which are expressed in specific tissues, and for specific disease states. In addition, the latest version of miRDB (Chen and Wang, 2019), allows users to implement a context specific expression filter for miRNA targets. Both of these databases provide a useful service to the miRNA research community. Limitations however, include, that CSmiRTar treats gene expression in different biological context as being binary, and therefore does not allow the user to implement a specific user-defined expression filter for targets. Conversely, miRDB, whilst allowing users to implement their own expression filter, only provides target predictions for the miRDB target prediction algorithm, and does not provide target predictions for other commonly used target prediction algorithms. In addition, both of these database report miRNA-gene predicted interactions exclusively, though miRNAs can more accurately be stated to act on mRNA transcripts rather than genes.

Despite these advances in the field, there is still the necessity for tools which enable the use of expression data to inform flexible use of target prediction workflows for a number of different prediction algorithms. In addition, there is also a need to allow the user to filter predicted results based on expression information, and in doing so, address the problem of the large number of false positive predictions made by existing target prediction algorithms.

Chapter 3: FilTar and FilTarDB design and development

3.1 Contributions

Simon Moxon: Initial idea of implementing an expression filter for miRNA target prediction, as well as the initial prototypical design of the web user-interface. Project supervision.

Leighton Folkes: Help with beta testing the completed FilTar tool

Thomas Bradley: Software design, development and programming of the FilTar tool. Design and development of the FilTarDB database. Design and development of the FilTarDB web application. System development and administration of FilTarDB system environment. FilTar and FilTarDB performance testing. Administration of FilTar and FilTarDB version control, user documentation, and online repositories. Origination and development of the idea of implementing 3'UTR reannotation for miRNA target prediction. Data analyses.

3.2 Introduction

FilTar is a combination of software utilities, existing in the form of a command line tool and a web application, in which RNA-Seq data is utilised in order to tailor miRNA target predictions for a specific biological context, such as a given cell type or tissue, and thereby increase prediction accuracy.

In this chapter, we will discuss the motivation for starting the FilTar project, aims specified before and during development, the design and final implementation of FilTar as a command-line application, and a database and web application (*i.e.* FilTarDB), both of which depend on the same core FilTar backend pipeline. In the following chapter we will discuss the analysis and interpretation of results generated using FilTar, including an examination of FilTar's prediction accuracy and the implications of those results for computational miRNA target prediction.

3.3 Motivation

As discussed in the introduction to this thesis, the aim of the PhD project more generally was to utilise bulk RNA-Seq high-throughput sequencing data in order to improve the accuracy of miRNA target prediction in animals.

FilTar represents the component of the broader PhD project in which we realise this approach in the form of an application or a collection of applications which is freely accessible and usable by members of the research community. My motivation for developing FilTar was to provide a means by which the general biological researcher could interrogate a database of animal microRNA target predictions which had been annotated, augmented or improved in some way using expression information. This would provide them with a more physiologically relevant and thereby accurate set of target predictions for their biological system of interest, in particular, reducing the larger number of false positive predictions commonly found in miRNA target prediction workflows. As miRNAs are believed to regulate a large number of key

developmental and physiological processes in most animal lineages (Bartel, 2018), providing researchers with a more accurate set of target predictions would be reasonably expected to aid investigation of these processes.

More specifically, expression information is used to remove lowly expressed mRNA transcripts from miRNA target prediction process. As both the miRNA and the target need to be expressed within the same biological context to interact, I predict that removal of non- or lowly expressed transcripts will likely lead to an increase in the specificity of the miRNA target prediction process. In addition, expression data is also used to generate context specific 3'UTR annotations. Alternative polyadenylation and cleavage events on the nascent precursors mRNA transcript lead to the establishment of 3'UTR isoforms within and between different cell types. Creating cell-type specific 3'UTR annotations is therefore expected to increase the accuracy of target predictions.

In addition, another aim of this project was to develop a tool which would allow more technically advanced users to apply the FilTar pipeline to their own RNA-Seq datasets, in order to generate target predictions of particular relevance to their own biological samples.

3.4 Aims statement

The aims statement for the FilTar project can then be stated as the following:

To provide some interface to biologist end-users, which would allow them to filter a database of computationally predicted canonical 3'UTR microRNA targets based on the expression of those microRNA targets within a given biological context of interest. Target predictions stored in the FilTar database are to be derived from 3'UTR models generated from a combination of the user's own sequencing dataset as well as reference 3'UTR models.

Secondly, to provide a command-line application for GNU/Linux operating systems which allows users to run the FilTar pipeline with locally stored RNA-Seq datasets.

3.4.1 Target user base

“...*biologist*...” is to be interpreted here as any user with standard knowledge and understanding of the basic principles of molecular biology, such as the relationship between DNA, RNA and protein, and how non-coding RNAs can be used by the cell to modulate these relationships. All graduates of standard biology bachelor's courses, and some existing undergraduate students would likely meet this requirement. Potential users motivated to use this tool but potentially lacking the prerequisite biological knowledge needed to navigate the user interface and understand returned results would likely benefit from consulting standard molecular biology textbooks, or published reviews in order to aid their understanding.

Bioinformatics knowledge is helpful but not required. In particular, as a major aim of the tool is the measurement of gene or transcript expression, it would be helpful if the user understood the methods and metrics used by FilTar when computing and reporting expression information. Briefly, it would be helpful to understand that the broad term ‘expression’ in this context refers specifically to a measurement of the relative abundance of a transcript within the context of which the RNA is sampled, and that this is measured in units of *transcripts per million* (TPM) (Li, et al., 2009). A brief explanation of the TPM unit would be provided as part of the GUI (graphical user interface). Nonetheless, even without an in-depth knowledge of the TPM unit, the user should still to be able to find the results generated from the tool to be broadly interpretable (in a manner which is biologically accurate) if they understand TPM as a scalar unit corresponding to the abundance of a transcript within a sample, with the caveat that the interpretation of TPM values compared between different samples can be more difficult (Pimentel, 2014).

By specifying a relatively low threshold of technical competence needed to use the tool, I hoped to increase the potential user base of the tool, and thereby increase the utility of the tool to biologists interested in animal miRNA biology.

3.4.2 MicroRNA Target Type

Computational microRNA targets (paraphrasing) refers to the precise

specification that this project be concerned with computationally predicted miRNA targets, and not miRNA targets which have been verified, predicted or inferred directly from experiments.

The reasoning for choosing to focus on computational miRNA targets for this project, is that because of the limited number of experiments performed to directly identify microRNA targets, computationally predicted microRNA targets would considerably increase the scope of applicable animal species and biological contexts for which FilTar could be applied and thereby increase the utility of the tool to miRNA biologists.

Canonically predicted microRNA targets (paraphrasing) refers to the fact that I only consider microRNA targets with a predicted target site containing full Watson-Crick base-pair complementarity to the seed region of the miRNA (Bartel, 2018), *i.e.* canonical miRNA target sites (Bartel, 2018).

There are a number of methods which have been developed which are able to identify non-canonical microRNA target interactions (Enright, et al., 2003; Gumienny and Zavolan, 2015; John, et al., 2004; Kiriakidou, et al., 2004). However, doubts have been raised concerning the functionality of non-canonical targets, even if they represent genuine binding events between the miRNA-AGO complex and the target molecule (Agarwal, et al., 2015). As the initial main aim of this project was to focus on post-processing steps (*i.e.* the filtering of predicted targets) after the identification of computational microRNA targets using

a core target prediction algorithm, it was decided to focus on high-confidence microRNA target predictions, and as a result, non-canonical target predictions were not considered.

Protein-coding (paraphrasing) refers to an exclusive focus on mRNA targets of miRNAs, and not the targeting by miRNAs of other non-coding RNA molecules, such as long non-coding RNA (lncRNA) or circular RNAs (circRNA). The competitive endogenous RNA theory (Denzler, et al., 2014) asserts that most ncRNA miRNA targets act as ‘molecular sponges’, modulating the number of unbound miRNA-AGO complexes in the cytoplasm which can bind and repress mRNA transcripts. However, the core targeting rules governing the targeting of ncRNA by miRNAs may be different than that governing mRNA-miRNA interactions, requiring different core prediction algorithms to predict ncRNA-miRNA targets, and as a result, examination of non-coding RNA targets of miRNAs was considered to be beyond the scope of this project.

The second component of this condition is that miRNA targets are to be restricted to the **3’UTR** only, and not any other features of the mRNA such as open reading frames (ORFs) and 5’ untranslated regions (5’UTR) despite it being known that miRNAs can target these regions of the mRNA (Reczko, et al., 2012). However, 3’UTR miRNA targets are the most effective and most abundant form of targeting (Bartel, 2009), and so 3’UTR targets are the focus of this project.

The last remaining ambiguous term in the stated aim is the term ***expression***. As mentioned previously, the term expression used in this context

refers to the relative abundance of transcripts within a given cell type or tissue measured in units of TPM.

3.5 Community needs addressed

The FilTar tools fulfils a number of key community needs. The major community need that it fulfils is to provide a software application that allows users to, within the same application: i) reannotate the 3'UTRs from protein coding transcripts using RNA-Seq data ii) perform miRNA target prediction on the sample specific 3'UTR annotations and iii) enables expression filtering of predicted miRNA targets.

In this way it addresses the need to perform context-specific miRNA target prediction. Within the general fields of developmental and molecular biology, at the time of writing, there is a current aim or need to increase the resolution at which different biological systems are examined *e.g.* the increasing emphasis on performing sequencing at single cell resolution. Although the FilTar tool does not enable miRNA target prediction analyses at the level of single cells, it does increase the specificity of analyses relative to the baseline or currently standard approach of conducting prediction analyses without specifying a cellular context.

By increasing the resolution at which target prediction is conducted, the accuracy of the analysis will likely increase, as context-specific miRNA targets are identified, providing the biological researcher with improved knowledge of their particular system of interest.

3.5.1 Current issues with existing software

The description of how FilTar will meet community needs exposes some limitations in existing tools which are used for the purpose of miRNA target prediction analyses. Some of these issues are described in more detail in the previous chapter, but just to briefly summarise: Existing tools do not allow users to perform miRNA target prediction on a set of reannotated 3'UTR transcripts specific to the particular biological context that they are investigating. As discussed previously, whilst some existing miRNA target prediction tools do incorporate expression information into the miRNA target prediction process, the accuracy of these tools has not been shown to equal that of current state-of-the-art methods for miRNA target prediction. FilTar crucially allows users to reannotate 3'UTRs and integrate expression information along with the use of the current existing best methods in miRNA target prediction.

3.6 General Design & Implementation

3.6.1 Workflow Management

Workflow management is performed using the dedicated *snakemake* (Köster and Rahmann, 2012) workflow management tool. Snakemake possesses many useful properties and features which are exploited by FilTar for the purposes of relatively simple and efficient workflow management, minimising cognitive overhead for both FilTar developers and end-users, for tool development and use respectively:

In brief, snakemake is a ‘target-based’ workflow management system, in which searches are performed for the absence or presence of *target* (*i.e.* destination file) in pre-defined paths to determine whether to complete a particular, discretely defined process or *rule*. This target-based workflow management approach leads to the implicit definition of directed acyclic graphs (DAGs) for target generation by the user whenever they define a set of inter-related snakemake rules. An implication of the target-based DAG approach is that each target file can be defined as the set of rules used to generate that target specifically, and recursively all rules used to generate the input files, which are needed to generate target files. The utility of this approach for FilTar developers and end-users is that the logic for generating specified destination files can automatically be scheduled and executed in the appropriate order, minimising any required manual work from users, and minimising the risk of human error. Secondly, in most cases, existing target files and intermediate files are not needlessly regenerated, minimising the inefficient use of available resources upon workflow execution. Additionally, the specification of wildcards substrings within the names of target files, means that there is a large degree of flexibility, generalisability and parallelisation in workflow execution. For example, FilTar commonly uses wildcards to generalise workflows and sub-workflows to different species, different biological contexts and different chromosomes.

Many other useful properties of using a workflow management system will become apparent as different aspects of FilTar design are discussed throughout this chapter.

3.6.2 General Schema

The basic, simplified schema for the FilTar backend pipeline is presented in figure 3.1.

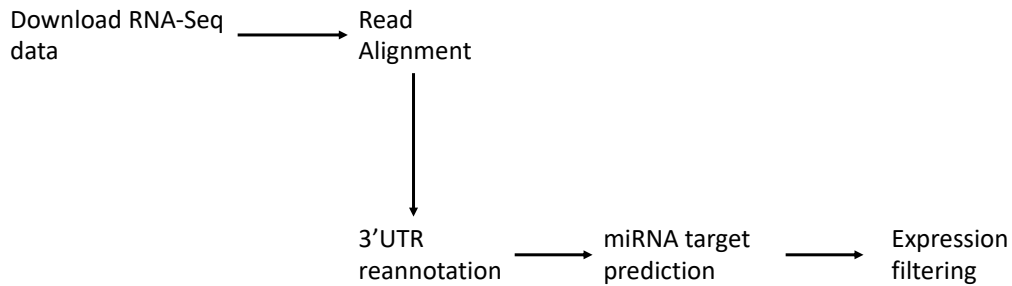


Figure 3.1 – A high-level overview of the FilTar workflow

These various steps and processes are managed using the Snakemake workflow manager. Figure 3.2 illustrates how Snakemake co-ordinates information relating to data, source code, configuration information and results. This schema relates the general structure of the FilTar repository, including files and subdirectories, and how those different repository components relate to each other in order to achieve a general function.

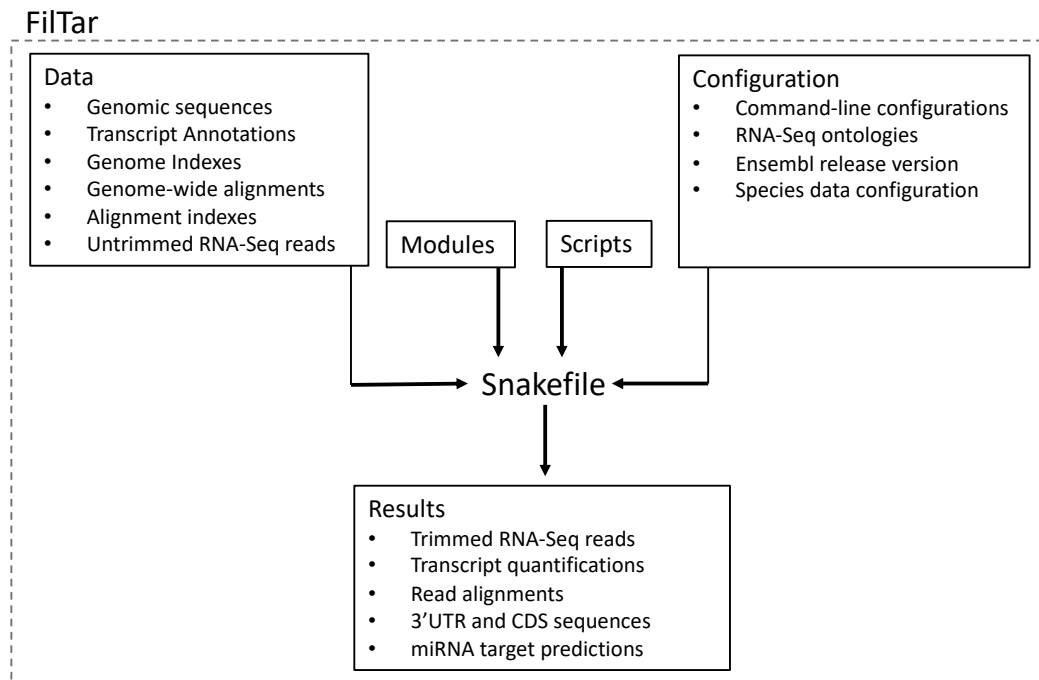


Figure 3.2 – Basic FilTar schema. The root snakefile acts as a central workflow management script and co-ordinates the activity of all subsidiary snakemake modules, represented by the ‘Modules’ box in the above schema. The data directory acts as a store of primary biological data sources, such as RNA-Seq reads and genome sequence files. The ‘Configuration’ directory contains specifications for the modulation of the behaviour of the FilTar pipeline according to user choice. The ‘Scripts’ directory contains a collection of scripts which are used by one or multiple modules. Black arrows represent an exchange of information between different FilTar components. All relationships are mediated and managed by the snakemake binary file.

As can be seen from figure 3.2, the repository is organised around a central snakemake workflow management script, called a *snakefile*, which is located at the root of the directory, and co-ordinates all other repository components. This includes raw data and script subdirectories which are required for the completion of some snakemake rules. This is mediated by the use of the snakemake binary which is able to read from user-defined subdirectories when performing workflow execu-

tion. The ‘modules’ and ‘configuration’ subdirectories, as shown in figure 3.2, are used to support and regulate workflow behaviour respectively, and require further explanation:

The data that the user has to supply to the pipeline is very simple. If the user wishes to analyse publicly available data, the user only has to supply the relevant ENA/SRA (*i.e.* European Nucleotide Archive/Sequence Read Archive) database run and sample accessions for their RNA-Seq datasets of interest. This data should be entered into a YAML configuration file contained within the FilTar repository. The user can choose to download data from either the SRA or ENA repositories.

For user-derived data, RNA-Seq data must be manually placed within the relevant ‘data’ directory within the FilTar repository. Users must assign their own identifiers to the RNA-Seq runs, and the samples from which the RNA-Seq runs derive, within the FilTar YAML configuration file. The filenames of the associated fastq data files must correspond with the assigned identifiers for that sequencing run.

There are some limitations to the FilTar tool in its current state. At the moment, FilTar only supports full analyses for human and mouse, including use of sequence orthology information when generating miRNA target predictions. FilTar is also supported for analyses of 20 other vertebrate species for which both Ensembl gene annotations and miRBase miRNA annotations are available. However, because orthology information in MAF format (multiple alignment format) is not

available for these 20 species, orthology information is not considered when generating miRNA target predictions for these species.

Compatibility with use of data deriving from non-vertebrate bilaterian species (*e.g. D. melanogaster* or *C. elegans*) would require substantial reformatting of non-Ensembl transcript annotation files to the format which is required by FilTar. This process is only advised for advanced users.

FilTar is not intended to be used as a general-purpose RNA-Seq quality control and analysis tool. There are many other tools which perform this function well (*e.g. (Davis, et al., 2013; Grüning, et al., 2017)*). As a result, quality control procedures for RNA-Seq datasets are not implemented as part of the FilTar workflow. It is strongly advised that users perform quality control on RNA-Seq datasets before using their data with FilTar.

3.6.3 Modular Design

The modules directory contains a number of subdirectories, each of which contains a snakefile at its root, and hence can be said to operate as a discrete snakemake *module*. The rules and internal logic of each module can be aggregated with that of other modules in order to generate larger workflows. The relationship between snakefile and snakemake module is recursive as some snakemake modules themselves contain subsidiary snakemake modules. This recursive relationship between snakemake modules and submodules is represented in figure 3.3:

This hierarchical arrangement achieves two core functions:

1) A clear, conceptual segregation of modules performing distinct, and clearly defined functions within the larger FilTar workflow, as well as a clear conceptual designation between modules performing top-level and subsidiary functions.

2) The top-level management of all subsidiary snakefiles by a single master snakefile, allows the operations of the entire workflow to be invoked by a single call to the snakemake binary at the root of the FilTar directory.

These two stated functions enable simple conceptualisation, development and maintenance of the FilTar pipeline, as well as ease of workflow execution for both developers and end-users.

FilTar

Snakefile

- Data
- Scripts
- Results

Modules

Module A

Snakefile

- Data
- Results
- Scripts

Module A.A

- Data
- Results
- Scripts

Module A.B

- Data
- Results
- Scripts

Module B

Snakefile

- Data
- Results
- Scripts

Module B.A

- Data
- Results
- Scripts

Module C

Snakefile

- Data
- Results
- Scripts

...

Figure 3.3 – The recursive relationship of snakemake modules and subsidiary modules exploited by FilTar for the purposes of efficient workflow management. In this schema, each module invokes the rules (contained within respective top-level snakefiles), data, results and scripts from subsidiary modules. Different FilTar modules possess a variety of recursion depths. An example of this type of

modular design is the read mapping module, which is contained within the 3'UTR reannotation module, as the mapping of RNA-Seq reads to a genome is only ever used within the context of 3'UTR reannotation when using FilTar.

Another implication of this schema, is that the FilTar pipeline possesses the properties of being easily *extensible* and *scalable*. The modular and recursive structure ensures for any new features integrated into the main project, additional modules can be added without disturbing the operation of other modules.

3.6.4 Module Configuration

Another important feature of the FilTar schema is the use of configurable options which can be used to select between different snakemake modules, in order to alter the behaviour of the pipeline. Examples of this would be switching between different modules on the basis of whether the user wishes to reannotate 3'UTR sequences or not, or based on the use of a particular miRNA target prediction algorithm. The general principle of utilising configuration information in order to switch between relevant modules is represented in figure 3.4

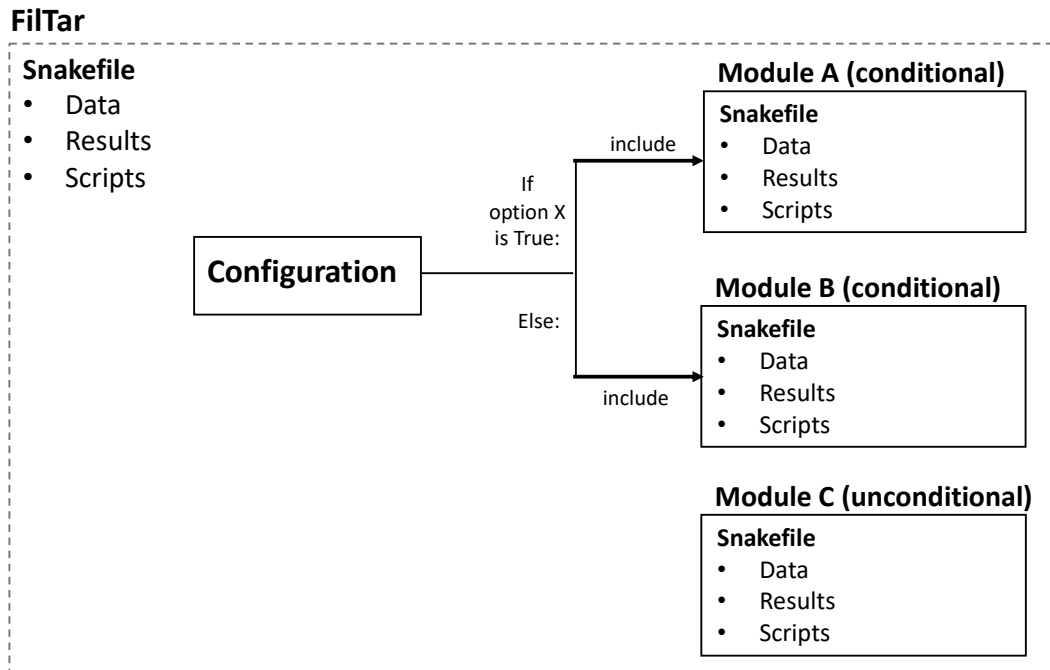


Figure 3.4 – The configurability of the FilTar pipeline architecture. End-users can configure pipeline behaviour using dedicated YAML configuration files and also directly when invoking the snakemake binary. This will cause the workflow manager to conditionally include some snakemake modules and hence determine pipeline behaviour. A specific example of this type of modular design are the sub-modules contained within the miRNA target prediction module, relating to each individual miRNA target prediction algorithm (see figure 3.5).

The utility of using configurable pipeline architecture such as this, is that it resolves any potential issues arising from conflicting snakemake modules by ensuring that conditionally irrelevant code is never processed when the pipeline is executed.

3.6.5 Modules Schema

Figure 3.5 is a schematic of the organisation and relationship between different FilTar modules:

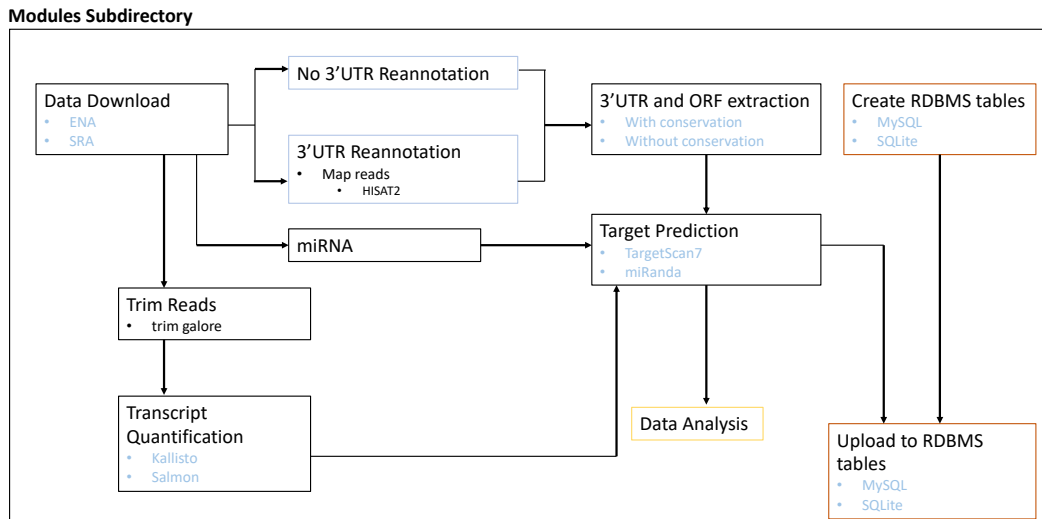


Figure 3.5 – Schema of FilTar modules. Black arrows represent the flow of information between different FilTar modules. Bullet points represent submodules contained within each module. Text or modules boxes highlighted in light blue represent an end-choice between two or more modules performing a similar function. Modules outlined in orange represent modules only contained within the ‘database’ branch of the FilTar relating to the FilTarDB database and web application. Modules outlined in gold are contained exclusively in the ‘validation’ branch of the FilTar repository in which extensive data analysis is performed in order to validate the FilTar’s methodology and approach to the problem of microRNA target prediction (see next chapter).

3.6.6 FilTar Modules and configuration

In the following section, a brief description of the functionality of each FilTar module will be provided:

3.6.6.1 Data Download

It is convenient to have a single module for the download of files needed for FilTar, for example, raw FASTQ files, cDNA files for transcript quantification, DNA files for read mapping *etc.* Within these modules,

rules relating to file download, file decompression and file pre-processing are connected via linked rule constructions. Data file pre-processing is used to remove any records from raw data files, which were not relevant for the larger FilTar workflow, for example the removal of records detailing mitochondrial transcripts within GTF annotation files. The aggregated effect of all the rules within the modules is to generate input files for core FilTar processing modules.

Utilities used for file download are *wget* and *rsync*, which respectively use *FTP* (File Transfer Protocol) and *rsync* transfer protocols. The *gzip* and *gunzip* utilities are used for file compression and decompression respectively. The specialised tool *fasterq-dump*, as part of the SRA toolkit (v2.9.6) is used for file download and decompression for RNA-Seq data stored at the Sequence Read Archive (SRA) (Leinonen, et al., 2010).

Concerning the download of raw sequencing data more specifically, it should be noted that it is optional for the user to either to choose to download data from the SRA or the European nucleotide archive (ENA) (Leinonen, et al., 2010), by making use of configuration options directly addressing this issue. The use of providing a choice to the user for data download, is that they can choose between download services which are faster or alternatively more robust.

Alternatively, if the user chooses to provide their own locally stored data files for processing, then this module is not used by FilTar as no data needs to be downloaded.

3.6.6.2 miRNA processing

The miRNA module enables filtering of the miRNA FASTA records, downloaded from miRBase (release 22) (Griffiths-Jones, et al., 2006; Kozomara, et al., 2018), in order to select miRNAs of the correct species, and also the user-selected miRNAs of interest. As miRBase is known to contain some incorrectly annotated miRNAs (Ludwig, et al., 2017), FilTar gives the user the option (set as default) to use miRBase's own high confidence set of miRNA annotations (Kozomara and Griffiths-Jones, 2014) when performing target predictions. Otherwise, users can opt to use the entire miRBase set of annotations for a given species for a more comprehensive, although, potentially less accurate, analysis.

This module is relatively simple as most miRNA rules are not contained within this module and are assorted into modules and sub-modules relating to specific miRNA target prediction algorithms, as different algorithms require different file formats for miRNA data.

3.6.6.3 Trimming of RNA-Seq Reads

Raw RNA-Seq reads undergo quality control and trimming using the *trim galore* tool (v0.5.0) (Krueger, 2015), a wrapper around *cutadapt* (v1.16) (Martin, 2011) which is run with default parameters, with the exception of the 'length' and 'stringency' parameters which were set to 35 and 4 respectively. Separate rules are defined for the processing of single-end and paired-end sequencing data using *trim galore*.

3.6.6.4 3'UTR reannotation

As may be expected, integrating this additional feature into the pre-existing FilTar pipeline required adjustment and re-organisation of existing code, and pipeline structure. In particular, there is an explicit bifurcation in workflow execution depending on whether the user chooses to reannotate 3'UTRs or not before the event of microRNA target prediction – which are respectively represented by the ‘with reannotation’ and ‘without reannotation’ snakemake modules. In the former module, RNA-Seq reads are mapped to the genome (GRCh38.p12 for human and GRCm38.p6 for mouse) using the splice-aware read-aligner *HISAT2* (Kim, et al., 2015) (v2.1.0). Using *HISAT2*, the location of exons and junction sites is determined by running the appropriate *HISAT2* scripts on the relevant species-specific GTF annotation file also obtained from Ensembl (release 94). The ‘hisat2-build’ binary is executed using the ‘--ss’ and ‘--exon’ flags indicating splice site and exon co-ordinates built from the previous step.

The indexed genome is used for FASTQ read alignment using the ‘hisat2’ command. The ‘rna-strandness’ option was used for strand-aware alignment. The strandedness of RNA-seq datasets is predicted using the ‘quant’ command of the *salmon* (v0.11.3) (Patro, et al., 2017) RNA-seq transcript quantification tool, by setting the ‘lib-type’ option to ‘A’ for automatic inference of library type. The samtools (v1.8) (Li, et al., 2009) ‘view’ and ‘sort’ commands were used to sort data from sam to bam format, and to sort the resultant bam files respectively.

Sorted bam files were converted to bedgraph format using the ‘genomeCoverageBed’ command of bedtools (v2.27.1) (Quinlan, 2014; Quinlan and Hall, 2010) using the ‘bg’, ‘ibam’ and ‘split’ options. Bedgraph files representing biological replicates of the same condition were merged using bedtool’s ‘unionbedg’ command. FilTar then calculated the mean average coverage value for each record in the merged bedgraph file. Existing gene models were produced by converting Ensembl GTF annotations files into genePred format using the UCSC ‘gtfToGenePred’ binary, and then from genePred format to bed12 format using the UCSC ‘genePredToBed’ binary (Kent, et al., 2002).

Alignment files from this mapping are then utilised by *APAttrap* (Ye, et al., 2018), the 3’UTR reannotation tool, using the ‘identifyDistal3UTR.pl’ perl script with default parameters, along with pre-existing reference transcript annotation files, in order to reannotate 3’UTR sequences. In brief, *APAttrap* implements 3’UTR reannotation using a ‘sliding window’ model. In this model, the annotated 3’UTR of transcripts with single-exon 3’UTRs is first considered; this initial space at the end of the pre-existing 3’UTR annotation is by default extended by 10kbp unless this extension hits upon a downstream gene. A sliding window (default size: 100bp) is then slid across this newly created space in 1bp increments in the 5’->3’ directions starting from the end of the existing 3’UTR annotation. A prerequisite for 3’UTR reannotation is that the first 100bp of this extended 3’UTR exceeds a mean coverage threshold of 10 reads – in order to ensure the transcript has been sampled at sufficient depth for 3’UTR reannotation. The sliding window traverses the extended 3’UTR space using *while* logic; the sliding window stops once less than 80% (default value) of bases in the sliding windows fails to exceed a coverage of 5% (default value) of the

mean coverage of the first 100bp window. For example, if the mean coverage for the first 100bp window was 30, then the sliding window would cease ‘sliding’ once the coverage of at least 21bp (using default parameter values) in the sliding window had a coverage of 0. At this point, comparisons are made between the existing window and the following window in order to prevent the calling of erroneous 3’UTR annotations from regions of locally poor coverage on the 3’UTR: It is required that two consecutive windows fail the criterion described above. The exact 3’UTR end is given as the first nucleotide in the sliding windows which fails the coverage criteria.

Custom scripts are then used in order to integrate novel 3’UTR annotations with pre-existing 3’UTR annotations (for transcripts in which the 3’UTR was not reannotated) in order to generate new annotation files. In the case in which the user chooses not to reannotate 3’UTR sequences, pre-existing GTF annotation files are used. Only truncations or elongations of single exon 3’UTR annotations were integrated into final 3’UTR annotations; novel 3’UTR predictions (*i.e.* prediction of 3’UTRs for transcripts without a previous 3’UTR annotation) were discarded and alterations of the 3’UTR start site were also not permitted, due to the reannotation of 3’UTR start sites by the APAtrap dependency as beginning at the start position of the final exon in standard Ensembl transcript models. No alterations to existing 3’UTR annotations spanning multiple exons were permitted, as this is not intended functionality of the APAtrap tool.

3.6.6.5 3’UTR and ORF extraction

The following module is responsible for deriving ORF and 3'UTR transcript sequences from genomic data, given transcript models contained within GTF annotation files. This sequence data is required for miRNA target prediction to occur.

At the base of this module there is a bifurcation in data processing on the basis of a user-configurable option, namely whether the user wants to obtain homology/conservation information for 3'UTRs or not in the form of multiple sequence alignments or whether to use single sequence 3'UTR models. Use of multiple sequence alignments will produce more accurate predictions using the core TargetScan7 algorithm, but with performance costs relating to data storage, memory usage and run time.

Multiple sequence alignments (MSA) are derived from 100-way (human reference) and 60-way (mouse reference) whole-genome alignments hosted at the UCSC genome browser (Kent, et al., 2002) generated using the *threaded blockset-aligner* (Blanchette, et al., 2004) stored in MAF (multiple alignment format) format. MAF files are indexed, and the relevant alignment regions corresponding to 3'UTR coordinates extracted using 'MafIO' functions contained within the biopython (v1.72) library (Cock, et al., 2009). For human MSAs, distantly related species, which are all fish species, are removed, due to the poor quality of their 3'UTR alignments with the reference genome resulting in 84-way multiple sequence alignments (Agarwal, et al., 2015).

If multiple sequence alignments are not used, single sequences are extracted from DNA files using relevant 3'UTR co-ordinates in bed format using the 'getfasta' command of *bedtools* (Quinlan, 2014; Quinlan

and Hall, 2010) with the ‘s’ option enabled. Custom scripts are used to process the output of this command in order to merge exon sequences, into a single contiguous 3’UTR sequence. Further scripting is required to convert miRNA and 3’UTR sequence and identifier information into a format which can be parsed by TargetScan algorithms.

Care is taken to ensure that strandedness information is accounted for, as well as the existence of multi-exonic 3’UTRs to ensure faithful representation of 3’UTR models. All sequence information is converted for compatibility with downstream target prediction algorithms.

3.6.6.6 RNA-Seq Data Ontology

There should be special consideration and thought about how FilTar processes and manages disparate RNA-Seq datasets for both use in the command-line tool and the FilTarDB database. In particular, the processing of RNA-Seq data has implications for how FilTar performs transcript quantification as well as 3’UTR reannotation.

In particular, when developing FilTar, there was a need to address the relationship between different RNA-Seq datasets, and how information from different RNA-Seq datasets could be integrated and labelled as to be of use to end-users and downstream applications, and the implications of this labelling for transcript quantification and 3’UTR reannotation.

The labelling of RNA-Seq datasets in FilTar corresponds somewhat with the hierarchical labelling of RNA-Seq data by the SRA and the

ENA. Namely, the BioSample designation (Barrett, et al., 2011) is used to group replicates of the same sample type under the same accession, and in addition, the run accession designation is used to label sequencing data relating to the same sequencing experiment.

FilTar integrates metrics derived from primary sequencing data by first grouping different sequencing runs which fall under the same BioSample, and secondly by combining BioSamples which are grouped under the same top-level label for a given species *e.g.* All human liver BioSamples would be grouped under the same label.

The precise ontology used dictates the manner in which data is merged and integrated. An arithmetic mean average method is used to determine average TPM values and base coverage values for individual sequencing runs falling under the same BioSample for the purposes of transcript quantification and 3'UTR reannotation respectively. The sample function is then applied recursively to determine appropriate average values for multiple BioSamples falling under the same top-level label.

3.6.6.7 Transcript Quantification

Transcript quantification is performed using an alignment-free 'pseudo-alignment' approach used by *kallisto* (v0.44.0) (Bray, et al., 2016). Human and mouse cDNA files were downloaded from Ensembl. cDNA files are indexed using the 'kallisto index' command with default parameters. Reads were pseudoaligned and relative transcript abundance values quantified using the 'kallisto quant' executable, using the 'bias'

option to correct for sequence-based biases. When kallisto was used with data derived from single-end RNA-sequencing experiments, 180nt and 20nt were used as required estimates of the mean average fragment length and standard deviation respectively.

Arithmetic mean average TPM values for each transcript are calculated using user-defined RNA-Seq data ontologies (see above section).

It is important to distinguish between two classes of aligners/pseudo-aligners which are used in FilTar. As discussed previously, kallisto and Salmon are used for the purposes of transcript quantification from RNA-Seq experiments. Because both of these tools, as discussed previously, do not conduct read alignment to the genome, or perform genuine alignments, they cannot be used for 3'UTR reannotation.

As a result, HISAT2, a splice-aware aligner is used for this purpose of 3'UTR reannotation whilst either Kallisto or Salmon is used for transcript quantification. A detailed discussion of the difference between these two tools follows:

As discussed previously, both Salmon and kallisto estimate fragment length distributions in order to process libraries containing cDNA fragments with a potentially great range of insert sizes. It is also important to note that both of these tools are able to process stranded and unstranded RNA-Seq libraries, as well as being able to distinguish between paired-end stranded libraries in which either the forward read or the reverse read is sequenced first.

When comparing the Salmon and kallisto tools, it is important to note that Salmon contains more information in its pseudo-alignments, including the orientation of pseudo-aligned reads – which allows Salmon to distinguish between paired-end RNA-Seq libraries in which reads have matching or differing orientations. As a result, it is possible for Salmon to automatically infer the read library type from an analysis of a relatively small number of pseudo-aligned reads, reducing the administrative burden on the end user. However, with Kallisto, the user must manually assign the library type when running the tool. It is possible to first use Salmon to infer the library type, and then use estimated library type when configuring Kallisto for use with a particular library or group of libraries.

It is also important to note that unlike kallisto, Salmon computes the conditional probability that each fragment derives from a transcript to which it maps, which allows it to estimate sample-specific parameters such as ‘positional biases in coverage, sequence-specific biases at the 5’ and 3’ ends of sequenced fragments, fragment-level GC bias, strand-specific protocols, and the fragment length distribution’.

Both Salmon and Kallisto perform similarly against state-of-the-art RNA-Seq genome alignment and quantification methods with substantially larger run-times to their computationally expensive methods of alignment and quantification. FilTar allows users to use either Salmon or Kallisto given their similar functions, and the broadly similar ‘quasi’ or ‘pseudo’ alignments methods they used in order to quantify relative transcript abundance from RNA-Seq data. The option is provided not

only due to the slightly different run-times, accuracies and bias correction models of the two tools, but also because of the varying APIs they provide to the user and the slightly different functionality of the two tools, as well as slightly different methods for reporting results, and logging metadata.

3.6.6.8 miRNA Target Prediction

The two core algorithms used for miRNA target prediction that the user can select from is the *TargetScan7 (v.7.0.1)* (Agarwal, et al., 2015) and the *miRanda (v3.3a)* (Enright, et al., 2003; John, et al., 2004) algorithms. The 3'UTR sequence data required for target prediction can either be provided as multiple sequence alignments or single sequences, with the former option enabling the computation of 3'UTR branch lengths and the probability of conserved targeting (P_{ct}) for putative miRNA target sites.

Each of these two options involves the pre-processing of upstream input files for use with either algorithm, as well as the post-processing of output files to remove records which did not exceed a given TPM threshold.

Generalised linear models are similarly used in TargetScan (with the special case of multivariate linear regression), in order to score the relative efficacy of predicted miRNA target sites using a metric referred to as the *context++ score*. However, whilst GLMs are used in both contexts to generally model gene expression outcomes, there is a particular difference in the sense that with DESeq2, log fold change is used as a

predictor variable coefficient – whilst in the case of TargetScan, log fold change is instead used as the (normally distributed) response variable.

More specifically, four different linear models are generated with TargetScan – one for each of the main canonical target site types (*i.e.* 8mer, 7mer-m8, 7mer-1a, 6mer). Each model is built from a linear combination of sequence-based, and contextual features – selected from using a hybrid strategy of identifying a superset of candidate features of interest from relevant literature, and then restricting this initial set to a set of 14 features – selecting the most informative features using the Akaike information criterion (AIC).

As the TargetScan model is comprised of four different linear models – one for each canonical target site type, then there are four distinct distributions of context⁺⁺ score values for each target site type. As can be seen in figure 3.6, each site type (7mer data has been pooled together due to their similarity) exhibits slightly different distributions, with slightly different means. Each distribution exhibits a strong floor effect indicating target sites which whilst possessing a canonical miRNA target site, is not predicted to be repressed by miRNA upregulation. The expected values of the 8mer, 7mer and 6mer context⁺⁺ score distributions are -0.27, -0.18 and -0.13 respectively. More negative values indicate greater repression using a logarithmic fold change scale.

Use of the *TargetScan7* algorithm involves reformatting of miRNA data, as well as the explicit automated chaining of TargetScan7 compu-

tations relating to target site scanning, 3'UTR branch length calculation, probability of conserved miRNA targeting, ORF lengths, ORF 8mer counts, affected isoform ratios (AIRs), as well as final context++ scores.

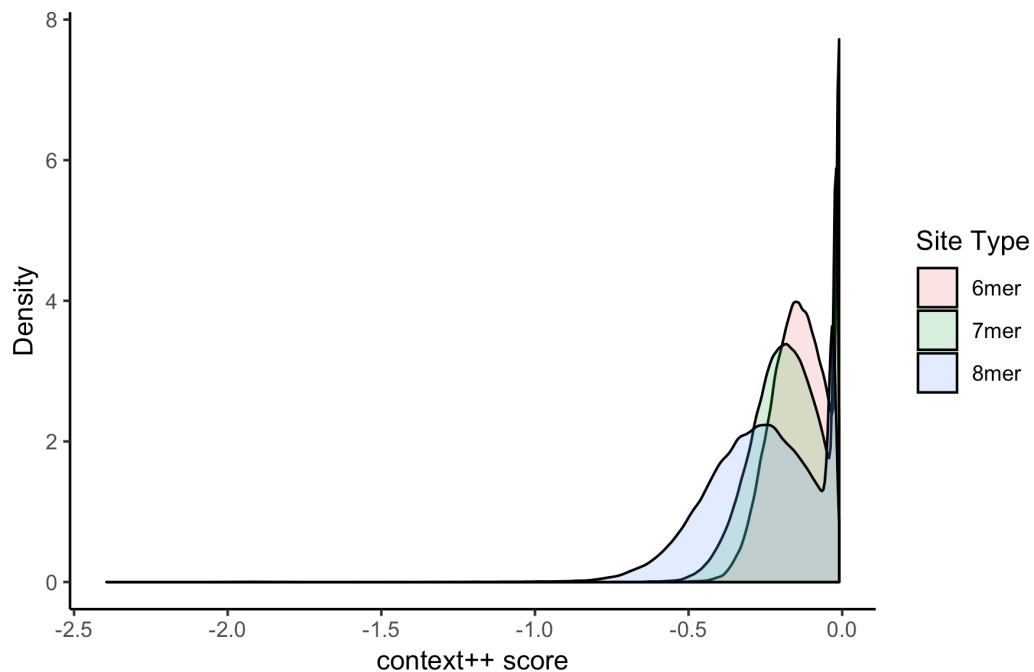


Figure 3.6 – The estimated distribution of the context++ scores of the TargetScan algorithm. Scores are taken from the official TargetScan web domain (*i.e.* targetscan.org). These scores are derived from an entire ‘all miRNA x all gene’ analysis for all human genes, and from conserved miRNA families from 84 vertebrate species. Expected values of -0.27, -0.18 and -0.13 for 8mers, 7mers and 6mers respectively.

Recalculation of AIRs by FilTar is of particular importance. As 3'UTRs isoforms exist within tissues, AIRs can be used to define 3'UTR isoform profiles for each annotated mRNA transcript (Nam, et al., 2014). However, the current set of 3'UTR profiles used with TargetScan7 are based on AIR scores derived from a small number of 3P-Seq (Poly-A position profiling) (Jan, et al., 2011) experiments conducted using only

four cell lines (Agarwal, et al., 2015). Using FilTar, we are able to post-process APAtap output to generate AIR scores, and hence distinct 3'UTR profiles for each biological context of interest, or indeed any potential biological context of interest to users.

A certain amount of caution is needed when interpreting the analyses of the cumulative distribution functions presented in this chapter. There are a number of potentially confounding variables to consider. Agarwal *et al.* (Agarwal, et al., 2015) identify three principal confounding variables when conducting these types of analyses: They demonstrated that in multiple transfection experiments, some of the mRNA which were perturbed upon miRNA transfection were unrelated to the transfected miRNA. Further investigation revealed that mRNA fold change was correlated to both the 3'UTR length and the AU content in the 3'UTR. In addition – Agarwal *et al.* in their analyses also identified a derepressive effect for the mRNA targets of miRNAs different to the miRNA perturbed in given experiments, potentially owing to an increase in competition for gene silencing protein machinery. Agarwal *et al.* also discovered the existence of batch effects potentially confounding observed mRNA fold changes, relating to different studies conducted by different laboratories and also relating to different transfection protocols.

FilTar provides users with the options of using either TargetScan7 or miRanda for miRNA target prediction. Target prediction is immediately downstream from the processing step in which 3'UTRs are reannotated.

The target score from the use of the miRanda algorithm represents the output of a dynamic programming alignment algorithm, in which alignment scores have a stronger weighting at the 5' end of the miRNA molecule reflecting the importance of the miRNA seed in targeting mechanisms. The expected value of the distribution of miRanda alignment scores (figure 3.7) is 147.1.

An expression filter is implemented on the results of miRNA target prediction in order to remove targets in which the predicted expression of the target does not exceed a given expression threshold. The target predictions, in the format corresponding to the respective tools with which the target predictions were generated, is then available to the user (in TSV format) with an additional column relating the relative transcript abundance of each transcript.

It is possible for users of the FilTar tool to take the union of results from TargetScan and miRanda in order to make more informed decisions about putative miRNA target interactions. One challenge with this approach is the difficulty in comparing results from two different predictions algorithms which score targets using different metrics. A naïve approach to standardising the two data types would be to take the Z-score of a given target prediction score for each prediction algorithm. However, this is only appropriate for cases in which both prediction score types are normally distributed. A more suitable approach for users would be to fit a model to each set of target predictions for each algorithm, and then the probability of observing a giving score can then be estimated from this model.

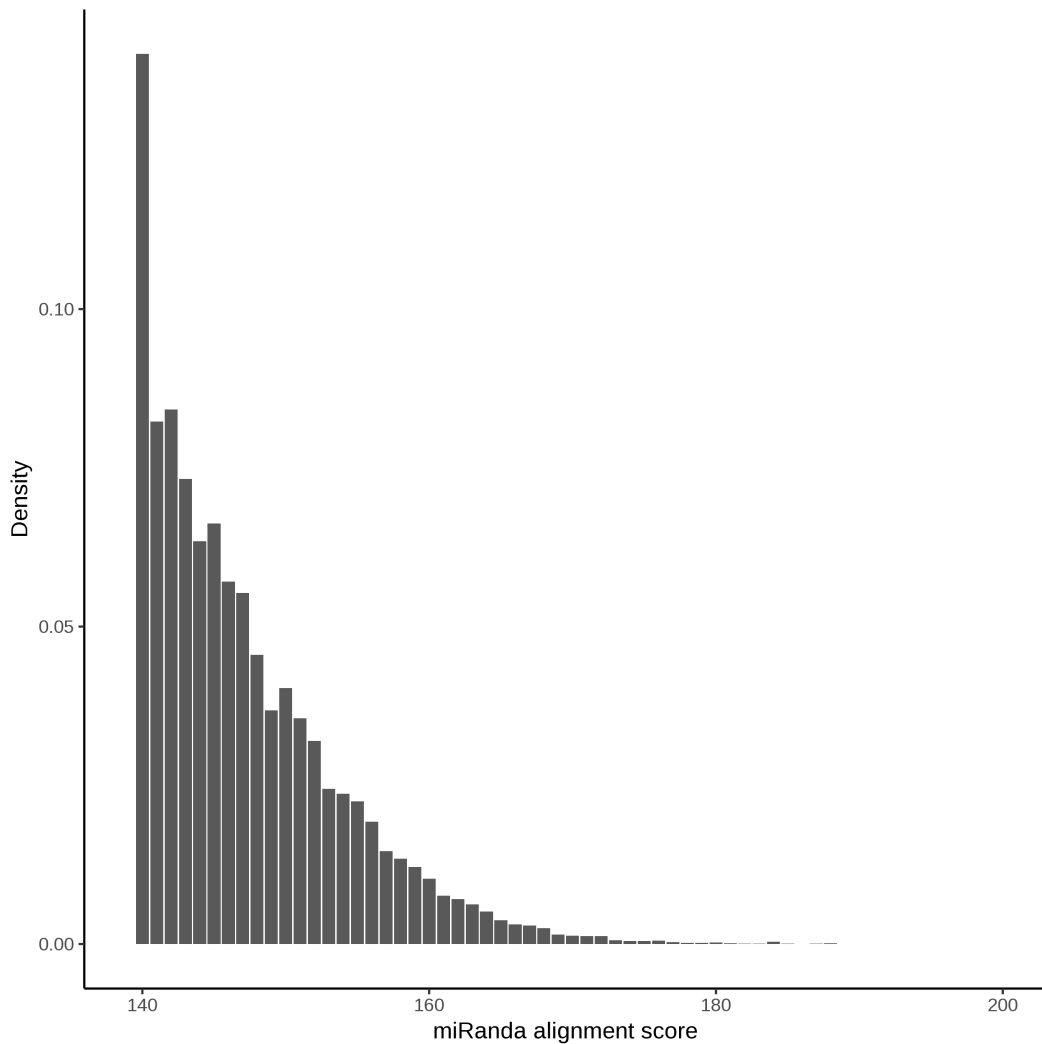


Figure 3.7 – The predicted probability mass function of miRanda alignment scores. Scores are taken from an ‘all miRNA x all gene’ target prediction analysis including all annotated mouse miRNAs and protein-coding genes. Expected value: 147.1

The probabilities would provide a common scale on which to compare scores outputted from both the TargetScan and miRanda algorithms. It would also be possible for users to take the intersection of target predictions using this approach for a more stringent, but potentially less sensitive analysis.

In summary, users can interpret the output of these target prediction algorithms together or individually. For TargetScan, there is a direct correspondence between the context++ score and the predicted logarithmic fold change of predicted targets as a result of miRNA upregulation. For miRanda, there is no direct interpretation of reported alignment scores, however, users can use score ranks or predicted score distributions in order to interpret the potential biological significance or reported alignment scores. This approach can also be used to interpret the relationship between TargetScan and miRanda prediction scores when taking the union or intersection of prediction scores from these two algorithms.

3.6.7 Dependency Management

The dependencies used by the FilTar application can be summarised in table 3.1:

<u>Dependency</u>	<u>Version used in FilTar</u>	<u>Software Type</u>	<u>Function in FilTar</u>
pigz	2.4	Command-line application	For parallel compression/decompression of FASTQ data files
SRAToolkit	2.9.1	Command-line application	For the download of FASTQ data from the sequence read archive (SRA)
bedtools	2.27.1	Command-line application	Extraction of ORF and 3'UTR sequences from transcripts models
biopython	1.72	Python library	Extract multiple sequence alignments from MAF files given a set of ORF and 3'UTR genomic co-ordinates
kallisto	0.45.1	Command-line application	Quantification of transcript read counts and relative abundance from bulk RNA-Seq data
salmon	0.13.1	Command-line application	Quantification of transcript read counts and relative abundance from bulk RNA-Seq data
miRanda	3.3a	Command-line application	miRNA target prediction by generating scores relating to miRNA-3'UTR alignment and thermodynamic stability
TargetScan	7.0.1	Multiple perl scripts	miRNA target prediction using a strict seed-pairing strategy. General linear models are implemented to score canonical seed targets of each type (i.e. 8mer, 7mer-m8, 7mer-1a, 6mer)
viennaRNA	2.4.9	Command-line application	A dependency for TargetScan. The RNAPfold utility as part of the viennaRNA suite of tools is used for scoring the structural accessibility feature of each putative target site
Trim Galore	0.5.0	Command-line application	A wrapper around CutAdapt. Used by FilTar for trimming RNA-Seq reads
HISAT2	2.1.0	Command-line application	Used by FilTar for splice-aware RNA-Seq read mapping. Alignments are used for 3'UTR reannotation
samtools	1.8	Command-line application	Used for converting SAM files (outputted by default by HISAT2) to BAM format, and then for sorting these BAM files.
UCSC-gtfto-genePred	366	Command-line application	Convert GTF files to genePred format – a necessary step in the 3'UTR reannotation pipeline
UCSC-genePredtoBed	366	Command-line application	Convert files in genePred format to bed format – a necessary step in the 3'UTR reannotation pipeline

APAttrap	Not specified	Multiple perl scripts	Used by FilTar for 3'UTR reannotation of previously annotated single exon 3'UTRs
Conda	4.6.14	Package and Environment Manager	Responsible for dependency management for the vast majority of dependencies used by FilTar
Snakemake	5.4.0	Workflow manager	Manages the entire FilTar workflow through linked rule constructions.
devtools	2.1.0	R package	Need to install the filter_R dependency
filter_R	0.1.0	R package	A library of core helper and general data manipulation functions for FilTar
CPANM	1.7044	Perl package management	For the installation and management of perl modules used by FilTar
Perl Modules: Bio::Perl, Statistics::Lite, Getopt::Long, Smart::Comments, experimental, List::Util, List::MoreUtils, Math::NumberCruncher, Exporter::Tiny		Perl Modules	A series of dependencies for perl scripts associated with the TargetScan and APAttrap utilities

Table 3.1 - A list of all dependencies needed to use the FilTar command line application

These dependencies will be discussed in greater detail in the remainder of this section.

3.6.7.1 Conda

Conda is the predominant method of dependency management within FilTar. Conda is a package and environment manager in which software packages can be accessed through conda *channels* which are distinct locations in which software are hosted, which are managed by conda

developers, individual conda users, or a community of conda users. A large suite of bioinformatics packages are hosted through a number of dedicated and specialist conda channels such as *bioconda* and *conda-forge*.

Conda can also be used to generate environments in which networks of dependent and co-dependent software can be bundled together and segregated from potentially conflicting software contained within the default user environment.

These features are utilised by FilTar to interactively download and install dependencies within rule-specific environments during FilTar workflow execution. The utility of this approach is that the installation process is fully automated, reducing the workload of end-users. In addition, the assignment of dependencies to individual rules prevents dependency conflicts between rules within FilTar.

3.6.7.2 The ‘filtar’ R Package

A dependency which is of fundamental importance to the main FilTar repository is the subsidiary ‘filtar’ R package which was also developed during the course of this project. This package contains a library of functions used by FilTar for the purposes of general data handling and manipulation. This dependency relationship is beneficial for FilTar as it results in a modularisation and abstraction of specific, data processing logic which occurs within scripts segregated away from the higher-level workflow logic which manages the relationship between scripts. This

is useful as it eases the development and maintenance of both data processing and workflow logic, and allows the relatively straight-forward invocation of discrete data processing functions at various locations within the FilTar source code. Storing user-generated R code within a separate R package also leads to benefits in terms of automated code testing and code documentation (see below), and also enables automated management of R package dependencies. This package is released as an open-source library, and is hosted on GitHub (https://github.com/TBradley27/filtar_R).

Not all core data processing logic is contained within this package. In some instances, custom python or shell code is used to perform data processing roles within FilTar. In these instances, such scripts are contained within the ‘scripts’ subdirectory at the root of the repository, or otherwise at the root of appropriate module subdirectories.

3.6.7.3 Miscellaneous

Perl modules are managed independent of conda using the CPANM (v1.7044) perl package manager. APATrap and TargetScan scripts are sourced from *SourceForge* online source code repository (sourceforge.net) and the TargetScan web domain (targetscan.org) respectively.

3.6.8 Automated Testing, Automated Building & Continuous Integration

Automated testing is a process by which the correctness of code is tested using other, external pieces of written code. This approach minimises the required manual supervision by developers or maintainers of software scripts and packages when testing the correctness of code, and is generally used to help minimise coding errors in software tools and packages.

Automated testing of the FilTar tool is performed using a combination of unit tests and integration tests. Unit testing is performed on functions of the `filtrar` R package using the *testthat* (v2.2.0) automated testing package. Integration testing is performed at the point in the FilTar workflow in which TargetScan7 microRNA target predictions are made by testing the correctness of FilTar-computed output against a reference value accessible via the official TargetScan website (targetscan.org - (Agarwal, et al., 2015)). As target prediction occurs at the end of the FilTar workflow, correctness of target prediction values entails correctness of all preceding processing steps.

Automated building is the process in which the workflow for building a software package is automated ‘from scratch’ so to speak in a clean environment, in order to ensure that a software package can be downloaded and installed in remote environments (with respect to the developers’ local environments) according to specified instructions and without additional unspecified dependencies. The process of automated building can be linked to that of automated testing by requiring an automated testing process at the end of the build, in order to ensure the correctness of the installation.

Continuous integration is the process by which changes made on development branches of a project, are, relatively speaking, frequently merged into production. The process of automated building and testing enables the process of frequent or continuous integration, by allowing developers to easily and rapidly test the stability of development branches, before those branches are merged into production – rapidly increasing the rate at which integration can occur.

Automated building and frequent, if not continuous, integration are enabled by use of the Travis CI (travis-ci.org) plug-in for GitHub. With this plug-in enabled, the stability of specified branches can be tested after every commit to GitHub-hosted remote repositories.

These processes help ensure the validity and correctness of FilTar with respect to specified aims, and enable the rapid development and integration of any future enhancements or additional features, greatly increasing the extensibility and utility of the FilTar tool for microRNA researchers.

3.7 Command-Line Application

3.7.1 User Interface

Although as mentioned previously, the FilTar command-line tool is built around the snakemake workflow management tool, the snakemake command-line syntax has been utilised in order to mimic, where possible the functionality of a stand-alone command-line utility. Using the command line snakemake ‘*-config*’ option users can specify a miRNA

of interest, or a gene of interest or also a species of interest. Users can also specify their preferred target prediction algorithm, and a particular biological context that they are interested in.

Additionally, users can specify generic options to the snakemake utility such as whether to execute a dry-run or to specify the number of cores that the user wishes to use. During a dry-run, a plan of workflow execution is reported to the user, but no data processing actually takes place.

3.7.2 Project Deployment, licensing & maintenance

The FilTar command line tool is deployed using the GitHub online repository (github.com) using GNU General Public License v3.0. This is a strong copyleft license which gives users the right to run, modify and share FilTar code under the condition that all derivative works are distributed using the same license. From GitHub, potential users of the FilTar tool can download any given release of the tool, as well as the latest development version. Users can also choose to create their own forks of the FilTar repository, in which they can make their own changes to the tool, and may choose to request that their changes are merged into the main FilTar repository via GitHub's 'pull request' feature.

These activities, enabled by the open source licensing of the FilTar project could potentially aid in the maintenance of the FilTar tool, as users could potentially identify any problems or issues with the repository. For example, if a server which FilTar relies on to source a dependency

is no longer functional, and this change is not detected during automated building which only occurs when new commits to the remote repository are made, then the hope would be that a user of the tool could detect this problem quickly and suggest a solution. Alternatively, if any problems do arise, users can alert the owner of the repository via the ‘issues’ feature of GitHub without suggesting a fix to that problem.

Having a direct mechanism by which repository owners can receive feedback from users is also beneficial in the sense that users can be helped and guided by repository owners or contributors in how to use the tool – which is a method by which a potential community of users can be helped and supported.

Version control, using git, is another feature which can be used to aid project management and maintenance. It allows project developers to easily track changes made to project code, through the course of the project history, and quickly switch and revert to different time points in developmental history.

Branching in the context of version control, refers to the creation of independent lines of development within the same repository. Branching is an important version control feature for the management of overlapping but different aspects of FilTar development. Separate branches used are branches for the FilTar command line tool, branches for the backend pipeline supplying data to the FilTar database and web application and also a branch for the analysis of data for the validation of the FilTar approach (see next chapter).

3.7.3 Documentation

Documentation of FilTar predominantly occurs at three different levels:

Inline documentation (i.e. ‘comments’): This form of documentation is used when a command is used, when the command is not expected to be intuitive to the casual observer.

Function documentation: Each function used either within the main FilTar repository or the subsidiary ‘faltar’ R repository is documented. This involves a statement of all required inputs of the function, expected output, usage instructions and usage examples, as well as a brief description of the purpose of the function.

Top-level documentation: This is a form of documentation which end-users can use as an instruction and a guide on how to download, install and use FilTar software. The documentation can be found at the following URL: <https://tbradley27.github.io/FilTar/>

3.7.4 Performance

3.7.4.1 Installation Time

FilTar can be installed in a relatively clean GNU/Linux operating system with python 3.6 pre-installed in approximately under 48 minutes (figure 3.8). The stated time includes the time necessary to install dependencies such as *gzip*, *miniconda*, *snakemake*, the *R* statistical and computing environment, and all other aforementioned dependencies,

and to perform a relatively simple run of the FilTar pipeline on a relatively small dataset without 3'UTR reannotation. However, this time does not include time necessary to download and install dependencies related to 3'UTR reannotation such as HISAT2.

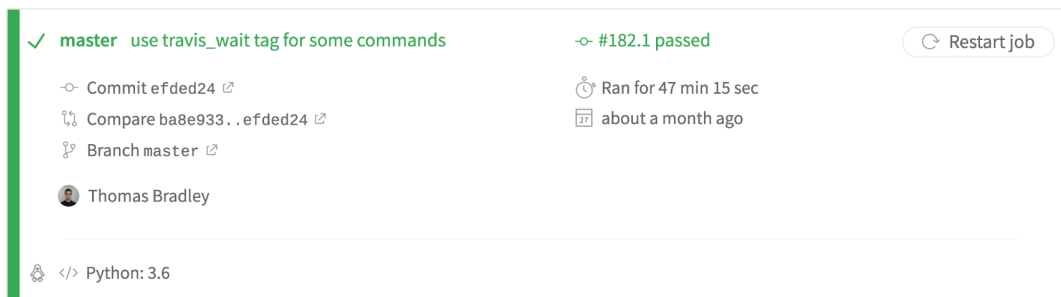


Figure 3.8 - Core installation duration for the FilTar command line tool. Installation time is determined by installing the application into a clean virtual environment with Travis CI. Automated tests are executed in order to test the correctness of the build. The application is built within an Ubuntu operating system with python 3.6 installed. Approximate installation time is 47 minutes.

3.7.4.2 Storage

The storage space requirements for the FilTar tool can be deconstructed into many components, such as space needed for source code, space needed for dependencies, and space for primary data required for FilTar to run. Some of these storage components may not be attributable exclusively to FilTar, for example, some users may already be running some dependencies on their operating systems. As a result, real storage costs are variable according to pre-existing user environments and also how the user intends to use the tool.

3.7.4.3 Performance Statistics

Source code: 4.1 Mb

Dependencies: ~1.2 GB

Data (excluding RNA-Seq):

Input Files:

- Without MSAs (default): 11 GB (mouse)
- With MSAs: 269 GB (mouse)
- Without MSAs (default): 14 GB (human)
- With MSAs: 790 GB (human)

Temporary/Intermediate Files:

- 240 GB

Output:

- Mouse: ~2 GB (all miRNA x all mRNA)
- Human: 2.4 GB (all miRNA x all mRNA)

Memory Usage:

- 40GB (default usage)
- 200GB (non-default usage)

It is important to note that for the memory usage for this tool, the non-default value of 200GB referenced above refers to the cases in which users decide to build their own splice-aware HISAT2 genome indices. Alternatively, for a limited set of species, users are able to download prebuilt genome indices (University). Otherwise, users can opt to build non-splice aware indices which requires significantly less memory (~40 GB).

An analysis of the run-time of FilTar has also performed (figure 3.9):

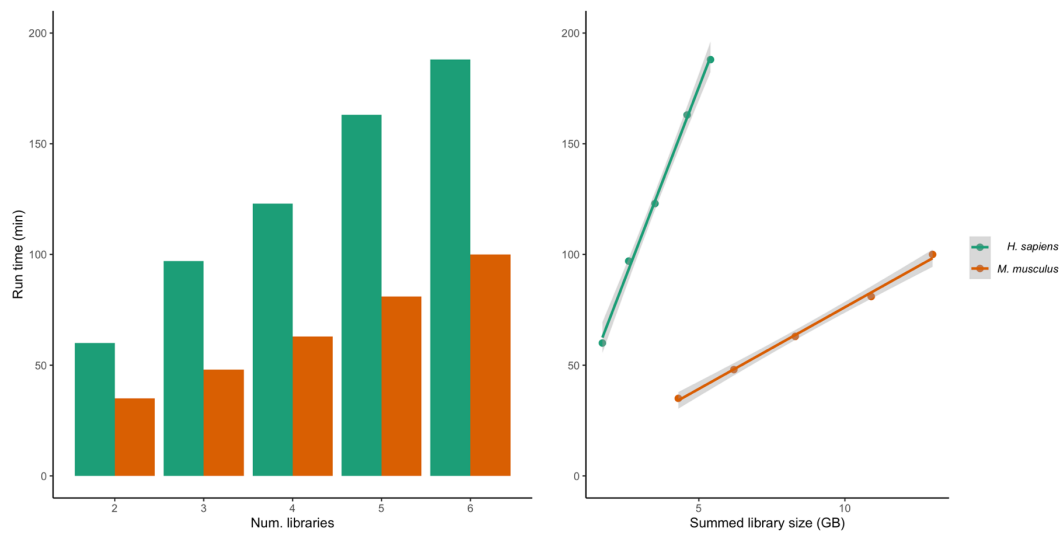


Figure 3.9 – The effect of library number and total library size on FilTar run time. Analyses are based on FilTar runs of a single miRNA (hsa-miR-188-5p or mmu-miR-188-5p) against all protein-coding genes of the corresponding species transcriptome using TargetScan. The time needed to index the genome, download the data and trim reads are not included in this analysis. This is because the user only has to index the genome once, and data download and read trimming are optional features of FilTar.

As can be seen from figure 3.9, the run time for FilTar for use with a single miRNA (excluding genome indexing, data download and read trimming), seems to increase linearly with the number of RNA-Seq libraries used in the analysis (left), with the caveat of this analysis potentially being confounded by the total amount of read information processed across all libraries (right). Human RNA-Seq libraries take longer to process than mouse libraries, which can probably be attributed to the fact of the larger size of the human genome, as well as the higher quality annotation of the human genome.

3.8 Web Application and Database

3.8.1 System Architecture

The web version of the FilTar tool, from here onwards, referred to as *FilTarDB*, was designed to be implemented as a basic modification of the standard LAMP (Linux-Apache-MySQL-PHP) web service stack, swapping the PHP component for the Django web framework.

GNU/Linux is the *operating system* used which allocates hardware resources to different software components, and generally mediates the relationship between hardware and software, and between different software components. It also provides a platform for which other stack components can be accessed and downloaded using networking protocols and package management systems, and also accessed via a file system.

The Apache HTTP server performs the role of the *web server* within the LAMP architecture. A web server is a piece of dedicated software for using established information transfer protocols (*e.g.* HTTP) for exchanging information between itself and clients on the World Wide Web.

The biologically relevant information that Apache serves to clients within this architecture ultimately derives from information stored within the MySQL (v14.14) relational database management system (RDBMS). The fundamental features of the relational database model are that data is organised into distinct data structures called *tables*, with

each *record* within each table being unique. Different attributes of the model are represented as table *columns*. Records of a data tables are uniquely identified using a *primary key* column. The relationship between different tables is established using a system of inter-relating columns, more specifically referred to as *foreign keys*. The RDBMS model when implemented carefully minimises data redundancy, and can be used to ensure data integrity within a database.

The Django (v1.11.7) web application development framework is used to build an application whose role it is to modulate, regulate and control the content sent to the client following a client request. Django achieves this using a model-template-view web framework. Within this framework, Django builds python data models derived from MySQL tables, modulates this data in dynamic response to user queries using the python scripting language, renders this information into HTML and sends this information to Apache, for Apache to then send this information to the client. These relationships are represented in figure 3.10. Although, the Apache web server performs the essential, but nonetheless technical tasks of receiving information from and sending content to the client, and the database acts as a store of biological information, the web application essentially performs the role of interpreting all client requests, and from this interpretation selecting an appropriate response to send to the client, via the web server. Within this architecture, Django uses the *mod_wsgi* (v4.6.7) python module in order to interface with the Apache web server, and the *mysqlclient* (v1.4.2.post1) module for interfacing with the FilTarDB database.

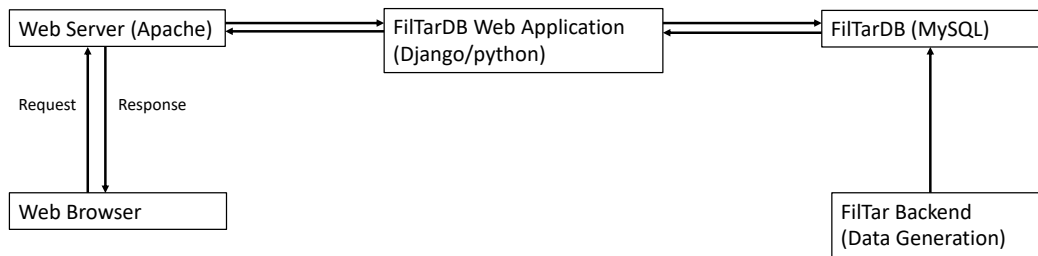


Figure 3.10 – The systems architecture of the FilTar web tool. The FilTar pipeline is used to generate data, which is deposited in FilTarDB, a database stored as a part of MySQL relational database management server. End-users can make requests upon the `filtradb.earlham.ac.uk` domain using web browser software and the HTTP protocol. HTTP requests are processed by an Apache web server associated with this domain. The Apache web server transmits the request to the FilTarDB web application, written using the Django web development framework. Logic within this application determines the correct response to a given HTTP request. The HTTP response is returned to the web server, which then, through web browser software, is able to serve the response to the end-user. In some instances, the FilTarDB web application will query the FilTarDB database in order to satisfy a given HTTP request.

3.8.2 Additional Backend Modules

In addition to all of the FilTar pipeline modules previously described and used for the command-line tool, the FilTarDB branch of the pipeline includes two additional modules which are specific to this branch:

3.8.2.1 Create MySQL Tables

This is a module for the creation of data tables within the FilTarDB MySQL database. Individual rules are devised for the creation of each table within the database, by sourcing relevant SQL files from the com-

mand line using the *mysql* command. Snakemake, figuratively speaking, cannot ‘look’ inside the database to test for table creation. As a workaround, empty text files with suitable file names are generated upon table creation, acting as a proxy for the completed table, and enabling the construction of an uninterrupted DAG workflow as part of the snakemake job scheduling system.

3.8.2.2 Upload to MySQL Tables

This is a module for the uploading of tabular formatted data in text files to relevant database tables in MySQL. Data is loaded using the *mysql* Unix command.

3.8.3 The FilTar Database

The database is designed in order to ensure a faithful representation of biological information, a minimisation of data redundancy, and low latency queries for FilTarDB end-users. Integer primary keys are used in order to ensure efficient computational search through tables with a large number of records. Unique keys are used in addition to primary keys for some columns, such as the name of a gene in a gene table, in order to ensure data integrity and to minimise redundancy. Compound unique keys are used for example in the target prediction table in instances in which the required level of uniqueness of records can only be encoded by multiple columns. A system of foreign keys (one-to-many column relationships) are implemented and enabled in order to ensure data integrity. The data types of each column are specified, and allocated space for each column value is selected order to minimise use

of storage space. A Many-to-Many relationship is formed between the *gene* and *species* tables, through an intermediary table, as an additional form of data normalisation. All relevant fields are indexed using the B-tree index type, to ensure low latency database queries. The FilTarDB database schema is represented in figure 3.11.

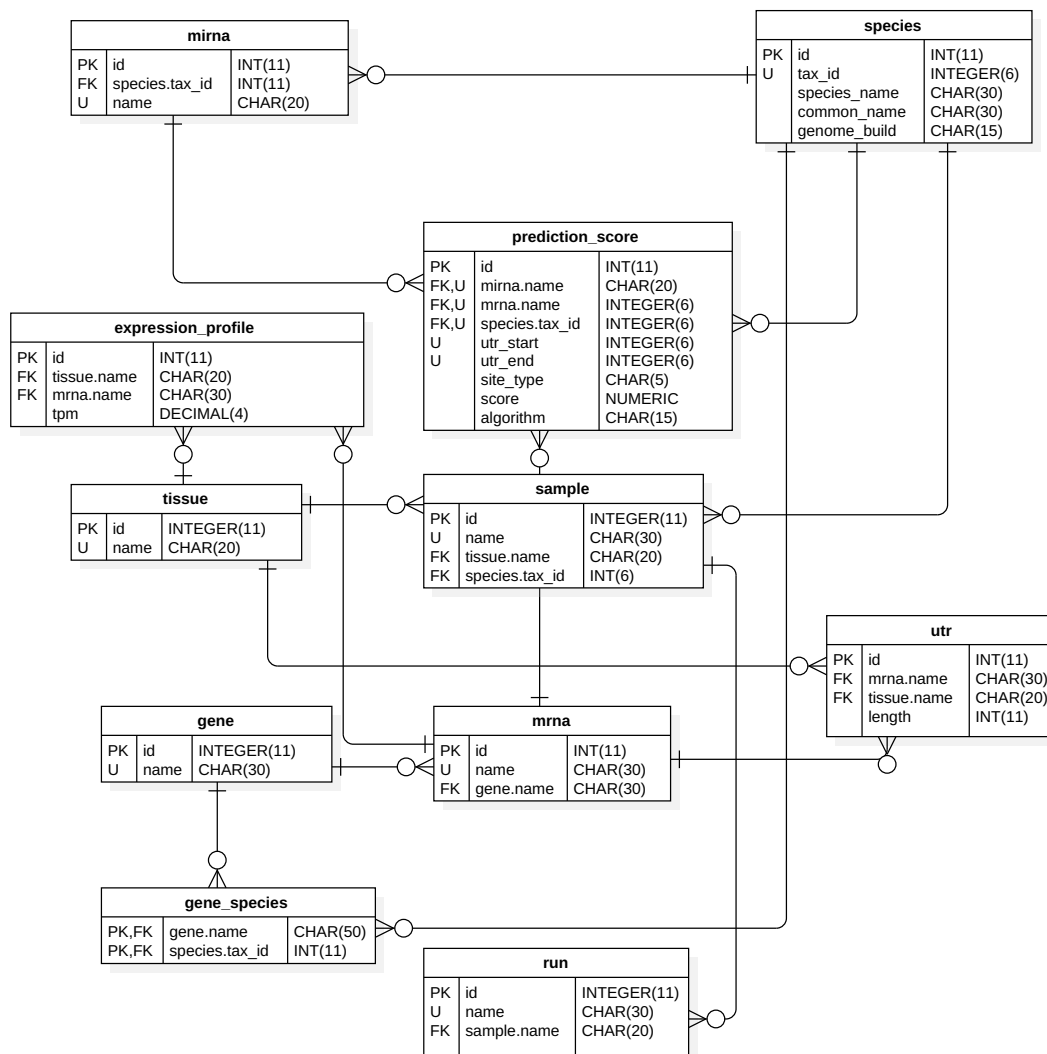


Figure 3.11 – FilTarDB database design. Each table in this diagram represents a table in the FilTarDB database. Each table record represents a column in the respective FilTarDB table. The first column of the tables in the diagram represents the key stage of the represented column (PK = primary key, FK = foreign key, U = unique key). The second table column denotes the name of the represented field. The third column denotes the represented field type. All primary key columns are auto-incremented. The five columns in the ‘prediction_score’ labelled with ‘U’ form a compound unique key over all of the labelled columns.

3.8.4 The FilTarDB web application

The FilTarDB web application exists as the scripting layer between the web server and the FilTarDB database. It is easiest to conceptualise the role of the web application by imagining a scenario in which the end user attempts to use the FilTar web tool:

From the FilTarDB domain name (`filtar.db.earlham.ac.uk`), the user navigates the FilTarDB website using a network of URLs. The relationship between the URLs entered by the user and content rendered by the web server is mediated by the web application. The *urls* document relates URL patterns and specific processing functions contained within the *views* document, which is the application document which mediates the processing of HTTP requests. Information from the *forms* and *models* documents are supplied to the *views* document. The *models* document contains representations of database data structures, encoded in the python language. The *forms* document utilises these models to create forms which are rendered through HTML templates, which allows end-users to construct their queries.

At the domain root, user requests are directed through the ‘home’ view function. Conditional logic is used to distinguish between HTTP requests of the type *GET* and of the type *POST*. Initial user requests are of the type *GET*. In this instance, information is imported from the *forms.py* document, and appropriate forms are rendered through a HTML template. Once the user has completed the forms, and submitted their query, the same home view function is invoked though this type using a *POST* request method. User form selections are subsequently

tested for their validity. After passing the validation test, form information is saved as session data, and the HTTP response, along with the user is redirected as a HTTP request to another URL. The urls document passes the request information to the corresponding view, which is the ‘results’ view in this instance.

Previously saved form session data is then invoked from within the results view. This function interrogates forms data in order to construct a string, which is used to directly query the MySQL database directly from python using the *Django.db* module. In all cases, the query requires a join of a target prediction table, the expression profiles table and the mRNA data table. The resultant data is stored in a named tuple structure, and is rendered through the appropriate HTML template, and served to the end-user. Importantly, this approach bypasses Django’s own models layer completely when querying database data, as the complexity of queries which could be constructed for Django data models was deemed to be insufficient.

The relationship between different components of the FilTarDB web application and the FilTarDB database and the Apache web server is represented in figure 3.12

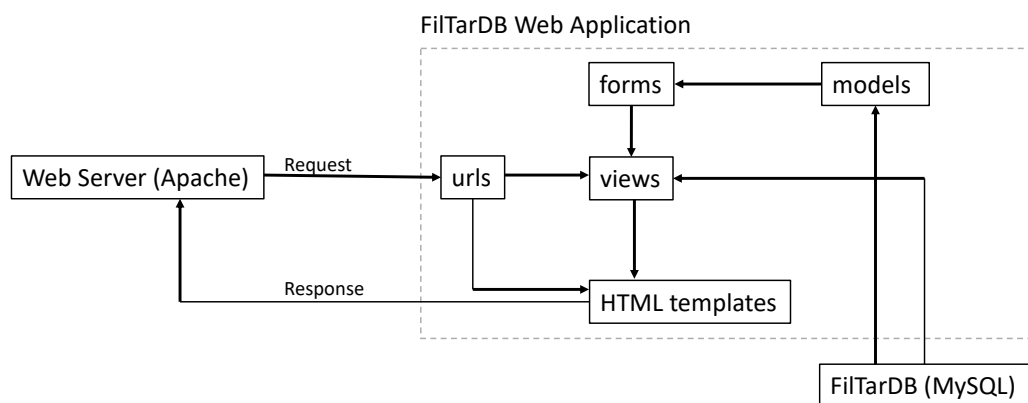


Figure 3.12 – The basic design of the FilTarDB web application. The FilTarDB web application is the scripting module used to determine how FilTarDB responds to requests from the web server. Requests are first interpreted by the *urls* module, which determines the appropriate response on the basis of the URL patterns of the incoming requests. For trivial requests, requiring static content exclusively, an appropriate HTML template is invoked directly, and the appropriate response is made to the web server. More complex requests are handled by the *views* module using python code. The views module makes use of forms from the *forms* module, enabling the entry of data by users on HTML web pages. Some forms are constructed from pythonic models of FilTarDB database data from the *models* module. The views model also queries the FilTarDB database directly, for more complex query types. An appropriate HTML template is selected after processing within the *views* module, and a response is sent to the web server.

3.8.4.1 Project Deployment, licensing & maintenance

The FilTarDB web application is hosted within a virtual machine on the CyVerse UK (cyverseuk.org) network, a service provided by the Earlham Institute (EI) National Capability in e-Infrastructure. The FilTarDB database itself is hosted within iRODS mounted storage space also managed by CyVerse UK. Content from the FilTarDB application is served via an Apache web server (v2.4.29) for the following domain name: filtar.db.earlham.ac.uk.

FilTarDB is currently released as a beta (v0.1-beta), and contains data relating to two biological species (Human and mouse), with five tissues or cell lines per species. Current target predictions and transcript quantifications stored within the FilTarDB database are for the gene and miRNA annotations associated with the 97th release of Ensembl. Maintenance of the FilTarDB project would involve recalculation of miRNA target predictions and transcript quantifications for updated gene and transcript models with each new Ensembl release. Maintainers of the FilTarDB project would also likely be interested in increasing the number of species and tissues/samples contained within the FilTarDB database, by running the FilTar pipeline on relevant datasets. Maintainers may also consider increasing the number of target predictions algorithms with which FilTarDB is associated, however, this may require considerable extension and development on the core FilTar application.

The source code for the FilTarDB application is released on GitHub (<https://github.com/TBradley27/FilTarDB>), licensed with version 3 of the GNU public license (GPL3). Documentation relating to the use of the FilTarDB application is available via the following URL: filtar.db.earlham.ac.uk/information.

3.8.4.2 User Interface

In contrast to the command-line tool, end-users of the FilTarDB web application interact with a GUI in order to retrieve information from a database of pre-computed miRNA target predictions (figure 3.13).

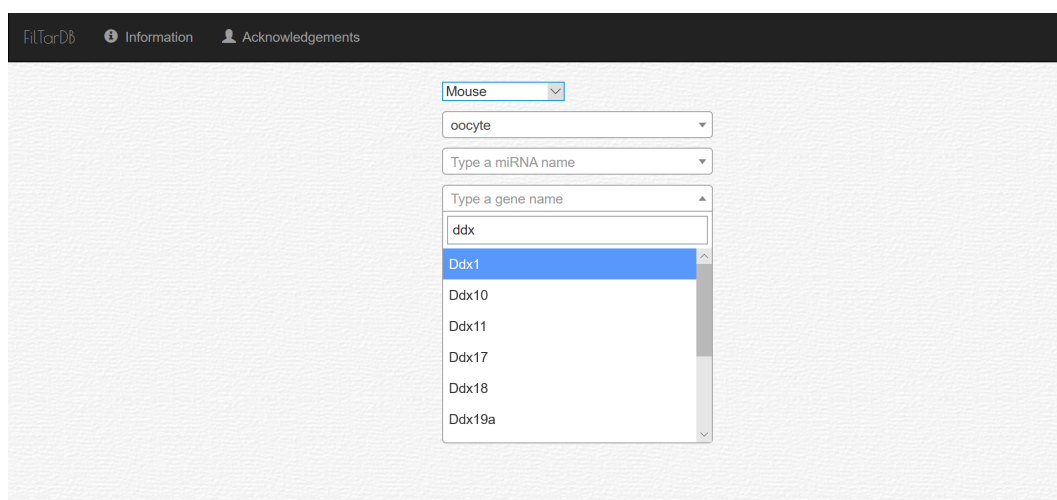
The screenshot shows the FilTarDB website interface. At the top, there is a navigation bar with 'FilTarDB', 'Information', and 'Acknowledgements' links. The main content area contains a query form with the following elements:

- A dropdown menu labeled 'Select a species'.
- A dropdown menu labeled 'Type a tissue name (required)'.
- A dropdown menu labeled 'Type a miRNA name'.
- A dropdown menu labeled 'Type a gene name'.
- A section titled 'Select a TPM Threshold:' with an input field containing '0' and a help icon '?'.
- A section titled 'Select one or multiple miRNA target prediction algorithms:' with two checkboxes:
 - TargetScan7
 - miRanda
- A 'Submit Query' button at the bottom.

Figure 3.13 – The home page of the FilTarDB website. The user specifies their query by completing a series of forms. First, they must select a species, then a biological context of interest within the second form. The user can choose to complete both or either of the miRNA and gene forms. The user then selects a TPM expression threshold for their query. Before submitting, users then have to select one or multiple miRNA target prediction algorithms. Instructions on how to use FilTarDB are available for the user to read at <http://fildatdb.earlham.ac.uk/information>.

The user query can be configured using a series of drop-down menus relating to species, miRNAs, genes and biological contexts which may be of interest to the user. The tissue, miRNA and gene fields (*i.e.* drop-down menus) chains from the top-level species field which means that available options for these respective fields are limited to those relating to the chosen species (*e.g.* you could not select a mouse miRNA for your query if you selected ‘human’ as your species of interest). This prevents the user from entering nonsensical queries, for example, requesting RNA-Seq data which does not exist in the FilTarDB database. Form chaining functionality is conferred via use of the Django Smart Selects plug-in (v1.5.4). Generic form functionality (*e.g.* form appearance, form scrolling) is conferred via use of the Django select 2 module (v7.1.0).

The tissue, miRNA and gene fields are also auto-complete fields, meaning that available field choices displayed to the user are superstrings of strings already entered into the field by the user (figure 3.14). As the user lengthens the entry field string by each character, the array of available choices to the user iteratively decreases, reducing the search space for the user, and easing recognition of intended target words. The utility of this type of field to users is that it allows them to quickly identify options of interest in fields which would otherwise contain thousands of possible options which the user would have to scroll through. FilTar autocomplete functionality is conferred via use of the Django autocomplete light plug-in (v3.2.10).



The screenshot shows the FilTar database interface. At the top, there is a navigation bar with 'FilTarDB', 'Information', and 'Acknowledgements'. Below this, a form is displayed with several fields. The first field is a dropdown menu labeled 'Mouse'. Below it is a text input field containing 'oocyte'. The next field is a dropdown menu labeled 'Type a miRNA name'. Below that is another dropdown menu labeled 'Type a gene name'. The text input field below this contains 'ddx'. A dropdown menu is open below the 'ddx' field, showing a list of gene names: 'Ddx1', 'Ddx10', 'Ddx11', 'Ddx17', 'Ddx18', and 'Ddx19a'. The 'Ddx1' option is highlighted in blue.

Figure 3.14 – Forms to be completed by the user exhibit field chaining and auto-complete functionality. Form auto-complete functionality restricts available form options to only those database entries which contain the user-entered characters as a substring. The tissue, miRNA, and gene forms all chain from the top-level species form – this ensures that only database entries specific to the relevant species is presented to the user. This minimises the possibility of the user generating spurious or ill-formed queries.

The TPM field is a numerical field, in which the user can choose to type in a given numerical value less than one million, or choose to positively increment the field value from a base of zero using an accompanying widget. An adjacent ‘help’ box can be clicked to give a brief description of the TPM unit.

Users select one or multiple prediction algorithms of interest from an unordered list of radio selection icons. Once this field and all aforementioned fields are completed then the user is ready to submit the query by interacting with the ‘submit’ widget.

Exception handling mechanisms are utilised to prevent the submission of illegal queries to the FilTarDB database (figure 3.15). Examples include the submission of queries without a selected species of interest, without a prediction algorithm of interest, or queries in which both the miRNA and gene fields have both been left empty. The blocking of the submission of ill-formed queries, helps to preserve resources with respect to the server and the client (*i.e.* end-users and end-user devices).

Figure 3.15 – Exception handling mechanisms prevents the user from submitting invalid queries. Exception handling procedures exist to ensure that the species, tissue and miRNA algorithm fields are completed. There is also a mechanism to ensure that the user completes either miRNA form or the gene form before submitting a query.

The HTML template returned to the user is variable according to the type of query they entered into their web browser, however a relatively large number of template features are featured irrespective to the type of user query entered (figure 3.16).

Transcript ID	Gene Name	3'UTR Start	3'UTR End	Site Type	Score	Average TPM
ENSMUST00000000080.7	Klf6	66	71	6mer	-0.028	0.00
ENSMUST00000000080.7	Klf6	432	438	7mer-m8	-0.020	0.00
ENSMUST00000000080.7	Klf6	2285	2290	6mer	-0.013	0.00
ENSMUST00000000122.6	Ngfr	845	850	6mer	-0.006	0.00
ENSMUST00000000122.6	Ngfr	637	642	6mer	0.000	0.00
ENSMUST00000000187.6	Fgf6	642	648	7mer-1a	-0.073	0.00
ENSMUST00000000187.6	Fgf6	2472	2478	7mer-m8	-0.020	0.00
ENSMUST00000000187.6	Fgf6	3583	3588	6mer	-0.017	0.00
ENSMUST00000000275.8	Gira3	146	153	8mer-1a	-0.188	0.00
ENSMUST00000000275.8	Gira3	4129	4135	7mer-1a	-0.010	0.00

Figure 3.16 – An example of a results page from the FilTar website. Relevant metadata is displayed above the results table. The user can toggle the number of

records they wish to view per web page. Widgets to iterate through result records are available below the table as well as widgets to download the data in a user selected format. Hyperlinks to external, relevant web domains are available for the user to select.

The template is divided in two segments. The upper segments act as a header with metadata relating to the data table displayed in the lower segment. Different metadata attributes can mostly be related to the initial query entered by the user – including the miRNA, gene, tissue and prediction algorithm(s) selected by the user. The values of some header attributes are hyperlinked to relevant web pages (*e.g.* names of miRNAs are hyperlinked to the relevant web page on miRBase).

As shown in figure 3.16, A series of buttons exist below the results table allowing the user to download the data in various formats. Once this data is downloaded, the user can use the main FilTar command line application in order to execute a function which joins genomic co-ordinate data to the results table, and calculates the precise genomic location of predicted miRNA target sites. Further information on how to execute the relevant functions to perform this operation is given in the official documentation for the FilTar tool.

In addition, a header entry not directly related to the user query is the ‘BioSample’ header attribute. This relates the number of BioSamples used from which expression information is aggregated, and relative transcript abundance values are computed. Specific BioSample accessions for used BioSamples are also provided, which hyperlink to relevant ENA web page entries.

miRNA: [mmu-miR-188-5p](#)
 Algorithm: [TargetScan7](#)
 Cell line/Tissue: [oocyte](#)
 BioSamples used: [1 - SRS540320](#)

Show entries

Search:

Transcript ID	Gene Name	3'UTR Start	3'UTR End	Site Type	Score	Average TPM
ENSMUST0000010941.5	Wnt2	417	422	6mer	-0.004	0.00
ENSMUST0000032180.6	Wnt7a	1212	1217	6mer	0.000	0.00
ENSMUST0000045747.4	Wnt4	478	483	6mer	0.000	0.00
ENSMUST0000067495.8	Wnt11	274	279	6mer	0.000	0.00
ENSMUST00000109424.3	Wnt7b	1038	1043	6mer	-0.025	0.00
ENSMUST00000167303.7	Wnt11	274	279	6mer	-0.017	0.00
ENSMUST00000167968.8	Wnt7b	1038	1043	6mer	-0.027	0.00
ENSMUST00000228546.1	Wnt10b	1427	1432	6mer	0.000	0.00
ENSMUST00000229495.1	Wnt7b	1038	1043	6mer	-0.027	0.00

Showing 1 to 9 of 9 entries (filtered from 9,682 total entries)

Previous Next

Figure 3.17 – The FilTar results data table can be searched using a search bar. When the search bar is used, only results records containing the search bar query as a substring of any of the available fields are displayed.

The second component of the results template is the data table. Each record denotes a particular target prediction record of interest given the user query. Each data column represents a discrete record attribute of potential interest to the user. Relevant attribute values are hyperlinked to relevant public database web pages. Data tables can be ordered by given fields by interacting with relevant column header names. All data tables returned exceeding a default value of 25 are paginated, with an accompanying widget at the right-hand side and below the data table, which can be used to select a relevant page of interest. The number of records displayed per page can be altered using a widget to the left of, and just above the data table. Information is displayed at the left-hand side of, and below the data table recording the total number of records returned in the data table, and the number of records currently displayed by the data table. A search field is used at the right-of and above the data table, which users can use to exclude records not containing any column values matching a given query string (figure 3.17). All returned data tables can be exported in plain text, excel, PDF and CSV formats

using appropriately labelled widgets at the left-hand side of and immediately below displayed data tables.

As mentioned previously, precise output template structure is variable depending on the nature of the user query. In particular, a gene name column will not exist in the data table if the user specifies a gene of interest, and similarly for when the user selects a miRNA of interest. There is a greater difference in template structure however, when the user selects multiple, instead of a single target prediction algorithm. Selection of multiple algorithms leads to the addition of an ‘algorithm name’ column in the data table denoting the appropriate prediction algorithm for each target prediction record. In addition, tables can be ordered by a particular column’s values ordered in ascending or descending order (figure 3.18)

miRNA: mmu-miR-188-5p
 Algorithm: TargetScan7
 Cell line/Tissue: oocyte
 BioSamples used: 1 - SRS540320

Show 10 entries

Transcript ID	Gene Name	3'UTR Start	3'UTR End	Site Type	Score	Average TPM
ENSMUST00000214948.1	Olfrl1385	1812	1818	7mer-m8	-0.028	564.86
ENSMUST00000214948.1	Olfrl1385	1610	1615	6mer	-0.018	564.86
ENSMUST0000024049.7	Bmp15	872	877	6mer	0.000	433.27
ENSMUST00000177943.7	Slc45a3	230	235	6mer	-0.021	330.39
ENSMUST00000221397.1	Nudt14	355	360	6mer	-0.046	322.33
ENSMUST0000027695.7	Slc45a3	230	235	6mer	-0.022	296.65
ENSMUST00000214364.1	Olfrl1039	1975	1980	6mer	-0.027	205.94
ENSMUST00000099547.3	Fam8a1	1907	1912	6mer	-0.018	190.52
ENSMUST00000101071.3	Tcd1	283	289	7mer-1a	-0.094	170.61
ENSMUST0000001513.7	Tubb6	259	264	6mer	-0.071	138.24

Showing 1 to 10 of 9,682 entries

Figure 3.18 – The ordering of columns of the results table. The results tab can be ordered by column for example to easily view the most abundant predicted miRNA targets, or those predicted miRNA targets with the greatest magnitude target prediction score.

3.8.5 Performance

FilTarDB performance can be assessed using a number of relevant metrics:

Number of species: 2

Number of biological contexts per species: 5

Number of miRNA target prediction algorithms: 2

Size of FilTarDB database: 23GB

Size of FilTarDB application & dependencies: 91 MB

Required Operating System: GNU/Linux family of operating systems

Query Speed: The query latency seems to follow a linear relationship with respect to the number of records from the database returned to the user (figure 3.19). In short, there seems to be a query latency of approximately 2.2 seconds per thousand records returned with the addition of a basal query latency of two seconds.

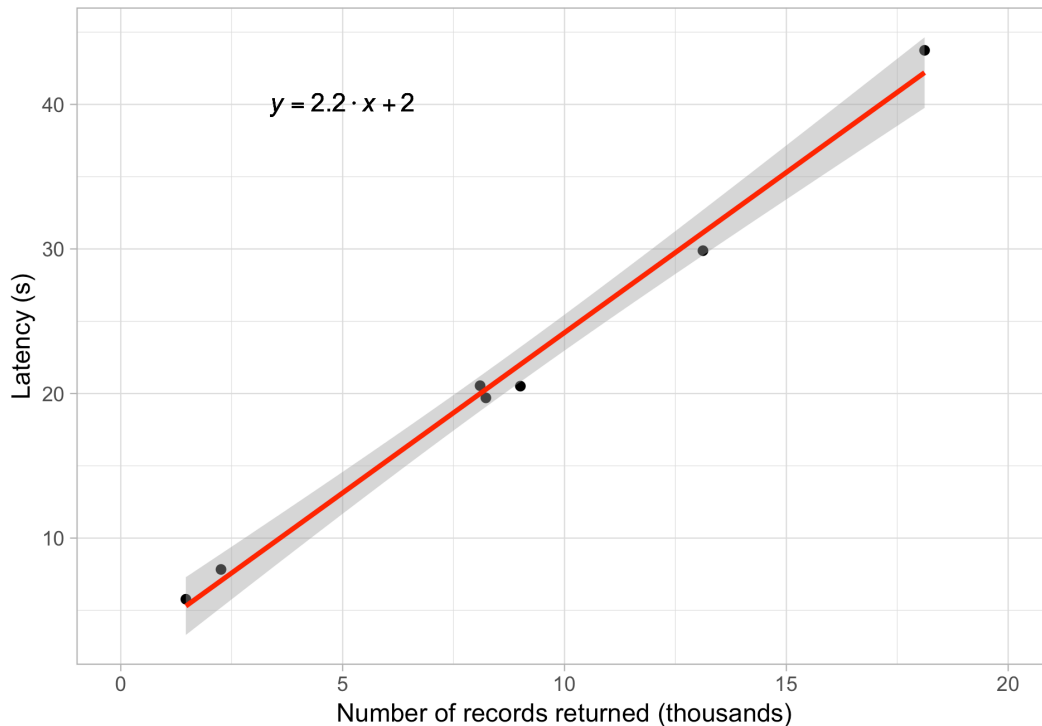


Figure 3.19 - Test of FilTarDB query latency. Tests were conducted on mouse miRNAs and transcripts within the ESC (embryonic stem cell) biological context. The expression threshold was set at 0 TPM. Target predictions were conducted using the TargetScan7 miRNA target prediction algorithm. Queries were performed using a specified miRNA, but without a specified gene. This relationship is assumed to be robust with respect to parameter choice.

3.9 Conclusion

In this chapter, I have described two distinct and complementary implementations, in the form of the FilTar and FilTarDB applications, of using the approach of utilising RNA-Seq data to hopefully improve the accuracy of miRNA target prediction in animals. Firstly, I have presented the FilTar command-line application, intended for detailed, thorough and investigative target prediction work by bioinformaticians. In contrast, the FilTarDB database and web application provides a means for users to utilise the FilTar approach by interacting with a graphical

user-interface allowing them to query and interrogate a database of pre-computed predicted miRNA targets, generated using the FilTar pipeline.

In the next chapter, we will explore the effect of FilTar on the accuracy of miRNA target prediction, and hence explore the potential utility of FilTar for users.

Chapter 4: Validation of the FilTar approach

4.1 Contributions

Simon Moxon: Initial idea of implementing an expression filter for miRNA target prediction. Project supervision.

Dagne Daskeviciute: Helped with a literature review to identify relevant miRNA mimic transfection studies and datasets which could be used for analysis

Thomas Bradley: Data selection, curation and quality control. Differential expression analysis. miRNA target prediction analysis. Additional analyses and data visualisation. Interpretation and discussion of results.

4.2 Introduction

In the previous chapter, I discussed the computational methods and processes involved in the design, development and implementation of the FilTar and FilTarDB applications. In this chapter, data analysis is performed in order to test the effect of using FilTar on miRNA target prediction accuracy. In order to benchmark the performance of the FilTar tool in a specific biological context versus general miRNA target prediction I used RNA-Seq data from miRNA mimic transfection experiments in mouse and human cell lines. Fold change values represent changes in relative mRNA abundance in samples transfected with a miRNA mimic compared to negative control samples transfected with non-miRNA nucleic acid molecules (*e.g.* plasmids).

As previously discussed, FilTar hopes to increase miRNA target prediction performance by implementing both pre-processing (3'UTR reannotation) and post-processing (expression filtering) steps with respect to the activity of a core miRNA target prediction algorithm; and as such, makes use of previously published prediction algorithms.

4.3 Methods

4.3.1 Data selection

For analysis of miRNA transfection experiments, FASTQ sequencing data generated from RNA-Seq experiments in human or mouse cell lines with at least two biological replicates were selected for further processing.

It was hoped that the selection of relevant datasets would allow the interrogation of the following questions:

- 1) Does expression filtering improve the accuracy of the miRNA target prediction process?
- 2) Does 3'UTR reannotation increase the accuracy of the miRNA target prediction process?
- 3) What is the effect of both expression filtering and 3'UTR reannotation on the total number of miRNA targets predicted?

Relevant datasets were selected using an unbiased procedure in which a literature search was conducted for publicly available RNA-Seq datasets in either mouse or human which derived from miRNA transfection datasets. No other selection criteria were applied other than this.

It is expected that samples transfected with a specific miRNA would lead to a reduction in expression of its target relative to the control sample. After differential expression analysis, if by inspection of cumulative plots the predicted miRNA targets could not be observed to be downregulated relative to non-target transcripts, then the transfection experiment was considered to have failed, and relevant datasets were not used for downstream analysis (Nam, et al., 2014; Polioudakis, et al., 2015; Zhang, et al., 2016) (figure A.2, table A.3).

A summary of datasets used with relevant database accessions is provided (table A.2) (Cao, et al., 2015; Diepenbruck, et al., 2017; Guo, et al., 2014; Liu, et al., 2017; Liu and Wang, 2019; Pua, et al., 2016; Stolzenburg, et al., 2016; Tamim, et al., 2014).

For subsampling experiments shown in figures 6 and 12 total reads were sampled using the seqtk tool (<https://github.com/lh3/seqtk>).

Specific assays are used for display in this chapter (*e.g.* figure 4.1) on the basis of the aim to select a sample set which included both human and mouse samples, in a diversity of different cellular contexts including both naturally occurring immortalised cell lines as would occur in stem cell populations – and immortalised cell lines deriving from tumorigenic or otherwise ‘cancer-like’ mutations.

4.3.2 Quality control and statistics

FASTQ data quality scores, GC-content, read lengths and similar statistics were generated using FASTQC (v0.11.5). Output from FASTQC was collated with data from the log files of other processes in order to produce a summary statistics report for each used BioProject using MultiQC (v1.6) (Ewels, et al., 2016) (table A.1). Information relating to the total number of reads for each library, as well as the number of mapped and pseudoaligned reads have been plotted (figure A.1). For each transfection assay experiment used in this project, the signal-noise ratio (see background chapter) was calculated along with the associated SNR reciprocal values (table A.4).

Considering all experiments together, arithmetic mean and standard deviation values for the SNR are 1.310 and 0.078 respectively. Evidence for the validity of this approach comes from considering experiments relating to U20S cell lines – in which the magnitude of the signal (including noise) and noise elements are shown to be approximately equal. This is concordant with a visualisation of the empirical cumulative distributions for the experiments (figure A.2) in which there is no observable difference in the cumulative proportion of downregulated transcripts (*i.e.* $LFC < 0$) between the two distributions.

4.3.3 Differential expression analysis

Differential expression analysis for miRNA transfection experiments was completed within the R (v.3.5.0) statistical computing environment (Team, 2013). Transcript-level read count data derived from RNA sequencing of miRNA mimic or negative control transfected cell lines were imported using the tximport package (v1.10.1) (Soneson, et al., 2015). Differential expression analysis on length and library size normalised read counts was performed using DESeq2 (v1.22.2) (Love, et al., 2014) comparing expression between negative control and miRNA mimic transfection conditions. Log₂ fold change values were subsequently shrunk using the default DESeq2 ‘normal’ shrinkage estimator (Love, et al., 2014) to account for the large uncertainty in predicted fold change values at low transcript expression values. For plotting, records corresponding to non-coding RNA transcripts were discarded. Transcript records were discarded when there was zero expression for all control and transfection replicates and fold change values could not be calculated. Target prediction data was used to label the remaining records as either predicted targets or non-targets of the transfected miRNA.

TargetScan is executed using both Ensembl 3’UTR annotations, and updated annotations produced using FilTar for the purposes of the differential expression analyses reported in this study.

For some differential expression analyses, null hypothesis significance testing was performed using two-sample, one-sided Kolmogorov-

Smirnov tests to test whether different fold change distributions were sampled from the same underlying distribution.

The effect size of the changes in gene expression between the mock transfection condition and the miRNA mimic transfection condition is assessed using *DESeq2*'s (Love, et al., 2014) log fold change metric. It is important to point out here that 'fold change' in this context does not denote a literal fold change as traditionally understood (*i.e.* the ratio between two point estimates of gene expression between two conditions). Rather, 'fold change' in this context refers to a beta parameter in the generalised linear regression model relating expression levels between different conditions.

An additional complexity when considering *DESeq2*'s 'log fold change' metric is that an empirical Bayes procedure is used to modify the maximum likelihood estimate of the fold change beta parameter (deriving from an initial round of GLM fits). A zero-centred distribution of the MLEs for all genes/transcripts is used as the prior distribution for this Bayesian procedure. In a further round of GLM fits, maximum *a posteriori* estimates are used to obtain a point estimate of the log fold change – which is taken as the mode of the posterior distribution of this Bayesian process. In the case of genes which are not differentially expressed, the expected value of the log fold change parameter would be zero.

The motivation for implementing a Bayesian approach in this instance is to 'shrink' log fold change estimators in cases in which there is too little information to make confident estimates – which could occur in

cases in which read counts are low (and is therefore associated with a large amount of uncertainty), there is a relatively large amount of variance in gene expression between biological replicates, or the sample size is too small. In these cases, the log fold change estimates will be shrunken towards the prior distribution (*i.e.* shrunken towards zero).

Volcano plots for each differential expression analysis conducted as part of this chapter can be found within figure A.3

4.3.4 Data Visualisation

All visualisations are produced using R's ggplot2 package (v3.1.0) (Wickham, 2016).

For figure 4.1, the filtered miRNA predicted targets curves represents protein-coding transcripts with a miRNA seed target site to the transfected miRNA mimic, which have been filtered at an expression threshold of 0.1 *transcripts per million* (TPM) (Li, et al., 2009).

For figure 4.5, the 'added seed sites' are identified as those transcripts which had not previously been labelled as predicted miRNA targets using target prediction results derived from existing Ensembl 3'UTR annotations, but had been identified as predicted miRNA targets using target prediction results derived from 3'UTR sequences reannotated using the FilTar workflow due to 3'UTR extension.

For figure 4.9, the 'removed seed sites' are identified as those transcripts which had previously been labelled as predicted miRNA targets

using target prediction results derived from existing Ensembl 3'UTR annotations, but had not been identified as predicted miRNA targets using target prediction results derived from 3'UTR sequences reannotated using the FilTar workflow due to 3'UTR truncation. Filtering for all groups occurred at an expression threshold of greater than or equal to 5 TPM. This was to reduce the number of false positive 3'UTR truncations (see discussion).

Additional plots for remaining datasets analysed are contained within figures A.2, A.3 and A.4 with the exception of cases where there was an insufficient number of added or removed target transcripts predicted ($n < 15$).

4.3.5 FilTar Implementation

All following steps were carried out using the FilTar tool. The workflow and parameters are described in detail below:

FilTar is a command line tool for GNU/Linux operating systems written predominantly in the python (v3.6.8) and R (v3.5.0) programming languages. Users can configure the tool to process available RNA-Seq datasets from public repositories such as the ENA (Harrison, et al., 2018; Leinonen, et al., 2010) and the SRA (Leinonen, et al., 2010), and also the user's own private sequencing data. All reported parameters are fully configurable within the FilTar tool. FilTar utilises Snakemake (v5.4.0) (Köster and Rahmann, 2012) for workflow management. Most FilTar dependencies are managed using conda (v4.6.6).

4.4 Results

4.4.1 Expression filtering

The first hypothesis to be tested was the hypothesis that implementing an expression filter for candidate miRNA targets would, as a whole, improve the accuracy of miRNA target prediction. Predicted miRNA targets filtered at $\text{TPM} \geq 0.1$ as a whole, exhibited stronger repression after miRNA transfection than the full miRNA target set without expression filtering. This is evident by the shift in the cumulative distribution for the filtered miRNA seed target set to proportionately more negative fold changes when comparing against the corresponding unfiltered set of transcripts (figure 4.1 and figure A.4).

Predicted miRNA targets removed by FilTar generally exhibited low absolute fold change values suggesting that these are false positive predictions in these specific cellular contexts (figure 4.2).

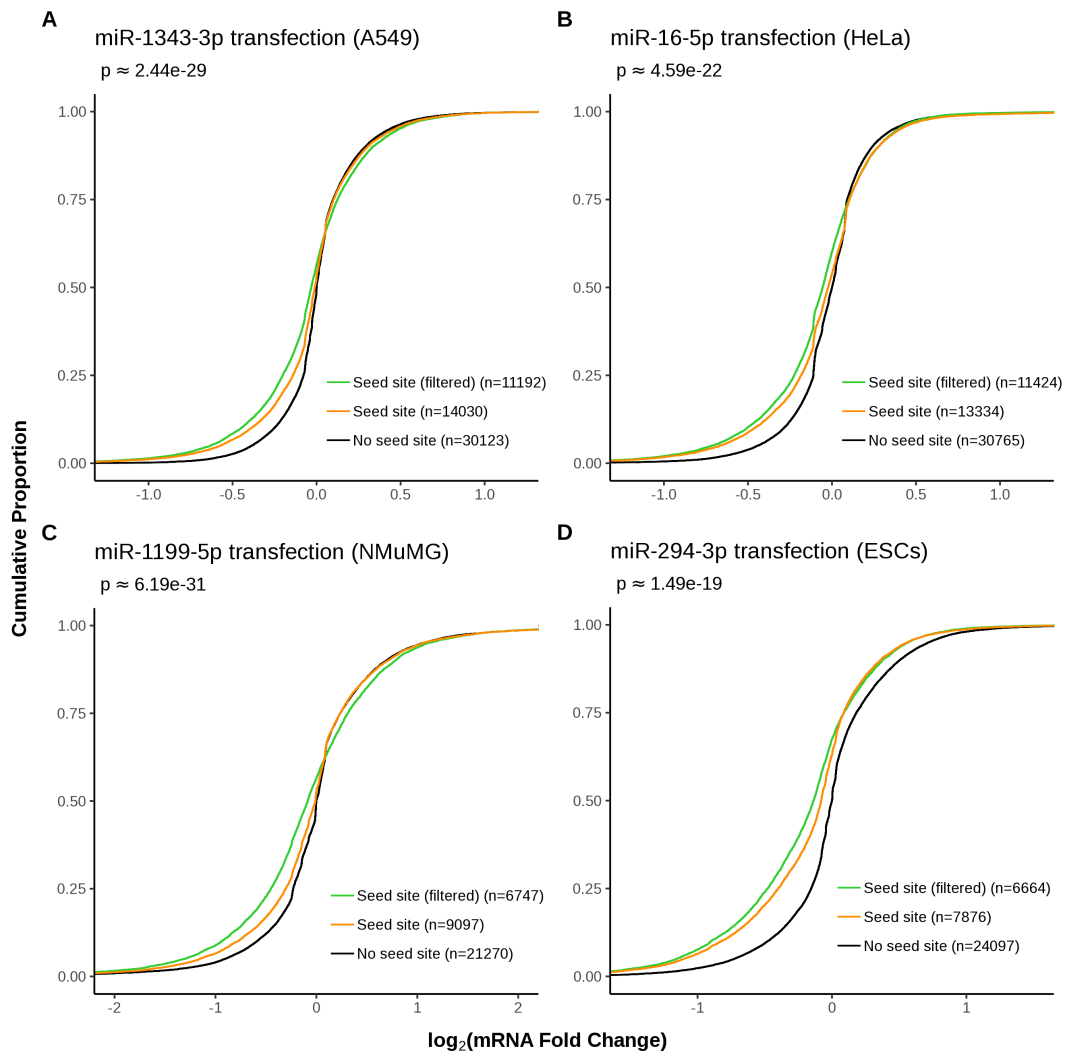


Figure 4.1 - Implementing an expression threshold on predicted miRNA targets improves miRNA target prediction accuracy. Curves show the cumulative \log_2 fold change distributions of i) protein-coding non-target transcripts (black) ii) protein-coding seed target transcripts (orange) and iii) expression filtered (TPM > 0.1) protein-coding seed target transcripts (green). Numbers in round brackets represent the number of mRNA transcripts contained in each distribution. Approximate p-values were computed using one-sided, two-sample, Kolmogorov-Smirnov tests between unfiltered and filtered target fold change distributions. Data presented for miRNA mimic transfection into **A)** A549 and **B)** HeLa cell lines, **C)** normal murine mammary gland (NMuMG) cells and **D)** mouse embryonic stem cells (ESCs).

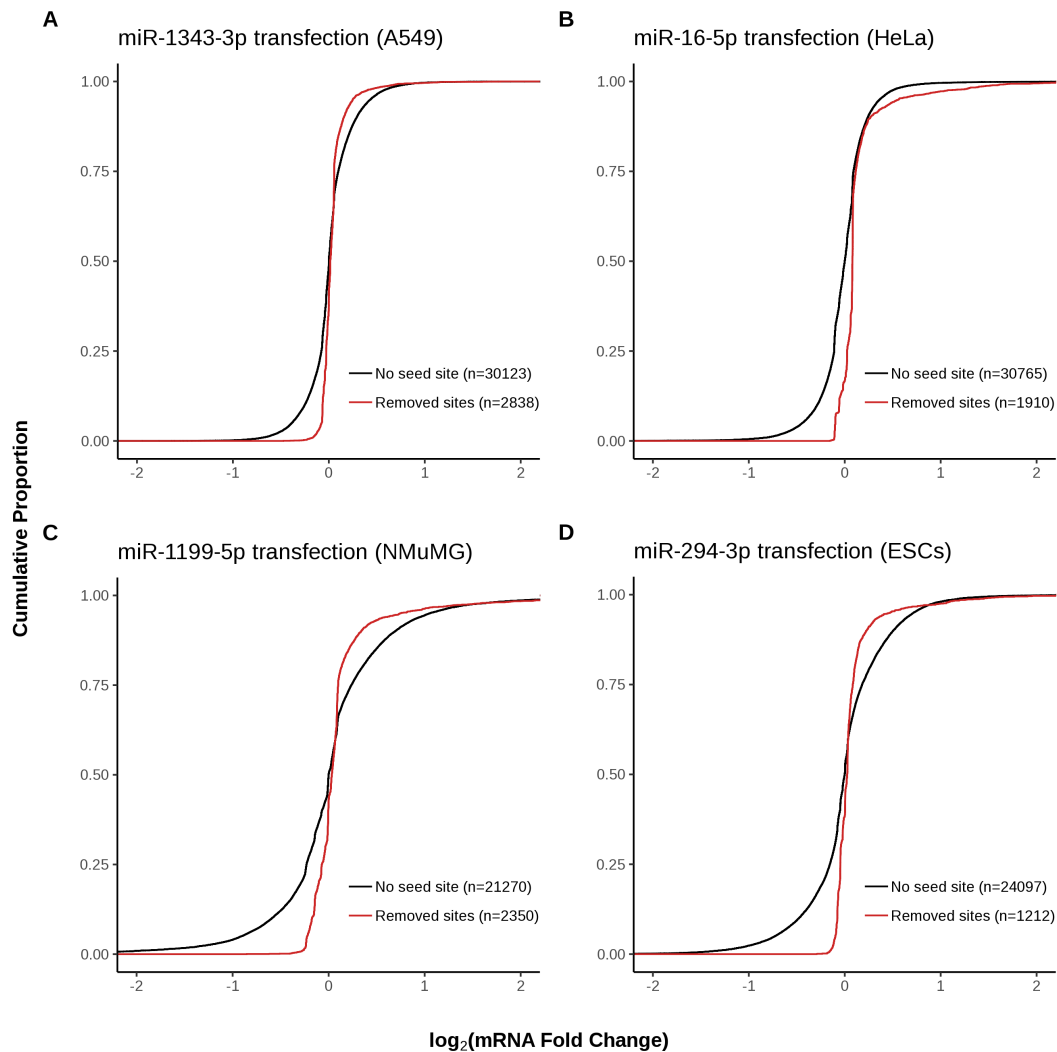


Figure 4.2 – Differential expression of lowly expressed predicted miRNA targets upon miRNA mimic transfection. For the analysis presented in figure 4.1, the cumulative \log_2 fold change distributions of lowly expressed transcripts (<0.1 TPM) with canonical seeds sites (dark red), in their 3'UTRs compared against the distribution of transcripts without a canonical seed site in their 3'UTRs (black).

Implementing expression filters for a range of different TPM values reveals that increasing this threshold results in retained transcripts which exhibit greater repression upon miRNA transfection (figure 4.3).

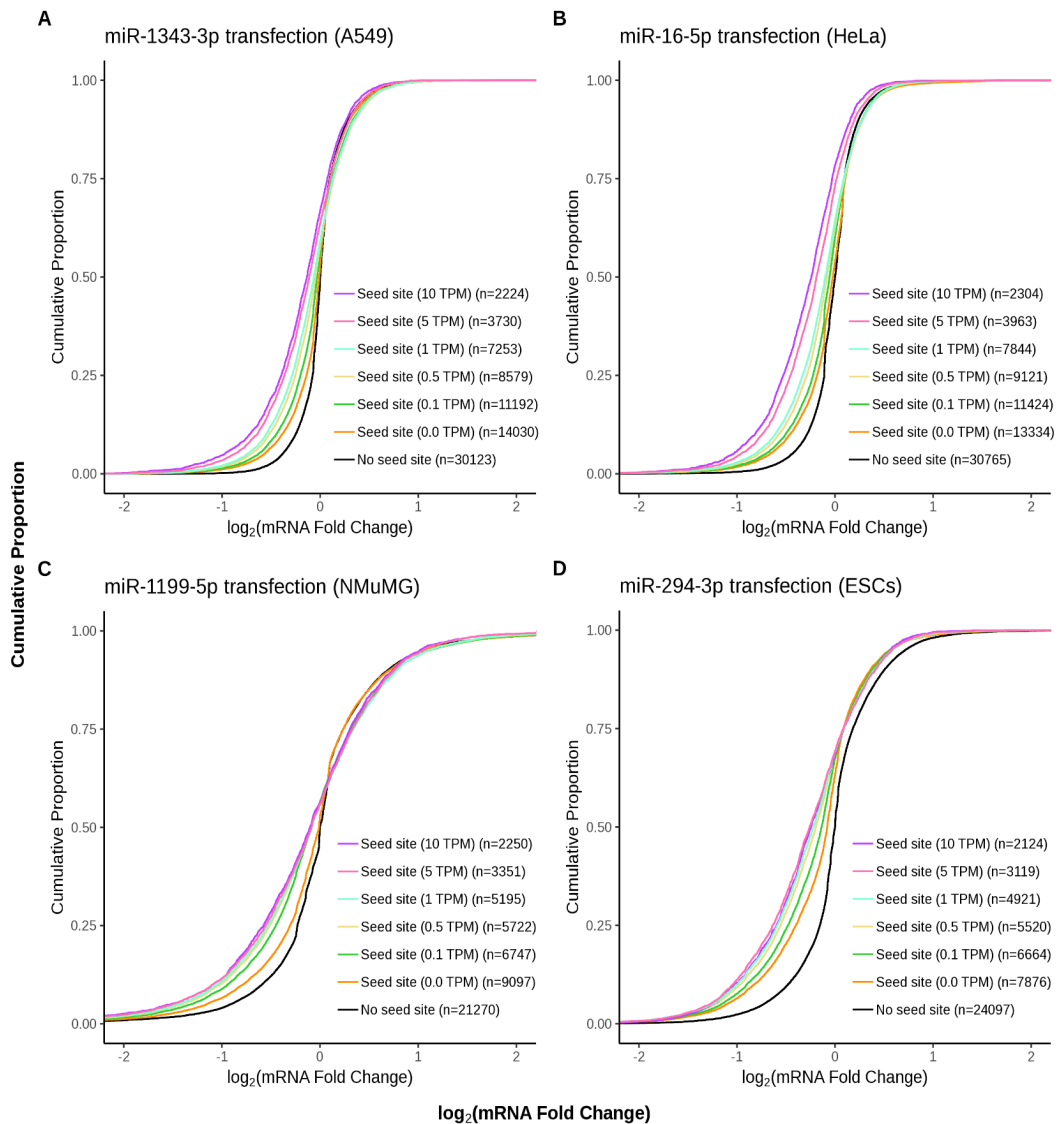


Figure 4.3 - The effect of expression filtering on retained protein-coding transcripts using multiple expression thresholds. Expression thresholds are implemented at TPM values of 10 (purple), 5 (pink), 1 (light blue), 0.5 (gold), 0.1 (green) and 0.0 (orange). Otherwise as in figure 4.1.

However, increasing the expression threshold beyond a particular point (between 1 – 10 TPM for experiments analysed) leads to the removal of a considerable number of mRNA transcripts which are repressed by the transfection of a miRNA mimic (figure 4.4).

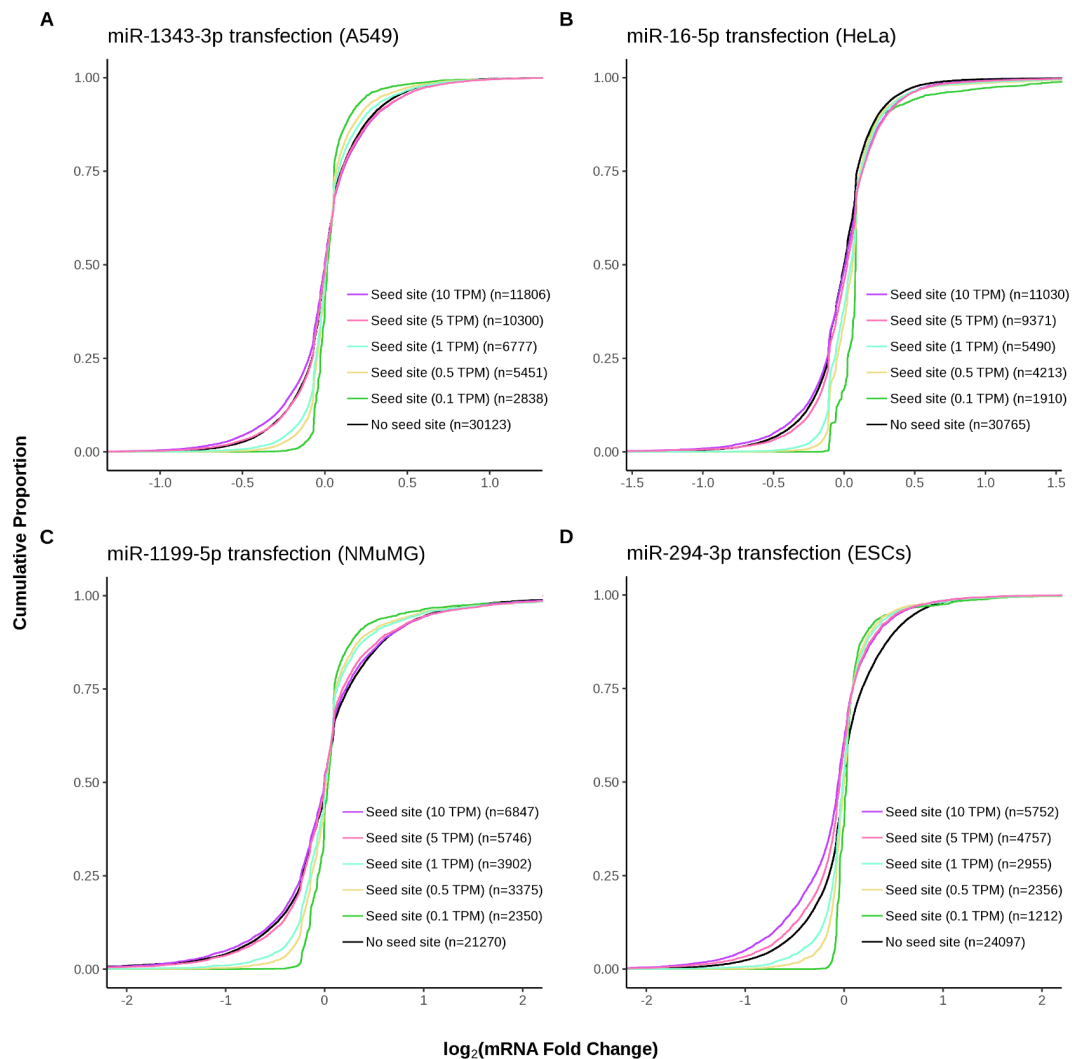


Figure 4.4 - The effect of expression filtering on removed protein-coding transcripts using multiple expression thresholds. Expression thresholds are implemented at TPM values of 10 (purple), 5 (pink), 1 (light blue), 0.5 (golden), and 0.1 (green). Otherwise as in figure 4.2

4.4.2 3' UTR extension

The next hypothesis to be tested was that 3'UTRs which had been elongated as a result of the 3'UTR reannotation process, and had acquired new predicted miRNA targets as a result, would behave similarly to previously annotated miRNA targets upon miRNA transfection. Newly gained miRNA target predictions deriving from FilTar's refined 3'UTR annotations of protein-coding transcripts (*i.e.* miRNA targets deriving from the elongation of existing 3'UTR annotations), generally exhibited similar levels of repression to miRNA target predictions deriving from Ensembl 3'UTR annotations. This can be discerned by observing the similarly shaped cumulative distributions between previously annotated miRNA seed targets, and newly annotated miRNA seed targets (figure 4.5 and figure A.5).

Anomalies were results deriving from the transfection of miR-107 and miR-10a-5p miRNA mimics into HeLa cells in which newly identified miRNA target predictions did not exhibit a log fold change distribution commensurate with that exhibited by already existing miRNA target predictions (figure A.5).

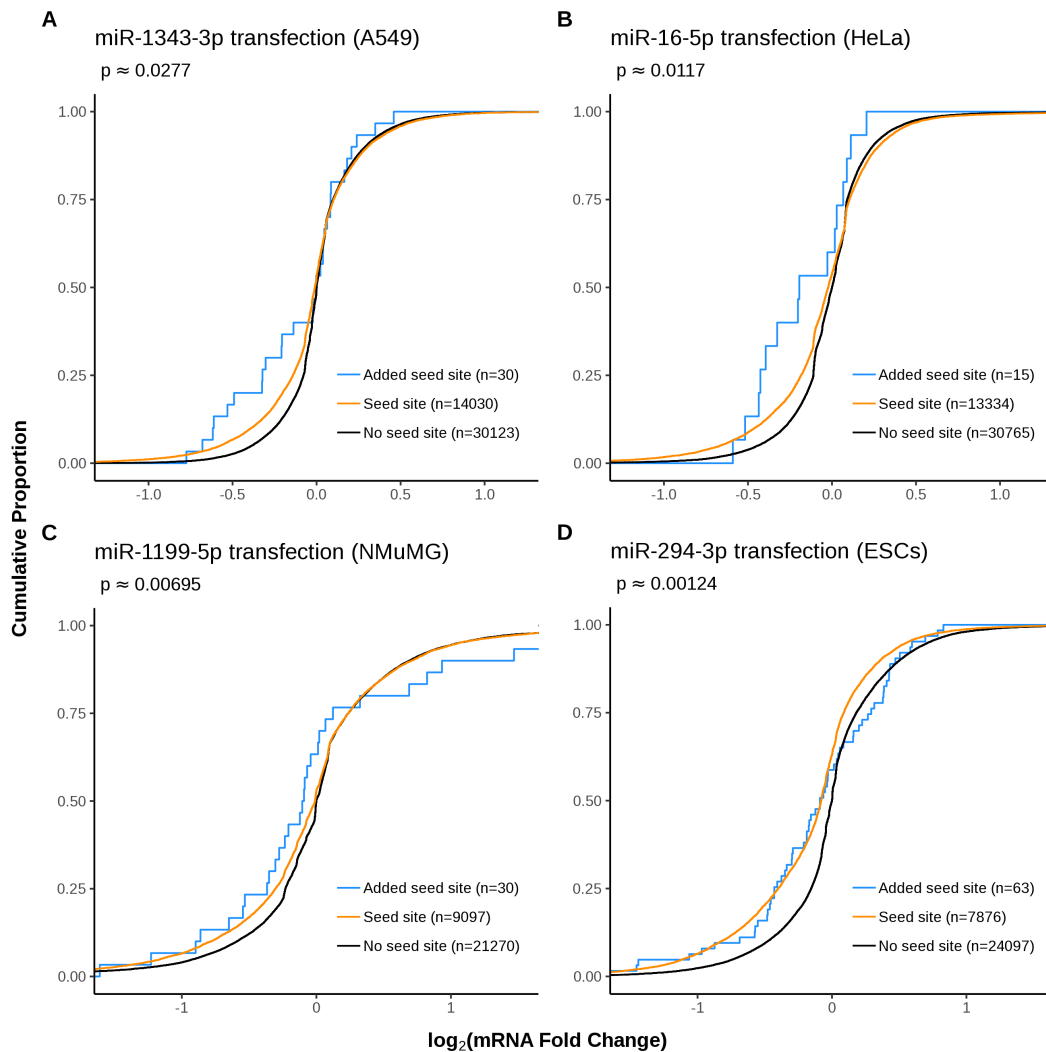


Figure 4.5 - 3'UTR elongation by FilTar leads to the identification of additional valid miRNA targets. mRNA transcripts contained in each distribution.

Approximate p-values were computed using one-sided, two-sample, Kolmogorov-Smirnov tests between unfiltered and filtered target fold change distributions. Data presented for miRNA mimic transfection into **A)** A549 and **B)** HeLa cell lines, **C)** normal murine mammary gland (NMuMG) cells and **D)** mouse embryonic stem cells (ESCs).

Next, it was important to determine the relationship between sequencing depth and the extent of 3'UTR reannotation occurring – and the implication of this for 3'UTR reannotation analyses. Completed analyses demonstrated a positive relationship, to a point of saturation between the number of RNA-Seq reads used for 3'UTR reannotation within a sample, and the number of 3'UTR bases gained (*i.e.* through 3'UTR elongation) during 3'UTR reannotation (figure 4.6).

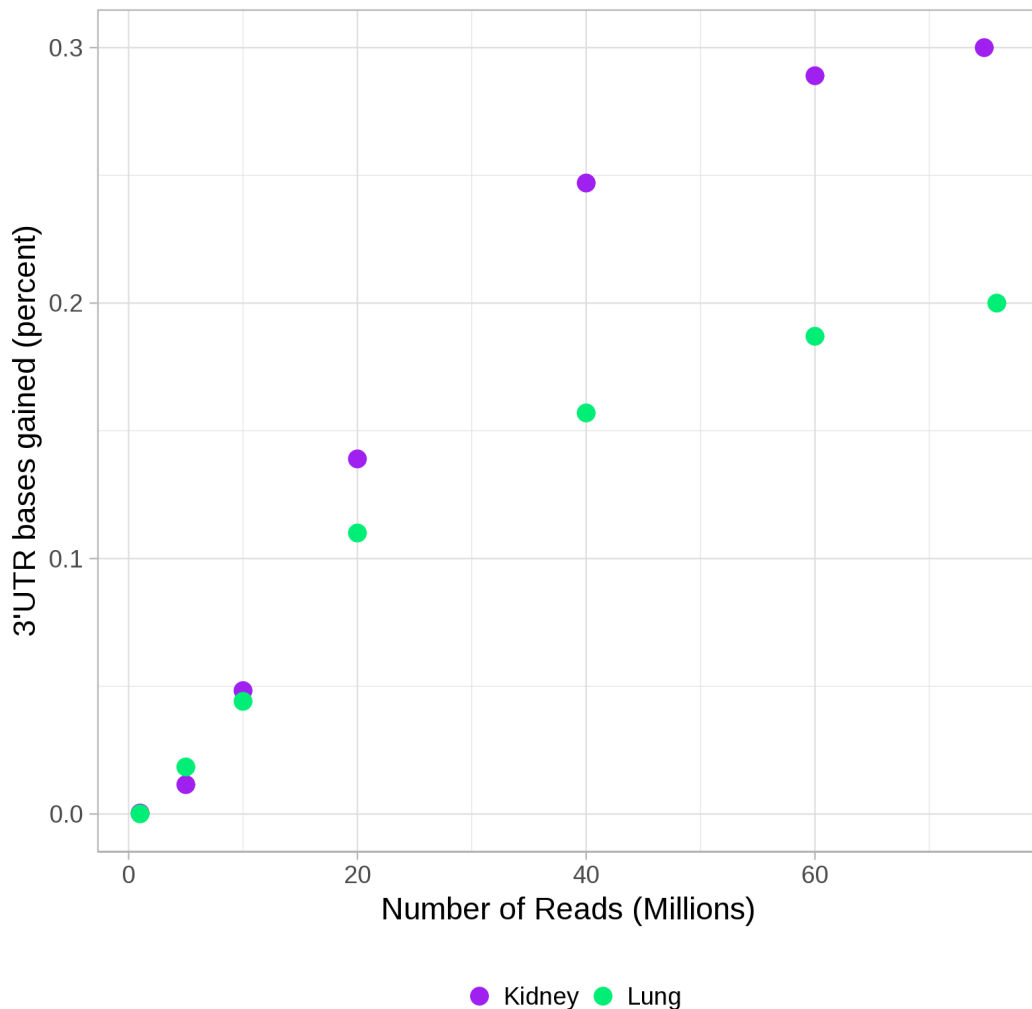


Figure 4.6 - Greater sequencing depth leads to greater 3'UTR elongation up to a point of saturation. The relationship between the number of reads sequenced and

the extent of 3'UTR elongation observed when using FilTar for human kidney (purple) and lung (green) datasets. Variable read counts generated by randomly sampling reads from the total.

Next, a similar analysis was performed, though this time testing for a potential between-samples effect for sequencing depth and 3'UTR elongation. This analysis would test the hypothesis that the extent of 3'UTR elongation in a transcriptome could be predicted from the sequencing depth, irrespective of sample-specific details such as cell type. When this analysis was performed, it was discovered that there was a weak positive relationship between the extent of 3'UTR elongation and the number of mapped RNA-Seq reads (figure 4.7).

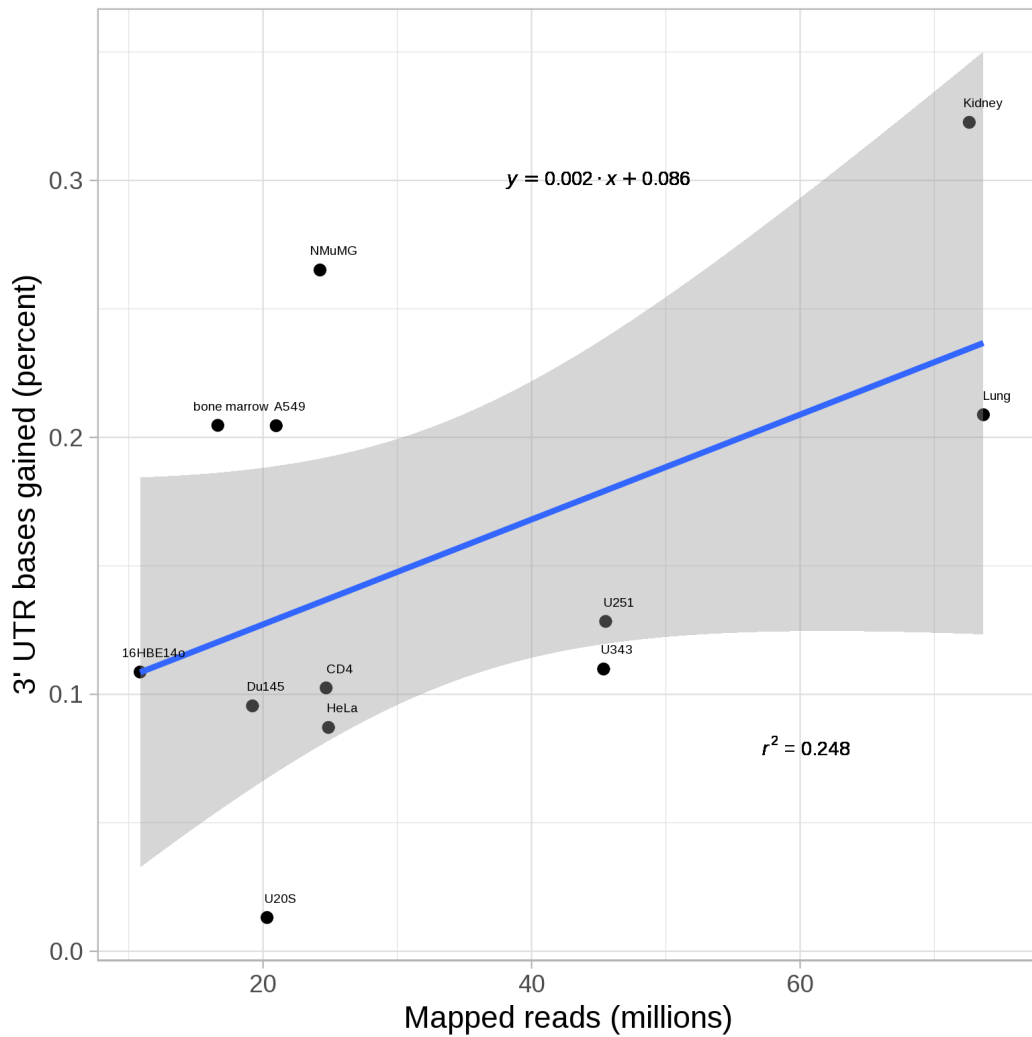


Figure 4.7 - The relationship between the number of mapped reads and the extent of 3'UTR elongation observed when using FilTar. Each point represents a different dataset analysed using FilTar. Refer to table A.2 for metadata relating to datasets analysed. Outlier values have been removed.

Additionally, it was also important to examine whether an increase in 3'UTR length due to 3'UTR reannotation led to a linear increase in the number of predicted target sites – as expected. A further analysis showed that the proportionate gain in miRNA target sites predicted as a result of 3'UTR reannotation corresponds linearly to the extent of 3'UTR elongation (figure 4.8).

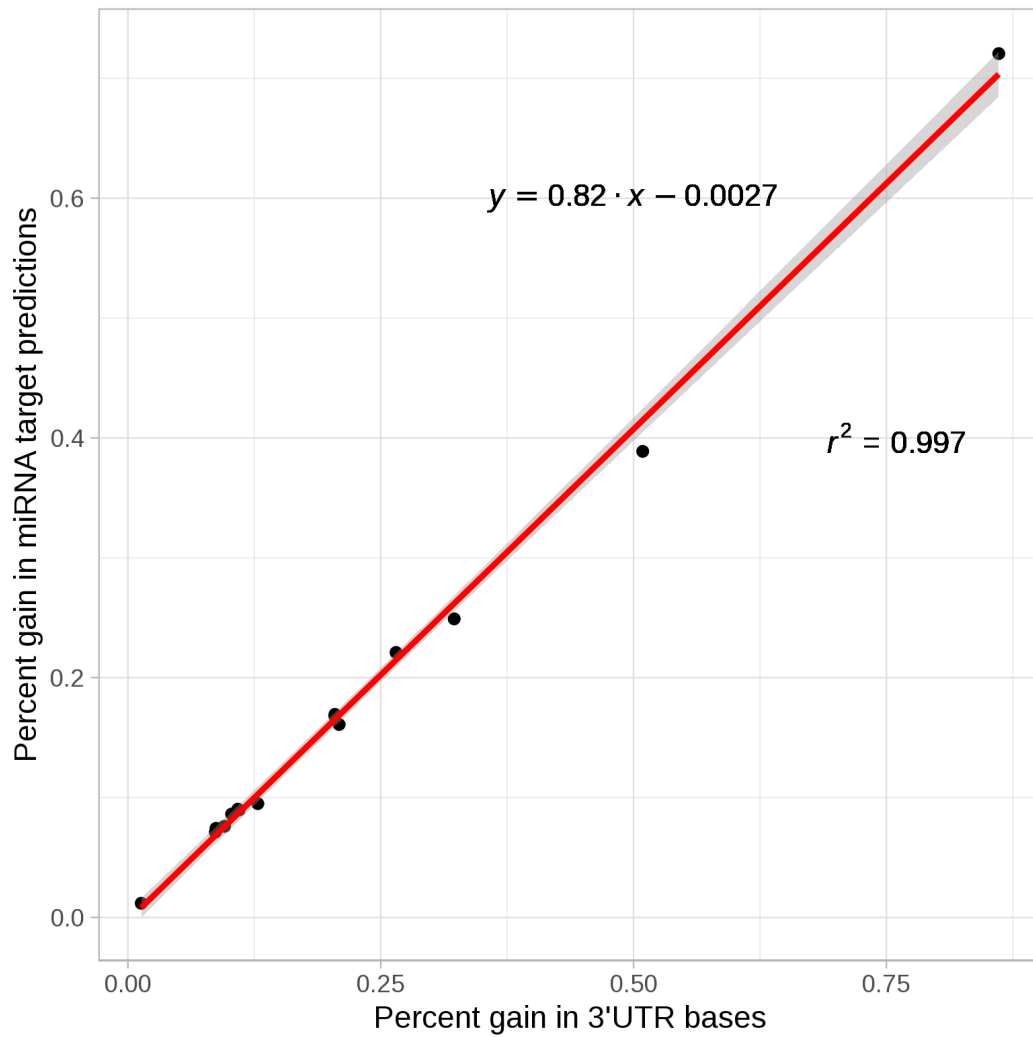


Figure 4.8 - A scatter plot of the percentage gain in total miRNA target site predictions vs. percentage gain in 3'UTR bases for a number of cell lines and tissue datasets analysed (black dots). A linear regression model was fitted using the 'lm' function of the R stats package (red) with a 95% confidence interval (grey). R-squared is derived from the Pearson correlation coefficient.

4.4.3 3' UTR truncation

Conversely, miRNA target transcripts that were removed as a result of FilTar truncating 3'UTR annotations relative to standard Ensembl annotations, exhibited repression similar to that of annotated non-target transcripts (figure 4.9 and figure A.6). The very similar CDFs of the 'Removed seed site' and 'No seed site' distributions in these figures indicate that predicted miRNA targets discarded as a result of 3'UTR reannotation behave very similarly to mRNA transcripts which were never predicted to be miRNA targets – indicating the efficacy of the 3'UTR reannotation and miRNA target predictions processes.

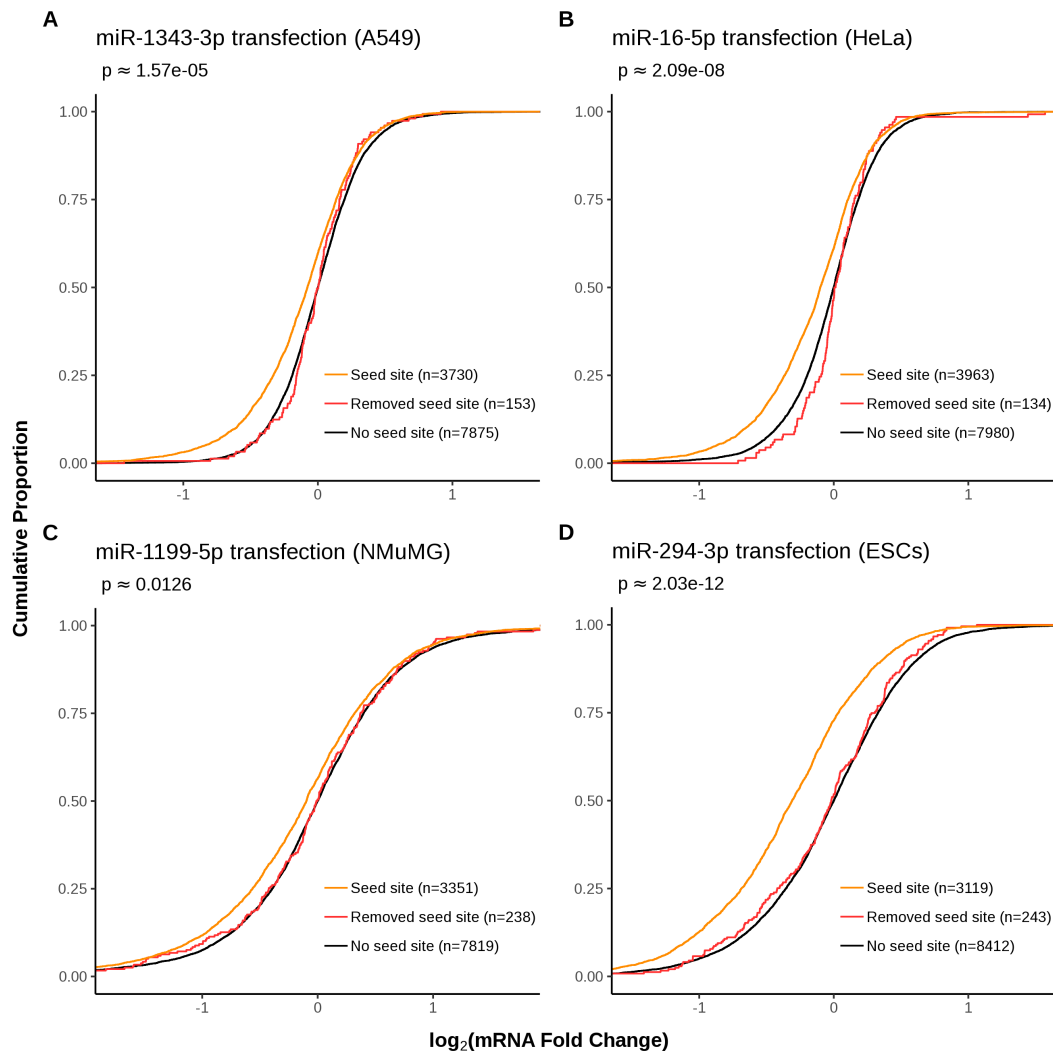


Figure 4.9 - 3'UTR truncation by FilTar leads to the removal of false positive miRNA target predictions. Curves are plotted of the cumulative log fold change distributions of expression filtered i) protein-coding non-target transcripts (black), ii) protein-coding seed target transcripts (orange) and iii) predicted target transcripts deriving from Ensembl 3'UTR annotations but not FilTar 3'UTR annotations (red). Approximate P-values were computed using one-sided, two-sample, Kolmogorov-Smirnov tests between non-target and discarded miRNA target fold change distributions. Otherwise as in figure 4.1.

An expression filter of > 5 TPM was implemented for transcripts to undergo 3'UTR truncation, as a preliminary analysis revealed that

3'UTR truncation without an expression filter, led to poor target prediction performance, indicating that some 3'UTRs had been truncated erroneously (figure 4.10). As can be seen in figure 4.10, without the low expression filter – discarded predicted miRNA targets tend to respond similarly to genuine miRNA targets, indicating erroneous 3'UTR truncation at these low expression levels.

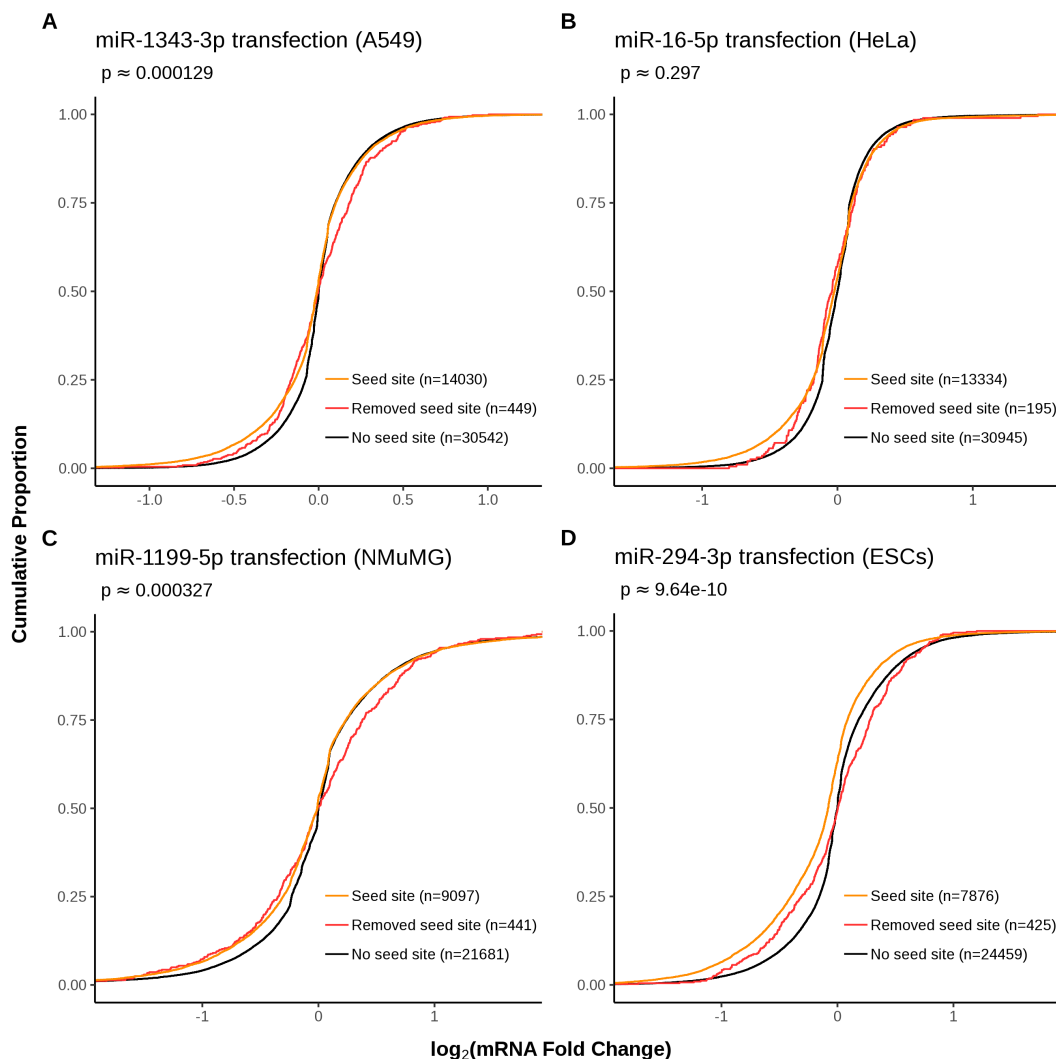


Figure 4.10 - As in figure 4.9, with the exception that no expression threshold has been implemented to filter data points contained with the removed seed site distribution.

This erroneous truncation was also observable from the alignment of RNA-seq reads to the genome, in which, for some lowly expressed genes, there was substantial read coverage downstream of the point at which 3'UTR truncations was called by the APAtap dependency (figure 4.11).

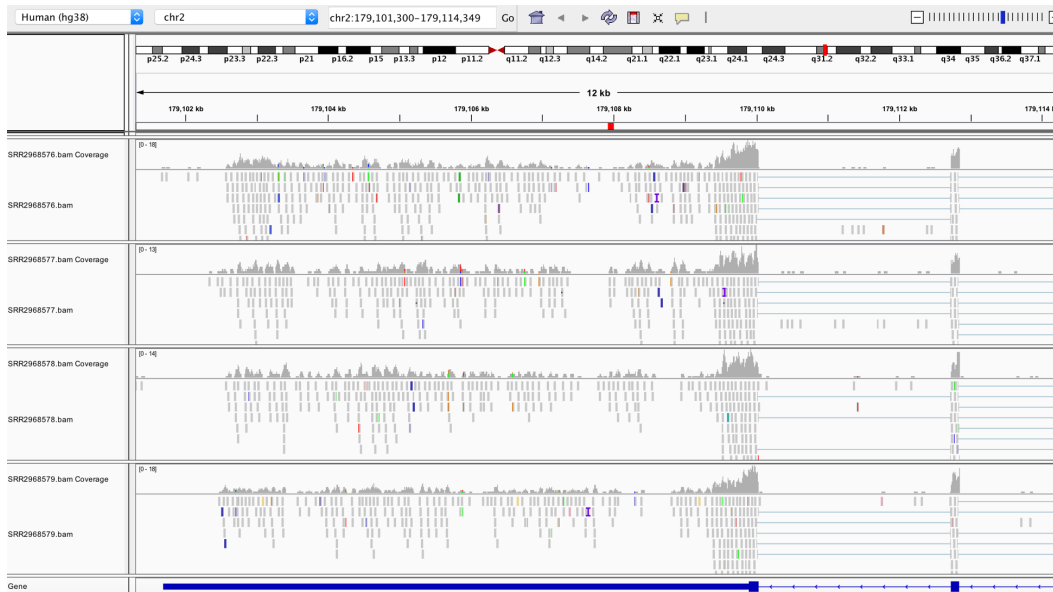


Figure 4.11 – An example of erroneous 3'UTR model predictions for lowly expressed genes. The alignment of RNA-Seq reads to the NKAP gene, from sequencing of A549 cell cultures treated with mock (*i.e.* negative control) transfections. The red rectangle represents the point of the 3'UTR in which the APAtap dependency called 3'UTR truncation for this gene and this example dataset. The maximum read coverage for the 3'UTR is less than 20 reads for all replicates, indicating that this is a lowly expressed gene. As can be observed, there is substantial read coverage downstream of the assigned truncation point. Four replicates were used for this analysis with the following run accessions: SRR2968576, SRR2968577, SRR2968578, SRR2968579. Alignments are visualised using the integrative genomics viewer (IGV) (v2.4.4) (Thorvaldsdóttir, et al., 2013).

In a minority of datasets analysed, removed target transcripts exhibited significantly less repression than target transcripts, but nonetheless exhibited greater repression than annotated non-target transcripts. In these datasets, the removed target log fold change distribution tended to align with the non-target distribution at the negative extremity, but not at small negative fold change value ranges - indicating that for a minority of datasets, labelled 'removed targets' may be mildly repressed by targeting miRNAs. It was important to test or not whether these removed targets constituted a weaker form of miRNA target interaction or not. Additional analysis demonstrated that for these datasets, such targets exhibited significantly weaker repression in response to miRNA transfection than 6mer targets, which are the weakest canonical miRNA target site type (Bartel, 2018) (figure 4.12) – suggesting that these removed targets were not just composed of weak miRNA target site types.

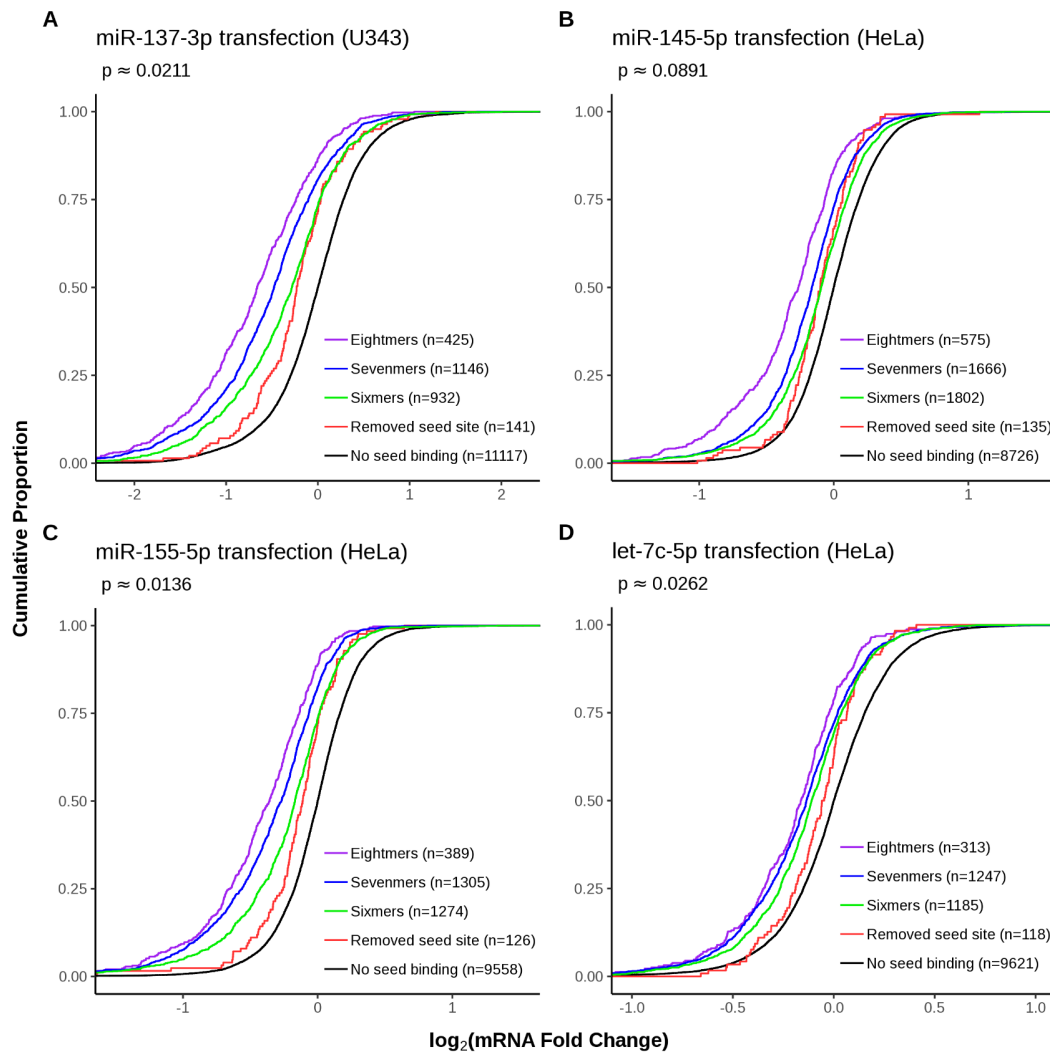


Figure 4.12 - Predicted targets removed by FilTar exhibit weaker repression in response to miRNAs than 6mer targets. In experiments in which removed predicted target transcripts exhibit evidence of low-level repression, repression is less than that observed by transcripts targeted by marginally effective 6mer seed sequences. As in figure 4.9, with predicted target transcripts divided by miRNA target site type into sixmer (green), sevenmer (blue) and eightmer (purple) subsets. Approximate P-values were computed using one-sided, two-sample, Kolmogorov-Smirnov tests between discarded miRNA target and sixmer target fold change distributions.

The relationship between sequencing depth and the extent of 3'UTR truncation is similar to that between sequencing depth and 3'UTR elongation. Results of the within-sample analysis in this case can be found within figure 4.13.

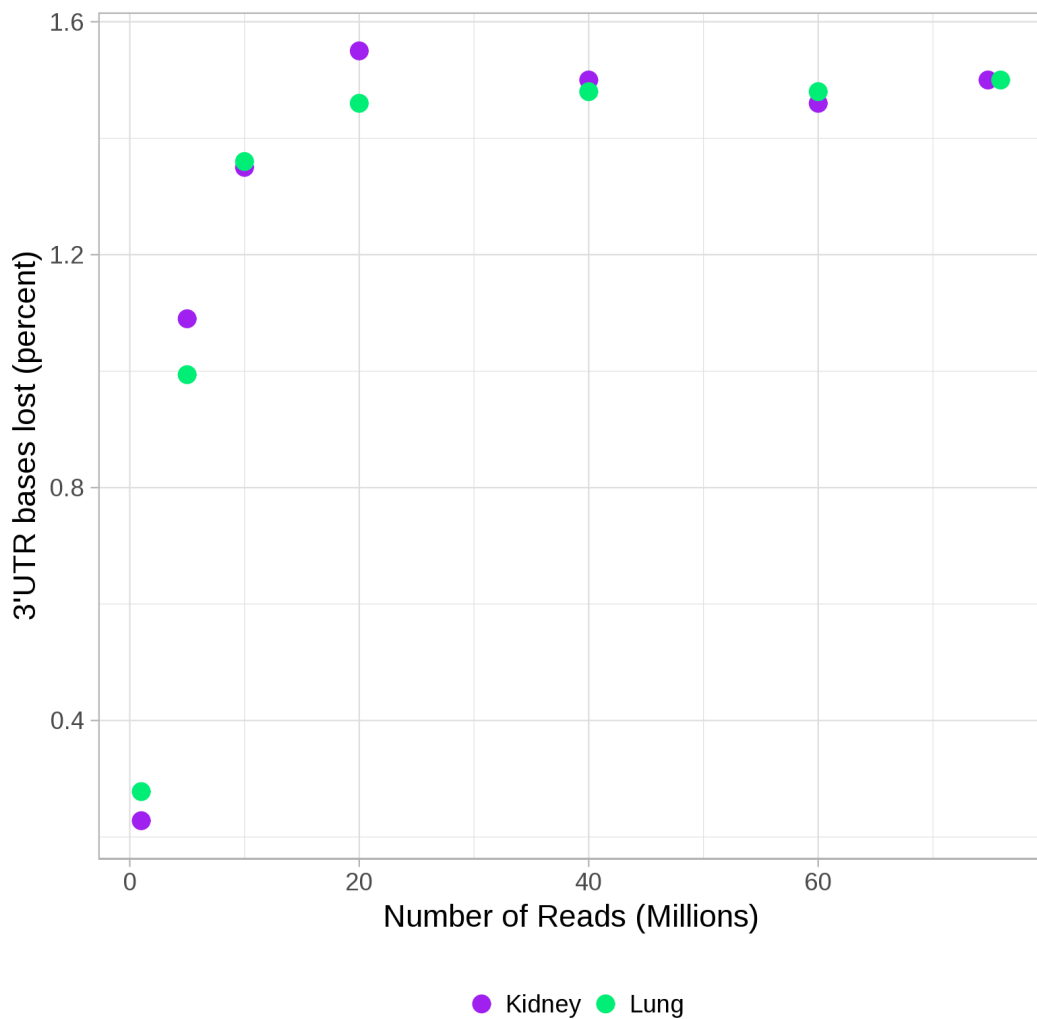


Figure 4.13 - Greater sequencing depth leads to greater 3'UTR truncation up to a point of saturation. The relationship between the number of reads sequenced and the extent of 3'UTR truncation observed when using FilTar within a given sample. Otherwise as in figure 4.6.

And the results of the between sample analysis can be found within figure 4.14.

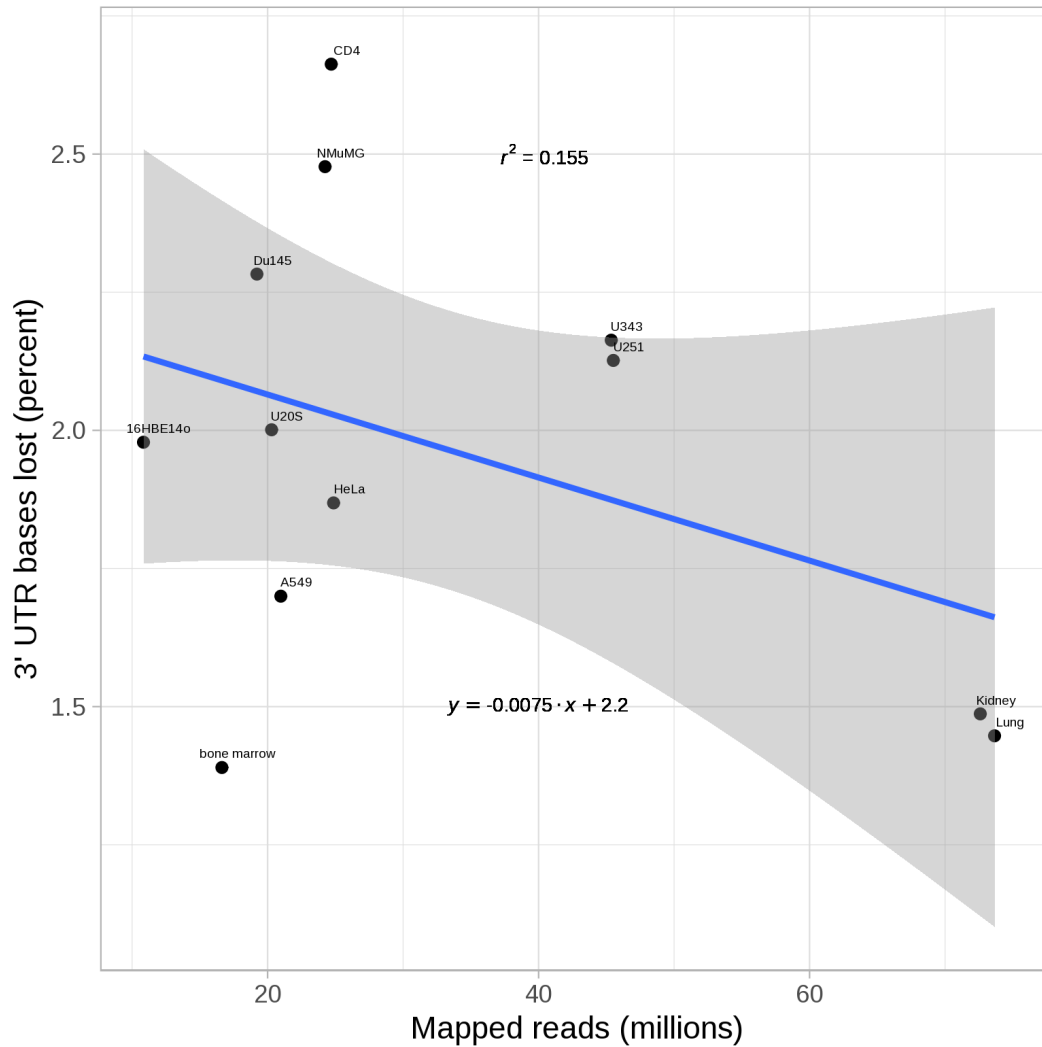


Figure 4.14 - The relationship between the number of mapped reads and the extent of 3'UTR truncation observed when using FilTar. Otherwise as in figure 4.7.

In addition, as with 3'UTR elongation, the extent of 3'UTR truncation was shown to correspond linearly with the proportionate decrease in the number of miRNA targets predicted after 3'UTR truncation (figure 4.15).

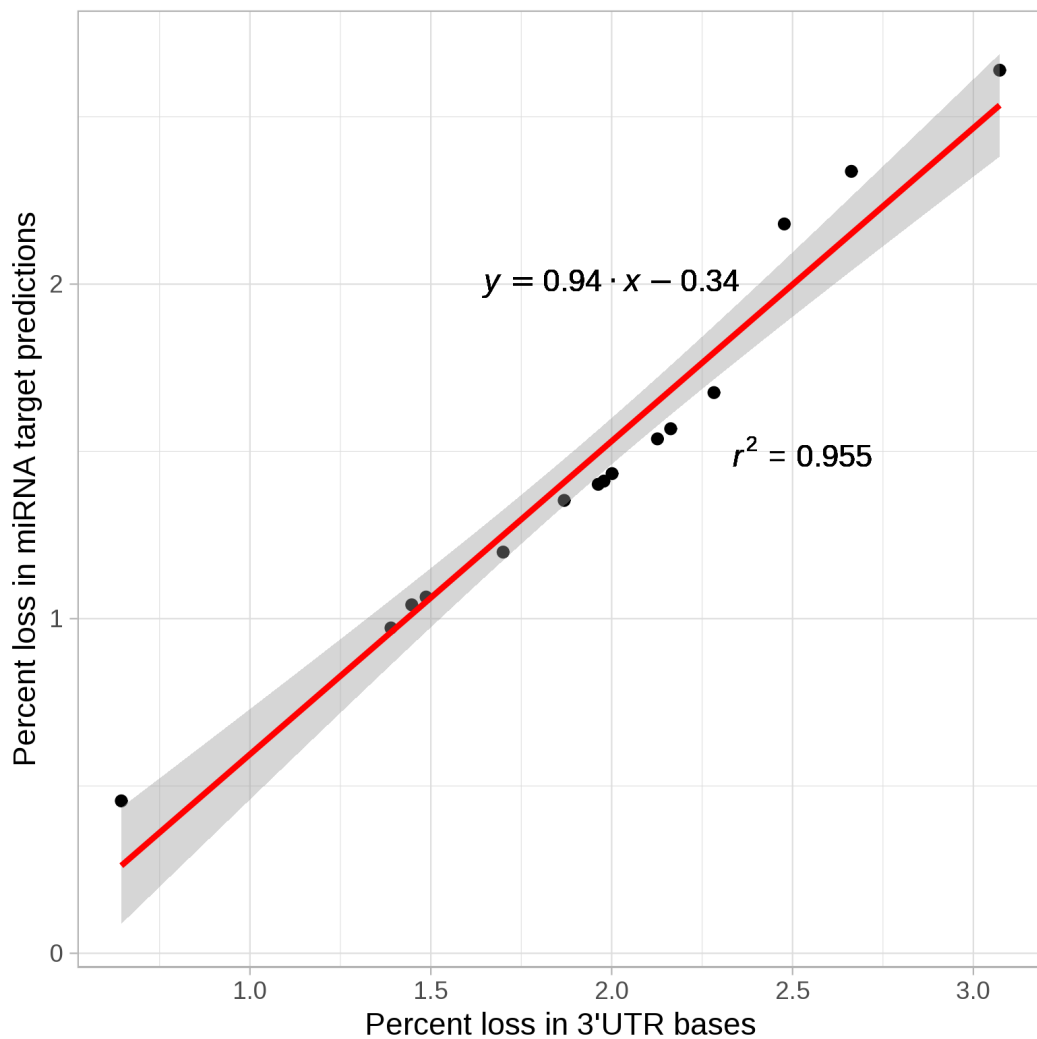


Figure 4.15 - A scatter plot of the percentage loss in total miRNA target predictions vs. percentage loss in total 3'UTR bases. Otherwise as in figure 4.8.

4.4.4 Cumulative effect of filtering and reannotation

Next, it was important to examine the extent to which 3'UTR reannotation would affect miRNA target predictions on a transcriptome-wide basis for a large number of cellular contexts. When the FilTar reannotation and miRNA target prediction workflow was applied transcriptome-wide, to multiple organs and cell lines, using all annotated miR-Base human miRNAs, there was a mean average gain and loss of miRNA target sites corresponding to 0.18% and 1.5% of the total original miRNA target sites predicted deriving from Ensembl 3'UTR annotations (figure 4.16).

As confirmed in the previous analyses of this chapter (figure 4.7 and figure 4.14), there does seem to be a relationship between sequencing depth and the extent of 3'UTR reannotation occurring and therefore the extent of changes in miRNA target predictions between original and reannotated 3'UTR models (figure 4.8 and 4.15). As a result, an important point to consider when interpreting the results presented in figure 4.16 is that to some extent the variability in results between samples may reflect the influence of technical rather than biological factors.

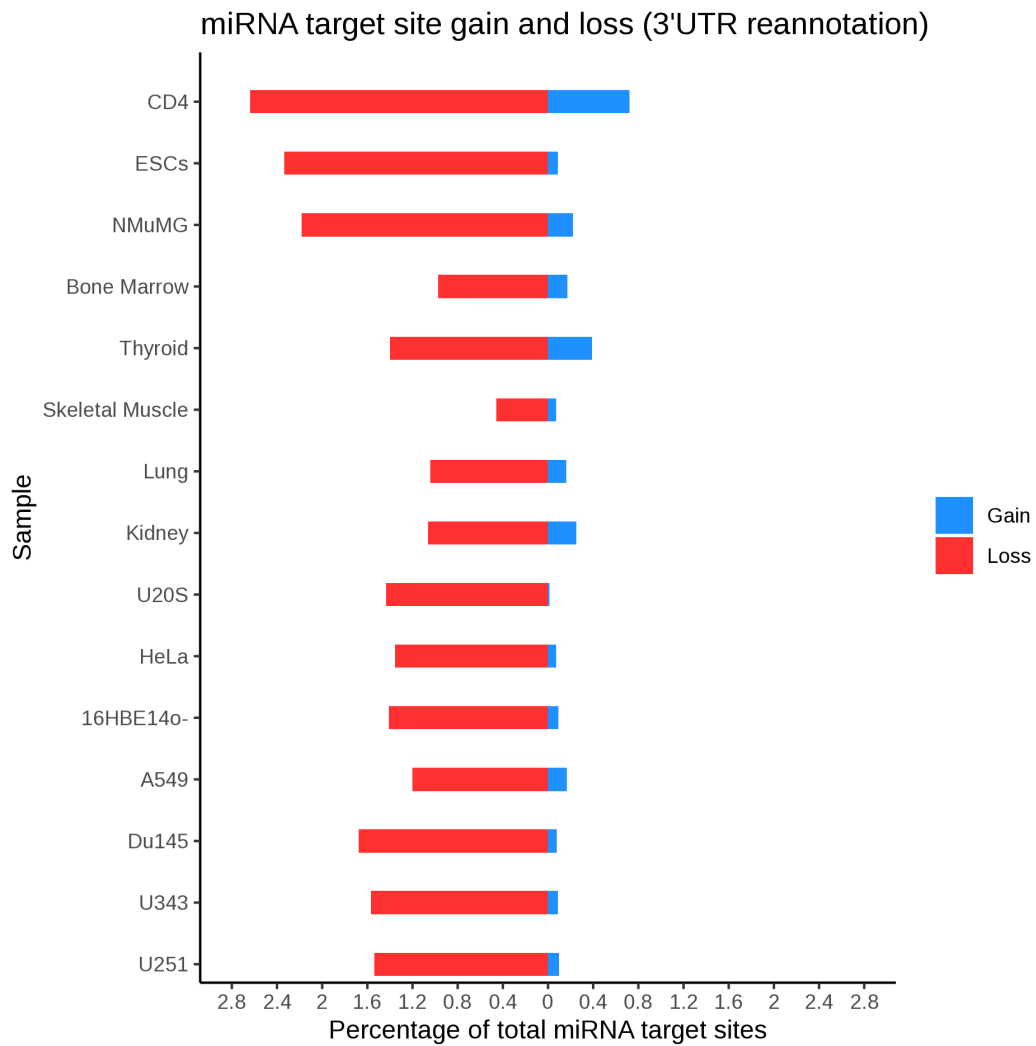


Figure 4.16 - Total miRNA target site gain and loss when applying FilTar to multiple sample types. FilTar is applied to the protein-coding transcriptome for all annotated human miRNAs for multiple tissues, organs and cell lines. Gained (blue) and lost (red) miRNA target sites is expressed as a percentage of the total number of target sites identified when deriving miRNA from Ensembl 3'UTR annotations.

A summary statistics table can be found within table 4.1 detailing the extent of 3'UTR truncation and elongation for each cellular context:

Species	Samples	Bases gained (Mb)	Bases gained (%)	Bases lost (Mb)	Bases lost (%)	3'UTRs elongated	3'UTRs elongated (%)	3'UTRs truncated	3'UTRs truncated (%)
<i>Human</i>	U251	0.08	0.1	1.30	2.1	352	0.7	5730	10.6
	U343	0.07	0.1	1.32	2.2	296	0.5	7395	13.7
	Du145	0.06	0.1	1.40	2.3	453	0.8	5342	9.9
	A549	0.13	0.2	1.04	1.7	281	0.5	6774	12.5
	16HBE14o-	0.07	0.1	1.21	2.0	213	0.4	6600	12.2
	HeLa	0.05	0.1	1.14	1.9	289	0.5	4087	7.6
	U20S	0.01	0.0	1.23	2.0	120	0.2	3614	6.7
	Kidney	0.20	0.3	0.91	1.5	708	1.3	5738	10.6
	Lung	0.13	0.2	0.89	1.4	538	1.0	5686	10.5
	Skeletal muscle	0.05	0.1	0.39	0.6	136	0.3	3018	5.6
Thyroid	0.31	0.5	1.20	2.0	460	0.9	7356	13.6	
Bone marrow	0.13	0.2	0.85	1.4	292	0.5	5444	10.1	
<i>Mouse</i>	NMuMG	0.13	0.3	1.18	2.5	454	1.1	6440	15.8
	CD4+	0.05	0.1	1.27	2.7	345	0.8	2447	6.0
	ESCs	0.41	0.9	1.46	3.1	493	1.2	7502	18.4

Table 4.1 - FilTar 3'UTR reannotation summary statistics for cell line and tissue data used in this study. Statistics are the total number or proportion of bases or transcripts gained or lost through 3'UTR reannotation respectively. All comparisons are made against a reference of Ensembl annotated 3'UTR sequences associated exclusively with protein-coding mRNA transcripts.

This corresponds to a gain and loss of total miRNA seed sides in the tens and hundreds of thousands respectively, as can be seen in table 4.2:

Species	Samples	Seed sites gained (3'UTR reannotation)	Seed sites lost (3'UTR reannotation)	Seed sites lost (expression filtering)
Human	U251	49345	800764	12942294
	U343	46701	816545	13657488
	Du145	39571	872804	12508511
	A549	87549	624503	15578814
	16HBE14o-	47031	735041	13193677
	HeLa	38712	704948	9792951
	U20S	6146	746686	12879630
	Kidney	129715	554534	11476758
	Lung	83821	542432	12057289
	Skeletal muscle	37028	237223	16615464
Thyroid	202504	730038	11682705	
Bone marrow	88212	506415	14632213	
Mouse	NmuMG	62367	615046	9858668
	CD4+	203359	744867	7358255
	ESCs	24318	659420	8947356

Table 4.2 - The total number of miRNA seed sites lost through expression filtering or 3'UTR reannotation of transcripts. Expression filtering occurs at TPM > 0.1. Total miRNA seed sites for human: 52084138 and mouse: 28216437

However, a much larger proportion of miRNA seed sites (mean average of 26.3%) are lost through expression filtering (figure 4.17), with expression filtering representing a loss of millions of miRNA seed sites (table 4.2).

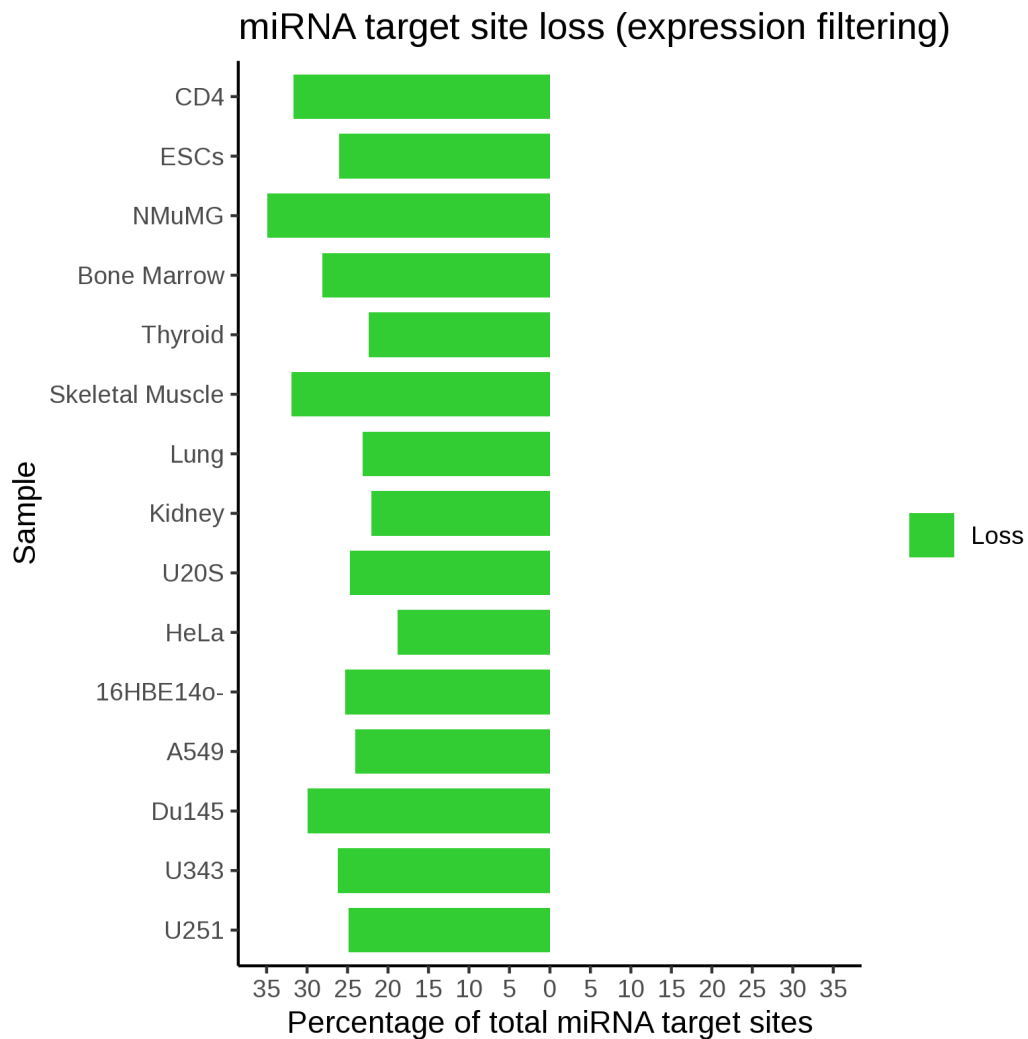


Figure 4.17 – The effect of expression filtering on multiple cell lines. The percentage of total miRNA targets removed through expression filtering at a threshold of 0.1 TPM in a set of different cell lines and tissue types for human and mouse species.

Species	Samples	Bases lost (Mb)	Bases lost (%)	3'UTRs removed	3'UTRs removed (%)
Human	U251	19.56	32.0	22653	42.0
	U343	21.07	34.4	21929	40.6
	Du145	24.06	39.3	25494	47.2
	A549	19.50	31.9	20783	38.5
	16HBE14o-	20.73	33.9	21221	39.3
	HeLa	15.09	24.6	18907	35.0
	U20S	19.96	32.6	22548	41.8
	Kidney	17.78	29.0	22476	41.6
	Lung	18.68	30.5	22647	42.0
	Skeletal muscle	25.86	42.2	28148	52.1
Thyroid	17.84	29.1	21529	39.9	
Bone marrow	22.78	37.2	23040	42.7	
Mouses	NMuMG	20.66	43.2	19592	47.7
	CD4+	18.61	38.9	19862	48.5
	ESCs	15.61	32.6	15505	37.9

Table 4.3 - Summary statistics of the effects of filtering protein-coding transcripts at an expression threshold of 0.1 TPM. Statistics are for the total number and proportion of bases and transcripts removed as a result of expression filtering.

This is commensurate with the mean average of 34.0% of 3'UTR bases lost when removing lowly expressed transcripts (< 0.1 TPM) from target predictions (table 4.3).

When considering the combined effect of expression filtering and 3'UTR reannotation, a mean average 36.1% of 3'UTR bases are lost, affecting a mean average of 53.4% of protein-coding 3'UTRs (table 4.4).

Species	Samples	Bases lost (Mb)	Bases lost (%)	3'UTRs affected	3' UTRs affected (%)
Humans	U251	20.86	34.1	28383	52.6
	U343	22.39	36.6	29324	54.3
	Du145	25.45	41.6	30836	57.1
	A549	20.54	33.5	27557	51.1
	16HBE14o-	21.94	35.8	27821	51.5
	HeLa	16.23	26.5	22994	42.6
	U20S	21.19	34.6	26162	48.5
	Kidney	18.69	30.5	28214	52.3
	Lung	19.57	32.0	28333	52.5
	Skeletal muscle	26.26	42.9	31166	57.7
	Thyroid	19.04	31.1	28885	53.5
Bone marrow	23.63	38.6	28484	52.8	
Mouse	NMuMG	21.84	45.7	25969	63.5
	CD4+	19.87	41.6	22309	54.5
	ESCs	17.07	35.7	23007	56.3

Table 4.4 – Combined statistics relating to 3'UTR reannotation and expression filtering. The sum of statistics from table 1 and table 3 relating to total combined 3'UTR bases and 3'UTRs affected by expression filtering and 3'UTR truncation.

4.5 Discussion

Results show that FilTar is successfully able to utilise RNA-Seq data to reannotate protein-coding 3'UTR sequences and filter based on expression data leading to a gain in specificity and sensitivity of target prediction evidenced through tests using experimental data.

That expression filtering target transcripts at even a modest expression threshold of 0.1 TPM leads to a loss of millions of seed sites in most datasets analysed (table 4.2) represents a radical reduction in the number of false positive predictions associated with miRNA target prediction and is indicative of the importance of considering the biological plausibility of candidate miRNA interactions. The positive relationship between the expression threshold chosen and the extent of repression of retained mRNA transcripts is evidence for the robustness of this effect (figure 4.3). The increase in specificity conferred by expression filtering does however seem to be accompanied by a corresponding loss of sensitivity of miRNA target prediction when large expression threshold values are chosen (figure 4.4), indicating that sufficient caution ought to be exercised by the user when choosing expression threshold values. However, even for larger expression thresholds, the reduction in sensitivity is less than the increase in specificity conferred by expression filtering (figure 4.3).

The number of newly predicted miRNA target sites deriving from FilTar elongated 3'UTR sequences is generally relatively low. For cell line datasets analysed, the maximum of number of newly predicted miRNA

targets made for any single miRNA was 67, with the majority of datasets analysed yielding less than 15 newly predicted targets (figure 4.5 and figure A.3). The number of newly identified target transcripts is commensurate with the universally low proportion of 3'UTRs extended, and the small proportion of bases added to the total of the 3'UTR annotation (table 4.1), even though this still represents a substantial increase in the number of miRNA seed target sites identified. This is in contrast to 3'UTR truncation in which the proportion of 3'UTRs truncated and bases removed from the 3'UTR annotation total are much greater. Analysis shows that there is a strong positive correlation between the number of 3'UTR bases reannotated, and the number of predicted miRNA target sites gained or lost through reannotation (figure 4.8 and figure 4.15). The bias in 3'UTR truncation as opposed to elongation can possibly be explained by either a pre-existing bias in standard Ensembl 3'UTR annotations to generate long 3'UTR models, or rather a bias in the FilTar reannotation workflow for 3'UTR truncation rather than elongation. A potential bias in the standard Ensembl annotation workflow could potentially be explained by the method of transcript annotation, in which, although transcript models are built on a tissue-specific basis, transcript models incorporated into the final Ensembl gene set typically only derive from the merging of RNA-sequencing reads from multiple different tissue samples (Aken, et al., 2016), therefore creating a bias towards the annotation of longer 3'UTRs. This effect may be exacerbated or supplemented by the existence of 3'UTR isoforms within a given sample and transcript - creating relatively low abundance isoforms towards the distal end of the 3'UTR, making annotation difficult, and likely generating a large amount of uncertainty, biases and variability in different methods used to model 3'UTRs.

Another possibility, is that the shortening and extension of existing 3'UTR annotations are qualitatively different problems requiring different respective sequencing depths. Within a given sample, a read sampling analysis demonstrates that there is a positive relationship, up to a point of saturation between sequencing depth and the number of bases used to elongate existing 3'UTRs (figure 4.6). In addition, the saturation point for the addition of bases to 3'UTRs is still substantially less than the proportion of bases removed at 3'UTRs even at relatively low sequencing depths indicating that the discrepancy between proportion of 3'UTR bases added or subtracted from the 3'UTRs cannot be explained by insufficient sequencing depth. A similar positive relationship is observed between sequencing depth and the number of bases truncated from existing 3'UTRs (figure 4.13), although far fewer reads seem to be required for saturation to occur, indicating a weaker reliance on sequencing depth for 3'UTR truncation compared to 3'UTR elongation.

In addition, the sequencing depth does seem to influence the extent of 3'UTR reannotation for a similar between sample analysis (figure 4.7 and figure 4.14). The weak positive correlation between sequencing depth and the proportion of 3'UTR bases gained is likely best explained by greater sequencing depth uncovering less abundant 3'UTR isoforms leading to an elongation of some 3'UTRs during the 3'UTR reannotation process. Conversely, greater sequencing depth seems to be somewhat negatively correlated with the extent of 3'UTR truncation – which can perhaps be explained by 3'UTR truncations occurring in error at low sequencing depths.

As mentioned previously, FilTar permits 3'UTR truncations only occurring on moderately-to-highly expressed transcripts, after discovery that the reannotation of the 3'UTRs of lowly expressed transcripts generated a relatively large number of what seemed to be false positive predictions (figure 4.10). The likely cause being that low transcript expression leads to sporadic and inconsistent coverage across the 3'UTR, in which there is insufficient information to correctly call 3'UTR truncation. The default behavior of the FilTar tool therefore is to only truncate the 3'UTRs of transcripts which are not poorly expressed (*i.e.* TPM ≥ 5).

When examining 3'UTR truncations further, for a minority of datasets analysed, some removed predicted miRNA targets seem to be marginally effective, with some transcripts exhibiting low levels of repression upon transfection of the miRNA mimic. Further analysis indicates that these marginally repressed transcripts exhibit even weaker repression than 6-mer targeted transcripts (figure 4.12), one of the least effective canonical miRNA target types (Bartel, 2018), indicating that the efficacy of these site types is marginal. A possible explanation for the existence of these site types is that, for some transcript annotations for which the 3'UTR was truncated, there may exist a small proportion of isoforms with longer 3'UTRs, which are too low in abundance to be detected by APAtap, but nonetheless still confer a marginal level of repression to the transcript, and hence are detectable when analysing experimental data.

Investigations into the effect of utilising expression data when making transcriptome-wide miRNA target predictions can be extended by closer examination of not only the refinement of 3'UTR annotations across different biological contexts, and its effects on miRNA target prediction, but more precisely the definition of specific 3'UTR profiles, incorporating information about 3'UTR isoforms within a given cellular context (Agarwal, et al., 2015), an existing feature in the current version of the FilTar tool. This enables the weighting of miRNA target prediction scores on the basis of sequencing data applied by the user themselves, enabling even further and extended tailoring of miRNA target prediction to the specific biological context being researched. Previous analyses indicate that the most effective target predictions occur when those predictions are weighted on the basis of 3'UTR isoform ratios (Nam, et al., 2014). In addition, the scope of FilTar's functionality can be increased by enabling the annotation of novel 3'UTR sequences for transcripts without a current annotated 3'UTR, and also for those 3'UTRs which themselves span multiple exons. In addition, both the configurability and precision of FilTar can be improved in the future by respectively, enabling use of additional tools for 3'UTR reannotation (Gruber, et al., 2018; Gruber, et al., 2018) and exploring the greater transcriptomic resolutions enabled by nascent single cell sequencing technologies.

4.6 Conclusion

FilTar utilises RNA-Seq data to increase the accuracy of miRNA target predictions in animals by filtering for expressed mRNA transcripts and reannotating 3'UTRs for greater specificity to a given cellular context

of interest to the researcher. In addition, the use of RNA-Seq data for the implementation of this approach as opposed to more specialist sequencing data, increases the accessibility and usability of FilTar for biological researchers. FilTar's compatibility with user-generated RNA-Seq data, confers functionality across a wide-range of potential biological contexts.

Chapter 5: The regulation of the post-mating response in *Drosophila melanogaster* by miRNAs

5.1 Contributions

Tracey Chapman: *Experimental design. Overall project supervision.*

Emily Fowler: *Conducted Experiments. Used domain-specific knowledge to write the introduction and discussion section of the publication associated with the study described in this chapter, which has been used for the formation of the introduction and discussion of fruit fly biology in this chapter. GO term enrichment analyses for the associated publication.*

Simon Moxon: *Experimental design, miRNA and mRNA quantification, differential expression analysis of miRNA and mRNA expression data (i.e. use of PaTMan, kallisto, sleuth and DESeq2). Overall project supervision.*

Thomas Bradley: *miRNA target prediction analysis, integrated analysis of mRNA and miRNA expression and differential expression data, miRNA-mRNA network analysis, exploratory data analysis (figures 5.2-5.7), discussion of the results of the integrated miRNA target prediction analysis. QC of differential expression analysis conducted by Dr. Moxon. Interpretation of the differential expression analysis, given the results of QC analysis.*

5.2 Introduction

As has been discussed in previous chapters, RNA-Seq data has been previously used as part of this PhD project in order to investigate miRNA-mediated gene regulation by attempting to infer the biological relevance of putative miRNA-mRNA interactions, and thereby increase the accuracy of miRNA target prediction in animals.

In this chapter, and the subsequent chapter, integrated analyses are undertaken using both RNA-Seq and sRNA sequencing datasets in order to better understand the extent of regulation exerted by individual microRNAs in specific biological processes. In this type of analysis, and the analyses presented in previous chapters, RNA-Seq data is fundamental in investigating the miRNA-mediated regulation of gene expression in any given biological context.

For this chapter in particular, the biological process being studied is the transcriptomic response of both male and female *D. melanogaster* flies to mating.

5.3 Background

The purpose of this study was to examine the post-mating response (PMR) in both male and female *Drosophila melanogaster*. Due to the differing reproductive and mating strategies of the two sexes, the post-mating response in male and females is expected to differ. Typically, mated females which mate multiply, exhibit a refractory response to

mating in order to support and sustain the production of fertile eggs. In contrast, the likely behavioural strategy of male fruit flies after mating is to minimise the length of any potential mating induced refractory periods, in order to re-commence mating (Fowler, et al., 2019).

Comparatively speaking, more research has been conducted on the post-mating response in female than male *D. melanogaster* (Ravi Ram and Wolfner, 2007; Sirot, et al., 2015). Female fruit flies are known to have behavioural and physiological responses to mating which is induced by seminal fluid proteins from the mating male (Sirot, et al., 2011; Wigby, et al., 2009), including increased oogenesis, ovulation and feeding (Carvalho, et al., 2006).

In males, once mating has occurred, both sperm and seminal fluid proteins must be replenished, which is a process which can take over 24 hours. There is also evidence to suggest that a more systemic post-mating response may occur in the male including changes in relation to the immune system (Winterhalter and Fedorka, 2009).

Due to the high similarity between male and female genomes for the fruit fly, PMRs are thought to be co-ordinated by changes in gene expression (Williams and Carroll, 2009), as well as other non-genomic responses, including differing patterns or neurotransmitter release (Heifetz, et al., 2014). miRNAs are known regulators of gene expression, and are known to regulate sex-related processes in *Drosophila* such as SFP production (Mohorianu, et al., 2018) and ovary morphology (Chen, et al., 2014). A comparative analysis of transcriptomic

changes in male and females in response to mating is lacking, including the particular role of miRNAs in regulating the post-mating response.

In this study, we hoped to test the hypotheses that there are significant changes in gene expression between virgin and mated flies for both sexes, as well as the hypothesis that the mode and nature of changes in gene expression related to PMR are different in the two sexes.

Similar studies have been conducted examining the expression profiles of the whole body of the female fruit fly in response to mating (Delbare, et al., 2017; Innocenti and Morrow, 2009; Lawniczak and Begun, 2004; McGraw, et al., 2008; McGraw, et al., 2004; Zhou, et al., 2014), whilst other studies have examined specific body parts (Dalton, et al., 2010; Kapelnikov, et al., 2008; Mack, et al., 2006; Prokupek, et al., 2009). Whilst similar studies have been performed on female insects of closely related species (Alfonso-Parra, et al., 2016; Gomulski, et al., 2012; Immonen, et al., 2017; Kocher, et al., 2008; Rogers, et al., 2008). A general outcome of these studies is that ‘PMRs can induce pervasive, genome-wide gene expression changes in reproductive, sensory and immune system genes’ with some gene expression changes being signatures of processes related to mating (Fowler, et al., 2019).

Comparatively fewer studies of this type have been completed in males. There are studies of whole body gene expression profiles of males of related species after mating (Gomulski, et al., 2012; Immonen, et al., 2017), and also an expression profile of the head of the male fruit fly after mating (Ellis and Carney, 2010).

Mating between fruit flies induces a *post-mating response* (PMR) in both males and females of this species. The behavioural PMR is different for both sexes, which is predictive of transcriptional differences in the PMR. As miRNAs are known regulators of developmental and physiological change, they are candidate regulators of the post-mating response in both male and female fruit fly. As miRNAs predominantly act on protein-coding transcripts, mRNAs are also likely to be included in PMR-associated gene regulatory networks, and are therefore investigated in this study. The transcriptional post-mating response is also likely to differ according to body part, due to the location of sex-related organs in different regions of the body, and so both the head-thorax and the abdomen of the fruit fly are sampled and sequenced during this study.

5.4 Methodology

5.4.1 Experimental Design

The treatment condition of interest in this study was the state of *matedness* (*i.e.* the virgin state or the mated state) of the fruit flies being examined. Other variables of interest were the sex and body part of the fly examined. The fly body parts examined as part of this study were the head/thorax (pooled), and the abdomen.

Combination of these three variables lead to the generation of eight experimental conditions of interest (table 5.1):

Condition	Sex	Matedness	Body part(s)
MMHT	male	mated	head/thorax
MMAb	male	mated	abdomen
MVHT	male	virgin	head/thorax
MVAb	male	virgin	abdomen
FMHT	female	mated	head/thorax
FMAb	female	mated	abdomen
FVHT	female	virgin	head/thorax
FVAb	female	virgin	head/thorax

Table 5.1 - The experimental conditions examined in this study. Conditions are based on a combination of the following variables of interest: sex, matedness, and body part(s). Abbreviations: MM (male-mated), MV (male virgin), FM (female-mated), FV (female virgin), HT (head/thorax), Ab (abdomen).

The relationship between all of the conditions listed above were explored using mRNA and sRNA transcriptomics through a process of RNA extraction, cDNA library preparation and subsequent sequencing. Two biological replicates were used for each condition. Each biological replicate represents RNA pooled (in order to generate sufficient RNA for sequencing) from 50 individual *D. melanogaster* organisms. Therefore, in total, 16 samples were sent for sequencing (8 conditions x 2 biological replicates per condition).

5.4.2 Sample preparation

Wildtype *D. melanogaster* flies were collected from a large laboratory population originally in the 1970s in Dahomey (Benin). Flies were reared on standard sugar yeast (SY) medium (100 g brewer's yeast powder, 50 g sugar, 15 g agar, 30 ml Nipagin (10% w/v solution), and 3 ml propionic acid, per litre of medium) in a controlled environment (25°C, 50% humidity, 12:12 hour light:dark cycle). Larvae were raised at a standard density of 100 per vial (glass, 75x25mm, each containing 7ml SY medium). Male and female adults were separated within 6 hours of eclosion using ice anaesthesia and stored in single sex vials at a density of 10/vial for 6 days. For the mated treatment, a single male was placed with a female and the time of mating was recorded. Immediately after mating the male was removed to a separate vial to prevent further mating. All mated flies were then flash frozen at 3 hours after start of mating in liquid N₂. For the virgin treatment, males and females were housed individually in vials for ~3-4 hours before flash freezing. Frozen flies were stored at -80°C until use.

5.4.3 RNA Extraction

To prepare tissue for RNA extraction, 50 flies from each sex, treatment and biological replicate were separated into HT and Ab tissues on dry ice, and the body parts were then pooled for RNA extraction (note that both body parts were intact, and thus the Ab contained the germline). Tissues were disrupted by grinding under liquid nitrogen, then total RNA was extracted using the miRvana miRNA isolation kit (Ambion, AM1561), according to the kit protocol. RNA was eluted in RNA storage solution (1 mM sodium citrate, pH 6.4 +/- 0.2, Ambion). Samples were DNase treated to remove residual genomic DNA (Ambion Turbo DNA-free kit, AM1907). RNA was assessed for quantity and quality using a NanoDrop 8000 spectrophotometer.

5.4.4 Library construction and sequencing

The 16 samples were sent to the Earlham Institute provider (Norwich Research Park, UK) for mRNA and sRNA library construction, and sequencing. Libraries were constructed using the Illumina TruSeq kit. For the sRNA libraries, a modified 'blocking oligo' was also used to preclude adapter ligation to the highly abundant 30nt 2S rRNA (Fowler, et al., 2018). Non-directional, single end RNA-seq was conducted using the Illumina HiSeq2500 platform with 50nt read length.

5.4.5 Sequence analysis and differential expression analysis

Kallisto version 0.46.0 (Bray, et al., 2016) was used to pseudoalign reads to the Berkeley *Drosophila* Genome Project 6 (BDGP6) cDNA sequences downloaded from Ensembl (release 89, (Zerbino, et al., 2018)). A kallisto index was created using the “kallisto index” command (k-mer size 31). Kallisto quant was used to obtain transcript count estimates and parameters were set to include 100 bootstrap samples and to perform sequence bias correction. Transcript to gene mappings were obtained using biomaRt (Durinck, et al., 2009) and transcript counts were aggregated in Sleuth (version 0.28.1) (Pimentel, et al., 2017) before calling pairwise differential expression between mated and virgin samples of the same body part and sex. Small RNA reads were converted from FASTQ to FASTA format and then processed to trim sequencing adaptors using a custom Perl script recognising the first 8 bases of the adapter sequence (‘TGGAATTC’). Trimmed reads were then aligned to miRBase (v22.0) *D. melanogaster* mature miRNA sequences using PatMaN (Prüfer, et al., 2008) (parameters -e 0 -g 0). A custom Perl script was used to parse the alignment files and generate an aligned read count table across all samples. DESeq2 (version 1.14.1) (Love, et al., 2014) was used for normalisation of counts between samples and calling differentially expressed miRNAs.

5.4.6 miRNA target prediction

This analysis is conducted predominantly using the R statistical computing environment (Team, 2013) (v3.5.1), with the additional use of custom shell scripts within a UNIX operating system environment.

All relevant miRNA data is downloaded from release 22 of miRBase. Shell commands were used to translate records from fasta format to tab-separated values (TSV) format. The three columns of the miRNA TSV file are the miRNA identifier, the miRNA seed (nucleotides 2-7) and the NCBI taxonomic ID of *D. melanogaster* (i.e. 7227).

As discussed previously, when using any of the TargetScan algorithms, target prediction is performed exclusively on the 3'UTRs of mRNA transcripts. Full-length cDNA sequences of mRNA molecules were obtained for the upstream process of transcript quantification using RNA-Seq data, however, these sequences are not delineated according to transcript feature (e.g. the 3'UTR), and so another source of sequence information is required.

A connection is made to the *Ensembl* (Zerbino, et al., 2018) biomaRt resource (Durinck, et al., 2009) (v2.38.0) from R in order to obtain transcript identifiers, and corresponding 3'UTR sequences from release 89 of Ensembl. In Ensembl, for *D. melanogaster*, transcript and gene annotations derive from release 6.02 of FlyBase (dos Santos, et al., 2014), which uses the 6th release of *D. melanogaster* genome from the Berkeley Drosophila Genome Project (GCA_000001215.4) (Hoskins, et al., 2015).

An issue with this data representation however, is that although upstream transcript quantification and differential expression analyses were conducted at the gene level, 3'UTR sequences as represented here, exist at the level of transcripts, and as such are not directly comparable

with gene-level analyses. Therefore, in order to enforce coherence between miRNA target prediction and upstream analyses, target prediction has to be conducted at the level of individual genes. However, the problem being, that the 3'UTR is not a gene-level sequence feature. As a result, some method had to be developed to assign gene-level models of the 3'UTR. There are different layers of complexity that can be considered when trying to develop a gene-level model of the 3'UTR. Each mRNA transcript (*i.e.* defined set of contiguously-joined exons) possesses its own abundance relative to other transcripts of that same gene. The transcript model, for each gene, with the longest 3'UTR was taken as being representative of that gene. This approach maximises the sensitivity of the target analysis, as the longest 3'UTR will contain all putative target sites, though potentially at the expense of prediction specificity, as the mRNA transcript with the longest 3'UTR, may not be the most abundant coding transcript for a gene within a given context. Genes which do not contain any transcripts containing 3'UTRs are not used for subsequent analysis. In cases in which multiple transcripts of the same gene all contain the maximum 3'UTR sequence of that gene, one of these transcripts is selected at random as being representative. Approximately 40% of fruit fly 3'UTRs contained multiple splice isoforms of that 3'UTR.

The next stage of the analysis was to perform miRNA target prediction, which involved the selection of a suitable target prediction algorithm. In particular, the decision had to be made whether to use an algorithm solely for the classification of mRNA transcripts as potential target or non-targets of a given miRNA, or alternatively, whether to use an algorithm which would use some form of regression model to score the predicted effectiveness of putative target sites. Preliminary analyses had

revealed that there was very poor or non-existent correlation between scores deriving from the context++ model (Agarwal, et al., 2015) (fig 5.1), and fold changes observed for predicted targets of miRNAs differentially expressed between two conditions. The primary assumption in this analysis being, that for a miRNA differentially expressed between two conditions, the expression of the direct targets of that miRNA would be expected to be perturbed as a result of the differential expression of the targeting miRNA.

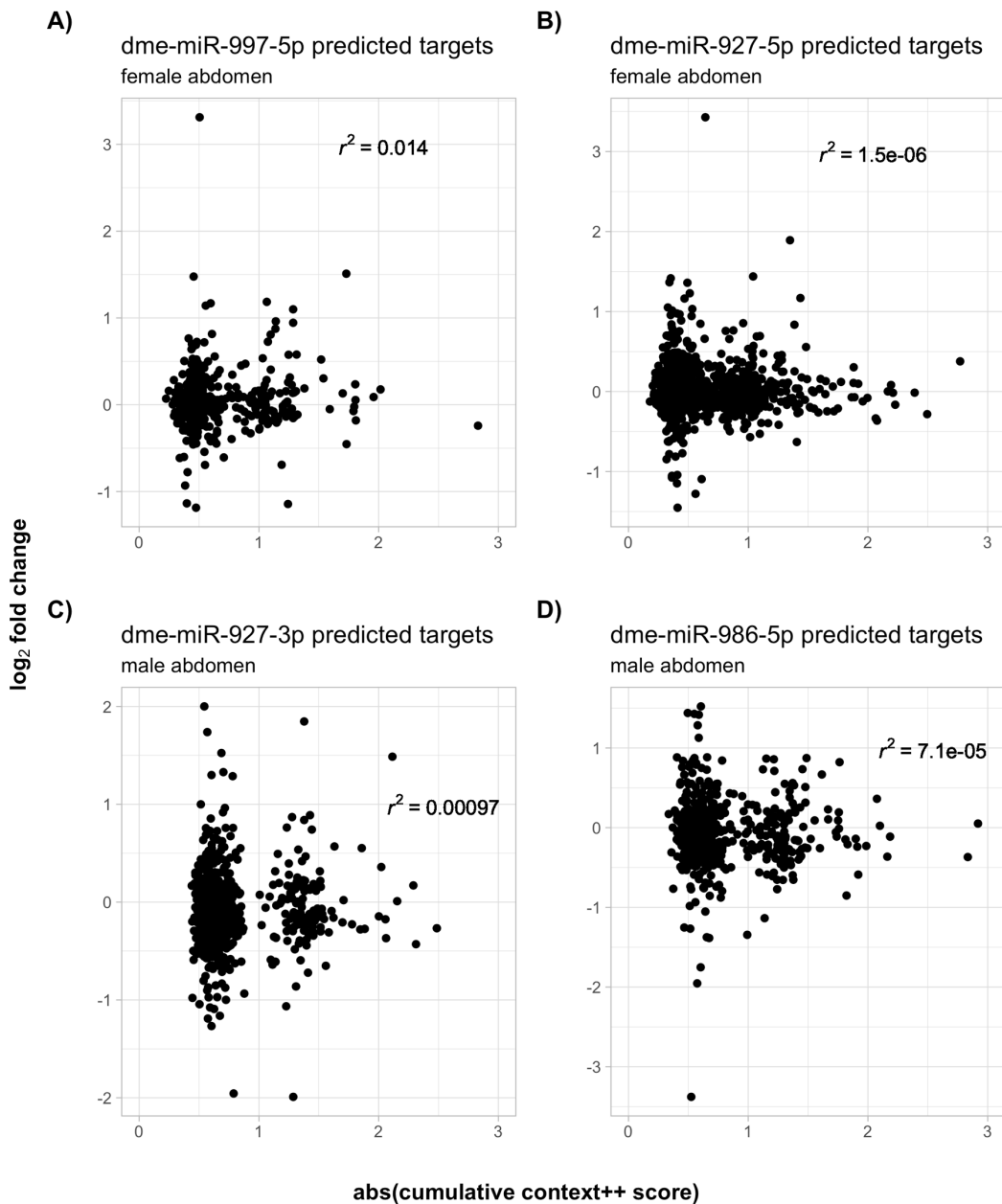


Figure 5.1 – The relationship between context++ scores and fold change between two conditions for the predicted targets of a differentially expressed miRNA. On the x-axis is plotted the cumulative context++ scores, which is the summed context++ score for multiple potential targets for each predicted target transcripts. All context++ scores are negative, but have been plotted with their absolute values (i.e. modulus values) for the sake of clarity. Plotted on the y-axis are log₂ fold changes relating to the post-mating response in a given comparison (given in subplot subtitles). The miRNA used in this analysis is are, for each respective subplot: **A)** dme-miR-997-5p for the female abdomen (LFC=6.83, p.adj=0.033); **B)** dme-miR-927-5p for the female abdomen (LFC=-3.77, p.adj=0.062); **C)** dme-miR-

927-3p for the male abdomen (LFC=-3.42, p.adj=0.0075) and **D**) dme-miR-986-5p for the male abdomen (LFC=6.26, p.adj=0.076). Outliers have been omitted.

As a result of this preliminary analysis, it was decided to proceed with a classification algorithm, without additional scoring of putative targets. As there is consistent evidence for the effectiveness of seed-based targeting rules (Agarwal, et al., 2015; Bradley and Moxon, 2017), and evidence that non-canonical target predictions can lead to an inflated number of false positive results (Agarwal, et al., 2015), the ‘TargetScanS’ algorithm (Lewis, et al., 2005) was chosen for this analysis, which identifies 7mer-1a, 7mer-m8 and 8mer target sites.

5.4.7 Data Pre-processing and Normalisation

Before null-hypothesis significance testing, a process of data pre-processing and cleaning was undertaken. The arithmetic mean average of TPM values between biological replicates was computed. Genes with average TPM values of 0 in either the virgin or the mated conditions were removed from further analysis for a particular comparison, as previous analysis has indicated that removal of lowly expressed mRNA increases the accuracy of miRNA target prediction (Bradley and Moxon, 2019). A pseudo-count of 1 was then added to each average TPM value in order to mitigate against the large stochasticity in transcript abundance typically observed at low expressions values.

As discussed earlier, during exploratory data analysis, it was discovered that 3’UTR length produced a strong confounding effect when evaluating fold change expression values between any two conditions. In order

to mitigate against this confounding effect, a procedure was implemented before statistical testing in order to normalise fold change values for 3'UTR length. The normalisation procedure is described as follows:

A histogram of 3'UTR sequence lengths was constructed separately for predicted target and nontarget datasets, starting from 0, in increments of 200nt, and to a maximum representing the maximum sequence length from both target and non-target datasets. Each break of the two histograms are iterated through, and for each iteration, log fold change values for predicted target and nontarget datasets are restricted to fall within the 3'UTR sequence length range given by the individual histogram breaks. Within this range, of the target and nontarget log fold change vectors, if vector sizes are unequal, the vector with the largest number of records is sampled to match the number of observations contained within the smaller vector. Log fold change values for both predicted target and nontarget datasets are concatenated for each iteration, in order to create log fold change distributions which are normalised for 3'UTR length. In the case of the Fisher exact test, an identical sampling procedure is implemented with the exception that transcript identifiers are sampled instead of log fold change values.

5.4.8 Integrated Analysis

A process of null hypothesis significance testing was undertaken in order to test whether the predicted targets of differentially expressed miRNAs differed significantly from predicted non-targets. The one-sided Kolmogorov-Smirnov (KS) and Fisher exact tests were used for this

purpose. The KS test is used to test for the equality between two continuous distributions, and in this instance, is used to test for the equality of predicted miRNA target and non-target distributions – resulting in use of the ‘two-sample’ form of the KS test. The Fisher Exact test is a test for enrichment, and was used test for the enrichment or depletion of miRNA target sites on downregulated and upregulated transcripts respectively. In addition, Fisher Exact and KS tests were also similarly conducted to test for potential combinatorial effects of different pairwise combinations of miRNAs which were differentially expressed in the same direction with predicted target sets designated as those mRNAs with predicted targets for both differentially expressed miRNAs.

Correction for multiple comparisons was conducted using the Benjamini & Hochberg correction (Benjamini and Hochberg, 1995), with an FDR value set at <0.05 . In order to counteract the stochasticity introduced into the analysis by the 3’UTR normalisation process described above, for each test, p-values were calculated 100 times, and the mean average p-value was taken as being representative.

5.5 Results

5.5.1 QC

Data relating to the numbers of reads generated from each library, and the proportion of the reads mapping or pseudoaligning to the reference genome or transcriptome is given in appendix B (figure B.1, B.2 and tables B.1, B.2)

In summary, for mRNA-seq, sequencing depth is generally consistent across duplicates, although inconsistent for samples of a different type (*e.g.* mated female abdomen). However, sequencing depth is consistent across duplicates and sample types for sRNA sequencing libraries.

Principal components analysis (figure 5.2) reveals that broadly speaking, the mRNA and the miRNA sequencing data cluster according to sex and body type, and also to some extent the mated status of the flies sequenced.

There are some exceptions to these broad trends however. Firstly, when examining the principal components analysis for the mRNA sequencing data, we can see that the body type (*i.e.* the head-thorax or the abdomen) are broadly separated along the first principal component, with the sex less clearly distinguished along the second principal component. However, mated female head-thorax cluster more closely to the corresponding male samples of this type – indicating potential issues with these samples.

With the miRNA PCA analysis we see again the body part samples are separated along the first principal component. Although abdomen samples are clearly separable by sex along the first principal component, all abdomen samples are closely clustered together irrespective of sex. This would indicate that there is sexual asymmetry in the transcriptional profile in the abdomen for miRNAs, but not in the head-thorax.

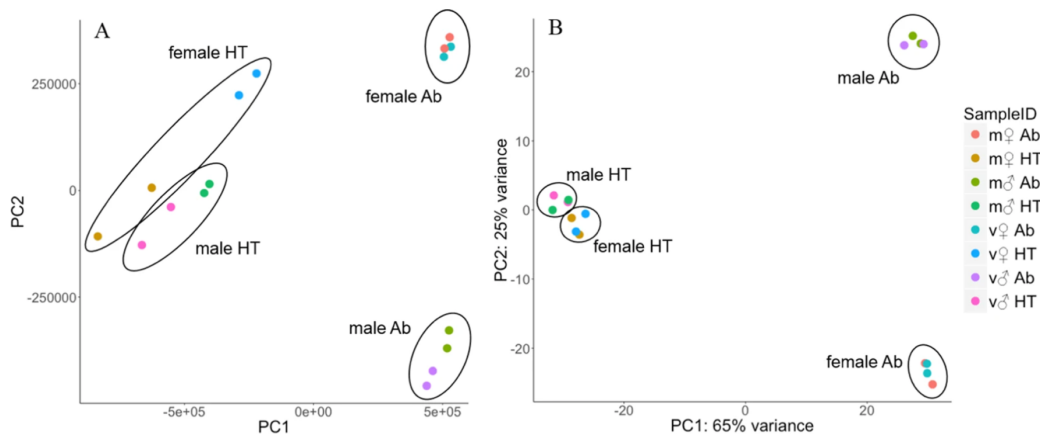


Figure 5.2 – Principle components analysis of expression data for both A) mRNA and b) miRNA in this study. Points are coloured by a combination of properties relating to the fly's sex, body part sequenced, and matedness status. Plots are manually annotated with groupings according to points with common sex and body part sequenced (black).

5.5.2 Differential Expression Analysis

Variable numbers of both protein coding genes and miRNAs were found to be differentially expressed between different comparisons (ta-

ble 5.2). In addition, patterns of protein-coding and miRNA gene up-regulation and downregulation were different between different comparisons.

Comparison	Gene Type	Num. upregulated	Num. downregulated	Total num. DE
female abdomen	Protein-coding	106	19	125
female head-thorax	Protein-coding	628	1412	2040
male abdomen	Protein-coding	1507	561	2068
male head-thorax	Protein-coding	0	0	0
female abdomen	miRNA	3	1	4
female head-thorax	miRNA	0	0	0
male abdomen	miRNA	0	2	2
male head-thorax	miRNA	0	0	0

Table 5.2 - The results of the differential expression analysis of miRNA and protein coding genes for comparisons relating to both sex and body part. In each comparison the respective mating and virgin conditions are compared (e.g. mated female abdomen vs. virgin female abdomen). Data is derived from mRNA and sRNA sequencing experiments with two biological replicates per condition.

It is important to examine potential reasons for the patterns of differential expression observed in table 5.2. Firstly, for protein-coding genes, volcano plots (figures B.3) reveal that although no genes are called as being differentially expressed for the male head-thorax, there are a large number of genes for this comparison which exhibit large changes in expression. It is unlikely that this can be explained by a large degree of technical variance from these samples as the technical variance distribution for these samples does not differ remarkably than from other comparisons (figure B.5). Whilst this comparison does contain genes with seemingly large expression changes, most of these genes are either poorly expressed (figure B.6), and have large standard errors for the beta effect size parameter (suggesting high uncertainty in the log-fold change estimates – figure B.7). This evidence together helps give a proximal understanding for the lack of differentially expressed genes for the male head/thorax. The same reasoning can be applied to explain the relatively low number of protein coding genes found to be differentially expressed in the female abdomen. However, it does not provide a biological reason for the lack of differentially expressed genes in these conditions.

A similar problem is found when trying to explain the very small number of miRNAs which are differentially expressed across all comparisons. Again, volcano plots reveal that the problem isn't a lack of miRNAs with large differences in expression levels between samples (figure B.4). Again, we can see that those genes tend to have low or moderate expression levels (figure B.8) and high log fold change standard errors (figure B.9). The high standard error values could be attributable to low expression values of these miRNAs (in which fold changes are increasingly variable) or high inter-replicate variability, or due to the small

sample size in this study. The performed principal components analysis (figure 5.2) suggest that perhaps high inter-replicate variability is not the dominant issue in the case. Rather, it is likely that there is insufficient statistical power to call differential expression in most cases with DESeq2 when the sample size is small (*e.g.* $n=2$), as unlike the case with the protein-coding genes, there is no or only a small number of miRNAs called as differentially expressed across all comparisons.

5.5.3 miRNA target prediction

5.5.3.1 miRNA-Gene Interaction Network & GO Term Enrichment Analysis

An initial result from the miRNA target analysis was the discovery of a number of genes which were differentially expressed in the opposite direction of a differentially expressed miRNA for a given comparison, and was also a predicted target of that same miRNA (table B.1). The majority of such interactions are found in the male abdomen, and network visualisations of these interactions highlights some important features of miRNA targeting found in this particular system (figure 5.3).

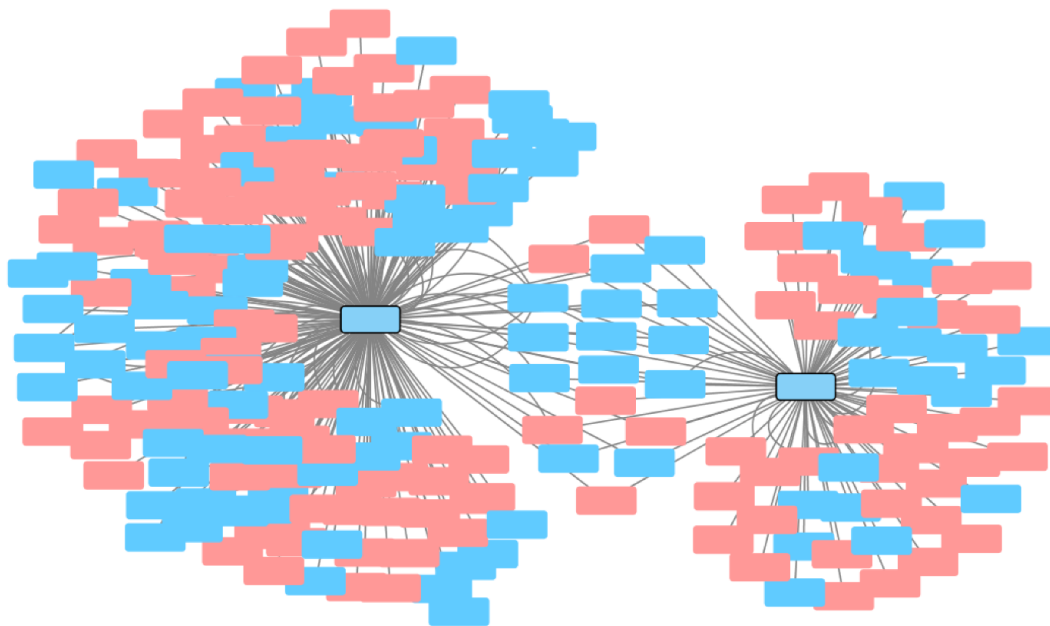


Figure 5.3 - Network visualisation of predicted miRNA interactions in the male abdomen: Nodes with thick borders denote miRNAs, whilst nodes without thick borders represent coding genes. Nodes coloured pale red denote coding genes upregulated in the mated male abdomen ($FDR \leq 0.05$), whilst nodes coloured pale blue denote genes downregulated in the mated male abdomen. Network edges denote a predicted targeting interaction between connected nodes. Network visualisations are not shown for other comparisons which either do not possess any differentially expressed miRNAs, or the number of differentially targets of differentially expressed miRNAs are too low to be informative.

The visualisation reveals the large number of predicted target genes differentially expressed in the opposite direction to the mRNA in this comparison. GO term enrichment analyses using the GOrilla web tool (Eden, et al., 2009), was used to test for the enrichment of GO terms processes in the upregulated predicted targets of dme-miR-927-3p and dme-miR-927-5p using a background reference set of all genes which were found to be expressed in the male abdomen. No GO terms were found to be enriched in these target sets ($FDR < 0.05$).

The identity of the genes which are co-targeted by dme-miR-927-3p and dme-miR-927-5p is given in table 5.3.

Gene name	FlyBase Gene ID	Num. predicted dme-miR-927-5p targets	Num. predicted dme-miR-927-3p targets	3'UTR sequence complexity
RpL37a	FBgn0030616	1	1	0.72
CG5707	FBgn0026593	1	1	0.80
CG17715	FBgn0041004	1	1	0.63
Myo95E	FBgn0039157	1	1	0.71
Nckx30C	FBgn0028704	1	6	0.68
CG13197	FBgn0062449	1	1	0.74
Cyp6a18	FBgn0039519	1	1	0.73
Sxl	FBgn0264270	1	2	0.60
Pdp1	FBgn0016694	1	2	0.59
CG42394	FBgn0259740	1	1	0.75
chrb	FBgn0036165	1	2	0.56
CG12567	FBgn0039958	1	1	0.74
Myc	FBgn0262656	2	1	0.64
Slh	FBgn0264978	1	1	0.74
twi	FBgn0003900	1	1	0.82
su(w[a])	FBgn0003638	1	2	0.67
CG31960	FBgn0051960	1	1	0.84
Pur-alpha	FBgn0022361	1	1	0.65
Nop60B	FBgn0259937	1	1	0.64

Table 5.3 - A table providing information relating to differentially expressed genes co-targeted by miR-927-3p and miR-927-5p. Complexity is calculated with word lengths of size 6 (*cf* miRNA seed length), using an adaptation (Orlov and Potapov, 2004; Troyanskaya, et al., 2002) of the linguistic complexity approach (Trifonov, 1990)

As can be seen from this table, a small subset of these 19 genes possess multiple predicted target sites to either of these miRNAs. There isn't any one gene with a particularly low 3'UTR sequence complexity, and there does not seem to be any clear relationship between the sequence complexity and the number of predicted target sites in the 3'UTR for these miRNAs. A GO term enrichment analysis using the GOrilla web application did not uncover any enriched gene functionality in this gene set.

5.5.3.2 Exploratory Data Analysis

Exploratory data analysis was performed with the intention of discovering patterns in miRNA target prediction data before proceeding onto more formal analyses. Firstly, the proportion of total predicted target sites which were attributable to each target site type was ascertained (figure 5.4):

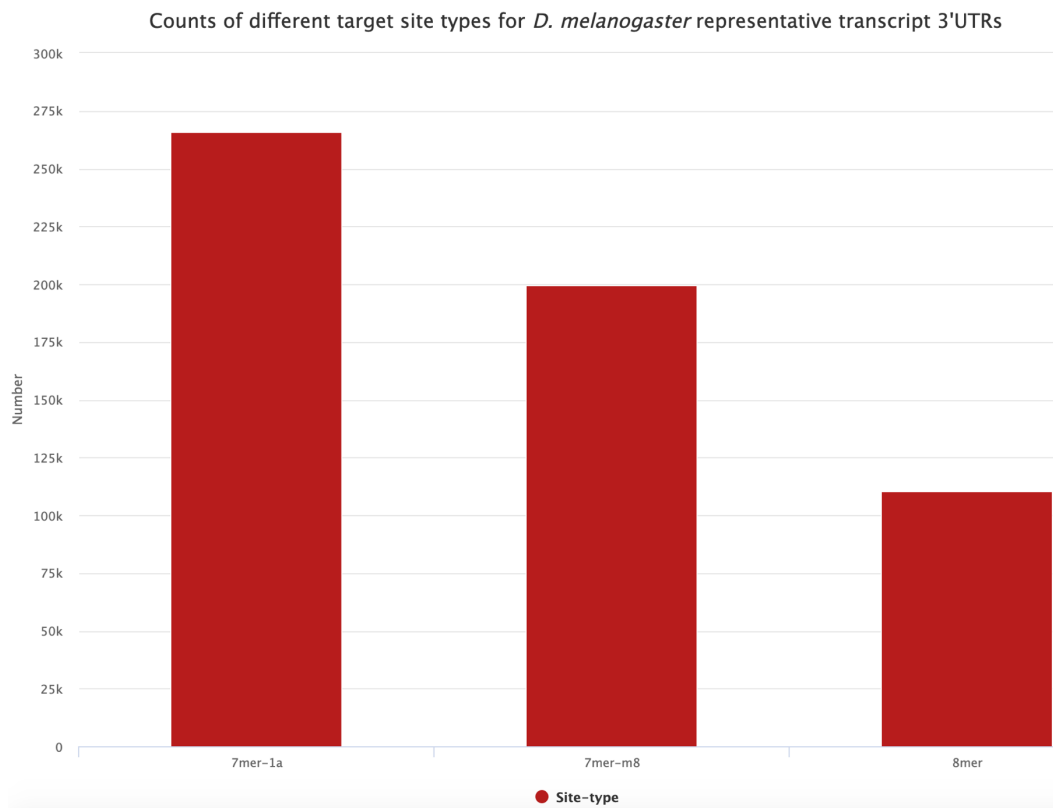


Figure 5.4 - The number of predicted miRNA seed target sites categorised by site type, after running the TargetScanS (Lewis, et al., 2005) algorithm with *D. melanogaster* miRNAs and 3'UTRs. The three types of miRNA target site type observed is the 7mer-1a, the 7mer-m8 and 8mer site types (Bartel, 2018).

As discussed earlier, miRNA target site types differ in their typical efficacy, with the 8mer sites being the strongest site type, followed by the 7mer-m8 site, and the 7mer-1a site (Bartel, 2018). This evidence would predict a somewhat heterogeneous response of predicted miRNA targets as a whole, according to how the different target site types are distributed across those targets. In addition, because some transcripts may be predicted to contain multiple different target sites to the same miRNA, not all of which may be of the same site type, which can act additively (Brennecke, et al., 2005; Doench and Sharp, 2004; Lai, et al., 2005) or synergistically when closely spaced (Grimson, et al., 2007), in

order to confer target repression, a heterogeneous response of predicted targets to miRNA differential expression may be expected.

In addition, not only will mRNA transcripts contain a variable number of predicted targets, to a single miRNA, but when the total ensemble of annotated miRNAs for *D. melanogaster* are taken as a whole, it can be observed that there is a large range in the number of predicted target sites for each mRNA (figure 5.5).

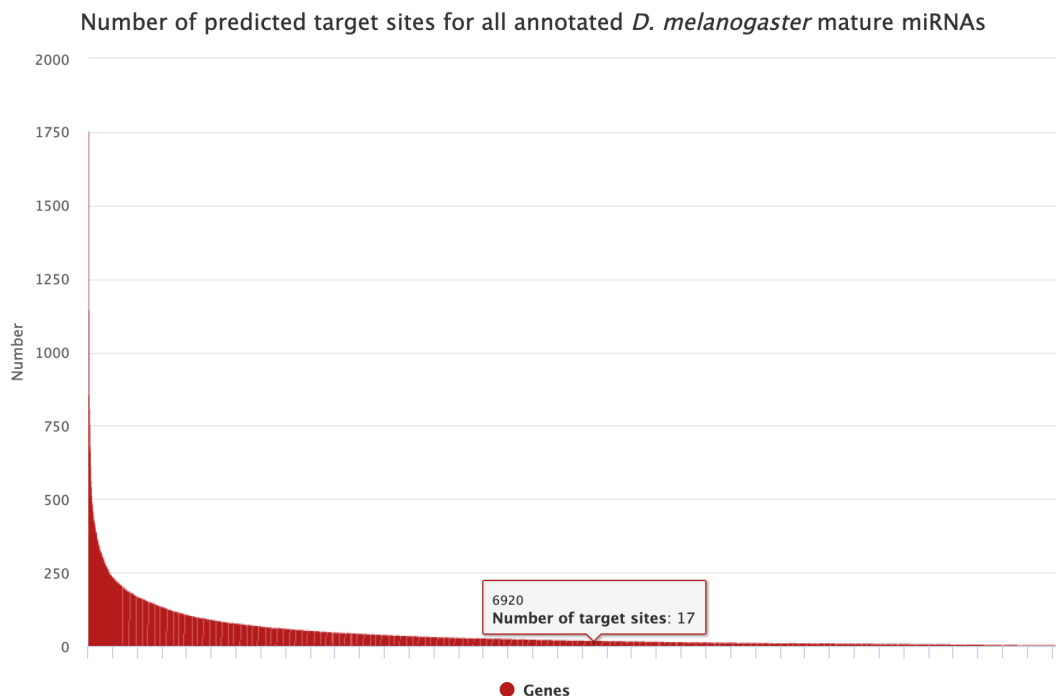


Figure 5.5 - The number of predicted seed miRNA target sites with respect to the gene in which those sites are found. Target predictions are made on the representative 3'UTRs of genes, which are designated as the longest 3'UTR splice isoform for a given gene.

As can be observed from figure 5.4, target site frequency on genes seems to follow a power law distribution, with a relatively small number of genes possessing a large number of predicted target sites, and a

relatively larger number of genes containing a small number of predicted target sites. The non-uniform distribution of total miRNA target sites on 3'UTRs, could perhaps help predict the response of sets of predicted miRNA targets to the combined differential expression of multiple miRNAs *i.e.* the combined differential expression of multiple miRNAs could cause uneven or variable responses of mRNA predicted to be targeted by those miRNAs.

In addition, there is no uniform distribution of the number of predicted target sites possessed by each annotated miRNA as can be observed in figure 5.6:

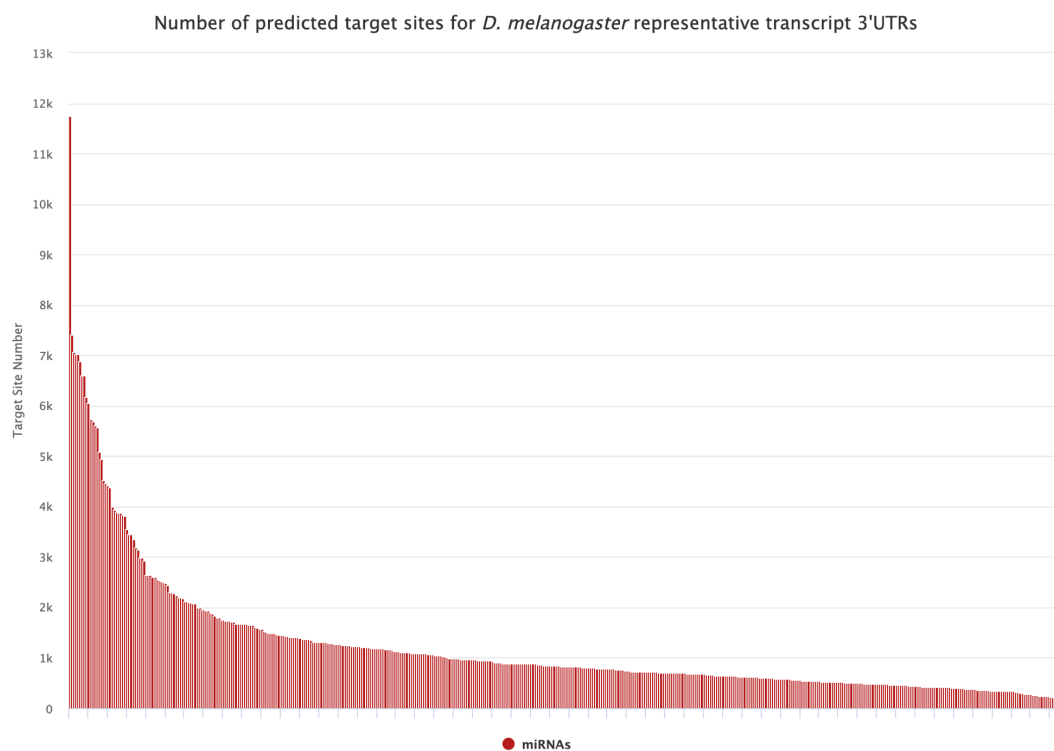


Figure 5.6 - The frequency of predicted seed miRNA target sites with respect to the targeting miRNA. miRNA sequences used encompass all fruit fly miRNAs stored in miRBase.

Given this evidence, it is likely that the extent of the transcriptomic response to miRNA differential expression could largely be dependent on the identity of the miRNA which is differentially expressed, as from figure 5.4, this could lead to a greater than 10-fold difference in the number of mRNA either repressed or derepressed as a result of miRNA differential expression. The caveat of this analysis being that some predicted sites may be non-functional, and that some miRNAs with a large number of predicted targets may possess a low complexity seed sequence which may align to a large number of pre-existing repetitive sequences in 3'UTRs. Inspection of the data reveals that this is likely to be case, with the miRNA with the largest number of target sites (*i.e.* dme-miR-4943-5p) containing very low seed sequence complexity: 'UUUAUUU'. However, previous research has shown that seed regions rich in AU-content lead to relatively unstable binding with targets, and as a result the ability of these miRNAs to repress targets is weaker (Garcia, et al., 2011).

The distribution observed in figure 5.3 may be partially explained by the observed distribution of 3'UTR sequence lengths, which appears to be log-normally distributed (figure 5.7):

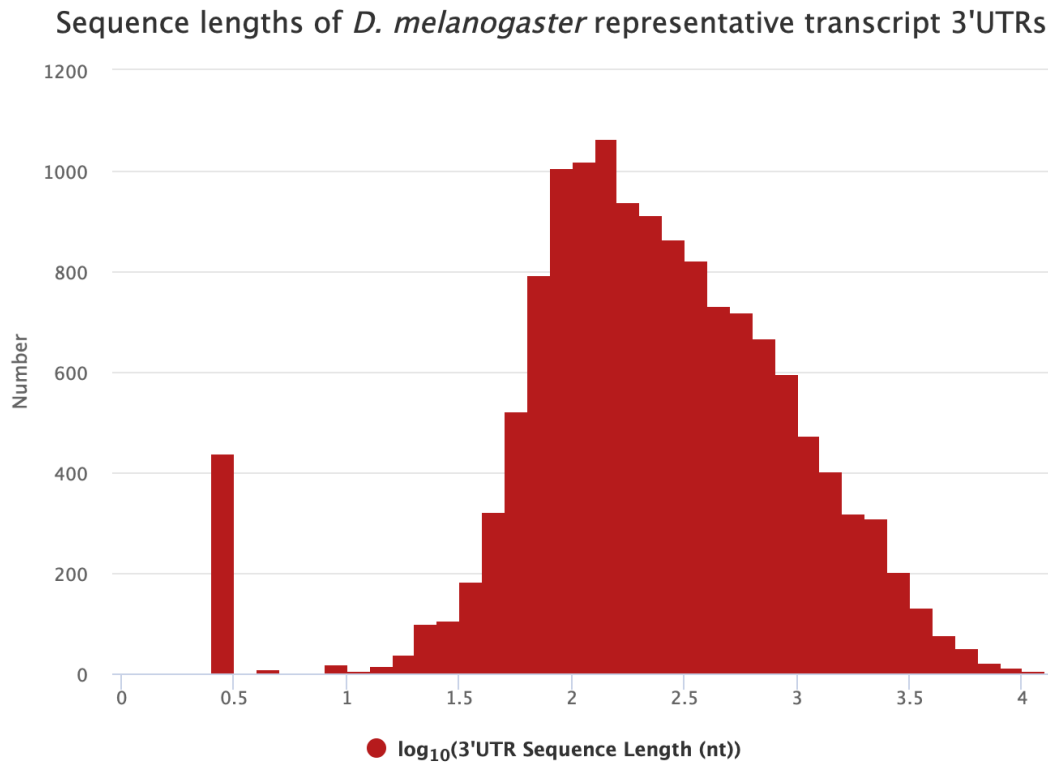


Figure 5.7 – The distribution of \log_{10} 3'UTR sequence lengths for *D. melanogaster* mRNA transcripts obtained from release 89 of Ensembl (Zerbino, et al., 2018). Representative 3'UTRs are designated as being the longest 3'UTR splice isoform for any given gene.

The mode of this distribution is at approximately at a value of 2.2, which corresponds to a 3'UTR sequence length of approximately 160nt. The distribution also appears to be negatively skewed. The substantial number of transcripts with 3'UTRs found with lengths between 1000-10,000nt, potentially explains the non-uniform distribution of the number of predicted miRNA targets on 3'UTRs if it is assumed that there is a positive correlation between observed predicted miRNA target site frequency and 3'UTR length. This correlation is to be expected if either a significant proportion of miRNA target sites are distributed in an unbiased manner, or alternatively if longer 3'UTRs are more extensively regulated. There is an observable peak corresponding to a

3'UTR sequence length of 3nt which is likely attributable to a ceiling effect owing to a minimum assigned 3'UTR sequence length of 3nt.

As observed in a previous study, miRNA target sites tend not to be uniformly distributed across the length of the 3'UTR, with a clear depletion of predicted target sites near the end of the stop codon and the start of the 3'UTR, and conversely an enrichment of miRNA target sites at the distal end of the 3'UTR (Grimson, et al., 2007) (figure 5.8). This enrichment of miRNA target sites towards the end of the 3'UTR can be explained by the reduction in sequence complexity at the distal end of the 3'UTR (figure 5.9).

There is also a very slight, but noticeable trough precisely at the half-way point at the 3'UTR, again corroborating observations made in a previous study (Grimson, et al., 2007).

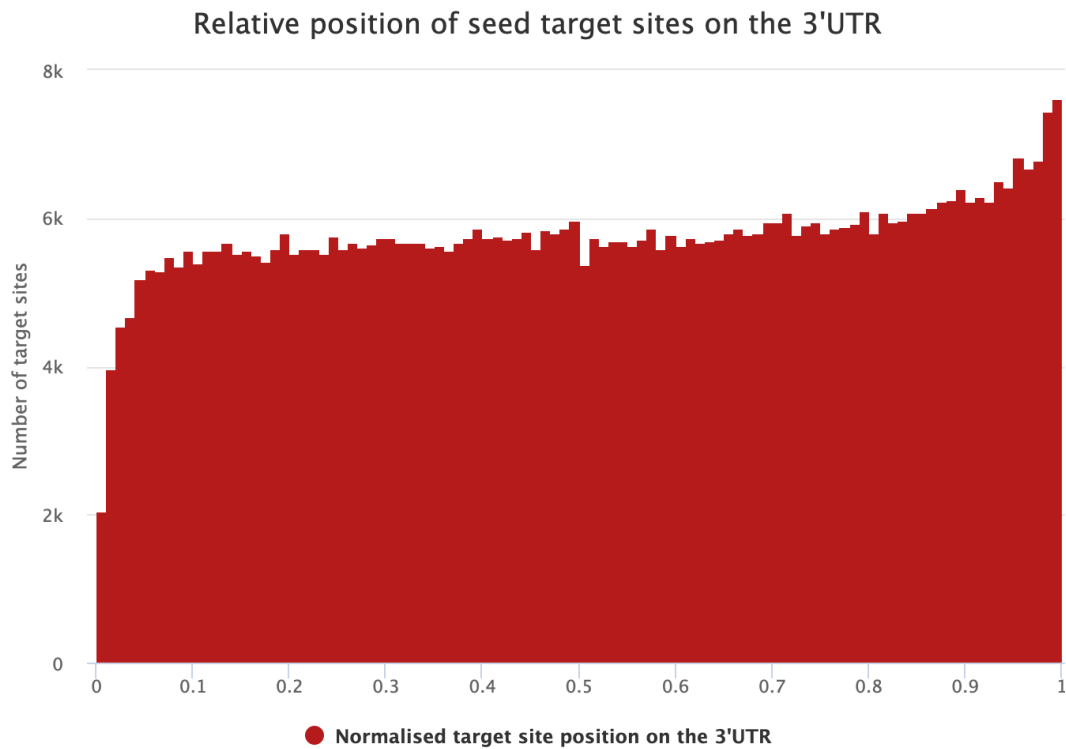


Figure 5.8 – A histogram of length normalised positions of predicted miRNA target sites along *D. melanogaster* 3'UTRs.

Although this observation did not directly impact the analysis, it is further evidence that general miRNA targeting rules and principles observed in this analysis, do not differ considerably from those which have been previously reported.

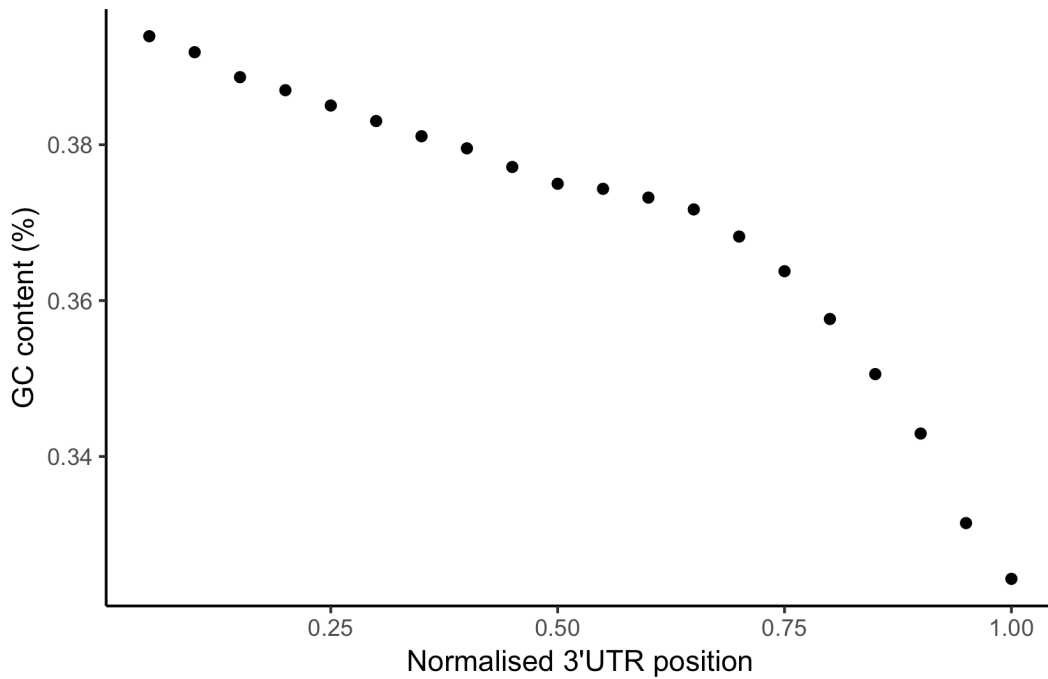


Figure 5.9 – The mean percentage GC content of *D. melanogaster* 3'UTRs along the normalised length of the 3'UTR.

It was also important to examine any potential differences in 3'UTR sequence length between the abdomen and head/thorax as 3'UTR length is known to be a confounder of miRNA mimic transfection analyses (Agarwal, et al., 2015). Analyses of both male and female sequence lengths reveals only a very small difference in 3'UTR lengths (figure 5.10).

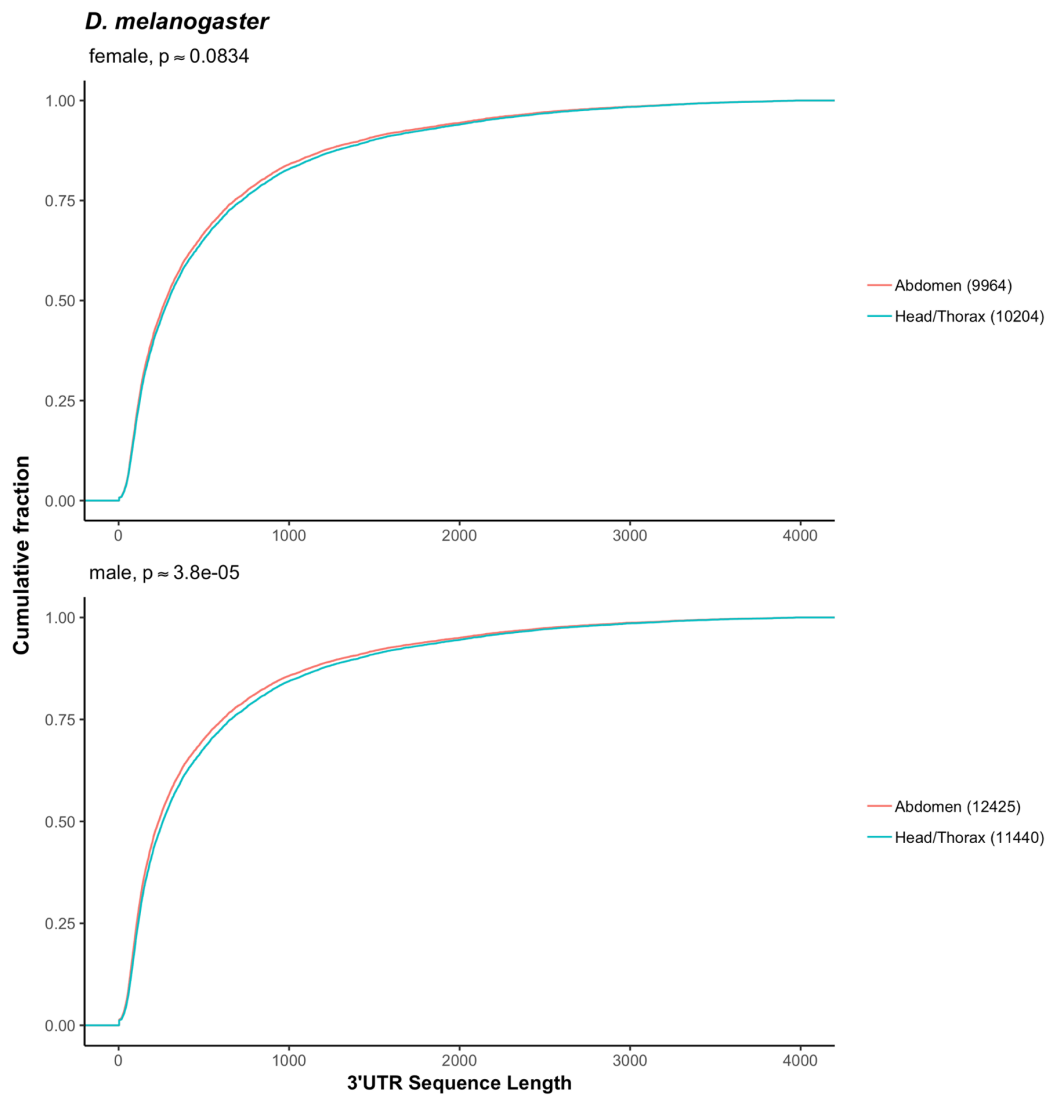


Figure 5.10 – An examination of the cumulative distributions of 3'UTR sequence length for both male and female fruit fly, grouped according to body type. Low abundance transcripts have been filtered from this analysis (less than 5 read counts for more than 47% of samples (Pimentel, et al., 2017)). P-values derive from two-tailed Kolmogorov-Smirnov tests.

Although the determined p-values are low, the effect size is small ($D = 0.018$ for females and $D=0.030$ for males) suggesting that this small difference in 3'UTR lengths is unlikely to be biologically significant.

From the initial exploratory data analysis, additional analysis was undertaken to investigate potential relationships between 3'UTR target site abundance, 3'UTR length, and the differential expression of genes between two different conditions. Analyses were conducted for the two comparisons with a sufficient number of differentially expressed coding genes for the analysis to be informative, namely, the comparison between the virgin male abdomen and the mated male abdomen (2068 differentially expressed coding genes; table 5.1) and also the comparison between the virgin female head/thorax and the mated female head/thorax (2040 differentially expressed coding genes; table 5.1).

In the first analysis of this type, the cumulative distributions of coding genes which are either upregulated, downregulated or not differentially expressed for the male abdomen are plotted with respect to target site frequency of the 3'UTRs of the representative transcripts of these genes (figure 5.11).

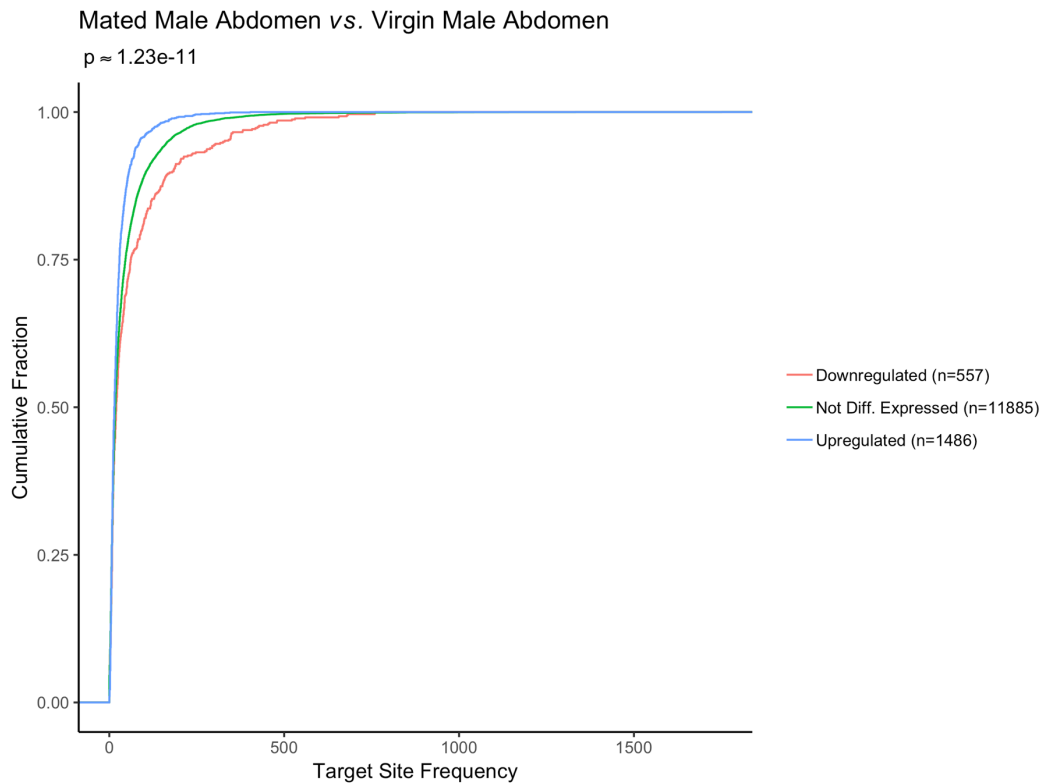


Figure 5.11 - Empirical cumulative distributions of coding genes which are either upregulated, downregulated or not differentially expressed, with respect to predicted miRNA target site frequency. Comparison: Mated male abdomen vs. virgin male abdomen

As can be observed from figure 5.11, upregulated, downregulated genes and genes which are not differentially expressed are not identically distributed with respect to target site frequency on their respective 3'UTRs. In the comparison, there appears to be an enrichment of predicted miRNA target sites in downregulated genes, and a depletion of predicted miRNA target sites in upregulated genes.

Conversely, in the female head/thorax, to some degree, the opposite trend is observed (figure 5.12):

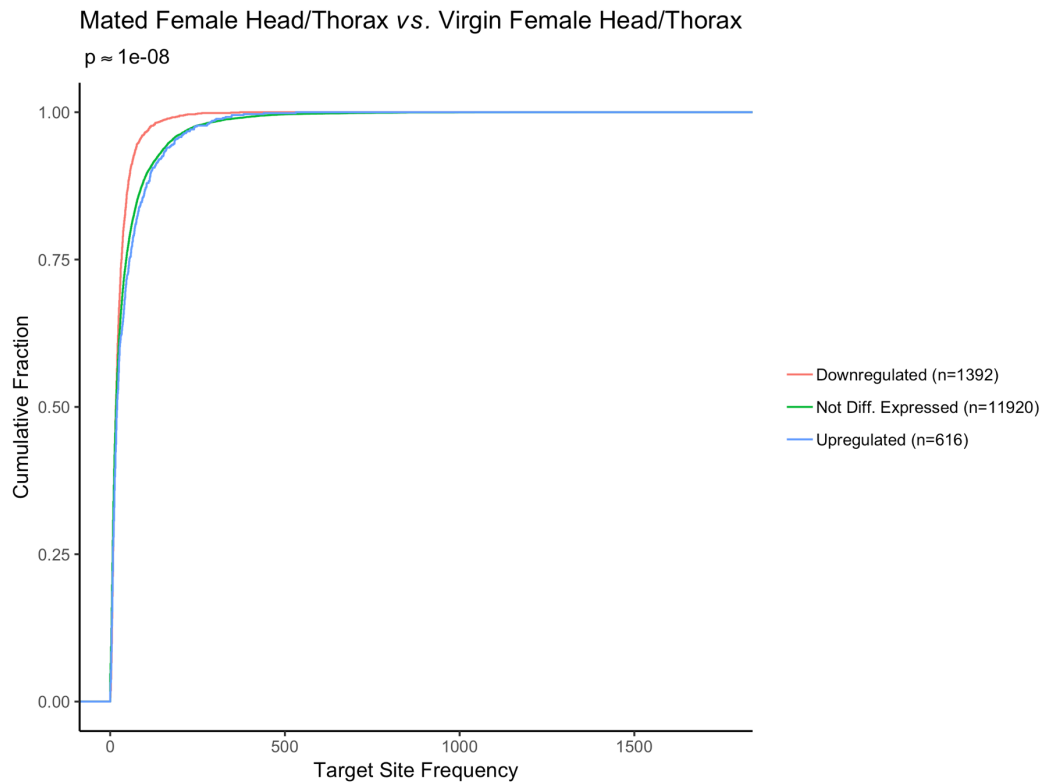


Figure 5.12 - Comparison: Mated female head/thorax vs. virgin female head/thorax – otherwise, as in figure 5.11.

In this comparison, there seems to be a depletion of predicted miRNA targets in the 3'UTRs of downregulated genes, and neither an enrichment nor depletion of predicted miRNA targets in the 3'UTRs of up-regulated genes in comparison to coding genes which are not differentially expressed. The reasons for the differences in these observed patterns are not altogether clear, and seem to be confounded by another variable, namely, 3'UTR length.

When examining similar cumulative plots, though, on this occasion with respect to 3'UTR sequence length, rather than predicted miRNA target site frequency, observed cumulative distribution patterns are similar as to those found in figures 5.11 and 5.12. For example, for the male abdomen, downregulated genes are generally enriched for long 3'UTR

sequences, whilst upregulated genes are generally enriched for shorter 3'UTR sequences (figure 5.13).

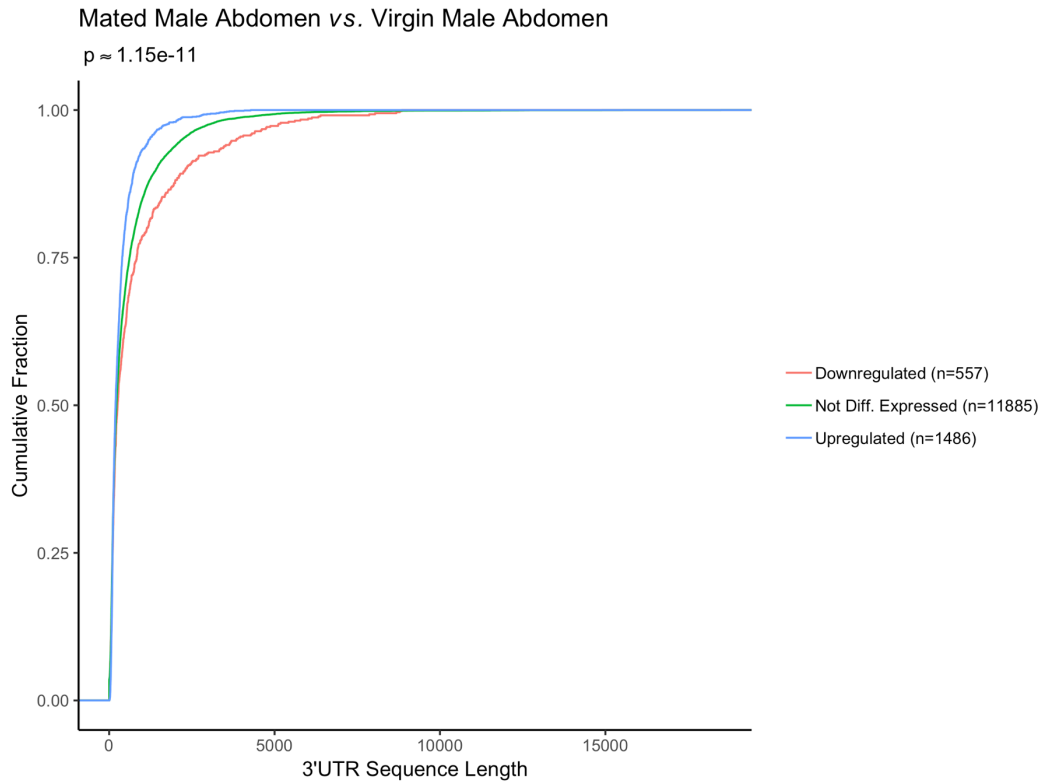


Figure 5.13 - Empirical cumulative distributions of coding genes which are either upregulated, downregulated or not differentially expressed, with respect to 3'UTR length. Comparison: Mated male abdomen vs. virgin male abdomen

In addition, for the female head/thorax, downregulated genes are enriched for shorter 3'UTR sequences (figure 5.14).

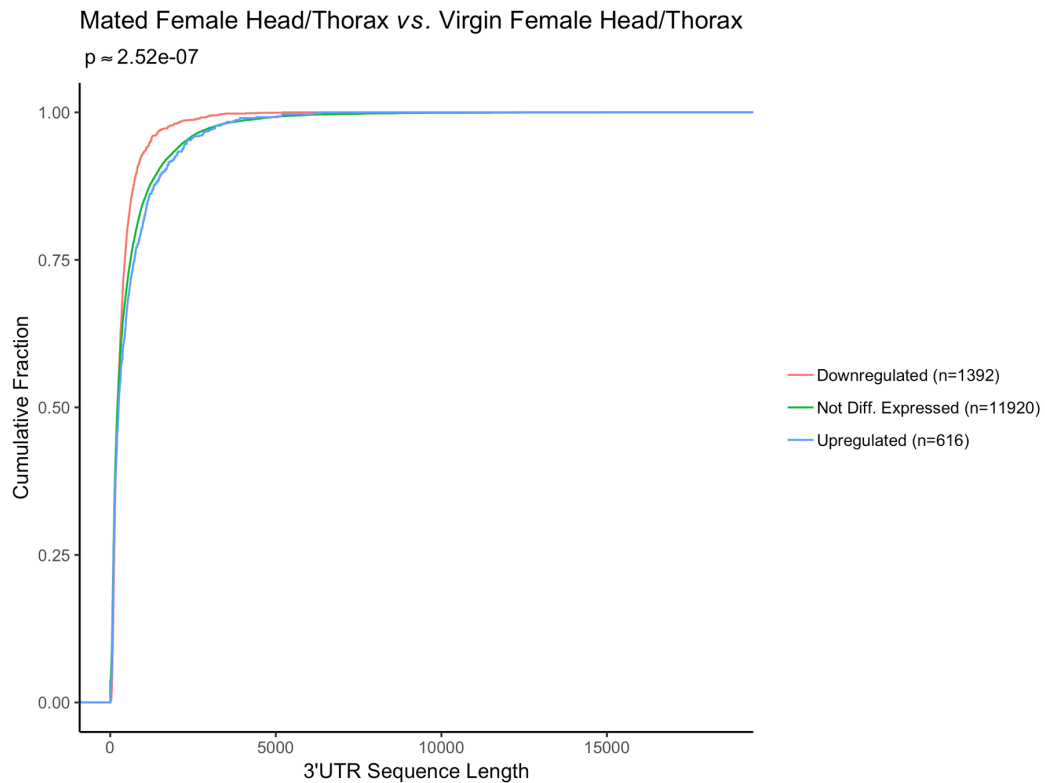


Figure 5.14 - Comparison: Mated female head/thorax vs. virgin female head/thorax. Otherwise, as in figure 5.13.

A relationship between 3'UTR length and predicted target site efficacy has been previously reported, with discoveries that effective miRNA target sites are enriched in shorter 3'UTRs (Agarwal, et al., 2015; Hausser, et al., 2009). This would perhaps explain why in the female head/thorax comparison, downregulated genes are enriched for shorter 3'UTRs.

To determine whether or not the distribution of total predicted miRNA target sites on 3'UTRs was potentially causative of patterns of gene dysregulation observed, a similar analysis to that presented in figures 5.8 and 5.9 was conducted with randomly generated miRNA seed sequences. Simulated seed sequences were generated by sampling (with replacement) seven bases from the list of RNA bases (*i.e.* U,C,A,G) and

concatenating the bases in the order in which they were sampled in order to form a seven letter string. This process was repeated for each seed sequence simulated. For the male abdomen, the cumulative distributions of coding genes which are either upregulated, downregulated or not differentially expressed with respect to the frequency of target sites for simulated seed sequences appears very similar to that for genuine miRNA seed sequences (figure 5.15).

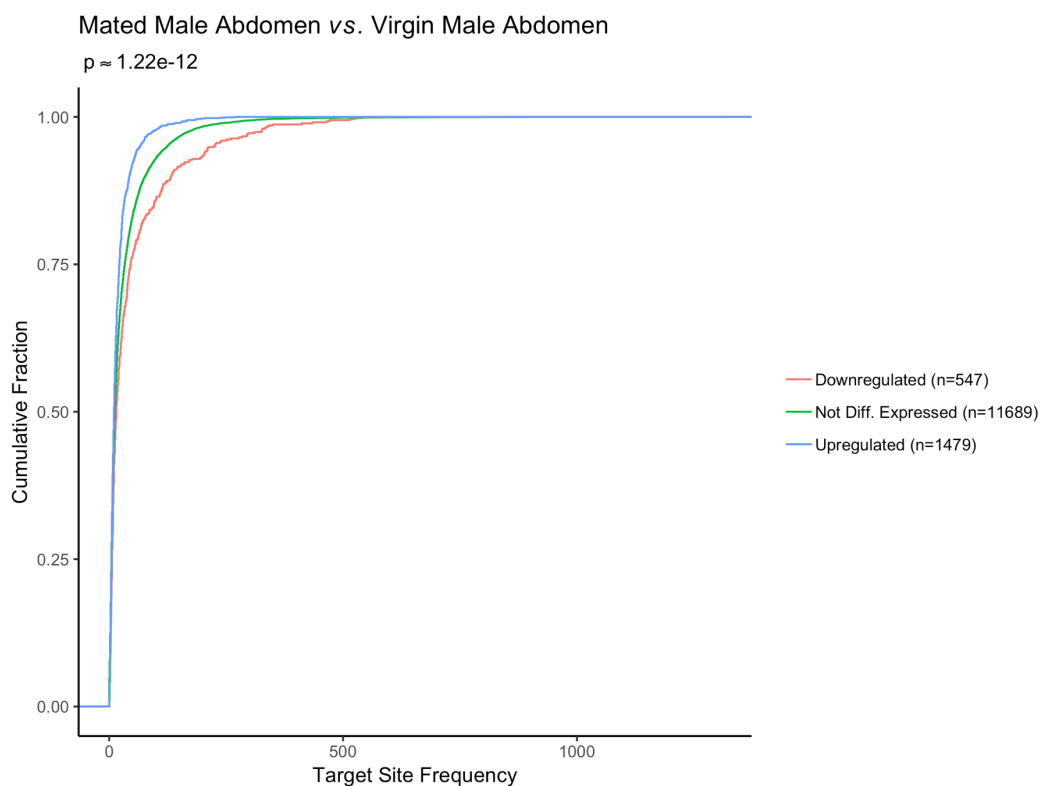


Figure 5.15 - Empirical cumulative distributions of coding genes which are either upregulated, downregulated or not differentially expressed, with respect to predicted target site frequency of randomly generated miRNA seed sequences. Comparison: Mated male abdomen vs. virgin male abdomen. Predicted target sites were generated by executing the TargetScan algorithm with the *D. melanogaster* 3'UTR set and randomly generated miRNA seed sequences.

Similar observations are also made for the female head/thorax comparison (figure 5.16).

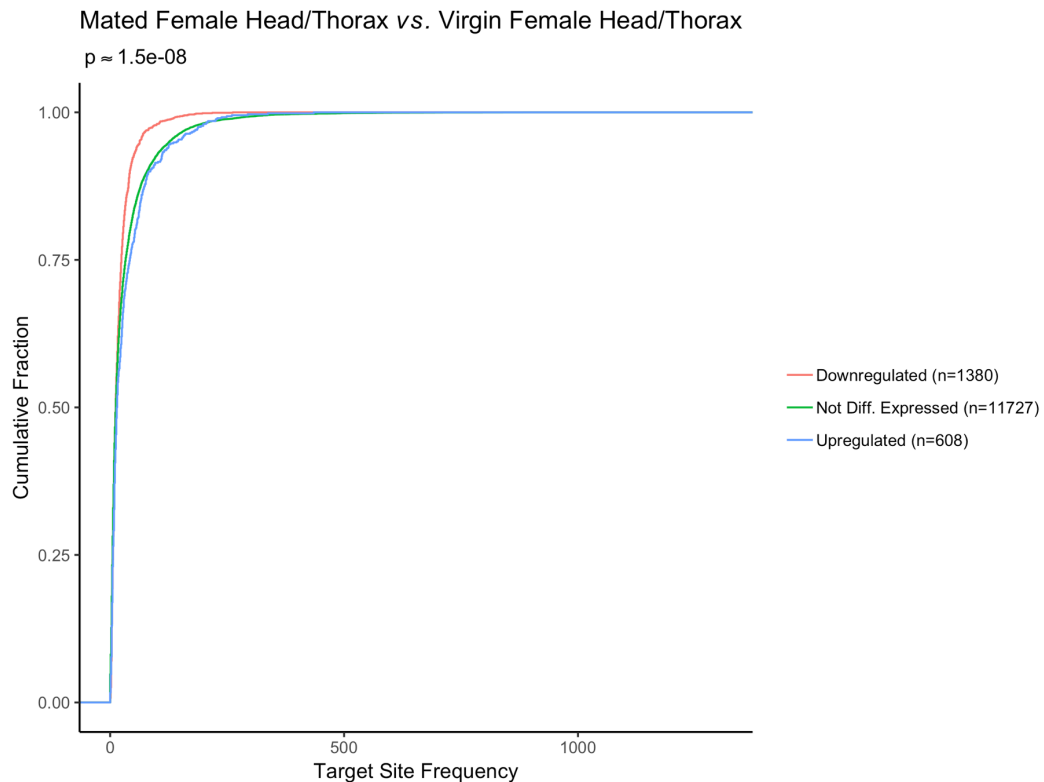


Figure 5.16 - Comparison: Mated female head/thorax vs. virgin female head/thorax. Otherwise, as in figure 5.15.

This evidence suggests that the observed patterns of enrichment of target sites on either upregulated or downregulated genes for the male abdomen and female head/thorax comparisons is unlikely to be adaptive, and that the observed patterns of dysregulation of coding genes is unlikely to be caused by the distribution of miRNA target sites on 3'UTRs. It is more likely in this case, that observed patterns of dysregulation are attributable to differences in 3'UTR lengths in different genes, though in a miRNA-independent manner. It cannot however be concluded that the differential expression of miRNAs in these comparisons has no influence on the observed expression values for the coding

transcriptome, but rather that any potentially existing effect is masked or confounded by the effect of 3'UTR length. As a result, a normalization procedure (see methods) is implemented to control for the effects of 3'UTR length during null hypothesis significance testing.

However, this account does not necessarily explain the large discrepancy in results observed between the male abdomen and the female head/thorax. As can be observed from figure 5.11-5.16, the observed patterns of results are almost inverted between these two conditions. This is particularly mysterious as Agarwal *et al.* (Agarwal, *et al.*, 2015) had noted that whilst the 3'UTR length confounding effect tends to change sign between different studies (*i.e.* is sometimes correlated with mRNA upregulation and also downregulation), it tends to be stable within the same study. One potential explanation is that transcript dysregulation in this study is not only confounded by 3'UTR sequence length, but also by the 3'UTR AU content (a prominent confounding variable in these types of experiments (Agarwal, *et al.*, 2015)). If there are discrepancies in the 3'UTR AU content of expressed genes in the female head/thorax compared to the male abdomen, that could potentially explain the results that are observed.

A similar analysis examining 3'UTR lengths was also attempted on data from the female abdomen (figure B.10) – however, because the small number of both upregulated and downregulated genes, it is difficult to derive meaningful conclusions from this analysis.

5.5.4 Integrated Analysis

For all tests conducted, all p- and adjusted p-values returned were above the chosen significance threshold of 0.05 (see table 5.4).

miRNA	Comparison	Direction	Test	p	Adjusted p
dme-miR-184-5p	Male Abdomen	Up	KS	0.375	0.981
dme-miR-286-3p	Female Abdomen	Down	KS	0.526	0.981
dme-miR-184-5p	Female Abdomen	Down	Fisher	0.607	0.981
dme-miR-997-5p	Female Abdomen	Down	Fisher	0.715	0.981
dme-miR-997-5p	Female Abdomen	Up	KS	0.730	0.981
dme-miR-927-3p	Male Abdomen	Down	KS	0.769	0.981
dme-miR-14-3p	Female Abdomen	Up	KS	0.771	0.981
dme-miR-927-5p	Male Abdomen	Down	KS	0.793	0.981
dme-miR-14-3p	Female Abdomen	Down	Fisher	0.854	0.981
dme-miR-997-5p	Female Abdomen	Up	Fisher	0.940	0.981
dme-miR-14-3p	Female Abdomen	Up	Fisher	0.947	0.981
dme-miR-184-5p	Female Abdomen	Up	Fisher	0.981	0.981

Table 5.4 - A statistics table for the integrated analysis of *D. melanogaster* sequencing data for single mature miRNAs. A table of p values and adjusted p values deriving from use of the Kolmogorov-Smirnov test and the Fisher Exact test on the targets of differentially expressed miRNAs in any given comparison. The ‘miRNA’ column denotes the name of the miRNA. The ‘comparison’ column denotes the context in which the comparison between mated and virgin flies was made. The ‘Direction’ column denotes the direction of differential expression of the miRNA along the virgin-mated conditions trajectory. The ‘test’ column denotes the type of test applicable for each record of the table.

This indicates that for differentially expressed miRNAs for any given comparison, the set of all predicted targets of those miRNAs did not differ significantly from the set of all predicted non-targets for that same miRNA.

When similar testing was conducted on gene sets which were predicted to be the target of multiple miRNAs, similar results were found (table 5.5):

1 st miRNA	2 nd miRNA	Comparison	Direction	Test	p	Adjusted p
dme-miR-14-3p	dme-miR-997-5p	Female Abdomen	Up	KS	0.366	1.000
dme-miR-14-3p	dme-miR-184-5p	Female Abdomen	Up	KS	0.455	1.000
dme-miR-997-5p	dme-miR-184-5p	Female Abdomen	Up	KS	0.732	1.000
dme-miR-927-3p	dme-miR-927-5p	Male Abdomen	Down	Fisher	0.791	1.000
dme-miR-927-3p	dme-miR-927-5p	Male Abdomen	Down	KS	0.825	1.000
dme-miR-14-3p	dme-miR-997-5p	Female Abdomen	Up	Fisher	1.000	1.000
dme-miR-14-3p	dme-miR-184-5p	Female Abdomen	Up	Fisher	1.000	1.000
dme-miR-997-5p	dme-miR-184-5p	Female Abdomen	up	Fisher	1.000	1.000

Table 5.5 - A table of p and adjusted p values testing for the combinatorial effect of multiple miRNAs differentially expressed in the same direction targeting the same set of targets. The first and second columns of the table denote the identifiers for the first and second miRNAs used for testing, respectively. Otherwise, as in table 3.

The plot of cumulative log fold changes values presented for predicted miRNA targets and non-targets of dme-miR-14-3p, for the female abdomen is typical for all differentially expressed miRNAs across all comparisons (figure 5.17)

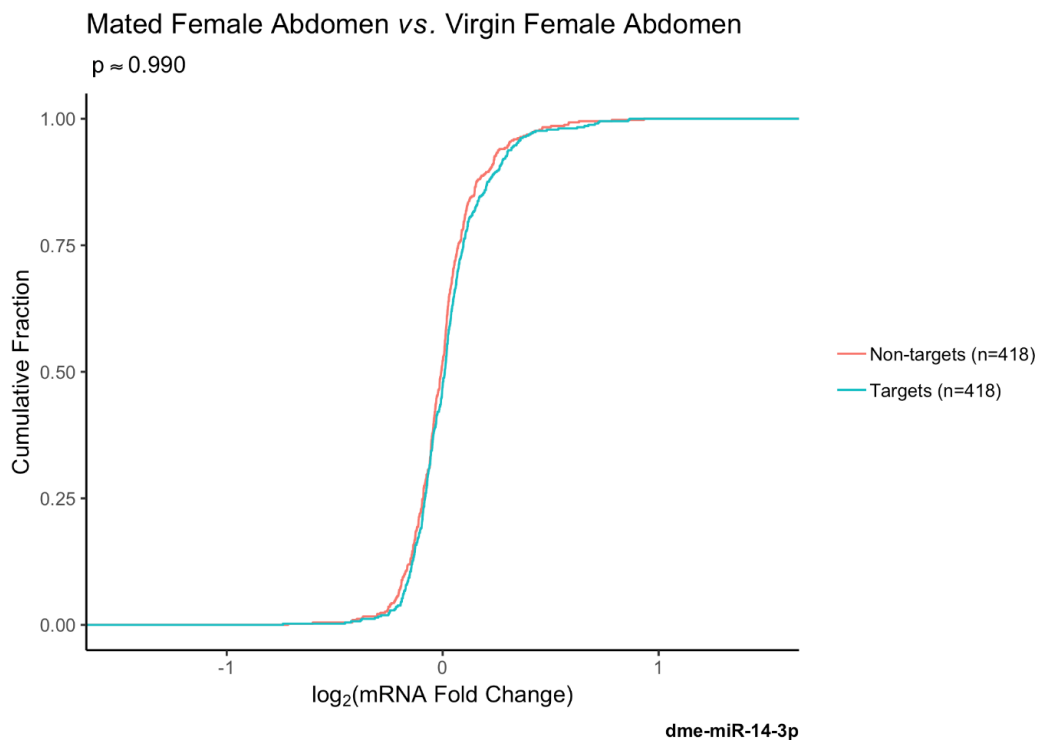


Figure 5.17 - Empirical cumulative distributions of the predicted target and non-targets of dme-miR-14-3p with respect to the \log_2 mRNA fold change. dme-miR-14-3p was chosen as a miRNA exhibiting typical behaviour of a differentially expressed miRNA in this comparison. The reported approximate p value refers to a one-sided, two-sample Kolmogorov-Smirnov testing for the equality between dme-miR-14-3p predicted target and non-target distributions. The number of observations for the predicted target and non-target distributions are identical, as a result of a sampling procedure implemented on both distributions to normalise for 3'UTR length, a potential confounding factor in this type of analysis. Further details of the sampling method can be found in the reported methods section of this study. Comparison: Mated female abdomen vs. virgin female abdomen.

5.6 Discussion

The network visualisation in figure 5.1 is indicative of the regulatory effects of miRNAs in this biological context. It can be observed that a large number of predicted targets of dme-miR-927-3p and dme-miR-927-5p are downregulated indicating that miRNAs could mediate physiological change by modifying expression changes for a relatively large proportion of the coding transcriptome. In addition, it can be observed that almost equal proportions of predicted targets of the differentially expressed targets are differentially expressed in the same direction as the miRNA. This pattern of differential expression may be explained by the existence of incoherent feedforward loop gene regulatory network architectures (Hornstein and Shomron, 2006), in which a transcription factor exerts positive regulation on a target, and also a miRNA repressing that target. The hypothesised purpose of such an architecture would be for the miRNA to fine-tune the expression of a gene activated by a given transcription factor. In this particular case, downregulation of a transcription factor may have caused downregulation of a miRNA, and a co-targeted mRNA. It is also possible that the observed patterns of differential expression are caused by more complex network architectures. Finally, the visualisation reveals a number of shared interactions between the two differentially expressed miRNAs, dme-miR-927-3p and dme-miR-927-5p. Six of the predicted co-targeted genes are up-regulated whilst 13 are downregulated. Both of the differentially expressed miRNAs in this comparison derive from the same precursor miRNA, dme-mir-927, indicating that miRNAs from both strands of this miRNA precursor are used to co-operatively regulate some of the same targets in this system.

As mentioned previously, there is a potential for multiple 3'UTR splice isoforms to confound the analysis in this study. The miRNA targeting analysis was conducted at the level of genes and so only a single 3'UTR was selected for each gene. As discussed at multiple points in this these alternative polyadenylation and cleavage is another source of 3'UTR isoforms along with splice isoforms. The implication of these 3'UTRs for this analysis is that gene 3'UTR models may not be fully accurate, leading to increased uncertainty in downstream analyses including miRNA target prediction.

A potential explanation for the findings in this study is that miRNA differential expression could influence physiological change in this system by mediating the repression of a restricted set of miRNA targets, which may be able to be distinguished from other targets due to their particularly high affinity to the miRNA, or due to some form of subcellular localization of these targets. Also worthy of examination as a potential explanation for observed results is the choice of the miRNA target prediction model used in this analysis – a more complex, regression-based model may have been able to account for some of the known variables acknowledged to be predictive of miRNA target site efficacy (discussed in chapter 2), which may have aided discrimination between predicted target and non-target mRNA transcripts through the implementation of a prediction score filter, which would presumably decrease the number of false target sites predicted. As it is possible that current target prediction methods which provide this functionality may be overfitted to microarray data, benchmarking of miRNA target prediction algorithms on data deriving from a diversity of experimental sources, including RNA-Seq would be beneficial. It is also possible that the 3'UTR models implemented at the gene level in this analysis lacked

sufficient specificity in order to properly distinguish between predicted targets and non-targets; as discussed earlier, selection of the longest 3'UTR isoform as being representative will lead to an inflation of the number of false positive miRNA target predictions in the analysis. Selection of the most abundant mRNA transcript splice isoform may have been a more appropriate choice, as to maximise miRNA target prediction accuracy. It should also be noted that typical effects of miRNA-mediated repression of coding transcripts, on average, is typically fairly modest (Baek, et al., 2008), especially if changes in cytosolic miRNA concentrations are relatively small. As a result, it may be expected that the discrimination between predicted target and non-target mRNA in fold change analyses such as this would be especially difficult. Analyses such as this investigating miRNA activity through assaying gene expression also cannot account for that component of miRNA-mediated gene regulation, which does not lead to mRNA destabilization (Pillai, et al., 2005), although this is not thought to be the predominant mode of action of miRNAs (Baek, et al., 2008). In addition, as alluded to earlier, a multi-omics approach examining the co-operative action of both transcription factors and miRNAs on targets may be necessary for a more complete understanding of regulatory relationships in this system. Finally, the confounding problem of 3'UTR length which was encountered during this analysis highlights the importance of implementing rigorous data cleaning and normalisation procedures before proceeding with more formal analyses for this type of analysis.

Another problem to consider in more detail is the small number of differentially expressed genes for the male head/thorax and the female abdomen. These results are somewhat counter-intuitive. It may be expected that there is a small amount or almost no transcriptomic change

in the male head/thorax as mating would not be expected to drastically alter male behaviour. However, a small degree of differential expression in the female abdomen is unexpected. QC analyses (appendix B) do not reveal any obvious technical issues with these samples (*e.g.* high inter-replicate variability or low read depth) which could potentially explain the low number of differentially expressed genes. The proximal explanation for the lack of differentially expressed genes is that most genes with large differential expression effect sizes in these two conditions occur for lowly expressed genes (figure B.6) or high standard errors of beta (figure B.7) and hence are not called as being differentially expressed (due to the logic of the sleuth differential expression algorithm).

It is difficult to discern the precise reason though why there are few large effects for moderately or highly expressed genes in the female abdomen in particular. The male specific expression of Y-chromosome male fertility factor genes such as *kl-2* and *kl-5* in the male abdomen libraries of this study would suggest that this is not an issue with sample mislabelling.

The results could in fact suggest that the female post-mating response predominantly occurs in the head/thorax. It has been noted (Fowler, et al., 2019) that these results corroborate previous studies which demonstrated that mating induces changes to ‘feeding behaviour, sleep patterns, sexual receptivity and aggression levels’ in female fruit fly (Bath, et al., 2017; Carvalho, et al., 2006; Fowler, et al., 2019; Isaac, et al., 2010) An alternative explanation is that the abdomen-specific PMR is potent but only requires differential expression of a small number of key genes, in contrast to the potentially large number of differentially

expressed genes needed in order to invoke behavioural changes. It has also been suggested (Fowler, 2019) that there may be insufficient spatial resolution in these types of bulk RNA-Seq studies to discern cell-type specific responses to mating in the fruit fly (Fowler, et al., 2019). For example, there is a ten-fold change in expression of some mating-responsive genes in different components of the female fruit fly reproductive system and organs (Prokupek, et al., 2009). Alternatively, these results could indicate that the abdominal female PMR requires a substantial amount of time to be properly invoked after mating (*i.e.* longer than a few hours after mating). Finally, it has also been noted (Fowler, et al., 2019) that the large number of differentially expressed genes in the male abdomen could simply reflect transcriptional changes needed to replenish seminal fluid proteins only a few hours after mating, which is a process which would occur in the abdomen (Sirot, et al., 2009).

Taking a closer look at specific genes and miRNAs which have been found to be differentially expressed in this study for particular conditions, reveals information about important transcriptional changes and biological processes underlying the post-mating response in male and female fruit flies. It has been noted (Fowler, et al., 2019) that certain sex-related genes such as *Send2*, which has functions in spermathecal secretory cells was found to be upregulated in the female after mating, as well as *fit* which is associated with feeding behaviour.

Examining more specifically the identities of differentially expressed miRNAs in this study, it has been shown previously that miR-927 has roles relating to adult fertility in the fruit fly (Chen, et al., 2014), that

miR-184 is essential for oogenesis (Iovino, et al., 2009) together indicating the validity of this study in being able to identify genuine markers of the post-mating response in the fruit fly.

Chapter 6: The regulation of sex transition in *Lates calcarifer* (Asian seabass) by miRNAs

6.1 Contributions

Simon Moxon: miRNA annotation and quantification.

Darrell Green: sRNA-Seq library preparation

Shubha Vij/Jolly Saju/Kathiresan Purushothaman: Fish husbandry, sample preparation, RNA extraction, mRNA-Seq library preparation, mRNA-Seq differential expression analysis, some aspect of mRNA-Seq QC (figure 6.1)

Laszlo Orban: Experimental design, Project supervision

Thomas Bradley: mRNA-Seq and sRNA-Seq QC, miRNA differential expression analysis, data visualisation (except figure 6.1), miRNA target prediction analysis, integrated analysis of expression and target prediction data, GO term enrichment analyses, discussion and interpretation of results

6.2 Introduction

In the previous chapter, I discussed how combined data from sRNA and RNA sequencing experiments had been used to infer the role of miRNAs in regulating the expression of protein-coding transcripts during a given developmental process. In particular, the fold change values of

the predicted targets of a given differentially expressed miRNA between two developmental points were compared to that of predicted non-targets of that miRNA in order to gauge the regulatory activity of that miRNA.

In this chapter, a similar, but slightly different strategy is used in order to assess miRNA activity within a different biological context. More specifically, combined sRNA and RNA sequencing experiments are performed on samples derived from Asian seabass (*Lates calcarifer*), as they undergo a naturally occurring sex transition developmental process in which the testis of adult males transform into ovaries (Guiguen, et al., 1994), in a process referred to as *sequential hermaphroditism*, or more specifically, *protandry*. It was determined in this analysis, and in the analysis performed in the previous chapter, that a substantial proportion of the differentially expressed predicted targets of differentially expressed miRNA are differentially expressed in the same direction as the miRNA. As a result, the adjusted p-values deriving from differential expression analyses, which is an unsigned indicator of differential expression, were used for comparison between predicted targets and non-targets instead of log fold change values. As discussed in more detail in the previous chapter, targets regulated in the same direction as the miRNA, could form part of an ‘incoherent feedforward loop’ network architecture (Hornstein and Shomron, 2006) along with a transcription factor targeting both the mRNA and miRNA genes, which would explain how the mRNA could be a genuine target of the miRNA despite the observed expression patterns.

By applying a modified version of the approach described in the last chapter in a novel biological context, I test the applicability of this approach across biological contexts, and also its robustness with respect to subtle alterations in analysis methodology. Interpretation of the results of this analysis, and that in the preceding chapter, can be used to assess the utility of using RNA-Seq data to infer miRNA regulatory activity for studied biological processes and contexts.

As this analysis was completed earlier than some of the other research reported in this thesis, not all of the tools and knowledge reported in previous chapters were implemented for this analysis. In particular, the FilTar tool had not been developed, and so 3'UTRs were not reannotated as part of this analysis.

6.3 Background

As briefly discussed previously, the biological context to this analysis is a developmental process which occurs in juvenile Asian seabass in which some males will transform into females. The transformation process can be divided into a series of intermediate stages between the fully developed male and the fully developed female. The most prominent morphological and histological markers for this process are present at or within the gonads of the Asian seabass.

The transforming gonads in particular can be divided into four different stages along this developmental trajectory, which are successively named as T1, T2, T3 and T4. The T1 gonad develops from the testis, whilst the T4 gonad develops into an ovary.

The different transforming gonad stages are temporally demarcated on the basis of histological and morphological criteria: For T1, the degeneration of male macular tissue is observable. In T2, ovarian and testicular tissue appear simultaneously within the gonad. In T3, a histological cross-section would reveal no testicular tissue, but would reveal ovarian tissues which comprises less than 50% of the gonad. In T4, the ovarian tissue would comprise more than 50% of the gonad. Oocytes are observable in early-mid stage ovaries, which distinguishes them from T4 transforming gonads.

There is evidence to suggest that miRNA are implicated in this developmental process. Sex-biased expression of miRNAs in gonadal tissue is exhibited in a number of closely related species. For example: The upregulation of miR-135b-5p in the Nile Tilapia (Xiao, et al., 2014), the upregulation of miR-19a and 19b in the ovary relative to the testis in Zebrafish (Vaz, et al., 2015), and the upregulation of miR-184-3p orthologue in the Chinese Mitten Crab ovary (He, et al., 2015).

Of more particular relevance to this study, is previous research examining miRNA expression profiles in developing gonads more specifically: Identified miRNAs of interest in this regard are again miR-135b-5p, miR-19a-3p, miR-19b-3p and miR-184-3p: miR-135b-5p miRNA in rainbow trout was shown to be upregulated in juvenile testis in comparison to mature testis (Farlora, et al., 2015), and demonstrated higher expression in prepubertal and pubertal compared to immature testis in Atlantic Salmon (*Skaftnesmo, et al., 2017*). Taken together, this information would suggest a correspondence between miR-135b-5p and the

male gonadal state. In particular, there may be a role for miR-135b-5p in signalling gonad masculinisation which would explain its upregulation during gonad masculinisation (Farlora, et al., 2015) and downregulation during gonad feminisation. Members of the miR-19-3p family, in particular, miR-19a-3p and miR-19b-3p, seem to have a more feminising influence on gonadal development. Consistent with our analysis, qRT-PCR evidence from another study (Liu, et al., 2015) demonstrated that miR-19a-3p and miR-19b-3p were upregulated in transitioning gonads, compared to testis in zebrafish. Additionally, it was shown through use of a luciferase assay conducted in the same study that *Dmrt1* is a direct target of both miR-19a-3p and miR-19b-3p. Upregulation of miR-19a-3p has been observed in female relative to male primordial germ cells in mouse (*Mus musculus*) (Fernández-Pérez, et al., 2018). miR-184-3p has also been shown to have a feminising influence in previous research, with reported upregulation in developing ovaries in the Chinese Mitten crab (He, et al., 2015), whilst deletion of miR-184-3p led to a loss of oogenesis in the fruit fly (Iovino, et al., 2009).

6.4 Experimental Design

Samples were taken from mature testis and mature ovary organs, as well as four intermediate stages along this developmental trajectory (T1, T2, T3 and T4) (Guiguen, et al., 1994). Samples underwent RNA extraction, cDNA library preparation and sequencing according to sRNA sequencing (testis n=5, T1 n=7, T2 n=1, T3 n=2, T4 n=2, ovary n=5) and RNA-Seq protocols (testis n=5, T1 n=4, T2 n=0, T3 n=1, T4 n=2, ovary n=5). More details of experimental procedures used can be found in the *methods* section of this chapter.

6.5 Methodology

6.5.1 Sample preparation

Asian seabass individuals were collected from the Marine Aquaculture Centre (Singapore). Asian seabass were reared in seawater conditions at a temperature range of 28-31 °C. All experiments and procedures were approved by Agri-food and Veterinary Authority (AVA) Institutional Animal Care and Use Committee (IACUC) (approval ID: AVA-MAC-2012-02) and performed according to guidelines set by the National Advisory Committee on Laboratory Animal Research (NACLAR) for the care and use of animals for scientific research in Singapore. Gonads at various stages of maturity were collected and staged as part of a previous study (Vij, et al., 2016), using previously defined morphological and histological criteria (Guiguen, et al., 1994).

6.5.2 RNA extraction

Total RNA was extracted using the RNeasy mini kit (Qiagen) and sRNA was purified using the mirVana miRNA isolation kit (Life Technologies). RNA concentrations and integrity were measured on the NanoDrop 8000 Spectrophotometer (Thermo Fisher Scientific) and visually assessed by agarose gel electrophoresis with ethidium bromide staining. RNA was stored at -80 °C.

6.5.3 Library construction and sequencing

cDNA libraries were generated using the TruSeq stranded total RNA prep kit (Illumina). The NextSeq 500 (Illumina) was used for 150bp paired end sequencing. For sRNA, libraries were constructed by ligating RNA to 3' and 5' HD adapters (Sorefan, et al., 2012). Ligated RNA products were reverse transcribed to cDNA and amplified by PCR. The cDNA products expected to contain 19-33 base pair inserts were purified by 8% polyacrylamide gel electrophoresis and ethanol precipitation (Xu, et al., 2015). 50bp single-end sequencing was performed on the HiSeq 2500 (Illumina).

6.5.4 miRNA sequence analysis, annotation and quantification

For sRNA the 3' adapter was trimmed using perfect sequence match to the first 8 nucleotides of the 3' HiSeq 2500 adapter (TGGGAATTC). The HD signatures (four assigned nucleotides at the ligating ends) of the reads were also trimmed. Reads longer than 17nt were kept for further analysis. Reads with low sequence complexity, *i.e.* those comprised of two or fewer distinct bases were removed from further analysis.

miRNA annotations derived from the sequencing data of testis, transforming gonads and ovary were combined. Annotations for miRNA that existed in the transforming gonads or ovary but not in the testis were generated using the same method as described in a previous study (Vij, et al., 2016). sRNA reads were mapped to annotated miRNA sequences

using PatMaN (Prüfer, et al., 2008). PatMaN output files were processed using a custom perl script to determine miRNA read counts.

A number of novel miRNAs were discovered during this process which are given temporary names for the purposes of this study, which were of the form (in regular expressions): ‘miR-nov[0-9*]-[3|5]p’. Each miRNA was assigned a number and a name according to this pattern. miRNAs deriving from the same miRNA precursor were assigned the same number; miRNAs not deriving from the same miRNA precursor were assigned different numbers. Name suffixes correspond to the arm of the miRNA precursor from which the mature miRNA derives. A full list of novel miRNA names and corresponding sequences are provided (table C.1).

6.5.5 Differential expression analysis

A testis and a T1 RNA seq sample was discarded for low read counts (28,991,835 reads) and a low mapping rate (33.28%), respectively. To ensure biological replicates existed for each group for both sRNA seq and RNA seq datasets, T1 and T2 datasets were pooled together creating the T1/T2 group (n=4 for RNA seq and n=8 for sRNA seq). Similarly, with T3 and T4 creating the T3/T4 group (n=3 for RNA seq and n=4 for sRNA seq). Sequenced reads were aligned to the Asian seabass scaffold genome assembly (GenBank accession: LLXD000000000) (Vij, et al., 2016) using TopHat (v2.0.13) (Trapnell, et al., 2009). Transcript abundance values (units: FPKM) were computed for annotated protein coding genes of the scaffold assembly and tested for differential expression using Cuffdiff 2 (v2.2.0) (FDR \leq 0.05) (Trapnell, et al., 2013).

DESeq2 (v 1.20.0) (Love, et al., 2014), a bioconductor (Huber, et al., 2015) package for the R statistical programming language and environment (Team, 2013) was used for the differential expression analysis of raw miRNA read count data. Default parameters were used, except for the ‘alpha’ parameter, which was set at 0.05 when calling the DESeq2 ‘results’ function. Note that for all differential expression analyses, for mRNA and miRNA, all comparisons made are unidirectional along the testis to ovary developmental trajectory.

6.5.6 miRNA target analysis

TargetScan (v7.0) (Agarwal, et al., 2015) was used to predict targets on 3’UTR sequences. 3’UTR sequences were predicted by extracting 1 kb of sequence downstream of annotated open reading frame for the scaffold genome assembly of Asian seabass (Vij, et al., 2016). The seed region, *i.e.* nucleotides 2-7 of the miRNA (Bartel, 2018; Lewis, et al., 2003), were extracted from annotated mature miRNA sequences of the same assembly and used as input for TargetScan to identify predicted targets. Only 8mer targets, which are predicted seed matches with the highest predicted efficacy, were used for downstream analysis due to the large number of false positive results associated with miRNA target prediction (Pinzón, et al., 2017). On the basis of target predictions, for each differentially expressed miRNA of each comparison, records from the differential expression analysis was divided into designated “target” and “non-target” sets. Cumulative distribution functions of adjusted p-values from the differential expression analysis of target and non-target sets were then constructed. The Kolmogorov-Smirnov test was used to

test the null hypothesis that target and non-target adjusted p-values derived from the same underlying distribution ($FDR \leq 0.05$).

6.5.7 Clustering and data visualisation

The `cummeRbund` (Goff, et al., 2013) R package for the manipulation of Cufflinks output was used for post-processing of mapped sequenced reads including hierarchical clustering of gonadal tissue types using the Jensen-Shannon distance (figure 6.1). The principle components analysis for clustering individual sRNA sequencing datasets (figure 6.2) was completed using the ‘`plotPCA`’ function defined within `DESeq2` package. Extensive use of ‘`tidyverse`’ packages were used for general data manipulation and plotting (Wickham).

6.5.8 GO Term enrichment analysis

The `BiNGO` plug-in (Maere, et al., 2005) for the Cytoscape tool (Shannon, et al., 2003) was used to perform GO term enrichment analysis using the hypergeometric test ($FDR \leq 0.05$). All functional annotations of protein coding loci for Asian seabass were used as the reference set (Vij, et al., 2016). The ontology used was the GO biological process set. The analysis was performed on both positively and negatively differentially expressed genes for each comparison determined from the RNA-seq analysis and the predicted targets of miRNAs determined to be of interest from the miRNA target analysis.

6.6 Results

6.6.1 Differential expression analysis

Normalised gene abundances showed distinct clustering between tissue groups. Testis exhibited clustering with T1/T2. Ovaries exhibited clustering with T3/T4. All replicates of each tissue group consistently clustered together (Figure 6.1).

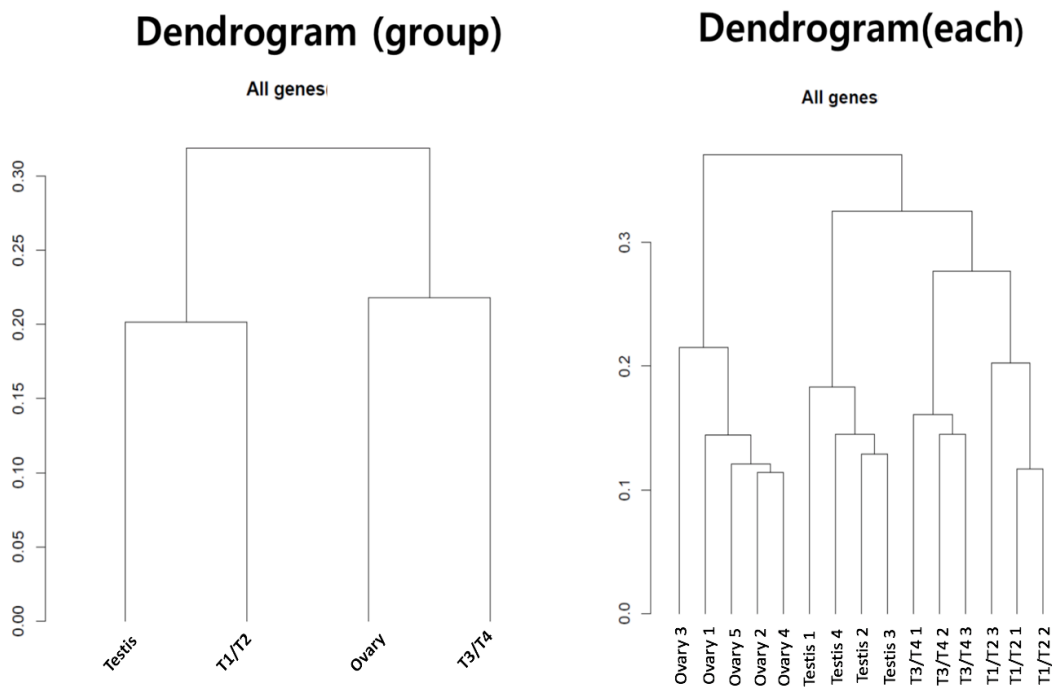


Figure 6.1 - Dendrogram shows hierarchical clustering of RNA-seq derived gene abundance data using the Jensen-Shannon distance. Clustering is performed on (A) gonadal tissue type (B) gonadal tissue type with additional labelling by biological replicate.

Clustering of sRNA sequencing datasets revealed a strong degree of clustering by gonadal tissue type. In particular, there is a visible separation of all gonadal tissue groups along the first principal component

(30% variance) and an additional separation between transforming (T1/T2 and T3/T4) and mature (testis and ovary) gonadal tissue along the second principal component (29% variance) (Figure 6.2). Similar patterns of clustering are observed for the principal components analysis of protein-coding genes.

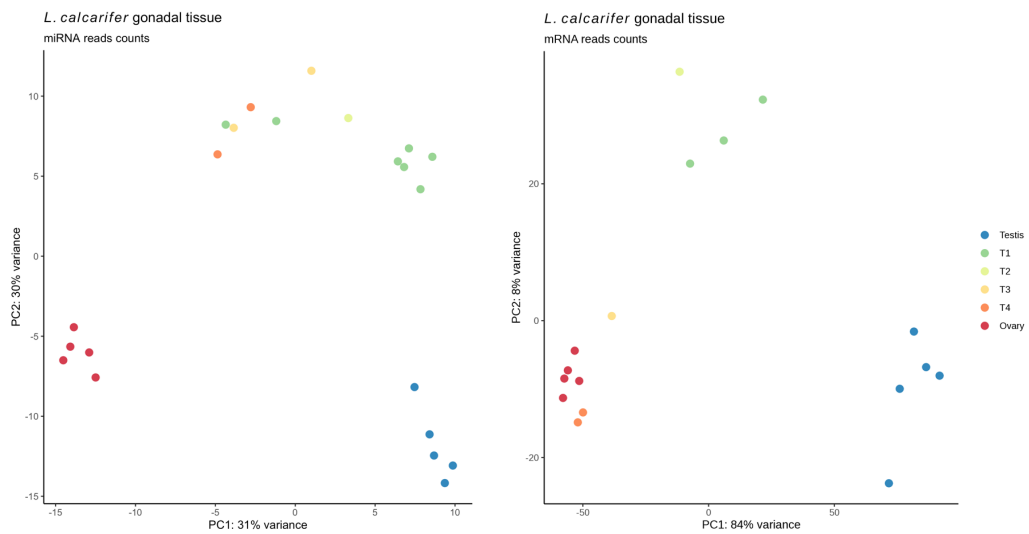


Figure 6.2 - Principal components analysis of the normalised miRNA read counts derived from sRNA sequencing of *L. calcarifer* gonadal tissue. The first principal component is plotted along the x-axis. The second principal component is plotted along the y-axis. Each datum point represents a single sRNA sequencing experiment. Colour labelling corresponds to gonadal tissue group associated with each experiment.

A large number of miRNAs and protein coding genes exhibit sexually dimorphic expression including differential expression between the gonads and transforming gonads. Global analysis of miRNA expression reveals a large degree of both positive and negative differential expression across all stages of gonadal transition (table 6.1 and figure 6.3).

Volcano plots helps visualise the large degree of differential expression observed for all comparisons (figure C.1)

Comparison	Upregulated	Downregulated	Total
Testis -> T1/T2	67 (0.15)	89 (0.20)	156 (0.35)
T1/T2 -> T3/T4	41 (0.09)	30 (0.07)	71 (0.16)
T3/T4 -> Ovary	55 (0.12)	67 (0.15)	122 (0.28)
Testis -> T3/T4	68 (0.15)	83 (0.19)	151 (0.34)
T1/2 -> Ovary	85 (0.19)	86 (0.19)	171 (0.39)
Testis -> Ovary	59 (0.13)	96 (0.22)	155 (0.35)

Table 6.1 - A summary of the results of the miRNA differential expression analysis with demarcations between the number of downregulated, upregulated and differentially expressed miRNA in each comparison. Numbers in parenthesis represent the proportion of miRNAs in a given instance relative to the total number of miRNAs found in *L. calcarifer*. Grey horizontal lines demarcate comparisons that span different relative developmental time spans. False discovery rate (FDR) \leq 0.05

There is a similarly large number of both positive and negative differential expression observed in protein coding transcripts with a relatively even distribution of total differentially expressed transcripts across all comparisons (Table 6.2 and figure 6.4). Again, the large degree of differential expression for all comparisons can be visualised with volcano plots (figure C.2).

Comparison	Upregulated	Downregulated	Total
Testis -> T1/T2	2315 (0.10)	863 (0.04)	3178 (0.14)
T1/T2 -> T3/T4	2411 (0.11)	1131 (0.05)	3542 (0.16)
T3/T4 -> Ovary	717 (0.03)	1940 (0.09)	2657 (0.12)
Testis -> T3/T4	2711 (0.12)	2074 (0.09)	4785 (0.22)
T1/T2 -> Ovary	3009 (0.14)	2504 (0.11)	5513 (0.25)
Testis -> Ovary	3546 (0.16)	1964 (0.09)	5510 (0.25)

Table 6.2 - A summary of the results of the RNA seq differential expression analysis. Otherwise as in Table 6.1.

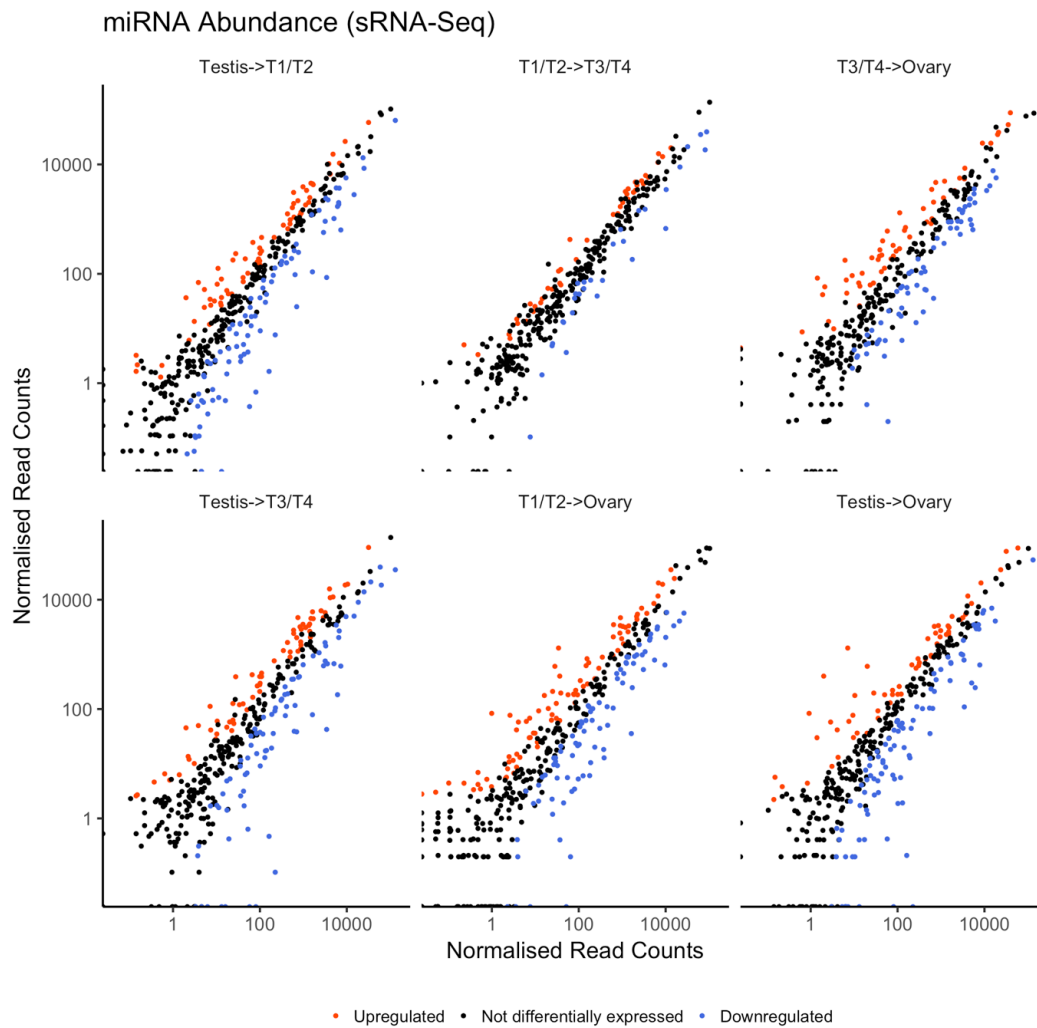


Figure 6.3 - miRNA transcript abundances recorded for (A) testis to T1/T2 comparison (B) T1/T2 to T3/T4 comparison (C) T3/T4 to ovary comparison (D) testis to T3/T4 (E) T1/T2 to ovary (F) testis to ovary comparison. The x-axis and y-axis denote normalised read counts for the gonadal stages indicated before and after the arrows in the subtitles. Red dots represent miRNAs that are positively differentially expressed. Black dots represent miRNAs with no observed differential expression. Blue dots represent miRNAs observed to be negatively differentially expressed. $FDR \leq 0.05$.

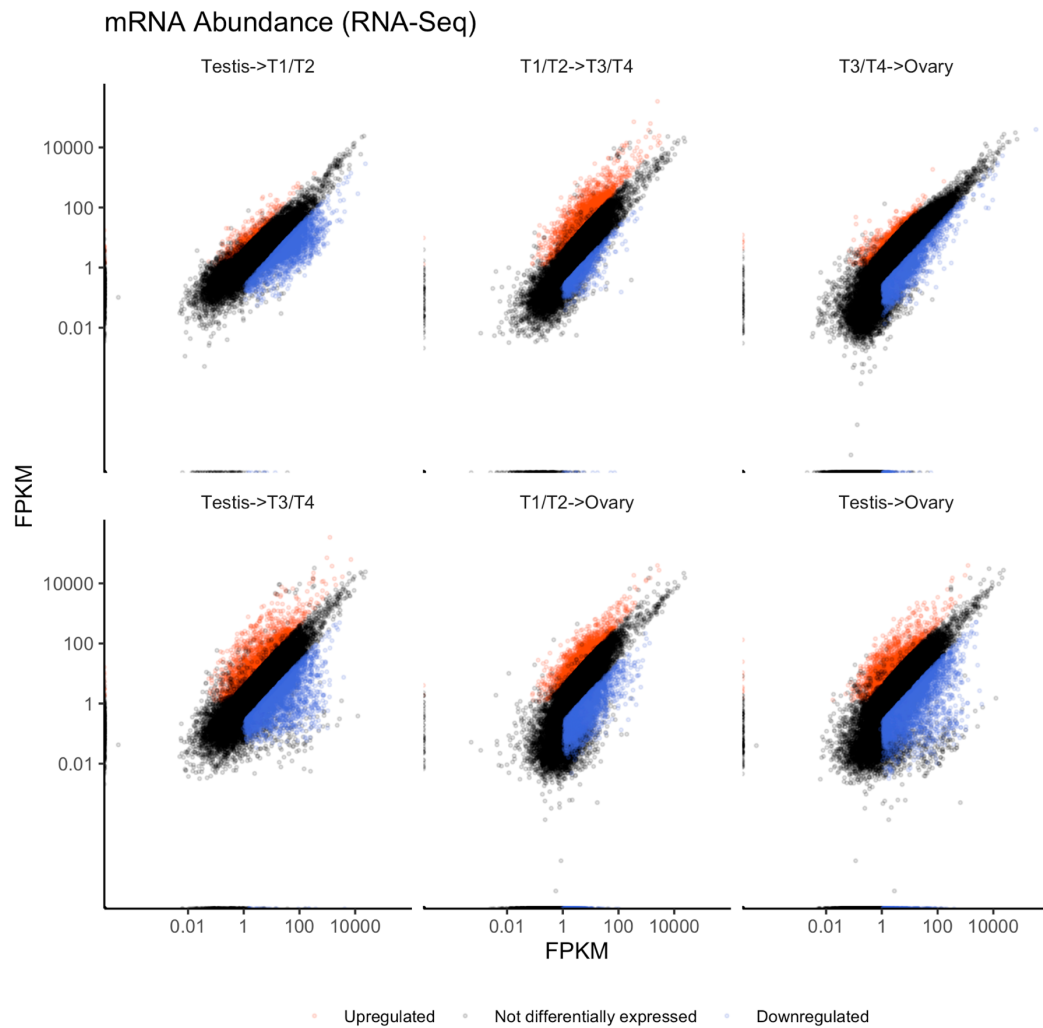


Figure 6.4 - mRNA transcript abundances recorded for (A) testis to T1/T2 comparison (B) T1/T2 to T3/T4 comparison (C) T3/T4 to ovary comparison (D) testis to T3/T4 (E) T1/T2 to ovary (F) testis to ovary comparison. The x-axis and y-axis denote fragments per kilobase of transcript per million mapped reads (FPKM) for gonadal stages indicated before and after the arrows in the subtitles, respectively. Red dots represent miRNAs that are positively differentially expressed. Black dots represent miRNAs with no observed differential expression. Blue dots represent miRNAs observed to be negatively differentially expressed. $FDR \leq 0.05$ and $|\log_2 \text{fold change}| > 2$.

6.6.2 miRNA targeting analysis

Our analysis revealed a number of miRNAs whose entire global predicted target set was shown to differ significantly from non-targets when comparing levels of differential expression. This finding implicates these miRNAs in the observed sequential hermaphroditism of Asian seabass. Cumulative distributions of the adjusted p-values inherited from the differential expression analysis were compared using the Kolmogorov-Smirnov test. More specifically, in this analysis, the adjusted p-values of a differentially expressed miRNAs predicted targets is compared with the adjusted p-values of the predicted non-targets of that same miRNA. A plot of a representative differentially expressed miRNA with clearly distinguishable target and non-target p-value distributions is shown (figure 6.5).

The test was applied to all differentially expressed miRNAs for each comparison (*e.g.* testis to T1/T2, T1/T2 to T3/T4 *etc.*). Table 6.3 shows a summary of significant miRNAs identified in this analysis.

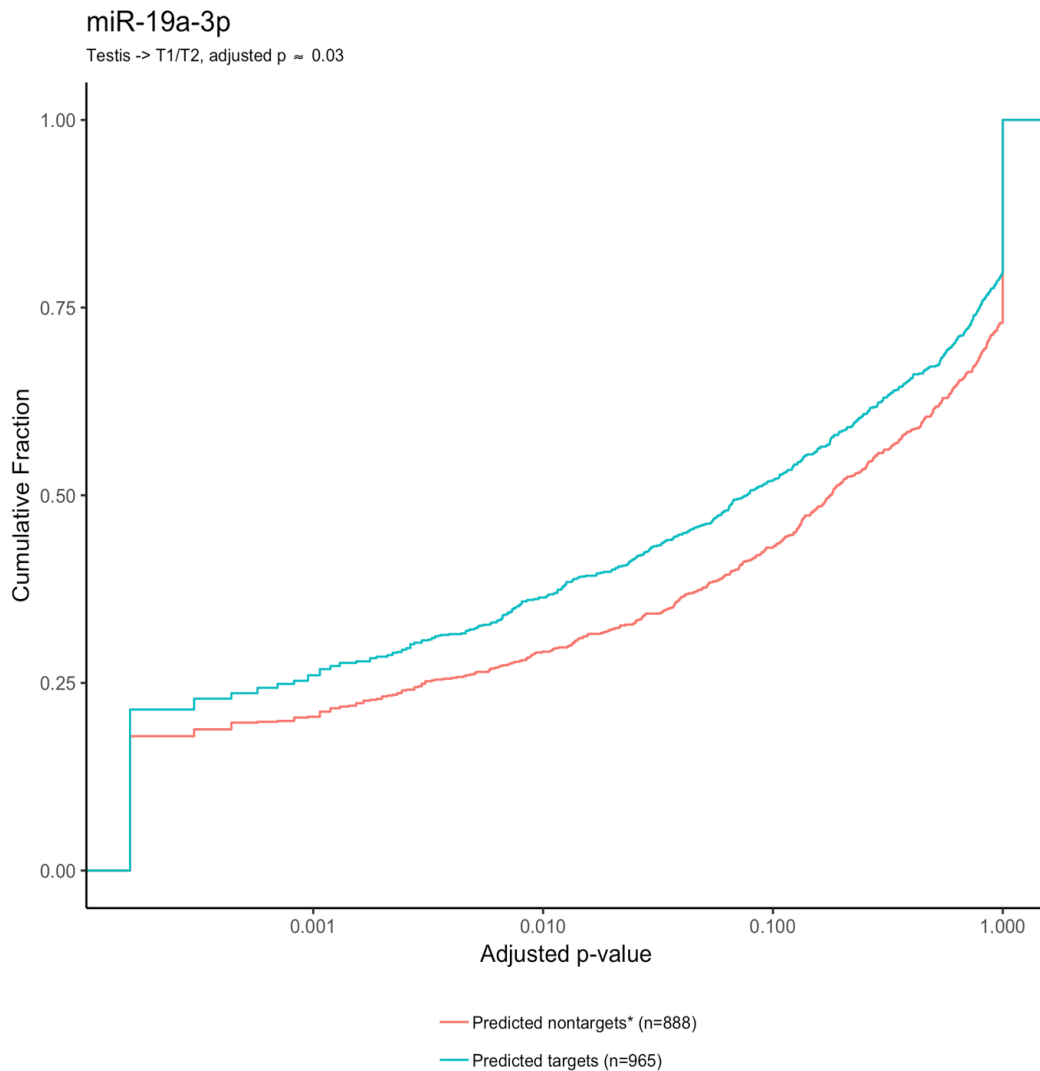


Figure 6.5 - A cumulative plot of the differential expression of predicted targets and predicted non- targets of miR-19a-3p when considering the testis to T1/T2 comparison. The x-axis on a \log_{10} scale represents adjusted p-values returned from the differential expression analysis. Observed floor and ceiling effects derive from the effects of cumulative minimum and cumulative maximum functions used within the Benjamini & Hochberg adjustment, respectively (Benjamini and Hochberg, 1995). Difference between distributions is tested using the Kolmogorov-Smirnov test, returning approximate p-values ($FDR \leq 0.05$). Non-targets in this instance refer to all protein coding transcripts not predicted to be targeted by any differentially expressed miRNA in this comparison.

miRNA	Comparison	Adjusted p-value	Direction
miR-nov29-5p	Testis->Ovary	0.030	upregulated
miR-135b-5p	Testis->T1/T2	0.030	downregulated
miR-135c-5p	Testis->T1/T2	0.030	upregulated
miR-nov14-5p	Testis->T1/T2	0.030	downregulated
miR-19a-3p	Testis->T1/T2	0.030	upregulated
miR-19b-3p	Testis->T1/T2	0.030	upregulated
miR-19d-3p	Testis->T1/T2	0.030	upregulated
miR-138-5p	T1/T2->Ovary	0.033	downregulated
miR-217-3p	Testis->T1/T2	0.039	downregulated
miR-24-3p	Testis->T1/T2	0.039	downregulated
miR-730-5p	Testis->T1/T2	0.039	downregulated
miR-184-3p	Testis->T1/T2	0.039	downregulated
miR-7132b-3p	T1/T2->Ovary	0.039	downregulated
miR-737-5p	Testis->T1/T2	0.039	downregulated

Table 6.3 - A summary of the results of the analysis of the predicted targets of differentially expressed miRNAs. miRNAs listed were found to exhibit a significant global effect on predicted targets. Also recorded is the associated comparison, adjusted p-value and the direction of differential expression.

6.6.3 GO term enrichment Analysis

A number of GO terms (Ashburner, et al., 2000; Consortium, 2016) linked to spermatogenesis are enriched in downregulated gene sets across multiple comparisons across the ‘testis to ovary’ sex transition including cilium morphogenesis, cilium assembly, and cell projection assembly. Identical GO terms are mostly enriched in downregulated genes from testis to T1/T2 and testis to T3/T4, whilst more specific chemotaxis and locomotion GO terms are downregulated from T1/T2 to ovary and from T3/T4 to ovary.

GO term enrichment analysis revealed that miR-184-3p targets, which is the only miRNA implicated in the transition from testis to T3/T4 in the previous analysis, is potentially enriched (adjusted p-value = 0.089) for GO terms in animal organ development (GO term ID: 48513). Targets for miR-184-3p that have been assigned this GO term include *lca5235*, which has also been attributed the GO term for ovarian follicle development (GO term ID: 0001541). The targets for miR-nov14-5p are potentially enriched for spermatogenesis related terms such as male gamete generation (GO ID: 48232) and spermatogenesis (GO ID: 7283) both of which have an associated adjusted p-value of 0.077.

A summary of some of these findings, with associated statistics is giving in table 6.4:

GO ID	GO term description	Condition	Direction	p-value	Adjusted p-values
60271	Cilium morphogenesis	testis->ovary	down	4.3508E-8	2.1696E-5
60271	Cilium morphogenesis	testis->T1/T2	down	4.3508E-8	2.1696E-5
60271	Cilium morphogenesis	testis->T3/T4	down	1.0312E-6	2.2343E-4
42384	Cilium assembly	testis->ovary	down	1.0248E-8	7.6654E-6
42384	Cilium assembly	testis->T1/T2	down	1.0248E-8	7.6654E-6
42384	Cilium assembly	testis->T3/T4	down	7.9815E-7	7.9815E-7
30031	Cell projection assembly	testis->ovary	down	1.0248E-8	7.6654E-6
30031	Cell projection assembly	testis->T1/T2	down	1.0248E-8	7.6654E-6
30031	Cell projection assembly	testis->T3/T4	down	7.9815E-7	2.0752E-4

Table 6.4 - GO term enrichment analysis. A summary of sex related GO terms found to be enriched for differentially expressed genes between different comparisons for a given direction of differential expression (FDR < 0.05).

6.7 Discussion

Predicted miRNA target analysis reveals that only a small proportion of differentially expressed miRNAs exhibit global effects on their predicted targets. When examining the distribution of implicated miRNAs across all six comparisons, a bias for the ‘testis to T1/T2’ comparison is apparent (10 out of 14 implicated miRNAs occur in this comparison), which reinforces findings that in a given developmental transition, miRNAs generally exert greater regulatory activity at a given stage of that transition, rather than act uniformly over the entire developmental timespan (Grishok, et al., 2001; Lee, et al., 1993; Wightman, et al., 1993). A caveat of this type of analysis however is that it is possible that differentially expressed miRNAs are not acting globally on all predicted targets. Firstly, because computational predictions lack direct experimental support, and therefore may include false positive predicted targets which will not interact with the differentially expressed miRNA *in vivo*. For this analysis, 1 kilobase windows downstream of open reading frames was used as a predictor of 3’UTR identity due to absence of more precise 3’UTR annotations. This is a coarse predictive model for 3’UTRs and will necessarily limit the accuracy of miRNA target predictions. It is also possible that miRNAs only act on a subset of known targets for any given process due to cellular and molecular constraints. For example, sub-cellular localisation of miRNA (Leung and Sharp, 2006) and targets (Holt and Bullock, 2009) can prevent the two types of RNA from interacting, and can lead to an overestimation of *effective* RNA relative abundance levels, which is not accounted for in differential expression analyses. In addition, the reported ‘sponging effect’ of competitive endogenous RNAs on miRNA (Salmena, et al., 2011),

would lead to a reduction in effective cytosolic miRNA expression levels, violating the assumptions of this analysis.

Key regulatory miRNAs identified during the course of this analysis, have elsewhere been implicated in sex development processes, providing evidence for the validity of this analysis approach: For example, miRNAs which have previously been reported to have sex biased expression in gonadal tissue of closely related species, include upregulation of miR-135b-5p in testis in Nile tilapia (*Oreochromis niloticus*) (Xiao, et al., 2014), the upregulation of miR-19a and miR-19b in ovary relative to testis in zebrafish (*Danio Rerio*) (Vaz, et al., 2015) and the upregulation of a miR-184-3p orthologue in the Chinese Mitten Crab (*Eriocheir Sinensis*) ovary (He, et al., 2015). Of more particular relevance to this study is previous research examining miRNA expression profiles in developing gonads more specifically. Identified miRNAs of interest in this regard are again miR-135b-5p, miR-19a-3p, miR-19b-3p and miR-184-3p: miR-135b-5p was shown to be downregulated from testis to T1/T2 in the current study, the same miRNA in rainbow trout (*Oncorhynchus mykiss*) was shown to be upregulated in juvenile testis in comparison to mature testis (Farlora, et al., 2015) and demonstrated higher expression in prepubertal and pubertal compared to immature testis in Atlantic Salmon (*Salmo salamar*) (Skaftnesmo, et al., 2017). Evidence for the involvement of the highly similar miR-135c-5p, which has also been implicated in this study is more sparing although this miRNA was found to be generally enriched in developing somatic zebrafish tissue compared to mature tissue (Soares, et al., 2009). In contrast, members of the miR-19-3p family, in particular, miR-19a-3p and miR-19b-3p identified in this study, seem to have a more feminising influence on gonadal development. Consistent with

our analysis, qPCR evidence from another study demonstrated that miR-19a-3p and miR-19b-3p were upregulated in transitioning gonads, compared to testis in zebrafish (Liu, et al., 2015). Additionally, it was shown through use of a luciferase assay conducted in the same study that the male biased *dmrt1* is a direct target of both miR-19a-3p and miR-19b-3p. The hypermethylation of *dmrt1* in ovary compared to testis is associated with downregulation of this gene suggesting that miR-19a-3p and miR-19b-3p may act in combination with epigenetic factors in order to regulate *dmrt1* activity (Domingos, et al., 2018). Upregulation of miR-19a-3p has been observed in female relative to male primordial germ cells in mouse (*Mus musculus*) (Fernández-Pérez, et al., 2018). miR-184-3p has also been shown to have a feminising influence in previous research with reported upregulation in developing ovaries in the Chinese Mitten crab (He, et al., 2015) whilst deletion of miR-184-3p led to a loss of oogenesis in the fruit fly (*Drosophila melanogaster*) (Iovino, et al., 2009). Conversely, in our study, miR-184-3p was shown to be highly expressed in testis and downregulated in the testis to T3/T4 comparison. The reasons for this discrepancy are unclear, although a species specific, or a protandry specific role for miR-184-3p in Asian seabass cannot be ruled out. Additionally, the results of the GO term enrichment analysis, revealing potential associations between miR-184-3p and miR-nov14-5p targets and organ development and spermatogenesis processes respectively is further evidence for the validity of the approach used in this analysis.

6.8 Conclusion

The use of the analysis approach discussed in this chapter to identify likely key regulators of developmental processes highlights the utility of using RNA-Seq data for investigating miRNA-mediated regulation of protein coding transcripts. In addition, the combined use of small RNA sequencing data as part of this analysis demonstrates how information from multiple sequencing experiments, including RNA sequencing can be integrated in order to further understand miRNA regulated developmental processes.

Chapter 7: Future Work and Conclusion

7.1 Future Work

FilTar and FilTarDB can be scaled and extended in many different ways in order to provide additional benefit to users. In its current versions, both applications allow the user to either generate or view results from two different core miRNA target prediction algorithms. It would be of benefit to researchers using these tools if more target prediction algorithms were included, with a particular emphasis on including algorithms which represent a wide diversity of valid methods for modelling the process of miRNA target recognition, as well as a diversity of data types for training and testing developed predictions algorithms. FilTar has been designed to be modular and scalable to facilitate further development, minimising the work needed to implement these types of proposed extensions. As well as a broader range of core target prediction algorithms, the FilTarDB application in particular would benefit from a broader range of available biological contexts available for the user to interrogate, and for a greater number of vertebrate species for which target predictions are available. In the current instance of the software, expression data is provided for only a small number of healthy, adult tissues for each available species. To increase the range of research areas in which this application can be used, it would be useful to include biological contexts relating to different developmental stages, as well as samples from diseased cells or tissues. Also, of potential use, would be attempts to generate new methods for combining

expression data from multiple different samples, sometimes deriving from different laboratories, or whose cDNA libraries are generated using slightly different protocols. Difference in reported gene expression values for such samples could possibly result from the existence of batch effects (Leek, et al., 2010). Correction for batch effects would therefore more likely lead to unbiased estimates of gene expression across multiple samples.

As discussed in this thesis, miRNA perturbation experiments have been used extensively in order to test the effectiveness of FilTar for improving miRNA target prediction accuracy. There are concerns associated with the use of such experiments for this purpose: Firstly, it is an indirect method for determining possible miRNA target interactions, as it gauges changes in mRNA expression levels as a result of miRNA perturbation, which in some cases may be the result of secondary effects of miRNA action – which could potentially explain the repression of some transcripts not containing a predicted target site to a transfected miRNA (chapter 4). A potential solution to this problem could have been to use CLIP and CLIP-ligation data in order to create a more reliable ‘non-target’ transcript sets, however, due to the reported high false negative rates associated with these studies (Agarwal, et al., 2015), this approach is unlikely to be beneficial. Alternatively, data from 3’UTR reporter assays could substitute or supplement that from miRNA perturbation experiments in order to more directly gauge the effects of expression filtering and 3’UTR reannotation on putative targets. There are additional potential confounds for this form of experimentation: In a previous study (Agarwal, et al., 2015), a comprehensive analysis of miRNA transfection microarray datasets, it was discovered that data

from transfection experiments was confounded by lab and protocol specific batch effects, 3'UTR AU content, 3'UTR length (as discussed in chapter 5 of this thesis), as well as a derepressive effect on the targets of naturally abundant miRNAs within the cell (which is speculated due to be increased competition between miRNAs for different components of the miRNA pathway). It is likely that implementing post-processing steps to clean experimental data in order to control for these effects, for example, the use of partial least squares regression by Agarwal and colleagues (Agarwal, et al., 2015), would reduce the signal-to-noise ratio in this type of analysis. An additional concern is that use of miRNA perturbation experiments only tests for the effects of miRNA on mRNA stability, but not on the known ability of miRNAs to directly inhibit the translation of mRNAs. This problem can be overcome via the use of ribosomal profiling in order to measure the occupancy of ribosomes on mRNA, or alternatively through the usage of quantitative proteomics in order to measure the effect of miRNAs on protein abundance levels directly. Both methods types have previously been used in miRNA research for this purpose (Baek, et al., 2008; Guo, et al., 2010).

More generally, future research could be used to develop a greater systems level understanding of how miRNAs operate within the cell; to develop an understanding of miRNA action somewhere between the level of direct RNA-RNA interactions, and the level of the organism at the opposite extremity. Due to the 'many-to-many' nature of miRNA interactions, this will likely require the modelling of molecular interaction and regulatory networks, as well as an understanding how the existence of these regulatory networks within the cell is used both to exert physiological change during processes of development and to otherwise

maintain homeostasis of transcript expression levels. As well as protein-coding transcripts, there is also evidence that miRNAs interact with some classes of non-coding RNA such as circular RNAs and long non-coding RNAs forming larger gene regulatory networks (Kleaveland, et al., 2018). Whilst, as has been discussed in this thesis, there has been a lot of research conducted detailing the principles of the recognition of mRNA targets by miRNAs, it may not be beneficial to assume that similar targeting principles apply for non-coding RNA transcripts, which may as a whole possess different secondary structures and structural constraints compared to protein-coding transcripts. Even if such targeting rules are identical for all types of RNA transcripts, existing target predictions do not model this type of miRNA interaction. Transcription factors and enhancers are additional components of GRNs involving miRNAs which function to determine or maintain cell states (Chakraborty, et al., 2019). As a result, combined multi-omics and systems approaches will likely be needed to develop an understanding of miRNA activity at the level of regulatory networks.

A consistent theme running through this thesis, is the use of differential expression analyses in order to gauge the effect of miRNA perturbation on protein coding transcripts. Within the context of miRNA activity, differential expression analyses can be performed either at the level of the gene, or the level of individual transcripts. Although generally gene-level estimation of differential expression is considerably more accurate than analyses conducted at the transcript-level (Soneson, et al., 2015), it does not always make sense to conduct analyses at this level given that the miRNA acts at the level of the transcript. As such, not all splice transcript isoforms of a given gene may contain the necessary target site to be targeted by a given miRNA. However, as mentioned, the relatively

poor accuracy of transcript-level differential expression analyses can hamper analyses made at this level. Such problems most likely arise from intrinsic issues with count-based transcript quantification methods from the alignment or pseudo-alignment of short RNA-seq cDNA reads to genomes and transcriptomes respectively: The use of short cDNA reads can lead to a large degree of uncertainty during read alignment, which is exacerbated during the quantification of transcript expression levels due to the typically high sequence identity shared between transcript splice isoforms. This uncertainty is compounded during transcript-level differential expression analyses when transcript abundance estimates are compared between two different conditions. This problem could possibly be mitigated by the use of long-read sequencing technologies. In particular, methods have been implemented using the nanopore sequencing approach for direct sequencing of full-length RNA molecules - bypassing fragmentation, reverse transcription and PCR amplification biases associated with short-read cDNA sequencing (Garalde, et al., 2018). In addition, use of longer reads when sequencing is likely to increase the accuracy of 3'UTR reannotation, resolving issues discussed earlier in this thesis in which local reductions in coverage across the 3'UTR which often led to spurious 3'UTR truncations during the reannotation procedure. However, the use of nanopore sequencing for differential expression analysis may not be currently feasible or advisable given the relatively low base-calling accuracy of approximately 85% associated with this sequencing approach (Jain, et al., 2017; Rang, et al., 2018). In addition, long-read cDNA sequencing methods have been developed, which allow researchers to sequence full-length mRNA transcripts, including their poly-A tails (Legnini, et al., 2019), which are known to affect mRNA expression levels (Jalkanen, et al., 2014; Nicholson and Pasquinelli, 2018), and may be

of particular use to researchers interested in the regulation of gene expression. Whatever the precise library preparation and sequencing strategy used, it would be hoped that long-read sequencing technologies could offer a less noisy approach to differential transcript expression analysis, making it easier to discern the regulatory effects of miRNAs on the transcriptome.

7.2 Conclusion

Through the course of this thesis, I have demonstrated the utility of using RNA-Seq data to investigate the activity of miRNAs in animals. In chapters 2 and 3, I presented two different, but related software applications which can be used to generate or view miRNA target predictions for putative targets which have had their 3'UTRs reannotated specifically for given biological contexts, and also have been filtered according to the expression of those targets within a given biological context. In chapters 4 and 5, I have also shown how RNA-seq data can be used to infer the effect of a given differentially expressed miRNA on the entire set of that miRNA's predicted targets, and as a result try to infer the role of that miRNA for a given developmental process. In this chapter, I have explored and discussed different ways in which the tools developed as part of this thesis could be extended, as well as more generally considering how future work could advance this research area.

In summary, I have determined that RNA-Seq data can improve investigations of miRNA activity in bilaterian animal species, firstly, by improving miRNA target prediction accuracy by a process of using this

data to (i) remove lowly expressed mRNA transcripts from miRNA target prediction workflows and to (ii) reannotate the 3'UTRs of mRNAs as a preprocessing step for miRNA target prediction. In addition, I have shown that RNA-Seq data can be used to help infer the regulatory strength of miRNAs acting across biological conditions by (iii) integrating data of this type with sRNA-seq data, in order to identify differentially expressed miRNAs whose entire set of predicted targets is detectably perturbed in comparison to predicted non-targets of this miRNA.

Definitions

Biological context – A particular biological state which may be distinguished from other states by a given attribute or set of attributes (*e.g.* cell type, sex, treatment *etc.*)

Comparison – In the context of this thesis, an evaluation of the differences in transcriptomic states of two different biological contexts (*e.g.* treated and control cell cultures)

FilTar – A command line application developed during the course of the examined studies, which enables users to use RNA-Seq in order to generate miRNA target predictions specific to a given biological context.

FilTarDB – A database and web application allowing users to access pre-computed results generated using FilTar

Module: In the context of discussion of developing snakemake workflows, ‘modules’ refers to a discrete self-contained directory, containing its own snakefile with associated scripts and data

T1/T2 – Pooled sample data from the T1 and T2 stages of the Asian seabass transforming gonads

T3/T4 – Pooled sample data from the T3 and T4 stages of the Asian seabass transforming gonads

Glossary

3'UTR reporter assay: A method of detecting miRNA targets in which the 3'UTR of a gene of interest is fused with that of a reporter gene

6mer: Predicted or validated miRNA target sites with complementarity to nucleotides 2-7 of the miRNA

7mer-1A: A 6mer target match, with an adenine base on the target in the 't1' position which corresponding to the first nucleotide of the miRNA

7mer-m8: Predicted or validated miRNA target sites with complementarity to nucleotides 2-8 of the miRNA

8mer: A 7mer-m8 target match, with an adenine base on the target in the 't1' position which corresponding to the first nucleotide of the miRNA

Affected isoform ratios (AIRs): The ratio of a particular 3'UTR segment in relation to the start of the 3'UTR. AIRs can be used to generate 3'UTR profiles

Akaike Information Criterion (AIC): A method of selecting between different statistical models, in particular, a method for optimising 'goodness of fit' whilst penalising the number of parameters used in a model

Alternative polyadenylation (APA): The choosing of the cellular machinery of different polyadenylation sites on mRNA, generating 3'UTR isoforms

Argonaute protein (AGO): The effector protein of the RISC protein in which the guide RNA is bound. Upon target recognition, argonaute

will catalyse cleavage of the target (in some homologues), or recruit proteins for translational inhibition or target degradation

Bilateria: The clade of animals, including all those with bilateral symmetry. The miRNA seed targeting mechanism is thought to have evolved within this clade or a recent ancestor

Bulk RNA-Seq: RNA-Seq protocols in which RNA is extracted from a large number of different cells and pooled together for sequencing

Canonical miRNA targets: Predicted or validated miRNA targets with perfect, contiguous complementarity between the miRNA seed and the target

cDNA microarray protocols: The quantification of RNA expression levels by RNA reverse transcription, fragmentation, radiolabelling, and subsequent hybridisation to an ordered array of oligonucleotide probes indexed by sequence and array location

Chimeric RNA: A single RNA molecule composed of RNAs of two different types or from different origins, *e.g.* the chimeric RNA resulting from the ligation of miRNA and its targets

Compensatory miRNA targeting: miRNA base pairing, in which base pairing at the 3' end of the miRNA compensates for mismatches or non-Watson-Crick base pairing at the miRNA seed region

Competitive endogenous RNA (ceRNA): Non-coding RNA targets of miRNAs which are theorised to 'sponge' cytosolic miRNAs, and thereby exert a derepressive effect on other miRNA targets

Configuration: In the context of software use, the changing of options of an application from default values, in order to support more specialised and particular use cases

Context++: The name of the multilinear regression model which is used in version 7 of the TargetScan project.

Co-operative miRNA targeting: A discovered an effect in which closely spaced miRNA target sites act synergistically to increase the total repressive effect on the target

Cross-linking and sequencing of hybrids (CLASH): A protocol developed in the lab of David Tollervey in which an extra ligation step was added to a standard CLIP protocol, generating chimeric RNA sequences for subsequent protein pulldown, and RNA sequencing

Cross-linking and immunoprecipitation (CLIP): A next-generation sequencing protocol for detecting RNA-protein interactions, involving RNA-protein crosslinking, protein immunoprecipitation, protein digestion, and RNA sequencing

DGCR8: A subunit of the microprocessor complex. It binds the pri-miRNA in preparation for cleavage by drosha. Orthologues of this protein in *D. melanogaster* and *C. elegans* are known as ‘pasha’

Dicer: An RNase III enzyme which facilitates the cleavage of pre-miRNAs and the formation of mature miRNA

Differential expression analysis: An analysis to be used to compare the expression levels of RNA transcripts, or a set of RNA transcripts between multiple conditions. Commonly used in the context of downstream analysis of data deriving from RNA-Seq experiments

Drosha: An RNase III enzyme which complexes with DGCR8 as part of the microprocessor complex in order to facilitate the cleavage of pri-miRNAs, and forming pre-miRNAs as a result

Extensibility: In the content of software development, the propensity of an existing piece of software to have new features added to it, or for existing features to be improved

Gene regulatory network (GRN): A network of interactions between different genetic components affecting the expression levels of all or

some nodes in this network. Commonly involves known classes of regulatory genes such as miRNAs and transcription factors

General linear model: A model in which one or multiple response/dependent variable is modelled by a linear combination of independent variables

Guide RNA (gRNA): An RNA molecule within a ribonucleoprotein complex which is used as a specificity determinant for potential targets of that RNP

High definition (HD) adapters: Adapters which are specialised for use for sRNA sequencing in which there are four degenerate base pairs on the ligating end of each adapter

Mature microRNA (miRNA): An approximately 22 nucleotide non-coding RNA molecule, produced from specific biogenesis pathways, involved in post-transcriptional regulation of gene expression

Microprocessor complex: A protein complex, containing DGCR8 and drosha which facilitates the conversion of pri-miRNA to pre-miRNA

miRNA mimic: Synthetic double-stranded mature miRNA molecules designed to mimic endogenous double-stranded mature miRNA molecules. Often used in miRNA transfection experiments

mirtron: miRNA molecules generated in a drosha-independent pathway during mRNA splicing

Non-canonical miRNA targeting: Validated or predicted miRNA targets which are not canonical (see above definition)

Nanopore sequencing: A novel sequencing approach in which DNA or RNA molecules are sequenced by analysis of characteristic base-specific current density signatures as a nucleic acid is transmitted through a nanopore through a process of electrophoresis.

Northern blotting: A method of assaying RNA expression levels by a process of RNA extraction, gel electrophoresis, transfer of RNA to a membrane, hybridisation to radioactive probes, and subsequent autoradiography

Offset 6mer: Predicted or validated miRNA target sites with complementarity to nucleotides 3-8 of the miRNA

Ontology: A system of concept, entity and attribute definitions, and the definitions of relationships between different concepts, entities and attributes within a particular field or domain of knowledge

Passenger strand: The strand of a precursor miRNA which is generally not incorporated into RISC

Photoactivatable ribonucleoside-enhanced CLIP (PAR-CLIP): A modification of the standard CLIP protocol, in which crosslinking is induced by UV irradiation of RNAs containing photoreactive ribonucleosides

Post-mating response (PMR): Behavioural, physiological, and molecular changes which occur in an organism after mating

Precursor microRNA (pre-miRNA): The hairpin-like RNA structure from which mature RNAs are generated in a process facilitated by the dicer enzyme

Primary microRNA (pri-miRNA): The RNA molecules from which precursor miRNA molecules are derived, in a process facilitated by the drosha enzyme

Protandry: The changing of a sex of an organism from male to female

Quantitative PCR (qPCR): The use of PCR (polymerase chain reaction) technology to quantify RNA expression levels

Transcript reannotation: The generation of transcript models differing from standard models existing in public scientific databases such as Ensembl

Transgene: A gene transferred from one organism to another. Transgenes can be engineered using genetic cloning procedures

Ribonucleoprotein (RNP): A complex of RNA and RNA-binding proteins, e.g. AGO-miRNA

RNA-induced silencing complex (RISC): A multiprotein complex in which guide RNAs are loaded, which destabilise and translationally repress RNA targets

RNA interference (RNAi): The molecular pathway by which guide RNA molecules, bound by RISC, destabilise and translationally repress RNA targets

RNA-Seq: A generic name for a family of protocols in which RNA is extracted, fragmented and reverse transcribed generating cDNA. cDNA molecules are adapter ligated for the purposes of PCR amplification and subsequent next-generation sequencing

Scalability: In the context of software development, the propensity of an existing piece of software to operate with increasing demands on resources

Seed region: In the context of miRNA biology, the seed region refers to the 5' end of the miRNA, typically from nucleotides 2-8, which is used as a specificity factor by argonaute for miRNA targeting

Small/short interfering RNA (siRNA): 20-25 base pair, double-stranded RNA molecules, derived from the cleavage of long double-stranded molecules, which is incorporated into RISC, and acts in the RNAi pathway

Snakefile: The component of a snakemake directory which is interpreted by the snakemake binary, and controls all associated scripts and data in that directory *cf.* makefiles

sRNA sequencing: RNA sequencing protocols specialised for the exclusive sequencing of small RNAs

Supplementary miRNA targeting: miRNA base pairing in which base pairing in the 3' region of the miRNA supplements canonical 5' base pairing

Support-vector machine (SVM): A supervised machine learning classification model. SVMs are binary, linear classifiers

Target-directed miRNA degradation (TDMD): Instances in which extensive complementarity between a miRNA and its targets can induce degradation of the miRNA, and generally increase the rate of turnover for that miRNA

Transfection: An experimental process by which purified nucleic acids are introduced into eukaryotic cells. Liposome vectors are often used for this purpose when transfecting cells with miRNA mimics

Workflow Management: In the context of software development, the use of applications, libraries and protocols in order to ease development and maintenance of computational workflows potentially containing many disparate components

Bibliography

- (2018) RNAcentral: a hub of information for non-coding RNA sequences. *Nucleic acids research*, 47(D1), D221-D229.
- Agarwal, V., *et al.* (2015) Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, 4
- Aken, B.L., *et al.* (2016) The Ensembl gene annotation system. *Database*, 2016
- Alfonso-Parra, C., *et al.* (2016) Mating-induced transcriptome changes in the reproductive tract of female *Aedes aegypti*. *PLoS neglected tropical diseases*, 10(2),
- Allen, E., *et al.* (2005) microRNA-directed phasing during trans-acting siRNA biogenesis in plants. *Cell*, 121(2), 207-221.
- Almeida, M.V., de Jesus Domingues, A.M. and Ketting, R.F. (2019) Maternal and zygotic gene regulatory effects of endogenous RNAi pathways. *PLoS genetics*, 15(2), e1007784.
- Alwine, J.C., Kemp, D.J. and Stark, G.R. (1977) Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proceedings of the National Academy of Sciences*, 74(12), 5350-5354.
- Ambros, V. (2001) Dicing Up RNAs. *Science (New York, N.Y.)*, 293(5531), 811-813.
- Ambros, V. (2003) MicroRNA pathways in flies and worms: growth, death, fat, stress, and timing. *Cell*, 113(6), 673-676.
- Ambros, V., *et al.* (2003) A uniform system for microRNA annotation. *RNA (New York, N.Y.)*, 9(3), 277-279.
- Ameres, S.L., *et al.* (2010) Target RNA-directed trimming and tailing of small silencing RNAs. *Science (New York, N.Y.)*, 328(5985), 1534-1539.
- Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Nature Precedings*, 1-1.
- Asangani, I.A., *et al.* (2008) MicroRNA-21 (miR-21) post-transcriptionally downregulates tumor suppressor Pcd4 and stimulates invasion, intravasation and metastasis in colorectal cancer. *Oncogene*, 27(15), 2128.
- Ashburner, M., *et al.* (2000) Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1), 25.
- Axtell, M.J., Westholm, J.O. and Lai, E.C. (2011) Vive la différence: biogenesis and evolution of microRNAs in plants and animals. *Genome biology*, 12(4), 221.
- Baccarini, A., *et al.* (2011) Kinetic analysis reveals the fate of a microRNA following target regulation in mammalian cells. *Current biology*, 21(5), 369-376.
- Baek, D., *et al.* (2008) The impact of microRNAs on protein output. *Nature*, 455(7209), 64.
- Barrett, T., *et al.* (2011) BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic acids research*, 40(D1), D57-D63.
- Bartel, D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116(2), 281-297.
- Bartel, D.P. (2009) MicroRNAs: Target Recognition and Regulatory Functions. *Cell*, 136(2), 215-233.

Bartel, D.P. (2018) Metazoan MicroRNAs. *Cell*, 173(1), 20-51.

Bath, E., *et al.* (2017) Sperm and sex peptide stimulate aggression in female *Drosophila*. *Nature ecology & evolution*, 1(6), 1-6.

Bazzini, A.A., Lee, M.T. and Giraldez, A.J. (2012) Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. *Science (New York, N.Y.)*, 336(6078), 233-237.

Beckers, M., *et al.* (2017) Comprehensive processing of high-throughput small RNA sequencing data including quality checking, normalization, and differential expression analysis using the UEA sRNA Workbench. *RNA (New York, N.Y.)*, 23(6), 823-835.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1), 289-300.

Berezikov, E., *et al.* (2007) Mammalian mirtron genes. *Molecular cell*, 28(2), 328-336.

Bernstein, E., *et al.* (2001) Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, 409(6818), 363.

Betel, D., *et al.* (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome biology*, 11(8), R90.

Billi, A.C., Freeberg, M.A. and Kim, J.K. (2012) piRNAs and siRNAs collaborate in *Caenorhabditis elegans* genome defense. *Genome biology*, 13(7), 164.

Billmeier, M. and Xu, P. Small RNA Profiling by Next-Generation Sequencing Using High-Definition Adapters. In, *MicroRNA Detection and Target Identification*. Springer; 2017. p. 45-57.

Bisognin, A., *et al.* (2012) MAGIA2: from miRNA and genes expression data integrative analysis to microRNA–transcription factor mixed regulatory circuits (2012 update). *Nucleic acids research*, 40(W1), W13-W21.

Blake, W.J., *et al.* (2003) Noise in eukaryotic gene expression. *Nature*, 422(6932), 633.

Blanchette, M., *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome research*, 14(4), 708-715.

Bohnsack, M.T., Czaplinski, K. and GÖRLICH, D. (2004) Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. *RNA (New York, N.Y.)*, 10(2), 185-191.

Bradley, T. and Moxon, S. An assessment of the next generation of animal miRNA target prediction algorithms. In, *MicroRNA detection and target identification*. Springer; 2017. p. 175-191.

Bradley, T. and Moxon, S. (2019) FilTar: Using RNA-Seq data to improve microRNA target prediction accuracy in animals. *BioRxiv*, 595322.

Braun, J.E., *et al.* (2011) GW182 proteins directly recruit cytoplasmic deadenylase complexes to miRNA targets. *Molecular cell*, 44(1), 120-133.

Bray, N., *et al.* (2015) Near-optimal RNA-Seq quantification. *arXiv preprint arXiv:1505.02710*,

Bray, N.L., *et al.* (2016) Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34525.

Bray, N.L., *et al.* (2016) Near-optimal probabilistic RNA-seq quantification. *Nature biotechnology*, 34(5), 525.

Brennecke, J., *et al.* (2003) bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in *Drosophila*. *Cell*, 113(1), 25-36.

Brennecke, J., *et al.* (2005) Principles of microRNA–target recognition. *PLoS biology*, 3(3), e85.

Cai, X., Hagedorn, C.H. and Cullen, B.R. (2004) Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA (New York, N.Y.)*, 10(12), 1957-1966.

Calcino, A.D., *et al.* (2018) Diverse RNA interference strategies in early-branching metazoans. *BMC evolutionary biology*, 18(1), 160.

Cao, Y., *et al.* (2015) miR-290/371-Mbd2-Myc circuit regulates glycolytic metabolism to promote pluripotency. *The EMBO journal*, 34(5), 609-623.

Carvalho, G.B., *et al.* (2006) Allochrine modulation of feeding behavior by the sex peptide of *Drosophila*. *Current Biology*, 16(7), 692-696.

Chakraborty, M., *et al.* (2019) Networks of enhancers and microRNAs drive variation in cell states. *bioRxiv*, 668145.

Chandradoss, S.D., *et al.* (2015) A dynamic search process underlies microRNA targeting. *Cell*, 162(1), 96-107.

Chen, C.-Y.A. and Shyu, A.-B. (1995) AU-rich elements: characterization and importance in mRNA degradation. *Trends in biochemical sciences*, 20(11), 465-470.

Chen, Y., *et al.* (2014) A DDX6-CNOT1 complex and W-binding pockets in CNOT9 reveal direct links between miRNA target recognition and silencing. *Molecular cell*, 54(5), 737-750.

Chen, Y. and Wang, X. (2019) miRDB: an online database for prediction of functional microRNA targets. *Nucleic Acids Research*,

Chen, Y.-W., *et al.* (2014) Systematic study of *Drosophila* microRNA functions using a collection of targeted knockout mutations. *Developmental cell*, 31(6), 784-800.

Chi, S.W., *et al.* (2009) Argonaute HITS-CLIP decodes microRNA–mRNA interaction maps. *Nature*, 460(7254), 479-486.

Choi, W.-Y., Giraldez, A.J. and Schier, A.F. (2007) Target protectors reveal dampening and balancing of Nodal agonist and antagonist by miR-430. *Science (New York, N.Y.)*, 318(5848), 271-274.

Chou, C.-H., *et al.* (2017) miRTarBase update 2018: a resource for experimentally validated microRNA–target interactions. *Nucleic acids research*, 46(D1), D296-D302.

Christov, C.P., *et al.* (2006) Functional requirement of noncoding Y RNAs for human chromosomal DNA replication. *Molecular and cellular biology*, 26(18), 6993-7004.

Cock, P.J., *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11), 1422-1423.

Conesa, A., *et al.* (2016) A Survey of Best Practices for RNA-seq Data Analysis.

Consortium, G.O. (2016) Expansion of the Gene Ontology knowledgebase and resources. *Nucleic acids research*, 45(D1), D331-D338.

Dalmay, T., *et al.* (2000) An RNA-dependent RNA polymerase gene in *Arabidopsis* is required for posttranscriptional gene silencing mediated by a transgene but not by a virus. *Cell*, 101(5), 543-553.

Dalton, J.E., *et al.* (2010) Dynamic, mating-induced gene expression changes in female head and brain tissues of *Drosophila melanogaster*. *Bmc Genomics*, 11(1), 541.

Darnell, R.B. (2010) HITS-CLIP: panoramic views of protein–RNA regulation in living cells. *Wiley Interdisciplinary Reviews: RNA*, 1(2), 266-286.

Davis, M.P., *et al.* (2013) Kraken: a set of tools for quality control and analysis of high-throughput sequence data. *Methods*, 63(1), 41-49.

De Felippes, F.F., *et al.* (2008) Evolution of *Arabidopsis thaliana* microRNAs from random sequences. *RNA (New York, N.Y.)*, 14(12), 2455-2459.

Delbare, S.Y., *et al.* (2017) Roles of female and male genotype in post-mating responses in *Drosophila melanogaster*. *Journal of Heredity*, 108(7), 740-753.

Denli, A.M., *et al.* (2004) Processing of primary microRNAs by the Microprocessor complex. *Nature*, 432(7014), 231.

Denzler, R., *et al.* (2014) Assessing the ceRNA hypothesis with quantitative measurements of miRNA and target abundance. *Molecular cell*, 54(5), 766-776.

Denzler, R., *et al.* (2016) Impact of microRNA levels, target-site complementarity, and cooperativity on competing endogenous RNA-regulated gene expression. *Molecular cell*, 64(3), 565-579.

Diepenbruck, M., *et al.* (2017) miR-1199-5p and Zeb1 function in a double-negative feedback loop potentially coordinating EMT and tumour metastasis. *Nature communications*, 8(1), 1168.

Doench, J.G., Petersen, C.P. and Sharp, P.A. (2003) siRNAs can function as miRNAs. *Genes & development*, 17(4), 438-442.

Doench, J.G. and Sharp, P.A. (2004) Specificity of microRNA target selection in translational repression. *Genes & development*, 18(5), 504-511.

Domingos, J.A., *et al.* (2018) Sex-specific dmrt1 and cyp19a1 methylation and alternative splicing in gonads of the protandrous hermaphrodite barramundi. *PLoS one*, 13(9), e0204182.

dos Santos, G., *et al.* (2014) FlyBase: introduction of the *Drosophila melanogaster* Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic acids research*, 43(D1), D690-D697.

Drewe-Boss, P., Wessels, H.-H. and Ohler, U. (2018) omniCLIP: probabilistic identification of protein-RNA interactions from CLIP-seq data. *Genome biology*, 19(1), 183.

Durinck, S., *et al.* (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature protocols*, 4(8), 1184-1191.

Ebert, M.S., Neilson, J.R. and Sharp, P.A. (2007) MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. *Nature methods*, 4(9), 721-726.

Ebert, M.S. and Sharp, P.A. (2010) Emerging roles for natural microRNA sponges. *Current Biology*, 20(19), R858-R861.

Eden, E., *et al.* (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC bioinformatics*, 10(1), 48.

Ellis, L.L. and Carney, G.E. (2010) Mating alters gene expression patterns in *Drosophila melanogaster* male heads. *BMC genomics*, 11(1), 558.

Ender, C., *et al.* (2008) A human snoRNA with microRNA-like functions. *Molecular cell*, 32(4), 519-528.

Enright, A.J., *et al.* (2003) MicroRNA targets in *Drosophila*. *Genome biology*, 5(1), R1.

Eulalio, A., *et al.* (2007) P-body formation is a consequence, not the cause, of RNA-mediated gene silencing. *Molecular and cellular biology*, 27(11), 3970-3981.

Eulalio, A., Huntzinger, E. and Izaurralde, E. (2008) GW182 interaction with Argonaute is essential for miRNA-mediated translational repression and mRNA decay. *Nature structural & molecular biology*, 15(4), 346.

Ewels, P., *et al.* (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, 32(19), 3047-3048.

Farlora, R., *et al.* (2015) Identification of microRNAs associated with sexual maturity in rainbow trout brain and testis through small RNA deep sequencing. *Molecular reproduction and development*, 82(9), 651-662.

Fernández-Pérez, D., *et al.* (2018) MicroRNA dynamics at the onset of primordial germ and somatic cell sex differentiation during mouse embryonic gonad development. *RNA (New York, N.Y.)*, 24(3), 287-303.

Fire, A., *et al.* (1991) Production of antisense RNA leads to effective and specific inhibition of gene expression in *C. elegans* muscle. *Development*, 113(2), 503-514.

Fire, A., *et al.* (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *nature*, 391(6669), 806-811.

Fowler, E.K., *et al.* (2019) Divergence in Transcriptional and Regulatory Responses to Mating in Male and Female Fruitflies. *Scientific Reports*, 9(1), 1-15.

Fowler, E.K., Bradley, T., Moxon, S., Chapman, T. (2019) Divergence in Transcriptional and Regulatory Responses to Mating in Male and Female Fruitflies. *Scientific Reports (accepted)*,

Fowler, E.K., *et al.* (2018) Small RNA populations revealed by blocking rRNA fragments in *Drosophila melanogaster* reproductive tissues. *PLoS One*, 13(2), e0191966.

Frankel, L.B., *et al.* (2008) Programmed cell death 4 (PDCD4) is an important functional target of the microRNA miR-21 in breast cancer cells. *Journal of Biological Chemistry*, 283(2), 1026-1033.

Friedersdorf, M.B. and Keene, J.D. (2014) Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs. *Genome biology*, 15(1), R2.

Friedländer, M.R., *et al.* (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nature biotechnology*, 26(4), 407.

Friedländer, M.R., *et al.* (2011) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic acids research*, 40(1), 37-52.

Friedman, R.C., *et al.* (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, 19(1), 92-105.

Fromm, B., *et al.* (2015) A uniform system for the annotation of vertebrate microRNA genes and the evolution of the human microRNAome. *Annual review of genetics*, 49, 213-242.

Fromm, B., *et al.* (2019) MirGeneDB 2.0: The metazoan microRNA complement. *databases*, 2019, 2847-51.

Gaidatzis, D., *et al.* (2007) Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics*, 8, 69-69.

Garalde, D.R., *et al.* (2018) Highly parallel direct RNA sequencing on an array of nanopores. *Nature methods*, 15(3), 201.

Garcia, D.M., *et al.* (2011) Weak seed-pairing stability and high target-site abundance decrease the proficiency of Isy-6 and other microRNAs. *Nature structural & molecular biology*, 18(10), 1139-1146.

Giraldez, A.J., *et al.* (2006) Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science (New York, N.Y.)*, 312(5770), 75-79.

Goff, L., Trapnell, C. and Kelley, D. (2013) cummeRbund: Analysis, exploration, manipulation, and visualization of Cufflinks high-throughput sequencing data. *R package version*, 2(0),

Gomulski, L.M., *et al.* (2012) Transcriptome profiling of sexual maturation and mating in the Mediterranean fruit fly, *Ceratitis capitata*. *PloS one*, 7(1),

Gregory, R.I., *et al.* (2004) The Microprocessor complex mediates the genesis of microRNAs. *Nature*, 432(7014), 235.

Griffiths-Jones, S., *et al.* (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic acids research*, 34(suppl 1), D140-D144.

Griffiths-Jones, S., *et al.* (2007) miRBase: tools for microRNA genomics. *Nucleic acids research*, 36(suppl_1), D154-D158.

Griffiths-Jones, S. (2004) The microRNA registry. *Nucleic acids research*, 32(suppl_1), D109-D111.

Grimson, A., *et al.* (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular cell*, 27(1), 91-105.

Grimson, A., *et al.* (2008) Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature*, 455(7217), 1193.

Grishok, A., *et al.* (2001) Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing. *Cell*, 106(1), 23-34.

Grosswendt, S., *et al.* (2014) Unambiguous identification of miRNA: target site interactions by different types of ligation reactions. *Molecular cell*, 54(6), 1042-1054.

Gruber, A.J., *et al.* (2018) Terminal exon characterization with TECtool reveals an abundance of cell-specific isoforms. *Nature methods*, 15(10), 832.

Gruber, A.J., *et al.* (2018) Discovery of physiological and cancer-related regulators of 3' UTR processing with KAPAC. *Genome biology*, 19(1), 44.

Gruber, A.J. and Zavolan, M. (2019) Alternative cleavage and polyadenylation in health and disease. *Nature Reviews Genetics*, 1.

Gruber, A.R., *et al.* (2014) Means to an end: mechanisms of alternative polyadenylation of messenger RNA precursors. *Wiley Interdisciplinary Reviews: RNA*, 5(2), 183-196.

Grüning, B.A., *et al.* (2017) The RNA workbench: best practices for RNA and high-throughput sequencing bioinformatics in Galaxy. *Nucleic acids research*, 45(W1), W560-W566.

Guan, D., *et al.* (2013) Switching cell fate, ncRNAs coming to play. *Cell death & disease*, 4(1), e464.

Guiguen, Y., *et al.* (1994) Reproductive cycle and sex inversion of the seabass, *Lates calcarifer*, reared in sea cages in French Polynesia: histological and morphometric description. *Environmental Biology of Fishes*, 39(3), 231-247.

Gumienny, R. and Zavolan, M. (2015) Accurate transcriptome-wide prediction of microRNA targets and small interfering RNA off-targets with MIRZA-G. *Nucleic Acids Research*, 43(3), 1380-1391.

Guo, H., *et al.* (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, 466(7308), 835.

Guo, J.U., *et al.* (2014) Expanded identification and characterization of mammalian circular RNAs. *Genome biology*, 15(7), 409.

Guo, S. and Kemphues, K.J. (1995) par-1, a gene required for establishing polarity in *C. elegans* embryos, encodes a putative Ser/Thr kinase that is asymmetrically distributed. *Cell*, 81(4), 611-620.

Hafner, M., *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1), 129-141.

Hall, A.E., Turnbull, C. and Dalmay, T. (2013) Y RNAs: recent developments. *Biomolecular concepts*, 4(2), 103-110.

Hamilton, A.J. and Baulcombe, D.C. (1999) A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science (New York, N.Y.)*, 286(5441), 950-952.

Hammond, S.M., *et al.* (2000) An RNA-directed nuclease mediates post-transcriptional gene silencing in *Drosophila* cells. *nature*, 404(6775), 293-296.

Hammond, S.M., *et al.* (2001) Argonaute2, a link between genetic and biochemical analyses of RNAi. *Science (New York, N.Y.)*, 293(5532), 1146-1150.

Hannon, G.J. (2002) RNA interference. *Nature*, 418(6894), 244-251.

Harrison, P.W., *et al.* (2018) The European Nucleotide Archive in 2018. *Nucleic acids research*, 47(D1), D84-D88.

Hausser, J., *et al.* (2009) Relative contribution of sequence and structure features to the mRNA binding of Argonaute/EIF2C-miRNA complexes and the degradation of miRNA targets. *Genome research*, 19(11),

He, L., *et al.* (2015) Profiling microRNAs in the testis during sexual maturation stages in *Eriocheir sinensis*. *Animal reproduction science*, 16252-61.

Heid, C.A., *et al.* (1996) Real time quantitative PCR. *Genome research*, 6(10), 986-994.

Heifetz, Y., *et al.* (2014) Mating regulates neuromodulator ensembles at nerve termini innervating the *Drosophila* reproductive tract. *Current Biology*, 24(7), 731-737.

Helwak, A., *et al.* Mapping the Human miRNA Interactome by CLASH Reveals Frequent Noncanonical Binding. *Cell*, 153(3), 654-665.

Helwak, A. and Tollervey, D. (2014) Mapping the miRNA interactome by cross-linking ligation and sequencing of hybrids (CLASH). *Nat. Protocols*, 9(3), 711-728.

Höck, J. and Meister, G. (2008) The Argonaute protein family. *Genome biology*, 9(2), 210.

Holt, C.E. and Bullock, S.L. (2009) Subcellular mRNA localization in animal cells and why it matters. *Science (New York, N.Y.)*, 326(5957), 1212-1216.

Hornstein, E. and Shomron, N. (2006) Canalization of development by microRNAs. *Nature genetics*, 38S20-S24.

Hoskins, R.A., *et al.* (2015) The Release 6 reference sequence of the *Drosophila melanogaster* genome. *Genome research*, 25(3), 445-458.

Hoyle, D.C., *et al.* (2002) Making sense of microarray data distributions. *Bioinformatics*, 18(4), 576-584.

Huber, W., *et al.* (2015) Orchestrating high-throughput genomic analysis with Bioconductor. *Nature methods*, 12(2), 115.

Humphreys, D.T., *et al.* (2005) MicroRNAs control translation initiation by inhibiting eukaryotic initiation factor 4E/cap and poly (A) tail function. *Proceedings of the National Academy of Sciences*, 102(47), 16961-16966.

Huntzinger, E., *et al.* (2010) Two PABPC1-binding sites in GW182 proteins promote miRNA-mediated gene silencing. *The EMBO journal*, 29(24), 4146-4160.

Huntzinger, E. and Izaurralde, E. (2011) Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nature Reviews Genetics*, 12(2), 99.

Huppertz, I., *et al.* (2014) iCLIP: protein–RNA interactions at nucleotide resolution. *Methods*, 65(3), 274-287.

Hutvagner, G., *et al.* (2001) A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science (New York, N.Y.)*, 293(5531), 834-838.

Immonen, E., *et al.* (2017) Mating changes sexually dimorphic gene expression in the seed beetle *Callosobruchus maculatus*. *Genome biology and evolution*, 9(3), 677-699.

Innocenti, P. and Morrow, E.H. (2009) Immunogenic males: a genome-wide analysis of reproduction and the cost of mating in *Drosophila melanogaster* females. *Journal of evolutionary biology*, 22(5), 964-973.

Iovino, N., Pane, A. and Gaul, U. (2009) miR-184 has multiple roles in *Drosophila* female germline development. *Developmental cell*, 17(1), 123-133.

Isaac, R.E., *et al.* (2010) *Drosophila* male sex peptide inhibits siesta sleep and promotes locomotor activity in the post-mated female. *Proceedings of the Royal Society B: Biological Sciences*, 277(1678), 65-70.

Izant, J.G. and Weintraub, H. (1984) Inhibition of thymidine kinase gene expression by anti-sense RNA: a molecular approach to genetic analysis. *Cell*, 36(4), 1007-1015.

Jackson, A.L., *et al.* (2006) Widespread siRNA “off-target” transcript silencing mediated by seed region sequence complementarity. *RNA (New York, N.Y.)*, 12(7), 1179-1187.

Jain, M., *et al.* (2017) MinION analysis and reference consortium: phase 2 data release and analysis of R9.0 chemistry. *F1000Research*, 6

Jalkanen, A.L., Coleman, S.J. and Wilusz, J. Determinants and implications of mRNA poly (A) tail size—Does this protein make my tail look big? In, *Seminars in cell & developmental biology*. Elsevier; 2014. p. 24-32.

Jan, C.H., *et al.* (2011) Formation, regulation and evolution of *Caenorhabditis elegans* 3'UTRs. *Nature*, 469(7328), 97-101.

Jin, H.Y., *et al.* (2015) Transfection of microRNA mimics should be used with caution. *Frontiers in genetics*, 6340.

John, B., *et al.* (2004) Human microRNA targets. *PLoS Biol*, 2(11), e363.

Jones-Rhoades, M.W., Bartel, D.P. and Bartel, B. (2006) MicroRNAs and their regulatory roles in plants. *Annu. Rev. Plant Biol.*, 5719-53.

Kalvari, I., *et al.* (2017) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic acids research*, 46(D1), D335-D342.

Kalvari, I., *et al.* (2018) Non-Coding RNA Analysis Using the Rfam Database. *Current protocols in bioinformatics*, 62(1), e51.

Kapelnikov, A., *et al.* (2008) Mating induces an immune response and developmental switch in the *Drosophila* oviduct. *Proceedings of the National Academy of Sciences*, 105(37), 13912-13917.

Karagkouni, D., *et al.* (2017) DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA–gene interactions. *Nucleic acids research*, 46(D1), D239-D245.

Kent, W.J., *et al.* (2002) The human genome browser at UCSC. *Genome research*, 12(6), 996-1006.

Kertesz, M., *et al.* (2007) The role of site accessibility in microRNA target recognition. *Nat Genet*, 39(10), 1278-1284.

Khorshid, M., *et al.* (2013) A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. *Nature methods*, 10(3), 253-255.

Kim, D., Langmead, B. and Salzberg, S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nature methods*, 12(4), 357.

Kiriakidou, M., *et al.* (2004) A combined computational-experimental approach predicts human microRNA targets. *Genes & development*, 18(10), 1165-1178.

Kishore, S., *et al.* (2011) A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nature methods*, 8(7), 559.

Kleaveland, B., *et al.* (2018) A network of noncoding regulatory RNAs acts in the mammalian brain. *Cell*, 174(2), 350-362. e317.

Klum, S.M., *et al.* (2018) Helix-7 in Argonaute2 shapes the microRNA seed region for rapid target recognition. *The EMBO journal*, 37(1), 75-88.

Kocher, S.D., *et al.* (2008) Genomic analysis of post-mating changes in the honey bee queen (*Apis mellifera*). *BMC genomics*, 9(1), 232.

Köster, J. and Rahmann, S. (2012) Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19), 2520-2522.

Kozomara, A., Birgaoanu, M. and Griffiths-Jones, S. (2018) miRBase: from microRNA sequences to function. *Nucleic Acids Research*, 47(D1), D155-D162.

Kozomara, A. and Griffiths-Jones, S. (2010) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic acids research*, gkq1027.

Kozomara, A. and Griffiths-Jones, S. (2010) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic acids research*, 39(suppl_1), D152-D157.

Kozomara, A. and Griffiths-Jones, S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic acids research*, 42(D1), D68-D73.

Krueger, F. (2015) Trim galore. *A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files*,

Kudla, G., *et al.* (2011) Cross-linking, ligation, and sequencing of hybrids reveals RNA–RNA interactions in yeast. *Proceedings of the National Academy of Sciences*, 108(24), 10010-10015.

Kulkarni, M., Ozgur, S. and Stoecklin, G. On track with P-bodies. In.: Portland Press Limited; 2010.

Kumar, P., *et al.* (2014) Meta-analysis of tRNA derived RNA fragments reveals that they are evolutionarily conserved and associate with AGO proteins to recognize specific RNA targets. *BMC biology*, 12(1), 78.

Kuscu, C., *et al.* (2018) tRNA fragments (tRFs) guide Ago to regulate gene expression post-transcriptionally in a Dicer-independent manner. *RNA (New York, N.Y.)*, 24(8), 1093-1105.

Laczny, C., *et al.* (2012) miRTrail-a comprehensive webserver for analyzing gene and miRNA patterns to enhance the understanding of regulatory mechanisms in diseases. *BMC bioinformatics*, 13(1), 36.

Lagos-Quintana, M., *et al.* (2001) Identification of novel genes coding for small expressed RNAs. *Science (New York, N.Y.)*, 294(5543), 853-858.

Lai, E.C., Tam, B. and Rubin, G.M. (2005) Pervasive regulation of Drosophila Notch target genes by GY-box-, Brd-box-, and K-box-class microRNAs. *Genes & development*, 19(9), 1067-1080.

Lall, S., *et al.* (2006) A genome-wide map of conserved microRNA targets in *C. elegans*. *Current biology*, 16(5), 460-471.

Lau, N.C., *et al.* (2001) An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science (New York, N.Y.)*, 294(5543), 858-862.

Lawnczak, M.K. and Begun, D.J. (2004) A genome-wide analysis of courting and mating responses in *Drosophila melanogaster* females. *Genome*, 47(5), 900-910.

Le Brigand, K., *et al.* (2010) MiRonTop: mining microRNAs targets across large scale gene expression studies. *Bioinformatics*, 26(24), 3131-3132.

Lee, H.-C., *et al.* (2010) Diverse pathways generate microRNA-like RNAs and Dicer-independent small interfering RNAs in fungi. *Molecular cell*, 38(6), 803-814.

Lee, R.C. and Ambros, V. (2001) An extensive class of small RNAs in *Caenorhabditis elegans*. *Science (New York, N.Y.)*, 294(5543), 862-864.

Lee, R.C., Feinbaum, R.L. and Ambros, V. (1993) The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell*, 75(5), 843-854.

Lee, Y., *et al.* (2004) MicroRNA genes are transcribed by RNA polymerase II. *The EMBO journal*, 23(20), 4051-4060.

Leek, J.T., *et al.* (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10), 733.

Legnini, I., *et al.* (2019) FLAM-seq: full-length mRNA sequencing reveals principles of poly (A) tail length control. *Nature methods*, 1-8.

Leinonen, R., *et al.* (2010) The European nucleotide archive. *Nucleic acids research*, 39(suppl_1), D28-D31.

Leinonen, R., Sugawara, H. and Shumway, M. (2010) The sequence read archive. *Nucleic acids research*, gkq1019.

Leung, A.K.L. and Sharp, P. Function and localization of microRNAs in mammalian cells. In, *Cold Spring Harbor symposia on quantitative biology*. Cold Spring Harbor Laboratory Press; 2006. p. 29-38.

Lewis, B.P., Burge, C.B. and Bartel, D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1), 15-20.

Lewis, B.P., *et al.* (2003) Prediction of mammalian microRNA targets. *Cell*, 115(7), 787-798.

- Li, B., *et al.* (2009) RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26(4), 493-500.
- Li, H., *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16), 2078-2079.
- Lim, L.P., *et al.* (2003) The microRNAs of *Caenorhabditis elegans*. *Genes & development*, 17(8), 991-1008.
- Lipp, J. Why sequencing data is modeled as negative binomial. In, *BIORAMBLE*. Wordpress; 2016.
- Litterman, A.J., *et al.* (2019) A massively parallel 3' UTR reporter assay reveals relationships between nucleotide content, sequence conservation, and mRNA destabilization. *Genome research*, 29(6), 896-906.
- Liu, C., *et al.* (2017) MicroRNA-141 suppresses prostate cancer stem cells and metastasis by targeting a cohort of pro-metastasis genes. *Nature communications*, 814270.
- Liu, H., *et al.* (2010) Improving performance of mammalian microRNA target prediction. *BMC bioinformatics*, 11(1), 476.
- Liu, J., *et al.* (2004) Argonaute2 is the catalytic engine of mammalian RNAi. *Science (New York, N.Y.)*, 305(5689), 1437-1441.
- Liu, J., *et al.* (2015) Dynamic evolution and biogenesis of small RNAs during sex reversal. *Scientific reports*, 59999.
- Liu, W. and Wang, X. (2019) Prediction of functional microRNA targets by integrative modeling of microRNA binding and target expression data. *Genome biology*, 20(1), 18.
- Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12), 550.
- Love, M.I., Huber, W. and Anders, S.J.G.B. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. 15(12), 550.
- Lu, J., Tomfohr, J.K. and Kepler, T.B. (2005) Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach. *BMC bioinformatics*, 6(1), 165.
- Lu, Y. and Leslie, C.S. (2016) Learning to predict miRNA-mRNA interactions from AGO CLIP sequencing and CLASH data. *PLoS computational biology*, 12(7), e1005026.
- Ludwig, N., *et al.* (2017) Bias in recent miRBase annotations potentially associated with RNA quality issues. *Scientific reports*, 7(1), 1-11.
- Mack, P.D., *et al.* (2006) Mating-responsive genes in reproductive tissues of female *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences*, 103(27), 10358-10363.
- Maere, S., Heymans, K. and Kuiper, M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21(16), 3448-3449.
- Majoros, W.H. and Ohler, U. (2007) Spatial preferences of microRNA targets in 3'untranslated regions. *BMC genomics*, 8(1), 152.
- Mapleson, D., *et al.* (2013) MirPlex: A Tool for Identifying miRNAs in High-Throughput sRNA Datasets Without a Genome. *Journal of Experimental Zoology Part B-Molecular and Developmental Evolution*, 320B(1), 47-56.

Marioni, J.C., *et al.* (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9), 1509-1517.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, 17(1), 10-12.

Martinez, J., *et al.* (2002) Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. *Cell*, 110(5), 563-574.

Mathonnet, G., *et al.* (2007) MicroRNA inhibition of translation initiation in vitro by targeting the cap-binding complex eIF4F. *Science (New York, N.Y.)*, 317(5845), 1764-1767.

Mathys, H., *et al.* (2014) Structural and biochemical insights to the role of the CCR4-NOT complex and DDX6 ATPase in microRNA repression. *Molecular cell*, 54(5), 751-765.

McGraw, L.A., Clark, A.G. and Wolfner, M.F. (2008) Post-mating gene expression profiles of female *Drosophila melanogaster* in response to time and to four male accessory gland proteins. *Genetics*, 179(3), 1395-1408.

McGraw, L.A., *et al.* (2004) Genes regulated by mating, sperm, or seminal proteins in mated female *Drosophila melanogaster*. *Current Biology*, 14(16), 1509-1514.

McJunkin, K. and Ambros, V. (2017) A microRNA family exerts maternal control on sex determination in *C. elegans*. *Genes & development*, 31(4), 422-437.

Meiri, E., *et al.* (2010) Discovery of microRNAs and other small RNAs in solid tumors. *Nucleic acids research*, 38(18), 6234-6246.

Meister, G., *et al.* (2004) Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Molecular cell*, 15(2), 185-197.

Miranda, K.C., *et al.* (2006) A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell*, 126(6), 1203-1217.

Mohorianu, I., *et al.* (2018) Control of seminal fluid protein expression via regulatory hubs in *Drosophila melanogaster*. *Proceedings of the Royal Society B: Biological Sciences*, 285(1887), 20181681.

Mohorianu, I., *et al.* (2012) FiRePat—finding regulatory patterns between sRNAs and genes. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(3), 273-284.

Mohorianu, I., *et al.* The UEA small RNA workbench: a suite of computational tools for small RNA analysis. In, *MicroRNA Detection and Target Identification*. Springer; 2017. p. 193-224.

Moran, Y., *et al.* (2017) The evolutionary origin of plant and animal microRNAs. *Nature ecology & evolution*, 1(3), 0027.

Mortazavi, A., *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7), 621.

Moxon, S., Moulton, V. and Kim, J.T. (2008) A scoring matrix approach to detecting miRNA target sites. *Algorithms for Molecular Biology*, 3

Moxon, S., *et al.* (2008) A toolkit for analysing large-scale plant small RNA datasets. *Bioinformatics*, 24(19), 2252-2253.

Mukherji, S., *et al.* (2011) MicroRNAs can generate thresholds in target gene expression. *Nature genetics*, 43(9), 854.

Nam, J.W., *et al.* (2014) Global analyses of the effect of different cellular contexts on microRNA targeting. *Molecular cell*, 53(6), 1031-1043.

Nam, S., *et al.* (2009) MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression. *Nucleic acids research*, 37(suppl_2), W356-W362.

Nelder, J.A. and Wedderburn, R.W. (1972) Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3), 370-384.

Neugebauer, K.M. (2002) On the importance of being co-transcriptional. *Journal of cell science*, 115(20), 3865-3871.

Nicholson, A.L. and Pasquinelli, A.E. (2018) Tales of detailed Poly (A) tails. *Trends in cell biology*,

Nicolas, F.E., *et al.* (2012) Biogenesis of Y RNA-derived small RNAs is independent of the microRNA pathway. *FEBS letters*, 586(8), 1226-1230.

Nielsen, C.B., *et al.* (2007) Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA (New York, N.Y.)*, 13(11), 1894-1910.

Niranjanakumari, S., *et al.* (2002) Reversible cross-linking combined with immunoprecipitation to study RNA–protein interactions in vivo. *Methods*, 26(2), 182-190.

Nobuta, K., *et al.* Bioinformatics analysis of small RNAs in plants using next generation sequencing technologies. In, *Plant MicroRNAs*. Springer; 2010. p. 89-106.

Nozawa, M., Miura, S. and Nei, M. (2010) Origins and evolution of microRNA genes in *Drosophila* species. *Genome biology and evolution*, 2180-189.

Nykänen, A., Haley, B. and Zamore, P.D. (2001) ATP requirements and small interfering RNA structure in the RNA interference pathway. *Cell*, 107(3), 309-321.

Okamura, K., *et al.* (2007) The mirtron pathway generates microRNA-class regulatory RNAs in *Drosophila*. *Cell*, 130(1), 89-100.

Okamura, K. and Lai, E.C. (2008) Endogenous small interfering RNAs in animals. *Nature reviews Molecular cell biology*, 9(9), 673.

Orlov, Y.L. and Potapov, V.N. (2004) Complexity: an internet resource for analysis of DNA sequence complexity. *Nucleic acids research*, 32(suppl_2), W628-W633.

Ovando-Vázquez, C., Lepe-Soltero, D. and Abreu-Goodger, C. (2016) Improving microRNA target prediction with gene expression profiles. *BMC genomics*, 17(1), 364.

Paicu, C., *et al.* (2017) miRCat2: accurate prediction of plant and animal microRNAs from next-generation sequencing datasets. *Bioinformatics*, 33(16), 2446-2454.

Pasquinelli, A.E., *et al.* (2000) Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, 408(6808), 86.

Patro, R., *et al.* (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, 14(4), 417.

Patro, R., Mount, S.M. and Kingsford, C. (2014) Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nature biotechnology*, 32(5), 462-464.

Pauli, A., Rinn, J.L. and Schier, A.F. (2011) Non-coding RNAs as regulators of embryogenesis. *Nature Reviews Genetics*, 12(2), 136.

Penso-Dolfin, L., *et al.* (2018) The evolutionary dynamics of microRNAs in domestic mammals. *Scientific reports*, 8(1), 17050.

Penso-Dolfin, L., *et al.* (2016) An improved microRNA annotation of the canine genome. *PloS one*, 11(4),

Peterson, K.J., Dietrich, M.R. and McPeck, M.A. (2009) MicroRNAs and metazoan macroevolution: insights into canalization, complexity, and the Cambrian explosion. *Bioessays*, 31(7), 736-747.

Pfeffer, S., Lagos-Quintana, M. and Tuschl, T. (2005) Cloning of small RNA molecules. *Current protocols in molecular biology*, 72(1), 26.24. 21-26.24. 18.

Pillai, R.S., *et al.* (2005) Inhibition of translational initiation by Let-7 MicroRNA in human cells. *Science (New York, N.Y.)*, 309(5740), 1573-1576.

Pimentel, H. What the FPKM? A review of RNA-Seq expression units. In.; 2014.

Pimentel, H., *et al.* (2017) Differential analysis of RNA-seq incorporating quantification uncertainty. *Nature Methods*, 14687.

Pinzón, N., *et al.* (2017) microRNA target prediction programs predict many false positives. *Genome research*, 27(2), 234-245.

Plaisier, C.L., Bare, J.C. and Baliga, N.S. (2011) miRvestigator: web application to identify miRNAs responsible for co-regulated gene expression patterns discovered through transcriptome profiling. *Nucleic acids research*, 39(suppl_2), W125-W131.

Polioudakis, D., Abell, N.S. and Iyer, V.R. (2015) miR-503 represses human cell proliferation and directly targets the oncogene DDHD2 by non-canonical target pairing. *BMC genomics*, 16(1), 40.

Prokupek, A.M., *et al.* (2009) Transcriptional profiling of the sperm storage organs of *Drosophila melanogaster*. *Insect molecular biology*, 18(4), 465-475.

Proudfoot, N.J., Furger, A. and Dye, M.J. (2002) Integrating mRNA processing with transcription. *Cell*, 108(4), 501-512.

Prüfer, K., *et al.* (2008) PatMaN: rapid alignment of short sequences to large databases. *Bioinformatics*, 24(13), 1530-1531.

Prüfer, K., *et al.* (2008) PatMaN: rapid alignment of short sequences to large databases. *Bioinformatics*, 24(13), 1530-1531.

Pua, H.H., *et al.* (2016) MicroRNAs 24 and 27 suppress allergic inflammation and target a network of regulators of T helper 2 cell-associated cytokine production. *Immunity*, 44(4), 821-832.

Quinlan, A.R. (2014) BEDTools: the Swiss-army tool for genome feature analysis. *Current protocols in bioinformatics*, 47(1), 11.12. 11-11.12. 34.

Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), 841-842.

Rajewsky, N. (2006) microRNA target predictions in animals. *Nature genetics*, 38(6s), S8.

Rajewsky, N. and Socci, N.D. (2004) Computational identification of microRNA targets. *Genome biology*, 5(2), P5.

Rang, F.J., Kloosterman, W.P. and de Ridder, J. (2018) From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome biology*, 19(1), 90.

Ravi Ram, K. and Wolfner, M.F. (2007) Seminal influences: *Drosophila* Acps and the molecular interplay between males and females during reproduction. *Integrative and comparative biology*, 47(3), 427-445.

Reczko, M., *et al.* (2012) Functional microRNA targets in protein coding sequences. *Bioinformatics*, 28(6), 771-776.

Rehwinkel, J., *et al.* (2005) A crucial role for GW182 and the DCP1: DCP2 decapping complex in miRNA-mediated gene silencing. *RNA (New York, N.Y.)*, 11(11), 1640-1647.

Reinhart, B.J., *et al.* (2000) The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. *nature*, 403(6772), 901-906.

Rhoades, M.W., *et al.* (2002) Prediction of plant microRNA targets. *Cell*, 110(4), 513-520.

Ritchie, W. and Rasko, J.E. (2014) Refining microRNA target predictions: sorting the wheat from the chaff. *Biochemical and biophysical research communications*, 445(4), 780-784.

Robinson, M.D. and Smyth, G.K. (2007) Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21), 2881-2887.

Rogers, D.W., *et al.* (2008) Molecular and cellular components of the mating machinery in *Anopheles gambiae* females. *Proceedings of the National Academy of Sciences*, 105(49), 19390-19395.

Ruby, J.G., Jan, C.H. and Bartel, D.P. (2007) Intronic microRNA precursors that bypass Drosha processing. *Nature*, 448(7149), 83.

Rutjes, S.A., *et al.* (1999) Rapid nucleolytic degradation of the small cytoplasmic Y RNAs during apoptosis. *Journal of Biological Chemistry*, 274(35), 24799-24807.

Sætrom, P., *et al.* (2007) Distance constraints between microRNA target sites dictate efficacy and cooperativity. *Nucleic acids research*, 35(7), 2333-2342.

Salmena, L., *et al.* (2011) A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell*, 146(3), 353-358.

Schena, M., *et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y.)*, 270(5235), 467-470.

Schirle, N.T. and MacRae, I.J. (2012) The crystal structure of human Argonaute2. *Science (New York, N.Y.)*, 336(6084), 1037-1040.

Schirle, N.T., *et al.* (2015) Water-mediated recognition of t1-adenosine anchors Argonaute2 to microRNA targets. *Elife*, 4e07646.

Schirle, N.T., Sheu-Gruttadauria, J. and MacRae, I.J. (2014) Structural basis for microRNA targeting. *Science (New York, N.Y.)*, 346(6209), 608-613.

Schmiedel, J.M., *et al.* (2015) MicroRNA control of protein expression noise. *Science (New York, N.Y.)*, 348(6230), 128-132.

Schwab, R., *et al.* (2005) Specific effects of microRNAs on the plant transcriptome. *Developmental cell*, 8(4), 517-527.

Selbach, M., *et al.* (2008) Widespread changes in protein synthesis induced by microRNAs. *nature*, 455(7209), 58.

Sethupathy, P., Corda, B. and Hatzigeorgiou, A.G. (2006) TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA (New York, N.Y.)*, 12(2), 192-197.

Shannon, P., *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13(11), 2498-2504.

Sharp, P.A. (2001) RNA interference—2001. *Genes & development*, 15(5), 485-490.

Shin, C., *et al.* (2010) Expanding the microRNA targeting code: functional sites with centered pairing. *Molecular cell*, 38(6), 789-802.

Simkin, A., *et al.* (2019) Evolutionary Dynamics of microRNA target sites across vertebrate evolution. *BioRxiv*, 693069.

Siro, L.K., *et al.* (2009) Seminal fluid protein depletion and replenishment in the fruit fly, *Drosophila melanogaster*: an ELISA-based method for tracking individual ejaculates. *Behavioral ecology and sociobiology*, 63(10), 1505-1513.

Siro, L.K., Wolfner, M.F. and Wigby, S. (2011) Protein-specific manipulation of ejaculate composition in response to female mating status in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences*, 108(24), 9922-9926.

Siro, L.K., *et al.* (2015) Sexual conflict and seminal fluid proteins: a dynamic landscape of sexual interactions. *Cold Spring Harbor perspectives in biology*, 7(2), a017533.

Skaftnesmo, K.O., *et al.* (2017) Integrative testis transcriptome analysis reveals differentially expressed miRNAs and their mRNA targets during early puberty in Atlantic salmon. *BMC genomics*, 18(1), 801.

Slutskin, I.V., *et al.* (2018) Unraveling the determinants of microRNA mediated regulation using a massively parallel reporter assay. *Nature communications*, 9(1), 529.

Smalheiser, N.R. and Torvik, V.I. (2005) Mammalian microRNAs derived from genomic repeats. *Trends in Genetics*, 21(6), 322-326.

Soares, A.R., *et al.* (2009) Parallel DNA pyrosequencing unveils new zebrafish microRNAs. *Bmc Genomics*, 10(1), 195.

Soneson, C., Love, M.I. and Robinson, M.D. (2015) Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, 4

Sorefan, K., *et al.* (2012) Reducing ligation bias of small RNAs in libraries for next generation sequencing. *Silence*, 3(1), 4.

Stam, M., *et al.* (1997) Post-transcriptional silencing of chalcone synthase in *Petunia* by inverted transgene repeats. *The Plant Journal*, 12(1), 63-82.

Stark, A., *et al.* (2003) Identification of *Drosophila* microRNA targets. *PLoS biology*, 1(3), e60.

Stocks, M.B., *et al.* (2018) The UEA sRNA Workbench (version 4.4): a comprehensive suite of tools for analyzing miRNAs and sRNAs. *Bioinformatics*, 34(19), 3382-3384.

Stocks, M.B., *et al.* (2012) The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. *Bioinformatics*, 28(15), 2059-2061.

Stolzenburg, L.R., *et al.* (2016) miR-1343 attenuates pathways of fibrosis by targeting the TGF- β receptors. *Biochemical Journal*, 473(3), 245-256.

Tamim, S., *et al.* (2014) Genomic analyses reveal broad impact of miR-137 on genes associated with malignant transformation and neuronal differentiation in glioblastoma cells. *PLoS one*, 9(1), e85591.

Tarver, J.E., *et al.* (2015) microRNAs and the evolution of complex multicellularity: identification of a large, diverse complement of microRNAs in the brown alga *Ectocarpus*. *Nucleic acids research*, 43(13), 6384-6398.

Tarver, J.E., Donoghue, P.C. and Peterson, K.J. (2012) Do miRNAs have a deep evolutionary history? *Bioessays*, 34(10), 857-866.

Tarver, J.E., *et al.* (2018) Well-annotated microRNAomes do not evidence pervasive miRNA loss. *Genome biology and evolution*, 10(6), 1457-1470.

Taylor, R.S., *et al.* (2017) MicroRNA annotation of plant genomes– Do it right or not at all. *BioEssays*, 39(2), 1600113.

Team, R.C. (2013) R: A language and environment for statistical computing.

Thattai, M. and Van Oudenaarden, A. (2001) Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences*, 98(15), 8614-8619.

Thomson, D.W., *et al.* (2013) On measuring miRNAs after transient transfection of mimics or antisense inhibitors. *PLoS one*, 8(1), e55214.

Thomson, D.W., *et al.* (2014) Assessing the gene regulatory properties of Argonaute-bound small RNAs of diverse genomic origin. *Nucleic acids research*, 43(1), 470-481.

Thorvaldsdóttir, H., Robinson, J.T. and Mesirov, J.P. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics*, 14(2), 178-192.

Trapnell, C., *et al.* (2013) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature biotechnology*, 31(1), 46-53.

Trapnell, C., Pachter, L. and Salzberg, S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9), 1105-1111.

Trifonov, E.N. (1990) Making sense of the human genome. *Structure and methods: proceedings of the Sixth Conversation in the Discipline Biomolecular Stereodynamics held at the State University of New York at Albany, June 6-10, 1989/edited by RH Sarma & MH Sarma*,

Troyanskaya, O.G., *et al.* (2002) Sequence complexity profiles of prokaryotic genomic sequences: A fast algorithm for calculating linguistic complexity. *Bioinformatics*, 18(5), 679-688.

Ule, J., *et al.* (2003) CLIP identifies Nova-regulated RNA networks in the brain. *Science (New York, N.Y.)*, 302(5648), 1212-1215.

University, J.H. HISAT2 manual. In.

Van Dongen, S., Abreu-Goodger, C. and Enright, A.J. (2008) Detecting microRNA binding and siRNA off-target effects from expression data. *Nature methods*, 5(12), 1023.

Van Nostrand, E.L., *et al.* (2016) Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature Methods*,

Vaz, C., *et al.* (2015) Deep sequencing of small RNA facilitates tissue and sex associated microRNA discovery in zebrafish. *BMC genomics*, 16(1), 950.

Vij, S., *et al.* (2016) Chromosomal-level assembly of the Asian seabass genome using long sequence reads and multi-layered scaffolding. *PLoS genetics*, 12(4), e1005954.

Vitsios, D.M., *et al.* (2017) Mirnovo: genome-free prediction of microRNAs from small RNA sequencing data and single-cells using decision forests. *Nucleic acids research*, 45(21), e177-e177.

Wang, X. (2016) Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-Ligation studies. *Bioinformatics*, btw002.

Waterhouse, P.M., Graham, M.W. and Wang, M.-B. (1998) Virus resistance and gene silencing in plants can be induced by simultaneous expression of sense and

antisense RNA. *Proceedings of the National Academy of Sciences*, 95(23), 13959-13964.

Weick, E.-M. and Miska, E.A. (2014) piRNAs: from biogenesis to function. *Development*, 141(18), 3458-3471.

Wickham, H. tidyverse: Easily Install and Load “Tidyverse” Packages (2017). URL <https://CRAN.R-project.org/package=tidyverse>. *R package version*, 1(1), 51.

Wickham, H. ggplot2: elegant graphics for data analysis. Springer; 2016.

Wigby, S., *et al.* (2009) Seminal fluid protein allocation and male reproductive success. *Current Biology*, 19(9), 751-757.

Wightman, B., Ha, I. and Ruvkun, G. (1993) Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. *Cell*, 75(5), 855-862.

Williams, T.M. and Carroll, S.B. (2009) Genetic and molecular insights into the development and evolution of sexual dimorphism. *Nature Reviews Genetics*, 10(11), 797-804.

Winterhalter, W.E. and Fedorka, K.M. (2009) Sex-specific variation in the emphasis, inducibility and timing of the post-mating immune response in *Drosophila melanogaster*. *Proceedings of the Royal Society B: Biological Sciences*, 276(1659), 1109-1117.

Wong, N. and Wang, X. (2014) miRDB: an online resource for microRNA target prediction and functional annotations. *Nucleic acids research*, 43(D1), D146-D152.

Wu, W.-S., *et al.* (2017) CSmiRTar: condition-specific microRNA targets database. *PloS one*, 12(7), e0181231.

Xiao, F., *et al.* (2008) miRecords: an integrated resource for microRNA–target interactions. *Nucleic acids research*, 37(suppl_1), D105-D110.

Xiao, J., *et al.* (2014) Identification and characterization of microRNAs in ovary and testis of Nile tilapia (*Oreochromis niloticus*) by using solexa sequencing technology. *PloS one*, 9(1), e86821.

Xu, N., *et al.* (2009) MicroRNA-145 regulates OCT4, SOX2, and KLF4 and represses pluripotency in human embryonic stem cells. *Cell*, 137(4), 647-658.

Xu, P., *et al.* (2015) An improved protocol for small RNA library construction using high definition adapters. *Methods in next generation sequencing*, 2(1),

Ye, C., *et al.* (2018) APAtrap: identification and quantification of alternative polyadenylation sites from RNA-seq data. *Bioinformatics*, 34(11), 1841-1849.

Yekta, S., Shih, I.-h. and Bartel, D.P. (2004) MicroRNA-directed cleavage of HOXB8 mRNA. *Science (New York, N.Y.)*, 304(5670), 594-596.

Yi, R., *et al.* (2003) Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes & development*, 17(24), 3011-3016.

Zerbino, D.R., *et al.* (2018) Ensembl 2018. *Nucleic Acids Res*, 46(D1), D754-d761.

Zhang, C., *et al.* (2016) Primate-specific miR-603 is implicated in the risk and pathogenesis of Alzheimer's disease. *Aging (Albany NY)*, 8(2), 272.

Zhao, S., *et al.* (2014) Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PloS one*, 9(1), e78644.

Zhou, S., Mackay, T.F. and Anholt, R.R. (2014) Transcriptional and epigenetic responses to mating and aging in *Drosophila melanogaster*. *BMC genomics*, 15(1), 927.

Index

- 3'UTR reannotation, 112, 125, 187, 298, 310
- 3'UTR reporter assay, 40, 53, 73, 81, 298, 305
- affected isoform ratio, 135
- AGO. *See* argonaute
- AIC. *See* Akaike information criterion
- Akaike information criterion, 76, 305
- alternative polyadenylation, 97, 112, 172
- APA. *See* alternative polyadenylation
- argonaute, 34, 63, 82, 305
- bilateria, 30, 35, 39, 48, 306
- canonical. *See* seed
- ceRNA, 52, 70, 110, 293, 306
- chimeric RNA, 58, 77, 306
- CLASH, 40, 76, 307
- CLIP, 40, 58, 75, 298, 307, 309
- compensatory miRNA targeting, 64, 68, 306
- competitive endogenous RNA, 52, *See* ceRNA
- conservation, 60, 61, 68, 69, 72, 73, 128
- co-operative miRNA targeting, 52, 68, 266, 307
- DAG. *See* directed acyclic graph
- development, 35, 46, 47, 271, 291, 294
- DGCR8, 36, 307
- dicer, 34, 307
- directed acyclic graph, 113
- drosha, 36, 307
- ENA. *See* European nucleotide archive
- Ensembl, 125, 130, 161, 224
- European nucleotide archive, 123, 166
- evolution, 30, 35, 44, 47
- gene regulatory network, 102, 264, 300, 307
- HD adapter. *See* high definition adapter
- high definition adapters, 88, 276, 308
- microarray, 55, 76, 85, 265, 298, 306
- microprocessor complex, 36, 308
- miRBase, 42, 166, 200, 224
- mirGeneDB, 44
- miRNA mimic, 55, 69, 84, 172, 308
- mirtron, 37, 308
- nanopore sequencing, 301, 308
- non-canonical miRNA targeting, 72, 109, 229, 308
- Northern blotting, 93, 309
- ontology, 130, 309
- PMR. *See* post-mating response
- post-mating response, 219, 309
- precursor miRNA, 36, 41, 309
- primary miRNA*, 36, 309
- qPCR, 309, *See* quantitative PCR
- quantitative PCR, 55, 295
- ribonucleoprotein, 39, 58, 310
- RISC, 32, 50, 82
- RNA destabilisation, 45
- RNA interference*, 32, 39, 310
- RNAi. *See* RNA interference
- RNP. *See* ribonucleoprotein, *See* ribonucleoprotein
- seed*, 31, 48, 61, 73, 74, 76, 77, 82, 100, 203, 229, 255, 278
- sequence read archive, 123
- short interfering RNA, 39, 75, 310
- siRNA. *See* short interfering RNA
- snakemake, 112
- sRNA sequencing, 41, 87, 101, 216, 270
- supplementary miRNA targeting, 64, 73, 311
- support-vector machine, 75, 311
- target-directed miRNA degradation, 52, 311
- translational inhibition, 48, 55, 306

Appendix A

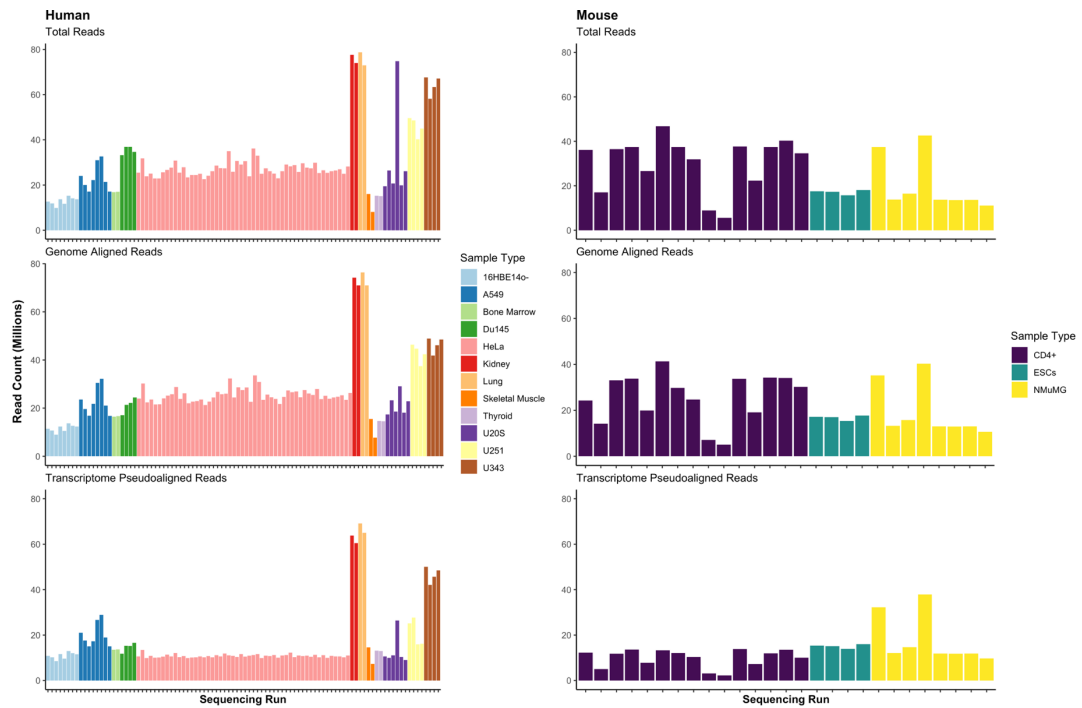


Figure A.1 – The total number of reads, as well as the number of aligned and pseudoaligned reads for all datasets analysed as part of chapter 4 of this thesis. Left: Human samples. Right: Mouse samples.

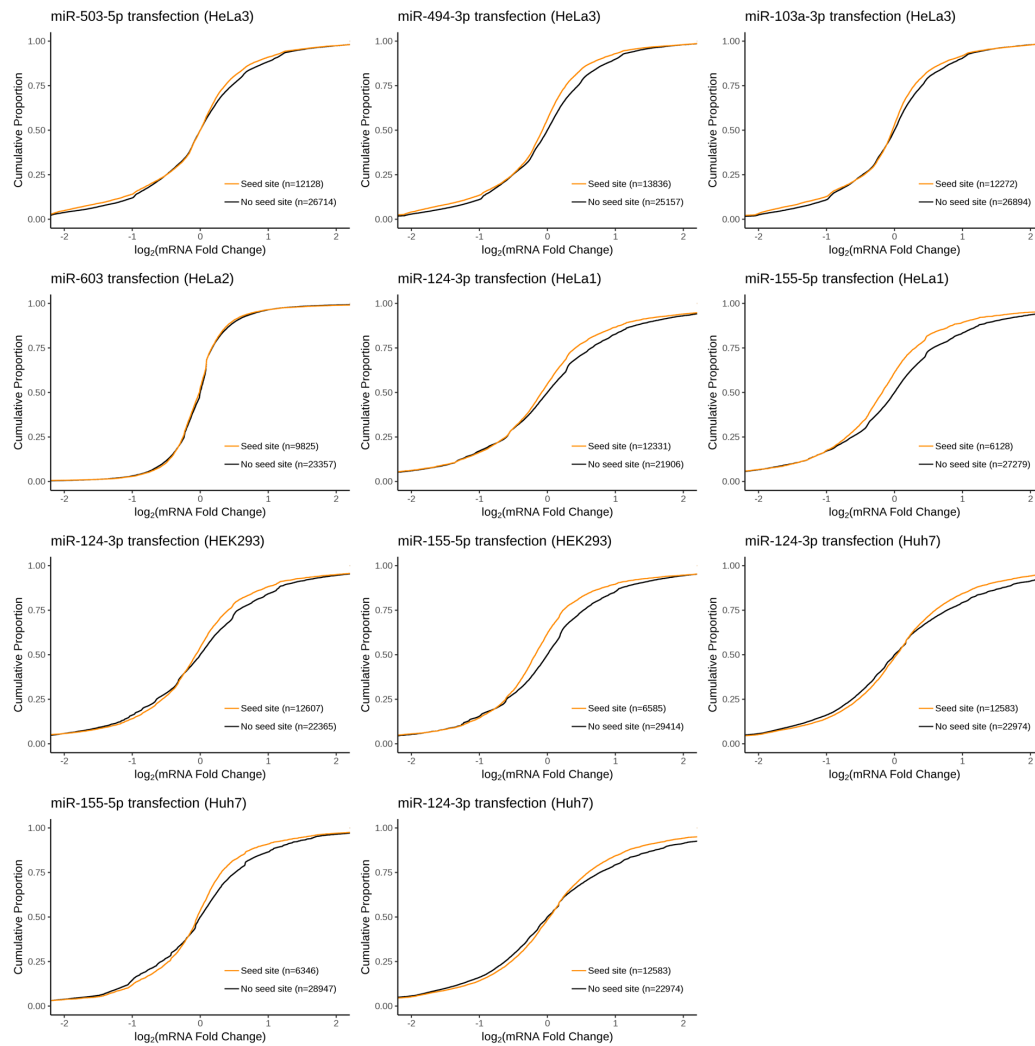


Figure A.2 – Preliminary analyses conducted on datasets, which were judged unsuitable for further analysis, due to extensive similarity between predicted target and non-target distributions, indicating a potential failure in transfection experiments used to generate the data. As in figure 4.1, with the exception that ‘filtered seed site’ distributions (light green) are not shown.

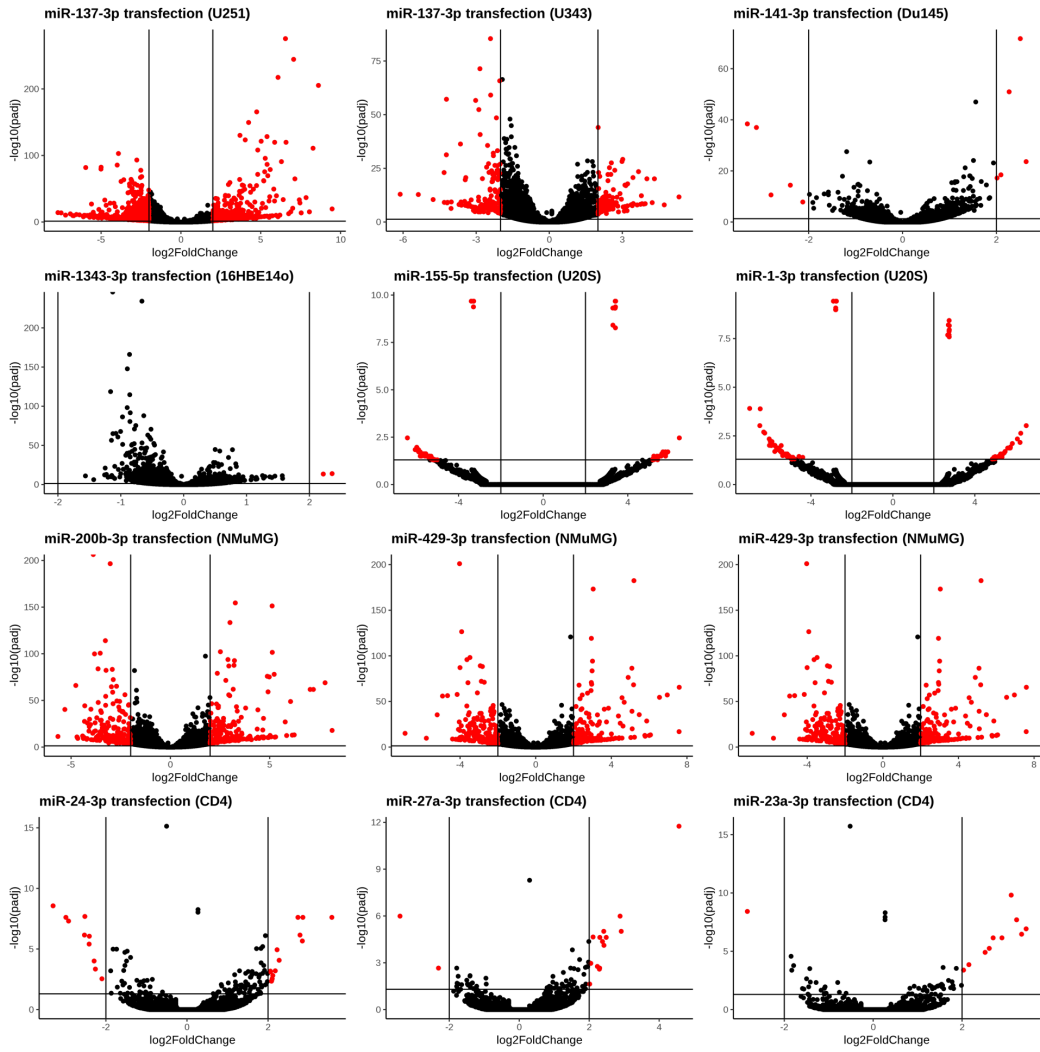
Accession	Frag. Length	% Kal-listo Aligned	% HISAT2 Aligned	% Trimmed	% Dups	% GC	Length	M Seqs
ERR030888	NA	83.20	96.60	4.1	44.4	47	73	74.6
ERR030893	NA	82.20	95.60	4.5	24.4	45	73	77.6
ERR030896	NA	87.90	97.00	4.4	47.3	46	74	78.7
SRR1047622	NA	50.70	93.40	0.9	51.5	46	36	49.6
SRR1047623	NA	57.00	92.00	1	48.4	45	36	48.6
SRR1047624	NA	39.50	92.80	0.8	45	46	36	40.3
SRR1047625	NA	35.90	94.20	0.9	53.6	48	36	45
SRR1047630	NA	74.00	72.30	0.6	59	51	36	67.6
SRR1047631	NA	72.30	71.90	0.4	59.6	51	36	58.2
SRR1047632	NA	72.00	72.70	0.6	60.9	52	36	63.4
SRR1047633	NA	72.20	72.30	0.7	59.1	51	36	67.1
SRR1598955	NA	54.30	89.20	10	44.4	48	36	19.5
SRR1598970	NA	37.90	87.90	17.9	77.8	49	36	26.4
SRR1598972	NA	53.70	89.70	10.9	47.9	48	36	20.7
SRR1598973	NA	35.30	38.90	23.2	79.2	50	49	74.8
SRR1598976	NA	52.20	91.00	9.9	59.2	48	36	19.9
SRR1598977	NA	34.60	87.60	16.9	79.7	49	36	26.1
SRR8382192	NA	41.50	94.00	0.6	36	47	50	25.5
SRR8382193	NA	42.40	94.90	0.6	40.8	46	50	31.8
SRR8382194	NA	41.50	93.70	0.8	33.9	47	50	23.9
SRR8382195	NA	43.30	94.00	1.4	40.5	47	50	25.1
SRR8382196	NA	44.00	94.00	0.5	34.4	46	50	22.9
SRR8382197	NA	44.20	94.20	5	55.5	47	50	22.9
SRR8382198	NA	41.00	93.90	0.6	33.1	46	50	25.6
SRR8382199	NA	42.60	94.30	0.8	36.6	47	50	26.7
SRR8382200	NA	38.40	93.10	1.8	41.3	46	50	27.7
SRR8382201	NA	39.40	93.50	1	43	46	50	30.8
SRR8382202	NA	40.50	93.50	0.5	35	46	50	25.5
SRR8382203	NA	38.50	93.80	0.9	39.4	46	50	27.9
SRR8382204	NA	42.40	94.00	0.4	33.8	46	50	23.4
SRR8382205	NA	41.40	92.90	1	49.6	48	50	24.4
SRR8382206	NA	41.90	93.90	0.4	34.4	46	50	24.4
SRR8382207	NA	42.30	94.30	0.7	39	46	50	25
SRR8382208	NA	45.00	94.20	0.4	35	47	50	22.6
SRR8382209	NA	44.60	94.20	0.7	38	48	50	24.1

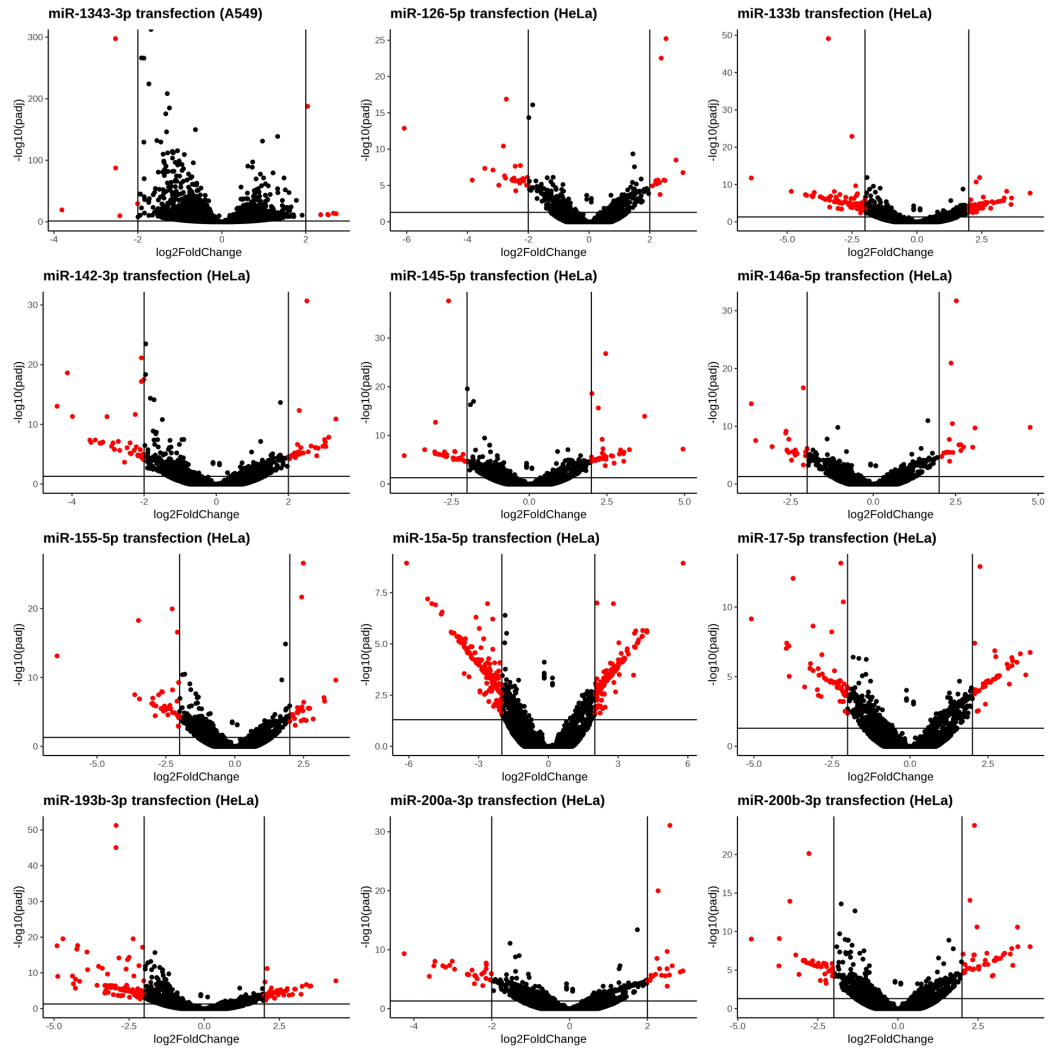
SRR8382210	NA	39.20	93.60	0.6	34.1	45	50	26.1
SRR8382211	NA	39.00	93.60	1.2	40.9	46	50	28.6
SRR8382212	NA	38.70	93.70	0.5	34	46	50	27.5
SRR8382213	NA	43.30	94.90	0.4	37.2	46	50	27.4
SRR8382214	NA	31.70	92.40	0.6	32.8	46	50	35
SRR8382215	NA	41.80	94.40	1.2	37.3	46	50	25.9
SRR8382216	NA	33.90	93.30	1.2	32.2	45	50	30.7
SRR8382217	NA	40.10	94.60	0.7	36.2	45	50	29.1
SRR8382218	NA	34.40	93.20	0.5	31.7	45	50	30.6
SRR8382219	NA	45.70	94.60	0.8	37.8	47	50	23.9
SRR8382220	NA	31.00	92.70	0.7	36.3	45	50	36.2
SRR8382221	NA	35.30	93.70	0.5	37.6	45	50	33
SRR8382222	NA	39.60	93.70	0.4	34.2	46	50	25
SRR8382223	NA	39.80	93.50	2.5	41	47	50	27.4
SRR8382224	NA	41.40	94.00	0.5	33.8	46	50	26.1
SRR8382225	NA	44.70	95.00	1	37.8	46	50	25.1
SRR8382226	NA	43.80	94.50	0.8	36.4	47	50	23
SRR8382227	NA	42.10	94.50	0.4	34.4	46	50	26.1
SRR8382228	NA	38.70	94.00	0.5	32.3	46	50	29.1
SRR8382229	NA	43.40	94.10	1.6	48.7	46	50	28.3
SRR8382230	NA	35.70	93.50	0.6	31.2	46	50	28.8
SRR8382231	NA	42.70	94.90	0.9	36	46	50	25.8
SRR8382232	NA	36.30	92.80	1.6	34	46	50	29.6
SRR8382233	NA	39.40	94.10	1.1	40.6	45	50	27.7
SRR8382234	NA	38.10	93.10	1.6	34.7	46	50	27.4
SRR8382235	NA	38.00	93.50	1	41	45	50	29.9
SRR8382236	NA	40.30	93.90	0.6	33.3	46	50	25.3
SRR8382237	NA	42.00	94.80	0.6	37.5	45	50	26.5
SRR8382238	NA	39.40	94.00	0.5	35.6	45	50	25.4
SRR8382239	NA	41.60	93.70	1.2	44	46	50	26.1
SRR8382240	NA	40.20	93.60	0.5	37	46	50	26.4
SRR8382241	NA	39.40	93.70	1.1	41.3	47	50	27
SRR8382242	NA	40.80	93.80	0.5	35	46	50	25
SRR8382243	NA	39.00	93.30	1.4	43.7	47	50	28.2
SRR3112237	NA	33.90	67.30	1.7	39.1	45	51	36.1
SRR3112238	NA	29.70	83.40	1.1	25.5	41	51	17
SRR3112239	NA	32.30	90.70	1.1	35.3	43	51	36.5
SRR3112240	NA	36.40	90.40	1	22.9	42	51	37.4
SRR3112241	NA	29.40	74.80	1.3	36.4	39	51	26.6

SRR3112242	NA	28.50	88.40	1.1	40.3	40	51	46.8
SRR3112243	NA	32.30	79.70	1.4	37.7	45	51	37.4
SRR3112244	NA	32.60	77.50	1.3	30.5	44	51	31.9
SRR3112245	NA	35.20	80.30	1.2	17.2	43	51	8.9
SRR3112246	NA	40.80	90.50	1	17.4	42	51	5.6
SRR3112247	NA	36.80	89.30	1.1	53.1	43	51	37.7
SRR3112248	NA	32.50	85.80	1.2	29	40	51	22.3
SRR3112250	NA	32.00	91.60	1	44.3	40	51	37.4
SRR3112251	NA	33.70	84.60	1.1	55.7	40	51	40.3
SRR3112252	NA	29.00	87.50	1.1	39.6	39	51	34.6
SRR4054984	NA	86.20	94.20	0.7	57.7	48	51	37.4
SRR4054985	NA	87.90	96.20	0.4	48.6	49	51	13.8
SRR4054992	NA	88.70	95.50	0.4	53	49	51	16.5
SRR4054995	NA	88.90	94.70	0.9	61.7	49	51	42.6
SRR4054996	NA	86.90	95.20	0.8	47.4	48	51	13.7
SRR4054999	NA	87.60	95.80	0.4	47.7	49	51	13.5
SRR4055002	NA	87.60	95.80	0.4	47.9	49	51	13.6
SRR4055005	NA	87.80	95.80	0.4	46.5	49	51	11.1
ERR030879	173.1	89.00	97.20	4.75	54.9	46	49.5	73
ERR030880	255.2	82.80	96.70	3.55	56.2	47	49.5	71.9
ERR030885	212.2	81.70	95.90	4.05	54.4	45	49.5	74
ERR315358	322.8	86.30	96.30	3.55	38.35	45	99	15.2
ERR315404	204.7	80.80	98.20	4	39.9	48	97.5	16.8
ERR315406	205.2	80.80	98.20	4.05	40.4	48	97.5	17
ERR315422	323.3	86.30	96.30	3.4	38.65	45	99	15.1
ERR579142	120.8	90.90	96.70	47.75	53.5	50	89.5	16
ERR579143	125.6	91.00	95.70	40.95	45	49	89.5	8.1
SRR2146408	179.6	35.60	51.40	3.05	66.5	44.5	75	33.2
SRR2146409	176.8	41.40	57.90	2.7	63.6	45	75	36.9
SRR2146410	171.7	41.20	60.00	2.6	63.9	44.5	75	36.9
SRR2146411	171.6	47.80	70.30	2.8	59.85	45	75	34.7
SRR2968576	179.8	87.60	98.30	1.55	46.4	49	48	24
SRR2968577	173.4	88.10	98.30	1.85	44.7	49	48	20
SRR2968578	170.8	88.20	98.30	1.7	42.65	49	48	17.1
SRR2968579	182.4	77.90	98.30	1.65	50.2	50	48	22.2
SRR2968580	180.8	86.00	98.40	1.3	50.65	49	48	31
SRR2968581	178.9	88.30	98.30	1.4	49	49	48	32.7
SRR2968582	184.8	88.70	98.30	1.4	44.05	49	48	21.4
SRR2968583	184.9	87.90	98.20	2.3	40.4	48	48	17.1

SRR2968584	225.0	85.70	90.00	8.55	29.25	49	47.5	12.7
SRR2968586	213.3	86.10	90.10	8.95	29.15	49	47.5	11.9
SRR2968588	206.5	86.30	90.30	7.75	27.15	49	47.5	9.9
SRR2968590	195.7	85.20	90.00	8.6	30.85	49	47.5	13.7
SRR2968592	202.5	82.80	90.10	8.75	29.25	50	47.5	11.7
SRR2968594	200.9	85.30	90.00	8.55	30.55	49	47.5	15.2
SRR2968596	209.6	85.30	89.80	9	29.15	49	47.5	14.1
SRR2968598	191.9	84.20	89.90	9.3	30.35	49	47.5	13.7
SRR1734389	182.4	88.10	98.40	3.1	46.3	49.5	98.5	17.5
SRR1734391	184.1	87.30	98.20	2.95	45.45	48	98.5	17.3
SRR1734393	188.6	88.90	98.00	2.85	48.1	49.5	98.5	15.7
SRR1734395	182.8	88.60	98.00	3.1	50.5	49	98.5	18.1

Table A.1 - A table of summary and quality control statistics for all sequencing runs used in this analysis. Statistics are given for the percentage of sequenced pseudo-aligned to the transcript aligned to the transcriptome using either kallisto. The estimated mean average fragment length is given for sequencing runs in which cDNA libraries are sequences using paired-end sequencing protocols. For single-end sequencing protocols, fragment length statistics cannot be inferred. The percentage of reads aligned to the relevant genome using HISAT2 (for the purposes of 3'UTR reannotation) is also given. QC statistics such as the percentage of reads trimmed, the percentage of reads which are duplicates, and the mean percentage GC content of reads is also reported. In the final two columns, the length of reads, and the number of reads sequenced for each sequencing run is given.





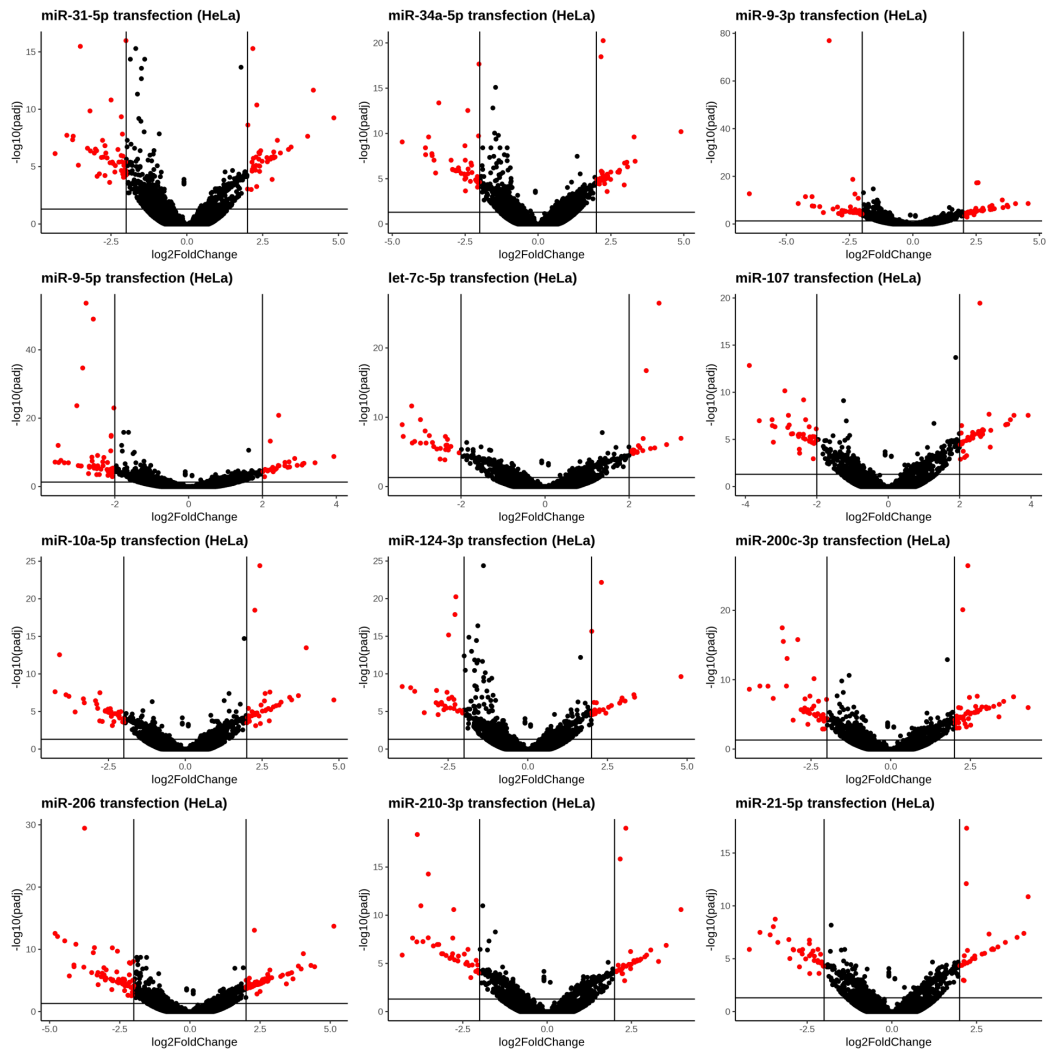
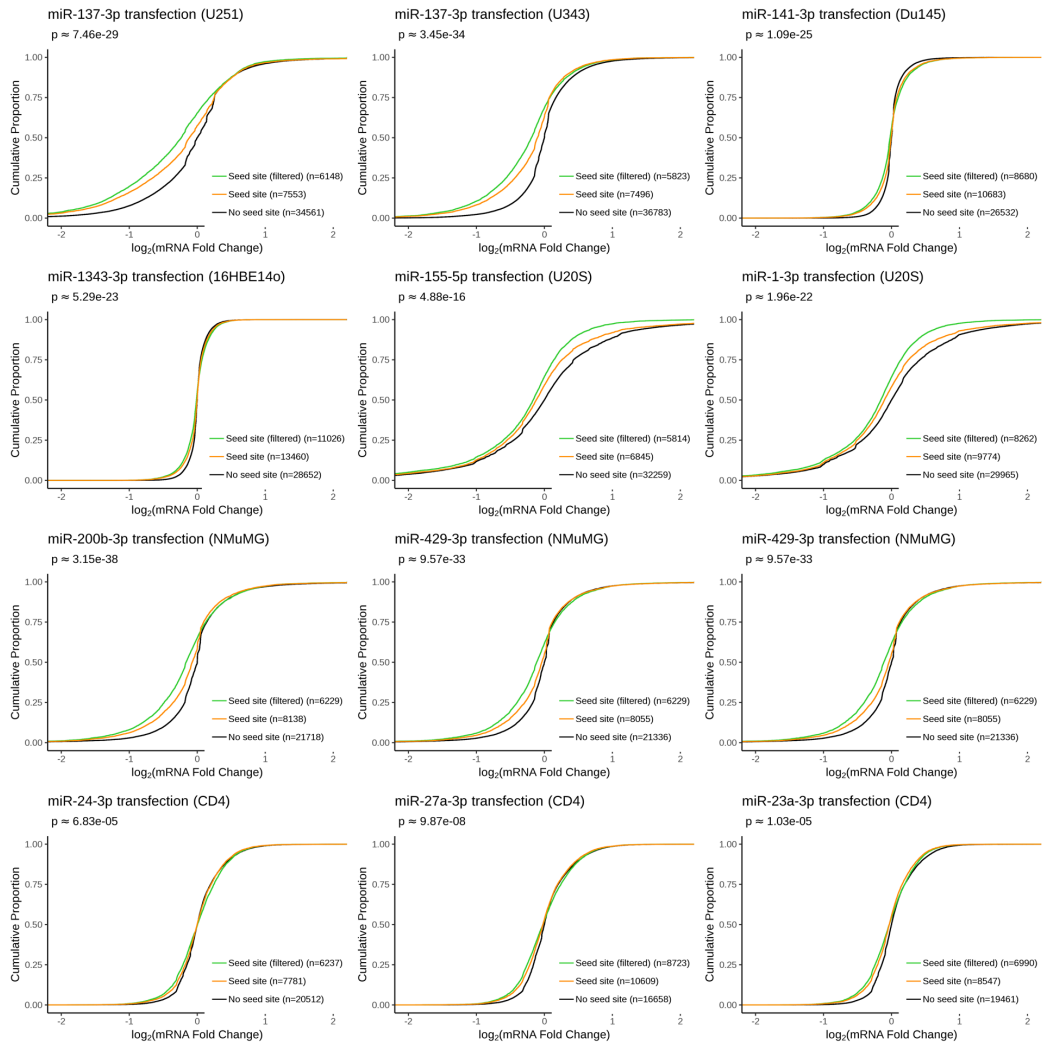
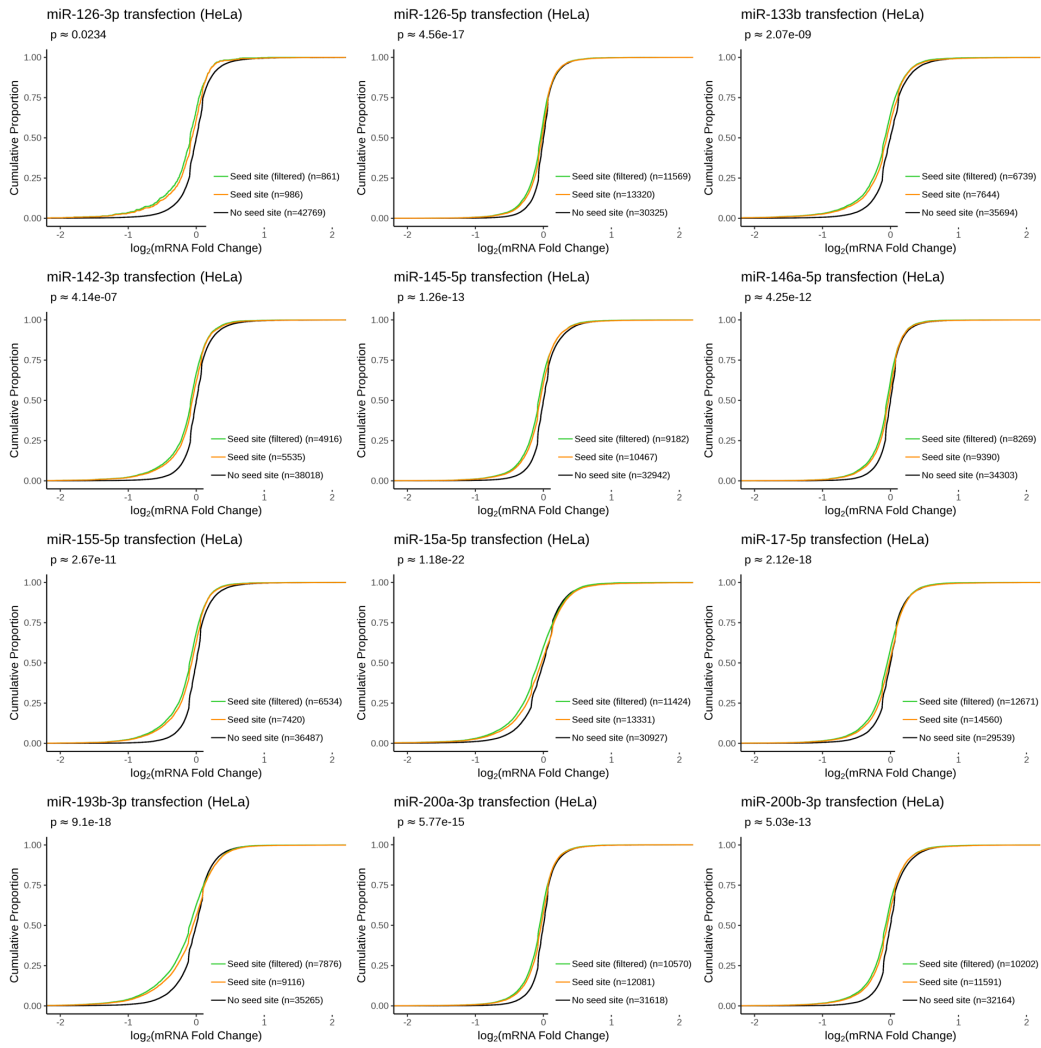


Figure A.3 – Volcano plots for all RNA-Seq transfection datasets analysed as part of chapter 4 of this thesis. Along the x-axis is \log_2 fold change (with shrinkage) as computed by the DESeq2 package. Along the y-axis are p-values associated with the log fold change parameters within a generalised linear model as computed using a Wald test within DESeq2. Differentially expressed transcripts are denoted in red ($|LFC| > 2$; adjusted p-value < 0.05), whilst non-differentially expressed transcript are denoted in black.





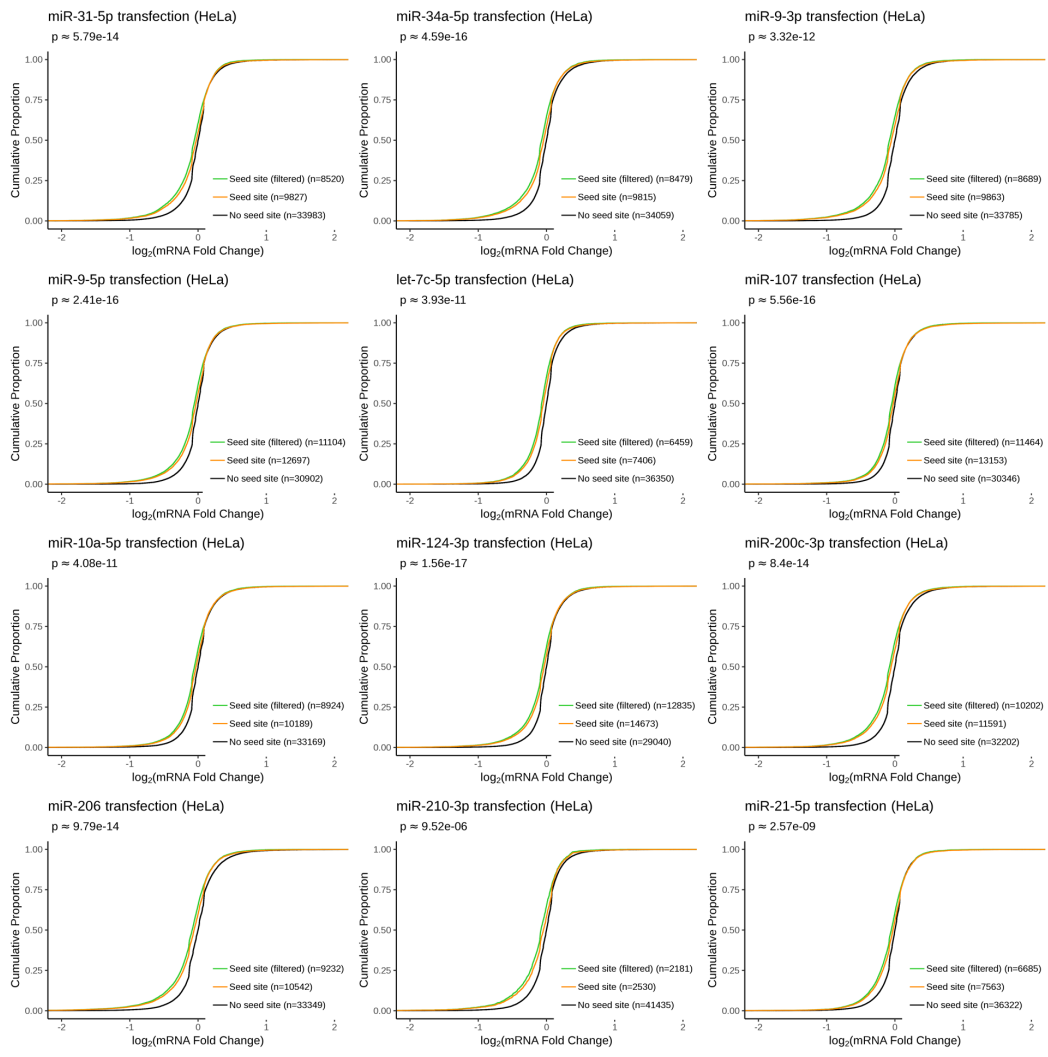


Figure A.4 - As in figure 4.1, though with a greater number of datasets analysed

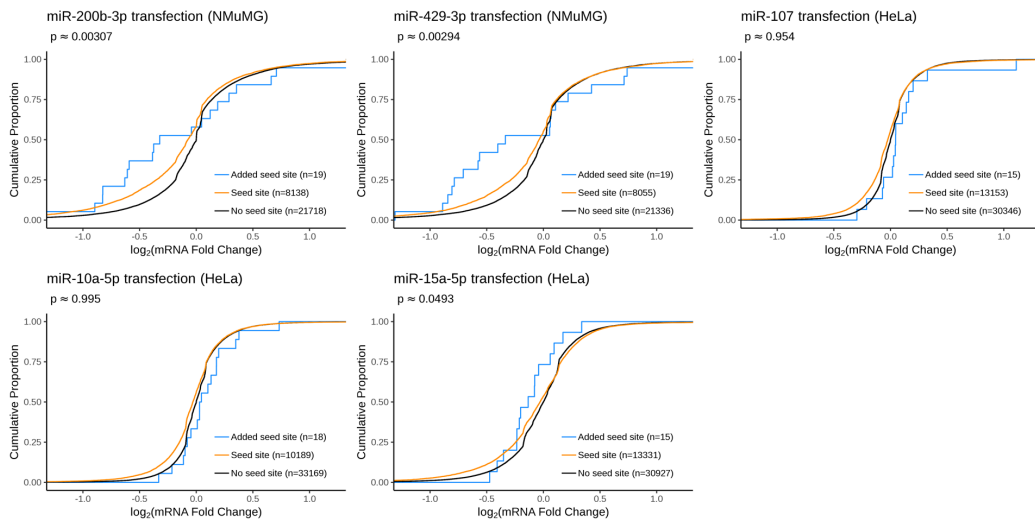
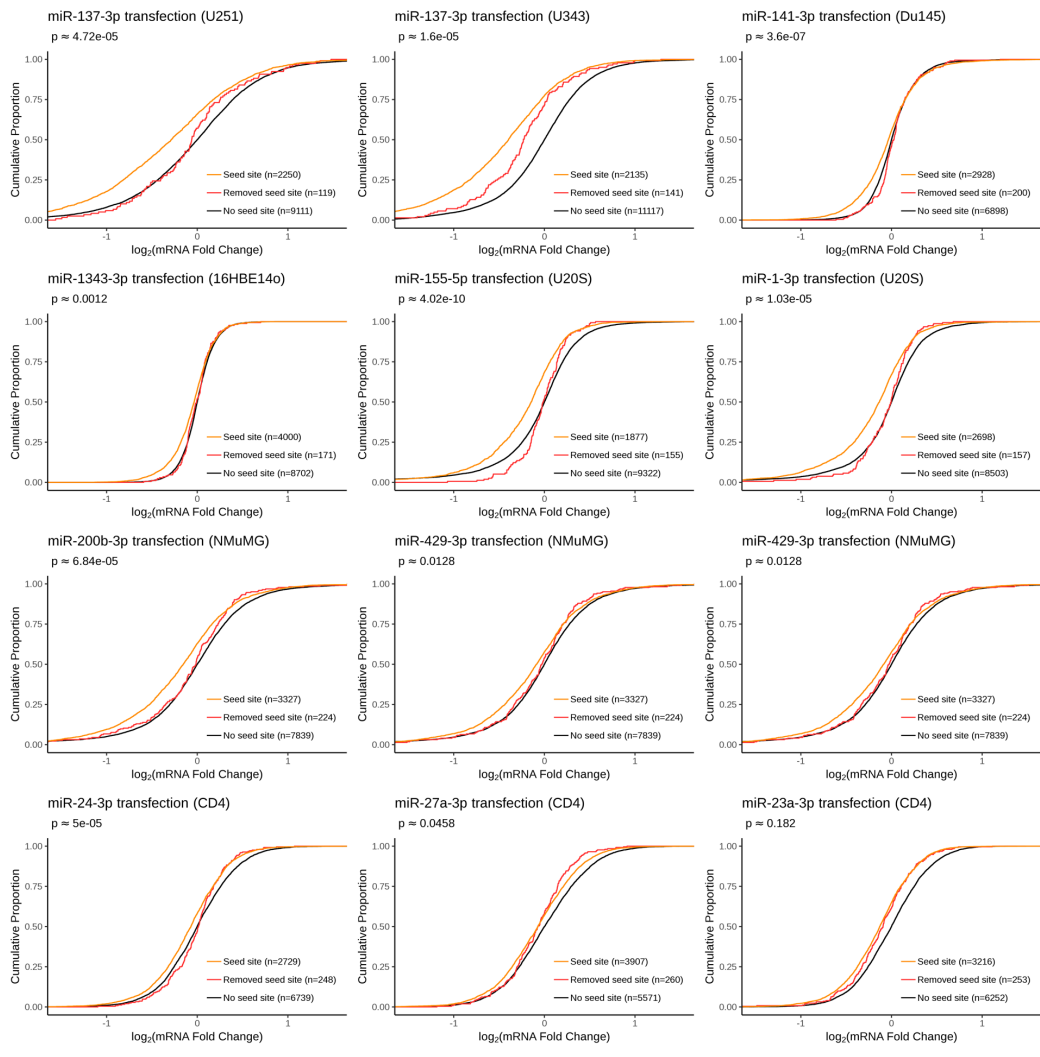
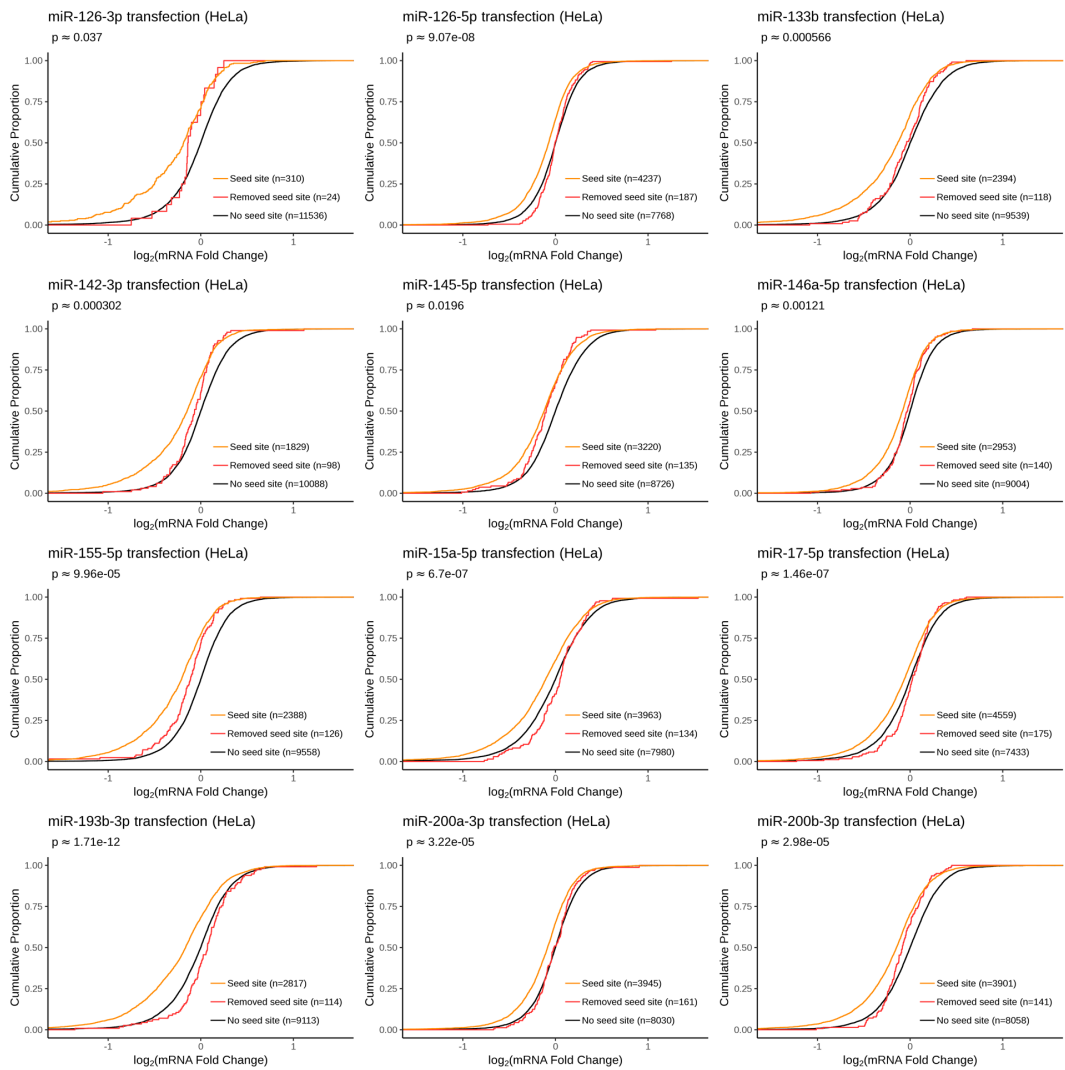


Figure A.5 – As in figure 4.5, though with a greater number of datasets analysed.





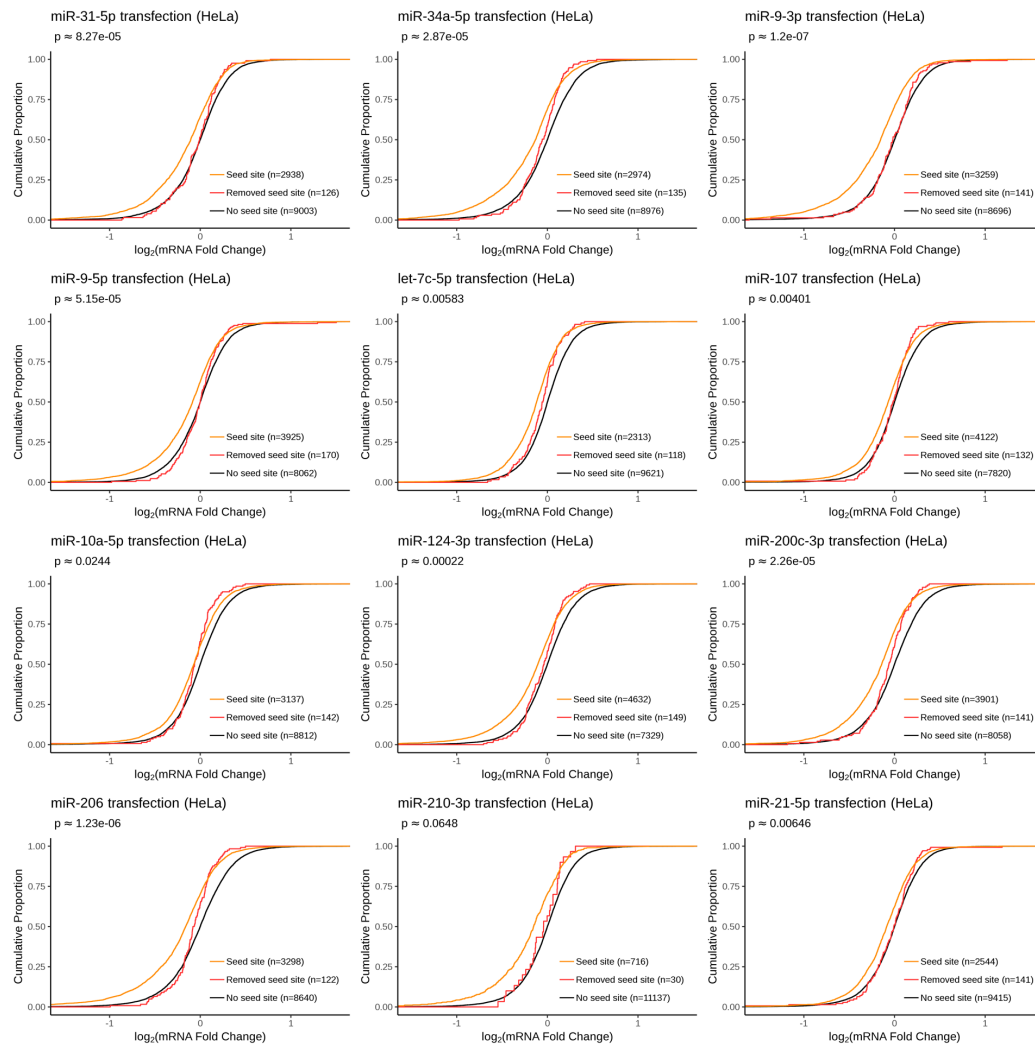


Figure A.6 – As in figure 4.9, though with more datasets analysed.

Species	BioProject Accession	Source/Study	Sample	Run Accessions
Humans	PRJNA231155	Tamim <i>et al.</i> 2014	U251	SRR1047622,SRR1047623, SRR1047624,SRR1047625
			U343	SRR1047630,SRR1047631, SRR1047632,SRR1047633
	PRJNA292016	Liu <i>et al.</i> 2017	Du145	SRR2146408,SRR2146409, SRR2146410,SRR2146411
	PRJNA304643	Stolzenburg <i>et al.</i> 2016	A549	SRR2968576,SRR2968577, SRR2968578,SRR2968579, SRR2968580,SRR2968581, SRR2968582,SRR2968583
			16HBE14o-	SRR2968584,SRR2968586, SRR2968588,SRR2968590, SRR2968592,SRR2968594, SRR2968596,SRR2968598
	PRJNA512378	Liu <i>et al.</i> 2019	HeLa	SRR8382192,SRR8382193, SRR8382194,SRR8382195, SRR8382196,SRR8382197, SRR8382198,SRR8382199, SRR8382200,SRR8382201, SRR8382202,SRR8382203 SRR8382204,SRR8382205, SRR8382206,SRR8382207, SRR8382208,SRR8382209, SRR8382210,SRR8382211, SRR8382212,SRR8382213, SRR8382214,SRR8382215 SRR8382216,SRR8382217, SRR8382218,SRR8382219, SRR8382220,SRR8382221, SRR8382222,SRR8382223, SRR8382224,SRR8382225, SRR8382226,SRR8382227, SRR8382228,SRR8382229,

				SRR8382230,SRR8382231, SRR8382232,SRR8382233, SRR8382234,SRR8382235, SRR8382236,SRR8382237, SRR8382238,SRR8382239 SRR8382240,SRR8382241, SRR8382242,SRR8382243
	PRJNA223608	Guo <i>et al.</i> 2014	U2OS	SRR1598955,SRR1598970, SRR1598976,SRR1598977, SRR1598972,SRR1598973
	PRJEB2445	Illumina BodyMap2 transcriptome	Kidney	ERR030885,ERR030893
			Lung	ERR030879,ERR030896
	PRJEB6971	Science for Life Laboratory, Stockholm	Skeletal Muscle	ERR579142,ERR579143
			Thyroid	ERR315358,ERR315422
			Bone Marrow	ERR315404,ERR315406
Mouse	PRJNA340017	Diepenbruck <i>et al.</i> 2017	NMuMG	SRR4054984,SRR4054985, SRR4054992,SRR4054995, SRR4054996,SRR4054999, SRR4055002,SRR4055005
	PRJNA309441	Pua <i>et al.</i> 2016	CD4+	SRR3112249,SRR3112250, SRR3112251,SRR3112252, SRR3112245,SRR3112246, SRR3112247,SRR3112248, SRR3112237,SRR3112238, SRR3112239,SRR3112240 SRR3112241,SRR3112242, SRR3112243,SRR3112244
	PRJNA270999	Cao <i>et al.</i> 2015	ESCs	SRR1734389,SRR1734391, SRR1734393,SRR1734395

Table A.2 - A summary of RNA-sequencing datasets analysed for chapter 4 of this thesis. The project accession, run accession, data source, as well as the biological context of each experiment is given for each dataset analysed

Species	BioProject Accession	Source/Study	Sample	Run Accessions
Human	PRJNA229375	Nam <i>et al.</i> 2014	HeLa1	SRR1032873, SRR1032874, SRR1032875, SRR1032876, SRR1032877, SRR1032878,
			HEK293	SRR1032879, SRR1032880, SRR1032881, SRR1032882 SRR1032883, SRR1032884
			Huh7	SRR1032885, SRR1032886, SRR1032887, SRR1032888 SRR1032890, SRR1032891, SRR1032892
			IMR90	SRR1032893, SRR1032894, SRR1032895, SRR1032896
	PRJNA284262	Zhang <i>et al.</i> 2016	HeLa2	SRR2031925, SRR2031926, SRR2031927, SRR2031928
	PRJNA271411	Iyer <i>et al.</i> 2015	HeLa3	SRR1737410, SRR1737413, SRR1737415, SRR1737416, SRR1737420, SRR1737421, SRR1737429, SRR1737430

Table A.3 - A summary of data considered during preliminary analysis, but were not used for further analysis. See figure A.1 for analyses of these datasets.

Sample	miRNA	Average LFC deviation (targets)	Average LFC deviation (non-targets)	SNR	1/SNR
U251	miR-137-3p	0.940	0.739	1.272	0.786
U343	miR-137-3p	0.609	0.455	1.338	0.747
Du145	miR-141-3p	0.264	0.185	1.428	0.700
A549	miR-1343-3p	0.320	0.251	1.276	0.784
16HBE14o-	miR-1343-3p	0.179	0.141	1.269	0.788
HeLa	let-7c-5p	0.278	0.235	1.184	0.845
HeLa	miR-107	0.307	0.240	1.281	0.781
HeLa	miR-10a-5p	0.322	0.266	1.210	0.827
HeLa	miR-124-3p	0.331	0.248	1.338	0.747
HeLa	miR-126-3p	0.398	0.298	1.336	0.748
HeLa	miR-126-5p	0.274	0.212	1.292	0.774
HeLa	miR-133b	0.432	0.310	1.394	0.717
HeLa	miR-142-3p	0.366	0.263	1.388	0.720
HeLa	miR-145-5p	0.325	0.258	1.257	0.796
HeLa	miR-146a-5p	0.286	0.222	1.285	0.778
HeLa	miR-155-5p	0.372	0.254	1.466	0.682
HeLa	miR-15a-5p	0.479	0.381	1.259	0.795
HeLa	miR-16-5p	0.372	0.279	1.336	0.749
HeLa	miR-17-5p	0.355	0.283	1.254	0.797
HeLa	miR-193b-3p	0.458	0.324	1.416	0.706
HeLa	miR-200a-3p	0.288	0.226	1.275	0.784
HeLa	miR-200b-3p	0.348	0.268	1.296	0.771
HeLa	miR-200c-3p	0.357	0.274	1.300	0.769
HeLa	miR-206	0.406	0.310	1.308	0.764
HeLa	miR-210-3p	0.366	0.268	1.364	0.733
HeLa	miR-21-5p	0.320	0.258	1.238	0.808
HeLa	miR-31-5p	0.352	0.280	1.258	0.795
HeLa	miR-34a-5p	0.362	0.278	1.302	0.768
HeLa	miR-9-3p	0.384	0.271	1.419	0.705
HeLa	miR-9-5p	0.337	0.254	1.328	0.753
U20S	miR-1-3p	0.972	0.984	0.988	1.012
U20S	miR-155-5p	1.089	1.076	1.012	0.988
NMuMG	miR-1199-5p	0.727	0.529	1.374	0.728
CD4+	miR-23a-3p	0.330	0.280	1.178	0.849
CD4+	miR-24-3p	0.349	0.292	1.194	0.838

CD4+	miR-27a-3p	0.368	0.296	1.242	0.805
ESCs	miR-294-3p	0.518	0.345	1.504	0.665

Table A.4 - An assessment of the signal-noise ratio in each miRNA mimic transfection experiment. ‘Average LFC deviation’ represents the mean average distance of log fold change values from 0. The signal-noise ratio (SNR) is calculated using the formula given in the background chapter of this thesis.

Appendix B

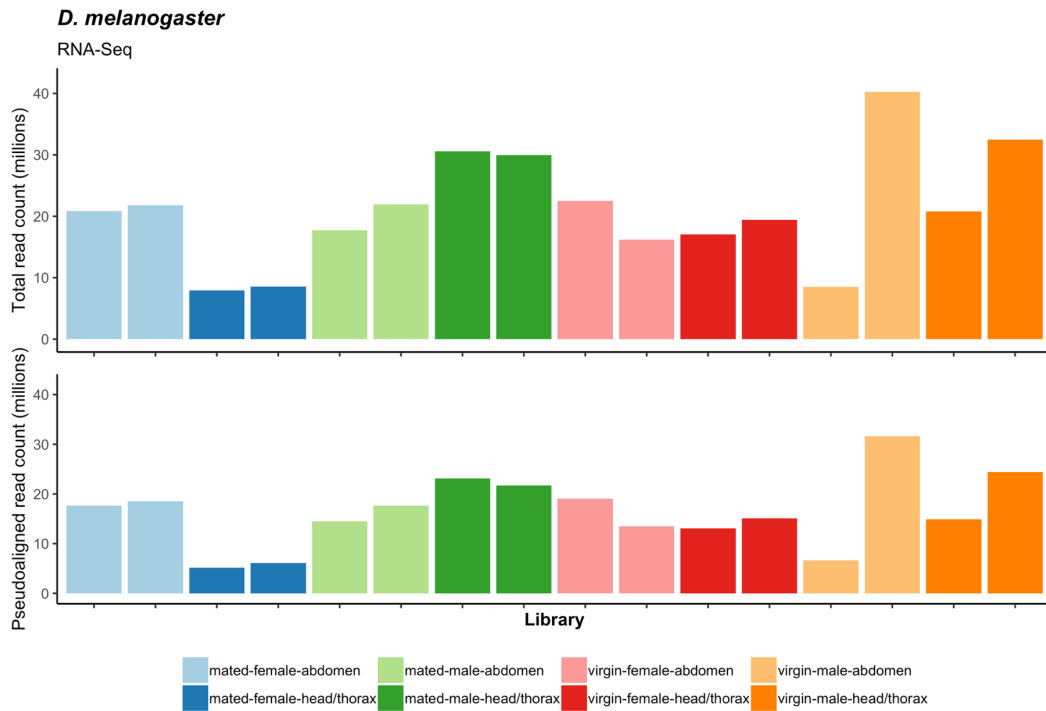


Figure B.1 – mRNA sequencing depth: Total (top) and pseudoaligned (bottom) read counts for mRNA sequencing libraries are given. Bars are colour coded according sample type.

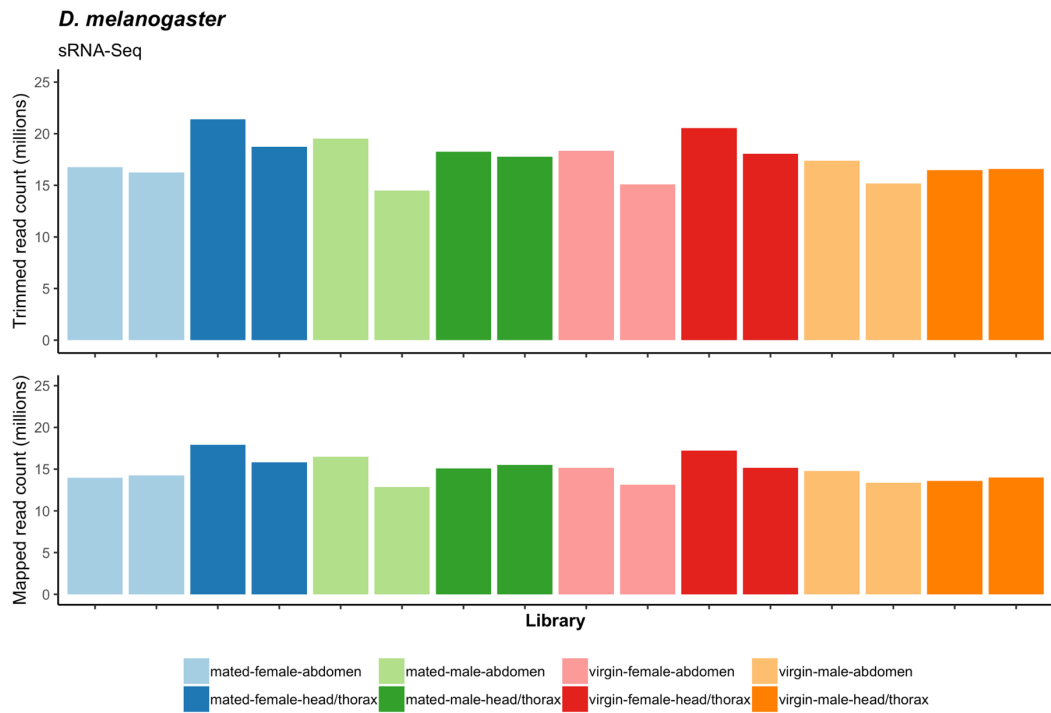


Figure B.2 – sRNA sequencing depth: Total (top) and mapped (bottom) reads counts for sRNA sequencing libraries are given. Bars are colour coded according to sample type.

Sample	Total reads	Mapped reads	Fraction mapped	Treatment	Sex	Body Part	Replicate
LIB21920	7936298	5129559	0.6463	mated	female	head/thorax	1
LIB21921	20841021	17656661	0.8472	mated	female	abdomen	1
LIB21922	17034527	13080104	0.7679	virgin	female	head/thorax	1
LIB21923	22483948	19088525	0.849	virgin	female	abdomen	1
LIB21924	30566395	23114969	0.7562	mated	male	head/thorax	1
LIB21925	17731066	14466364	0.8159	mated	male	abdomen	1
LIB21926	20795634	14905963	0.7168	virgin	male	head/thorax	1
LIB21927	8493430	6597732	0.7768	virgin	male	abdomen	1
LIB21928	8547418	6077181	0.711	mated	female	head/thorax	2
LIB21929	21781925	18536186	0.851	mated	female	abdomen	2
LIB21930	19403600	15091640	0.7778	virgin	female	head/thorax	2
LIB21931	16185107	13493870	0.8337	virgin	female	abdomen	2
LIB21932	29955574	21711602	0.7248	mated	male	head/thorax	2
LIB21933	21923237	17634202	0.8044	mated	male	abdomen	2
LIB21934	32456773	24419658	0.7524	virgin	male	head/thorax	2
LIB21935	40234601	31631526	0.7862	virgin	male	abdomen	2

Table B.1 - mRNA sequencing depth metrics and values, along with library metadata

Sample	Total reads after trimming	Genome mapping reads (perfect match)	% trimmed reads mapping	Treatment	Sex	Body Part	Replicate
LIB28804	21379806	17932126	83.8741287	mated	female	head/thorax	2
LIB28805	16763295	13950560	83.22087036	mated	female	abdomen	2
LIB28806	20535585	17213198	83.82131797	virgin	female	head/thorax	2
LIB28807	18334102	15158174	82.67748265	virgin	female	abdomen	2
LIB28808	18264596	15100638	82.67709836	mated	male	head/thorax	2
LIB28809	19516527	16494441	84.515247	mated	male	abdomen	2
LIB28810	16479791	13594843	82.49402556	virgin	male	head/thorax	2
LIB28811	17385491	14760849	84.90326215	virgin	male	abdomen	2
LIB28812	18743335	15811711	84.35911219	mated	female	head/thorax	1
LIB28813	16262939	14258152	87.67266482	mated	female	abdomen	1
LIB28814	18062535	15149644	83.87329907	virgin	female	head/thorax	1
LIB28815	15088226	13121058	86.96223135	virgin	female	abdomen	1
LIB28816	17779587	15521647	87.30037992	mated	male	head/thorax	1
LIB28817	14499606	12850728	88.62811858	mated	male	abdomen	1
LIB28818	16594161	14016780	84.46814515	virgin	male	head/thorax	1
LIB28819	15178474	13358604	88.01019127	virgin	male	abdomen	1

Table B.2 - sRNA sequencing depth metrics and values, along with library metadata

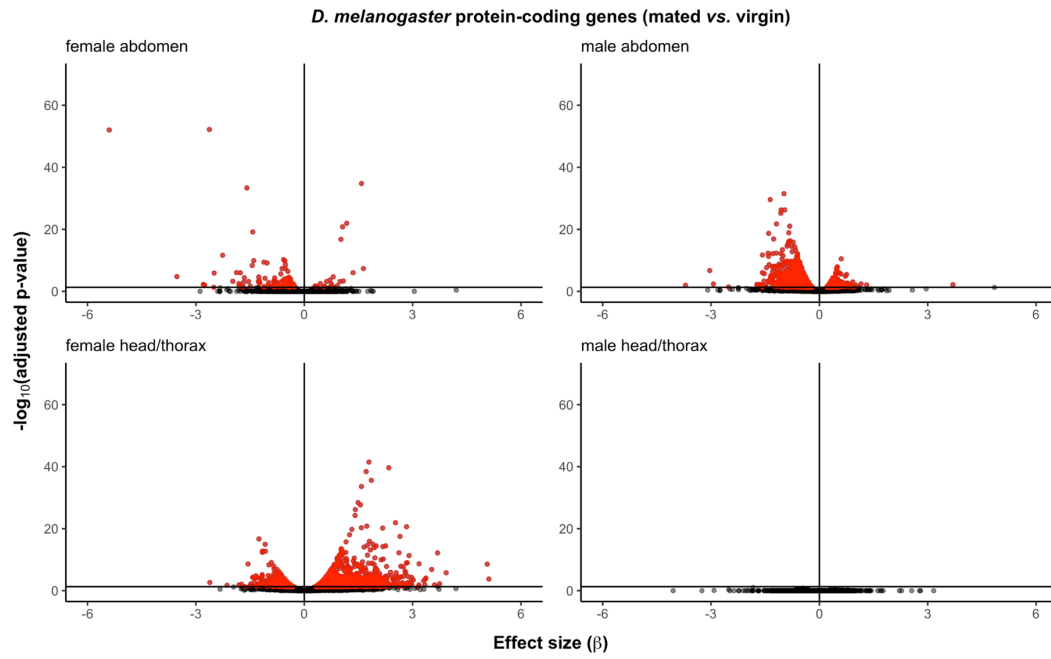


Figure B.3 – A volcano plot for protein-coding genes for the differential expression analysis presented in chapter 5 of this thesis. Colour legend: Red – differentially expressed genes; black – non-differentially expressed genes.

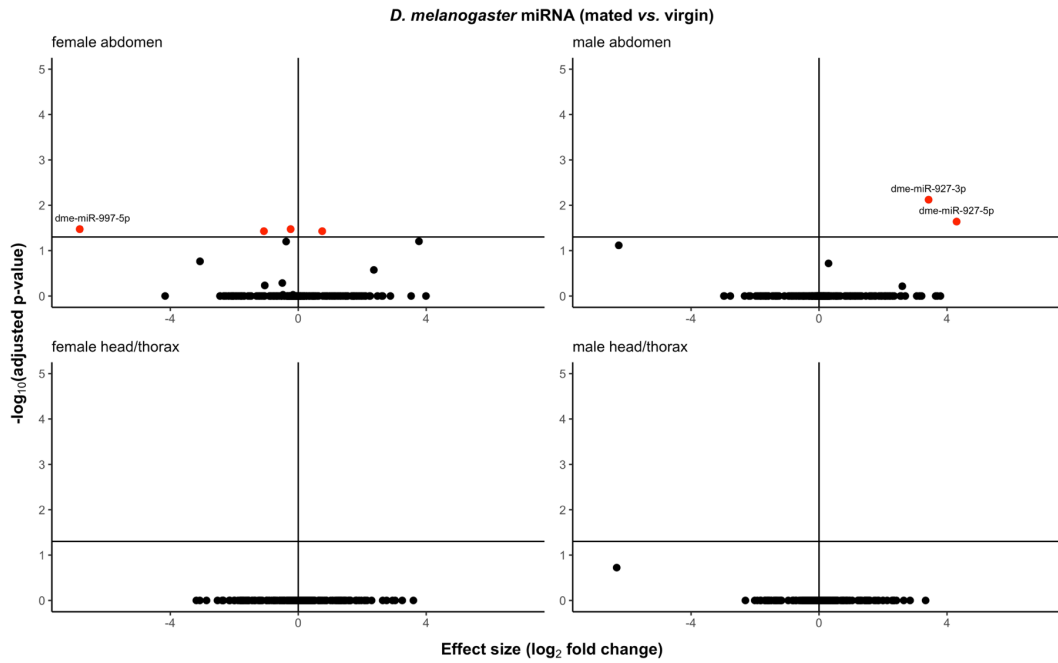


Figure B.4 – A volcano plot for miRNA for the differential expression analysis presented in chapter 5 of this thesis. Colour legend: Red – differentially expressed miRNAs; black – non-differentially expressed miRNAs.

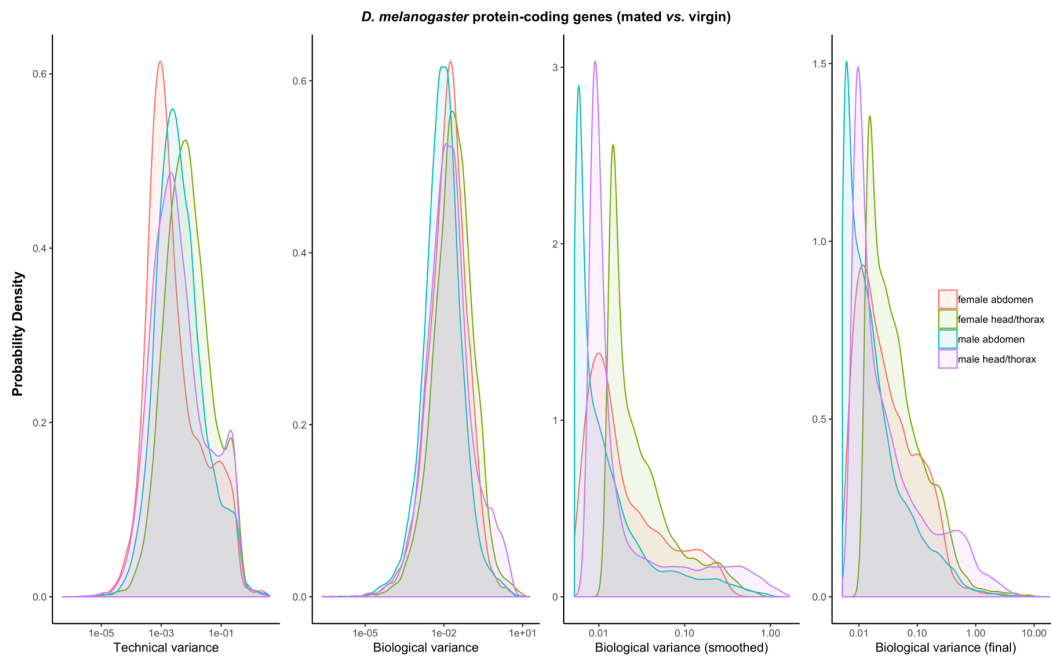


Figure B.5 – A decomposition of the variance observed between biological replicates for protein-coding genes for the analysis presented in chapter 5 of this thesis. The technical or inferential variance (far left) arises from ‘random sequencing and computational analysis of reads’ (Pimentel, et al., 2017). The biological variance (middle left) refers to the variance attributable to the difference in RNA content between samples as well as stochastic library preparation processes. The smoothed biological variance (middle right) is the biological variance after shrinkage in order to stabilise the variance. The final biological variance (far right) however is the maximum of the initial biological variance estimate and smoothed biological variance estimate.

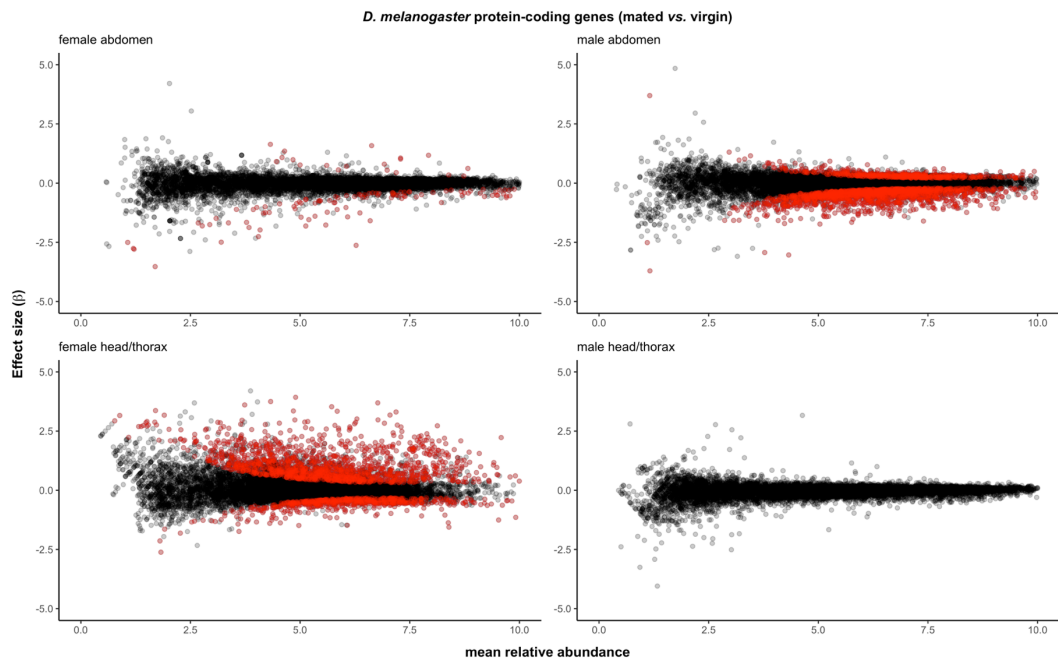


Figure B.6 – An MA plot for the *D. melanogaster* protein-coding genes. On the x-axis – the mean relative gene abundance across all replicates. On the y-axis – the effect size measured in the value of the beta parameter used within the generalised linear model constructed for the purpose of differential expression analysis. Colour legend: Red – differentially expressed genes; black – non-differentially expressed genes.

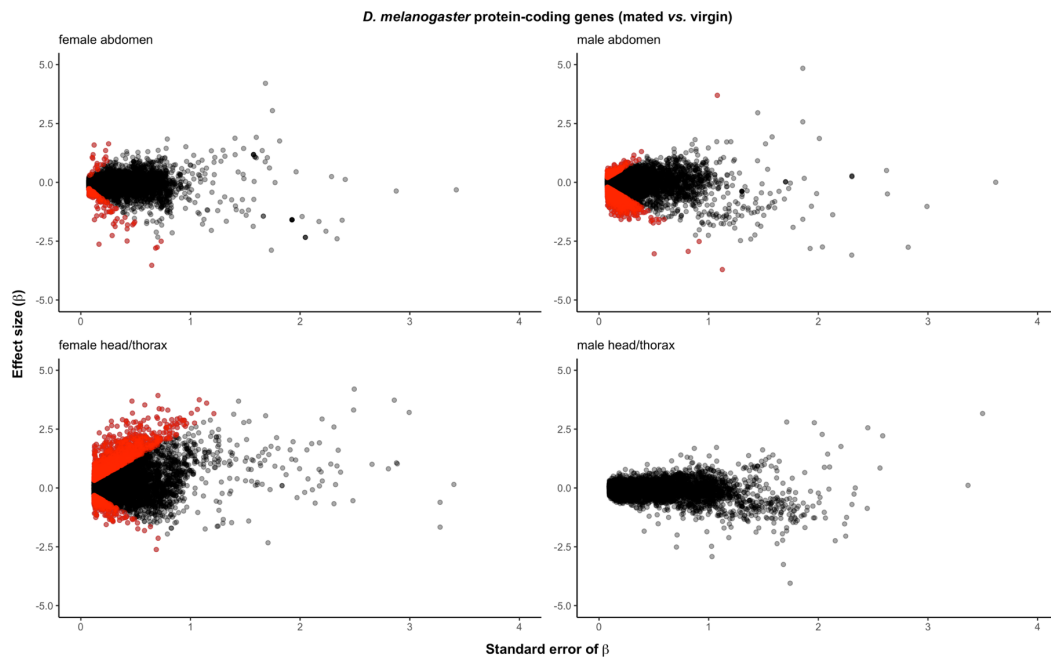


Figure B.7 – A comparison of the differential expression effect size with the uncertainty of the effect size estimate. On the y-axis: Effect size measured by the generalised linear model beta parameter for differential expression. On the x-axis: The standard error of beta. Colour legend: Red – differentially expressed genes; black – non-differentially expressed genes.

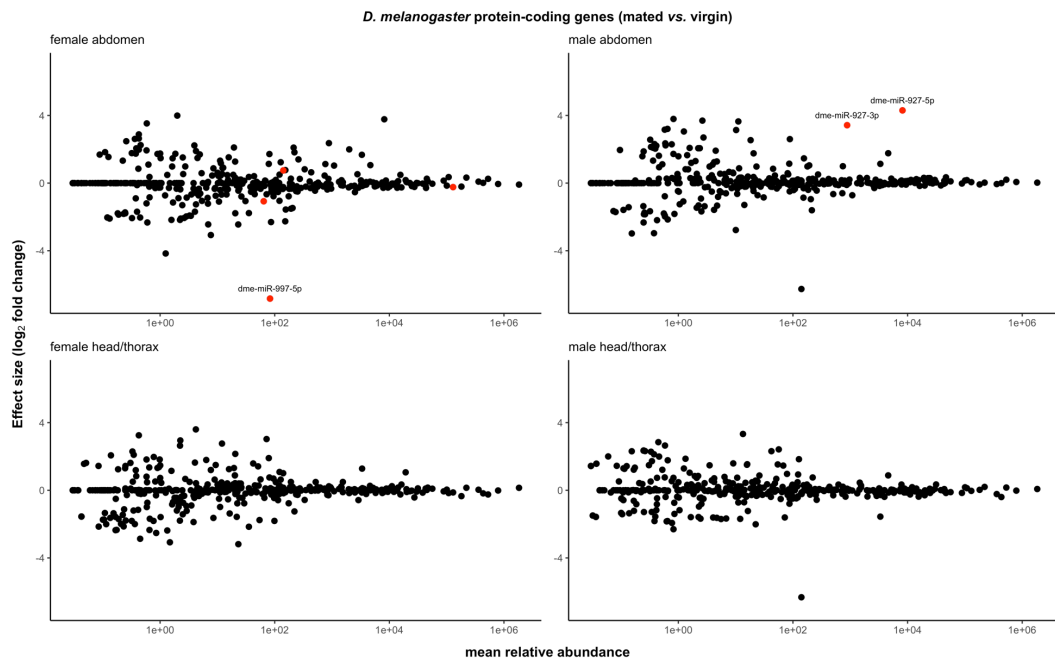


Figure B.8 – An MA plot for miRNA differential expression for the study presented in chapter 5. On the x-axis: mean relative abundance across all replicates for this comparison and for this miRNA. On the y-axis: log₂ fold change. Colour legend: Red – differentially expressed miRNAs; black – non-differentially expressed miRNAs.

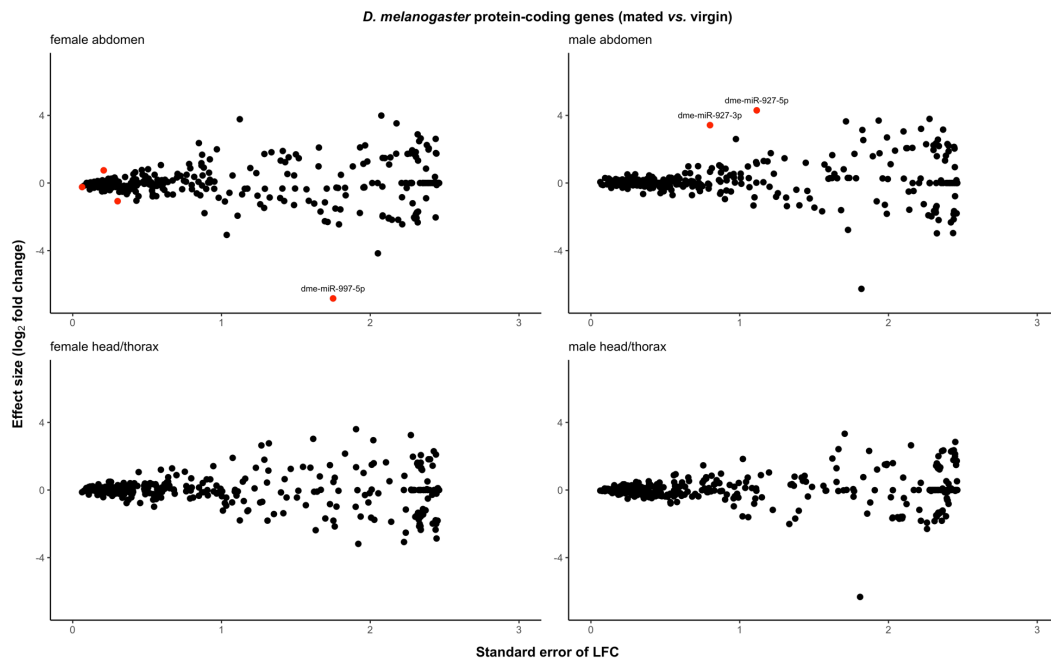


Figure B.9 – The relationship between the \log_2 fold change and its standard error for the differential expression analysis study presented in chapter 5 of this thesis. On the x-axis – the standard error of the \log_2 fold change. On the y-axis – \log_2 fold change. Colour legend: Red – differentially expressed miRNAs; black – non-differentially expressed miRNAs.

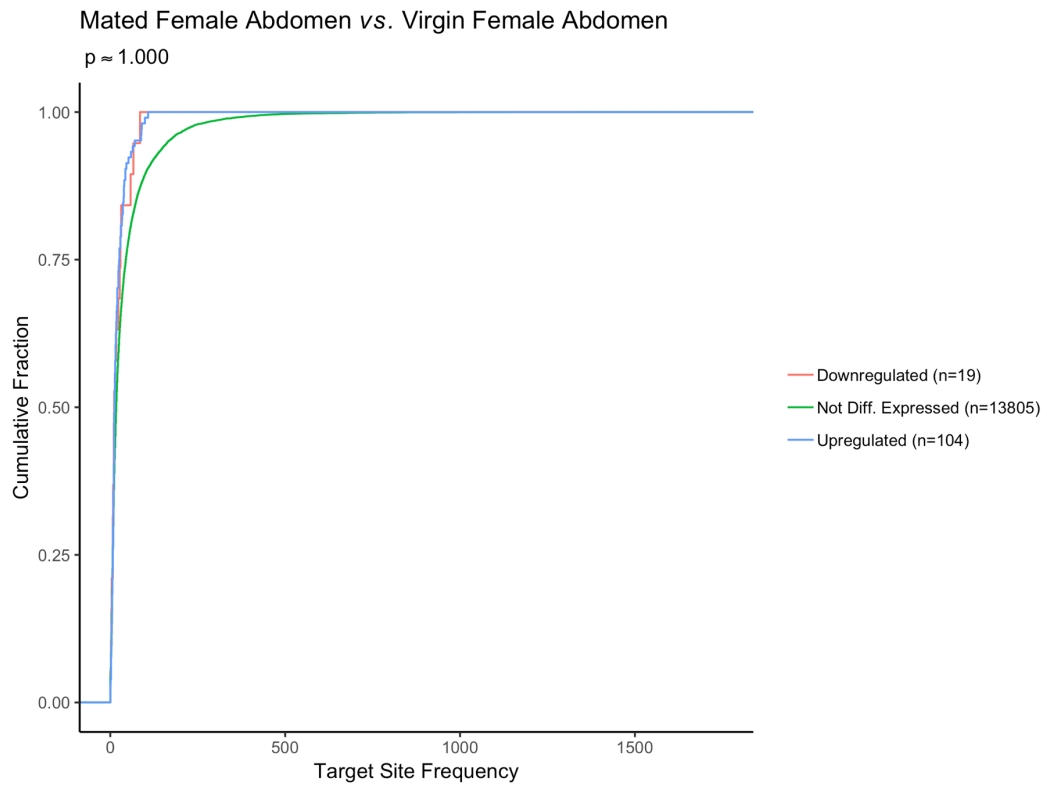


Figure B.10 – A comparison of the cumulative target site frequency distribution of downregulated, non-differentially expressed, and upregulated genes in the female abdomen when comparing mated and virgin fruit flies. Reported p-value derives from a two-sided Kolmogorov-Smirnov test between downregulated and upregulated transcripts.

Comparison	miRNA ID	miRNA direction	Gene ID	Gene direction
female abdomen	dme-miR-184-5p	up	FBgn0261989	down
female abdomen	dme-miR-14-3p	down	FBgn0033926	up
female abdomen	dme-miR-14-3p	down	FBgn0036778	up
male abdomen	dme-miR-927-3p	down	FBgn0000377	up
male abdomen	dme-miR-927-3p	down	FBgn0000592	up
male abdomen	dme-miR-927-3p	down	FBgn0001308	up
male abdomen	dme-miR-927-3p	down	FBgn0001308	up
male abdomen	dme-miR-927-3p	down	FBgn0003638	up
male abdomen	dme-miR-927-3p	down	FBgn0003900	up
male abdomen	dme-miR-927-3p	down	FBgn0003977	up
male abdomen	dme-miR-927-3p	down	FBgn0004414	up
male abdomen	dme-miR-927-3p	down	FBgn0004875	up
male abdomen	dme-miR-927-3p	down	FBgn0010504	up
male abdomen	dme-miR-927-3p	down	FBgn0015622	up
male abdomen	dme-miR-927-3p	down	FBgn0015808	up
male abdomen	dme-miR-927-3p	down	FBgn0020279	up
male abdomen	dme-miR-927-3p	down	FBgn0026415	up
male abdomen	dme-miR-927-3p	down	FBgn0026593	up
male abdomen	dme-miR-927-3p	down	FBgn0027358	up
male abdomen	dme-miR-927-3p	down	FBgn0027571	up
male abdomen	dme-miR-927-3p	down	FBgn0028978	up
male abdomen	dme-miR-927-3p	down	FBgn0030061	up
male abdomen	dme-miR-927-3p	down	FBgn0030616	up
male abdomen	dme-miR-927-3p	down	FBgn0031041	up
male abdomen	dme-miR-927-3p	down	FBgn0031174	up
male abdomen -	dme-miR-927-3p	down	FBgn0031536	up
male abdomen	dme-miR-927-3p	down	FBgn0031646	up
male abdomen	dme-miR-927-3p	down	FBgn0031646	up
male abdomen	dme-miR-927-3p	down	FBgn0031652	up
male abdomen	dme-miR-927-3p	down	FBgn0031723	up
male abdomen	dme-miR-927-3p	down	FBgn0032431	up
male abdomen	dme-miR-927-3p	down	FBgn0032685	up
male abdomen	dme-miR-927-3p	down	FBgn0032772	up
male abdomen	dme-miR-927-3p	down	FBgn0033268	up
male abdomen	dme-miR-927-3p	down	FBgn0033401	up
male abdomen	dme-miR-927-3p	down	FBgn0036285	up
male abdomen	dme-miR-927-3p	down	FBgn0036934	up

male abdomen	dme-miR-927-3p	down	FBgn0037512	up
male abdomen	dme-miR-927-3p	down	FBgn0038535	up
male abdomen	dme-miR-927-3p	down	FBgn0038535	up
male abdomen	dme-miR-927-3p	down	FBgn0038876	up
male abdomen	dme-miR-927-3p	down	FBgn0038876	up
male abdomen	dme-miR-927-3p	down	FBgn0039110	up
male abdomen	dme-miR-927-3p	down	FBgn0039464	up
male abdomen	dme-miR-927-3p	down	FBgn0039713	up
male abdomen	dme-miR-927-3p	down	FBgn0040236	up
male abdomen	dme-miR-927-3p	down	FBgn0040391	up
male abdomen	dme-miR-927-3p	down	FBgn0259937	up
male abdomen	dme-miR-927-3p	down	FBgn0260462	up
male abdomen	dme-miR-927-3p	down	FBgn0261563	up
male abdomen	dme-miR-927-3p	down	FBgn0261931	up
male abdomen	dme-miR-927-3p	down	FBgn0261931	up
male abdomen	dme-miR-927-3p	down	FBgn0262527	up
male abdomen	dme-miR-927-3p	down	FBgn0264978	up
male abdomen	dme-miR-927-5p	down	FBgn0000150	up
male abdomen	dme-miR-927-5p	down	FBgn0000181	up
male abdomen	dme-miR-927-5p	down	FBgn0000636	up
male abdomen	dme-miR-927-5p	down	FBgn0002284	up
male abdomen	dme-miR-927-5p	down	FBgn0002719	up
male abdomen	dme-miR-927-5p	down	FBgn0003053	up
male abdomen	dme-miR-927-5p	down	FBgn0003462	up
male abdomen	dme-miR-927-5p	down	FBgn0003638	up
male abdomen	dme-miR-927-5p	down	FBgn0003638	up
male abdomen	dme-miR-927-5p	down	FBgn0003900	up
male abdomen	dme-miR-927-5p	down	FBgn0004391	up
male abdomen	dme-miR-927-5p	down	FBgn0004396	up
male abdomen	dme-miR-927-5p	down	FBgn0010412	up
male abdomen	dme-miR-927-5p	down	FBgn0011016	up
male abdomen	dme-miR-927-5p	down	FBgn0011205	up
male abdomen	dme-miR-927-5p	down	FBgn0011205	up
male abdomen	dme-miR-927-5p	down	FBgn0011227	up
male abdomen	dme-miR-927-5p	down	FBgn0014455	up
male abdomen	dme-miR-927-5p	down	FBgn0014859	up
male abdomen	dme-miR-927-5p	down	FBgn0015010	up
male abdomen	dme-miR-927-5p	down	FBgn0015245	up
male abdomen	dme-miR-927-5p	down	FBgn0015600	up

male abdomen	dme-miR-927-5p	down	FBgn0015600	up
male abdomen	dme-miR-927-5p	down	FBgn0020386	up
male abdomen	dme-miR-927-5p	down	FBgn0020626	up
male abdomen	dme-miR-927-5p	down	FBgn0022708	up
male abdomen	dme-miR-927-5p	down	FBgn0023388	up
male abdomen	dme-miR-927-5p	down	FBgn0023512	up
male abdomen	dme-miR-927-5p	down	FBgn0023526	up
male abdomen	dme-miR-927-5p	down	FBgn0023526	up
male abdomen	dme-miR-927-5p	down	FBgn0023526	up
male abdomen	dme-miR-927-5p	down	FBgn0024314	up
male abdomen	dme-miR-927-5p	down	FBgn0024314	up
male abdomen	dme-miR-927-5p	down	FBgn0024509	up
male abdomen	dme-miR-927-5p	down	FBgn0025681	up
male abdomen	dme-miR-927-5p	down	FBgn0026593	up
male abdomen	dme-miR-927-5p	down	FBgn0026616	up
male abdomen	dme-miR-927-5p	down	FBgn0027329	up
male abdomen	dme-miR-927-5p	down	FBgn0027585	up
male abdomen	dme-miR-927-5p	down	FBgn0027605	up
male abdomen	dme-miR-927-5p	down	FBgn0027835	up
male abdomen	dme-miR-927-5p	down	FBgn0027835	up
male abdomen	dme-miR-927-5p	down	FBgn0027868	up
male abdomen	dme-miR-927-5p	down	FBgn0028292	up
male abdomen	dme-miR-927-5p	down	FBgn0028327	up
male abdomen	dme-miR-927-5p	down	FBgn0028474	up
male abdomen	dme-miR-927-5p	down	FBgn0028474	up
male abdomen	dme-miR-927-5p	down	FBgn0030050	up
male abdomen	dme-miR-927-5p	down	FBgn0030067	up
male abdomen	dme-miR-927-5p	down	FBgn0030096	up
male abdomen	dme-miR-927-5p	down	FBgn0030177	up
male abdomen	dme-miR-927-5p	down	FBgn0030242	up
male abdomen	dme-miR-927-5p	down	FBgn0030309	up
male abdomen	dme-miR-927-5p	down	FBgn0030316	up
male abdomen	dme-miR-927-5p	down	FBgn0030319	up
male abdomen	dme-miR-927-5p	down	FBgn0030331	up
male abdomen	dme-miR-927-5p	down	FBgn0030331	up
male abdomen	dme-miR-927-5p	down	FBgn0030465	up
male abdomen	dme-miR-927-5p	down	FBgn0030592	up
male abdomen	dme-miR-927-5p	down	FBgn0030616	up
male abdomen	dme-miR-927-5p	down	FBgn0030631	up

male abdomen	dme-miR-927-5p	down	FBgn0030631	up
male abdomen	dme-miR-927-5p	down	FBgn0030661	up
male abdomen	dme-miR-927-5p	down	FBgn0030761	up
male abdomen	dme-miR-927-5p	down	FBgn0030990	up
male abdomen	dme-miR-927-5p	down	FBgn0031078	up
male abdomen	dme-miR-927-5p	down	FBgn0031183	up
male abdomen	dme-miR-927-5p	down	FBgn0031260	up
male abdomen	dme-miR-927-5p	down	FBgn0031304	up
male abdomen	dme-miR-927-5p	down	FBgn0031364	up
male abdomen	dme-miR-927-5p	down	FBgn0031397	up
male abdomen	dme-miR-927-5p	down	FBgn0031420	up
male abdomen	dme-miR-927-5p	down	FBgn0031653	up
male abdomen	dme-miR-927-5p	down	FBgn0031869	up
male abdomen	dme-miR-927-5p	down	FBgn0032025	up
male abdomen	dme-miR-927-5p	down	FBgn0032536	up
male abdomen	dme-miR-927-5p	down	FBgn0032748	up
male abdomen	dme-miR-927-5p	down	FBgn0032897	up
male abdomen	dme-miR-927-5p	down	FBgn0033130	up
male abdomen	dme-miR-927-5p	down	FBgn0033692	up
male abdomen	dme-miR-927-5p	down	FBgn0033814	up
male abdomen	dme-miR-927-5p	down	FBgn0033844	up
male abdomen	dme-miR-927-5p	down	FBgn0034521	up
male abdomen	dme-miR-927-5p	down	FBgn0035490	up
male abdomen	dme-miR-927-5p	down	FBgn0035519	up
male abdomen	dme-miR-927-5p	down	FBgn0035947	up
male abdomen	dme-miR-927-5p	down	FBgn0035988	up
male abdomen	dme-miR-927-5p	down	FBgn0036024	up
male abdomen	dme-miR-927-5p	down	FBgn0036298	up
male abdomen	dme-miR-927-5p	down	FBgn0036467	up
male abdomen	dme-miR-927-5p	down	FBgn0036516	up
male abdomen	dme-miR-927-5p	down	FBgn0037137	up
male abdomen	dme-miR-927-5p	down	FBgn0037170	up
male abdomen	dme-miR-927-5p	down	FBgn0037249	up
male abdomen	dme-miR-927-5p	down	FBgn0037249	up
male abdomen	dme-miR-927-5p	down	FBgn0038321	up
male abdomen	dme-miR-927-5p	down	FBgn0038424	up
male abdomen	dme-miR-927-5p	down	FBgn0038598	up
male abdomen	dme-miR-927-5p	down	FBgn0039141	up
male abdomen	dme-miR-927-5p	down	FBgn0039419	up

male abdomen	dme-miR-927-5p	down	FBgn0039419	up
male abdomen	dme-miR-927-5p	down	FBgn0039562	up
male abdomen	dme-miR-927-5p	down	FBgn0039857	up
male abdomen	dme-miR-927-5p	down	FBgn0069354	up
male abdomen	dme-miR-927-5p	down	FBgn0086674	up
male abdomen	dme-miR-927-5p	down	FBgn0250789	up
male abdomen	dme-miR-927-5p	down	FBgn0259203	up
male abdomen	dme-miR-927-5p	down	FBgn0259209	up
male abdomen	dme-miR-927-5p	down	FBgn0259937	up
male abdomen	dme-miR-927-5p	down	FBgn0261068	up
male abdomen	dme-miR-927-5p	down	FBgn0261593	up
male abdomen	dme-miR-927-5p	down	FBgn0262146	up
male abdomen	dme-miR-927-5p	down	FBgn0262582	up
male abdomen	dme-miR-927-5p	down	FBgn0264296	up
male abdomen	dme-miR-927-5p	down	FBgn0264978	up
male abdomen	dme-miR-927-5p	down	FBgn0266464	up
male abdomen	dme-miR-927-5p	down	FBgn0266599	up

Table B.3 - A table of oppositely differentially expressed targets of differentially expression miRNAs. Columns denote, in order from left to right, the relevant comparison, miRNA identifier, direction of miRNA differential expression, the relevant gene identifier, and the direction of gene differential expression.

Appendix C

miRNA Name	miRNA sequence
miR-nov1-3p	TTCCCCTGTGCTGGTGGGGTTG
miR-nov1-5p	ACTCAACCCGCACAGAGGAGG
miR-nov2-3p	ATGGCGGCACGTTGAGTTTGC
miR-nov2-5p	AGGAACTCAACGTGCCGCCATG
miR-nov3-3p	TTACTCTGGACTGAAATCTTTC
miR-nov3-5p	TGTGAAGGGTTTCAGTCCAGACTGA
miR-nov4-3p	CCCAGAACTACCATCAGAGAAT
miR-nov4-5p	TTACTCTGGTGGTTGTTGTGT
miR-nov5-3p	ACTGAACATGCTCTCCAGACGA
miR-nov5-5p	TGTGGGGACTGTGTGTTTTGTGT
miR-nov6-3p	AGGCCAATGCCAAGGAAAGGAG
miR-nov6-5p	TAACTTTCCTTGTGTATTCCCA
miR-nov7-3p	TGGGACTGAGCAAACCTCATC
miR-nov7-5p	TAGAGTTTGCTCATTGTGCATG
miR-nov8-3p	TGGTCTGATCTGGTCTGATC
miR-nov8-5p	ACCAGACCAGACCAGACCTGAT
miR-nov9-5p	TTCCATAGTTCGGAGCTCTGA
miR-nov10-3p	GTGCCCAAGAAACTGCCTCAGT
miR-nov10-5p	CACTAGGCAGTTTTTTGGGTAA
miR-nov11-3p	TCAACACTGGAGTGGTCTCTGTCT
miR-nov11-5p	AGGACAGGAAAACCTGGACAGTATGGAC
miR-nov12-3p	TCTGAATGTTTGGTCCTGTTG
miR-nov12-5p	ACAGGACGCAGACGCTCAGAGG

miR-nov13-3p	TAATAGCAGTAGCAGCAGCAGT
miR-nov13-5p	TACTGCTACTACTTCTATCACT
miR-nov14-3p	TCCATATAAAATCAGCTGACAGG
miR-nov14-5p	TGTCAGCTGATTTATGTGGTAAC
miR-nov15-3p	CAGGCTTTCTTTGTGATGCACC
miR-nov15-5p	TGCATCACTAAGAAAGCCTGA
miR-nov16-3p	CTCTGATTGGCTGAGATGTGA
miR-nov16-5p	CCCACGTTTCCTCCAATCAGAGC
miR-nov17-3p	TGAAACTCTTCCCTCAGACCGA
miR-nov17-5p	AGCTTGAGGGAAGAGTTTCA
miR-nov18-3p	AGTTGCACACAAGCTGTCGGG
miR-nov18-5p	GACAGCTTGTGCACAACCTGGTT
miR-nov19-3p	TTCCACAGTCCAGCACACAGT
miR-nov19-5p	TGCTCTGCTGTCTGTGGAAATA
miR-nov20-3p	CCGCTGTCGCTCTGCCACACT
miR-nov20-5p	TGTGGGCAGAGAGTCAGACTGA
miR-nov21-3p	TAACGTTAGCCTCAGCTGCTGC
miR-nov21-5p	CTCAGCTGCCGCTAACGTTAGC
miR-nov22-3p	AGTCTGTGATCATGTGATTGAC
miR-nov22-5p	ACTTCACATGGTTACTGATCTTG
miR-nov23-3p	TGCACCTGCACCTCATGAGTCT
miR-nov23-5p	TCTCATGAGCTGCAGGTGGCGTT
miR-nov24-3p	ATTGGATAACTGATCACTGATC
miR-nov24-5p	TCAGTGTCCGTCCATCCTGTCA
miR-nov25-3p	TCTCATGGGAATTGTAGTTGCT
miR-nov25-5p	ACAGCTGCAACTCCCACGAGG
miR-nov26-3p	TACACGTTGCCGTCTTGCCAGGG

miR-nov26-5p	GCGGCGCTACGGTATCGTTACG
miR-nov27-3p	CCAGTATGATATGTGCTGCTCCT
miR-nov27-5p	TAGCAGCACATCATTACTGGTA
miR-nov28-3p	AACAAAGGTGGGCTTAGTCGA
miR-nov28-5p	ACACTAAAAGCATCTTTGTTCT
miR-nov29-3p	TCCACATAAATCAGCTGACTGG
miR-nov29-5p	TCCAGTCAGCTGATTTATGTGG
miR-nov30-3p	GGTGATGTTATTCAGAAGGACTTGG
miR-nov30-5p	GATACTTCAACATGAGTCATGAACA
miR-nov31-3p	TTATAAGGTGCCCGGAATGCTGGTT
miR-nov31-5p	TTGAGACCAACGAGCAAGAGGGGGG
miR-nov32-3p	GCAGCCTGTCATCAGTAGAGC
miR-nov32-5p	TCTATTGCTGACAAAGAGAAGC
miR-nov33-3p	TTGTTTCCAAATGGTGCCATGCACA
miR-nov33-5p	TGAGTTACTGGAGAGCCGCTGCTCT
miR-nov34-3p	TTCTGTTGATGTTTGGAGCAGAGAC
miR-nov34-5p	TTCTCACTTCTGAGGCAGCAGGGA
miR-nov35-3p	TCCACATAAATCAGCCGACAGG
miR-nov36-5p	AGCTTGAGGGAAGAGTTTCAAA
miR-nov37-3p	TATCATGAGCAGTTGAATGTT
miR-nov37-5p	TCATTAACTGCTTGTGGTACA
miR-nov38-3p	AGAGTGTGTGACAGAAACATC
miR-nov38-5p	TGTGTTTCTGGAACTACCACTCT
miR-nov39-3p	TCTGTTGTAGGTCTGTTGTGT
miR-nov39-5p	AACAACAGGCCAACAACAACCTGA
miR-nov40-3p	TTGAGCTGTCACATCCTGCTGC
miR-nov40-5p	TAGCGGATGAGTCAGACTCGC

miR-nov41-3p	TGCGGTAGCGTTAGCAACATGG
miR-nov41-5p	ATGTTGCTAACGCTGCCGCTAGCG
miR-nov42-3p	TATGTGATTGTTTCAGTAGACA
miR-nov42-5p	TGTGTAAGTAAAAGTCATATAT
miR-nov43-3p	GCTTCTTCACAACACCAGGGT
miR-nov44-3p	CATTTAGCCTTTGCCCTGTAG
miR-nov44-5p	ACAGAGCAAAGGACCAAATGCC
miR-nov45-3p	TTCCTCTGTGCTGGTGGATT
miR-nov45-5p	ACCCTACCTGCACAGAGGAG
miR-nov46-3p	AGTCTGGCACTGTCAGCTCAGA
miR-nov46-5p	TTTGAACCTTGACACTGCCATGCG
miR-nov47-3p	TGCGCACGGGGCCACGCCCTGC
miR-nov47-5p	TAGGCGTGTCACTGCGTGTACACA

Table C.1 - Identifiers of novel Asian seabass (*Lates calcarifer*) miRNAs discovered during the course of the research described in chapter 6. Novel miRNAs are successively named according to the following regular expression: miR-nov[0-9]+-(3|5)p.

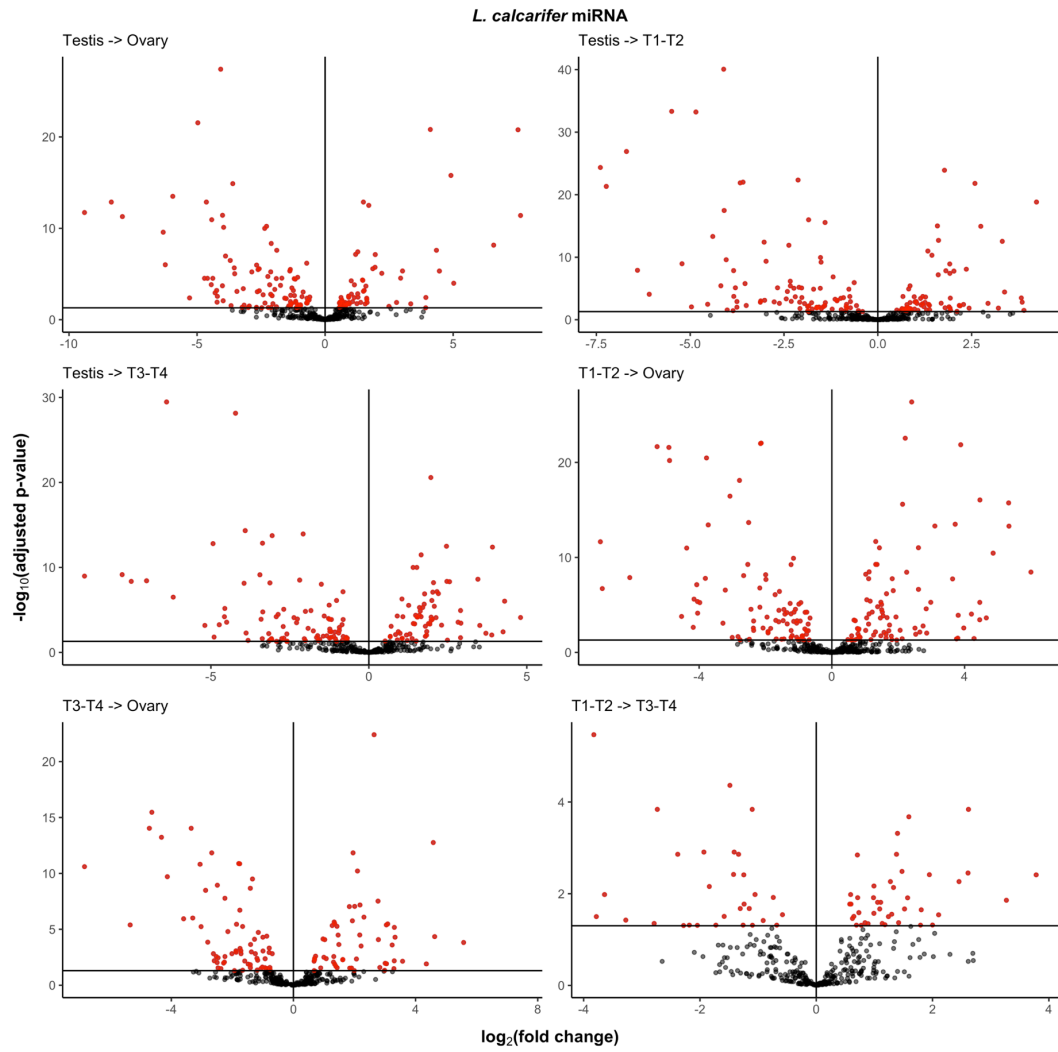


Figure C.1 – Volcano plot for miRNA for the differential expression analysis presented in chapter 6 of this thesis. Colour legend: Red – differentially expressed miRNAs; black – non-differentially expressed miRNAs.

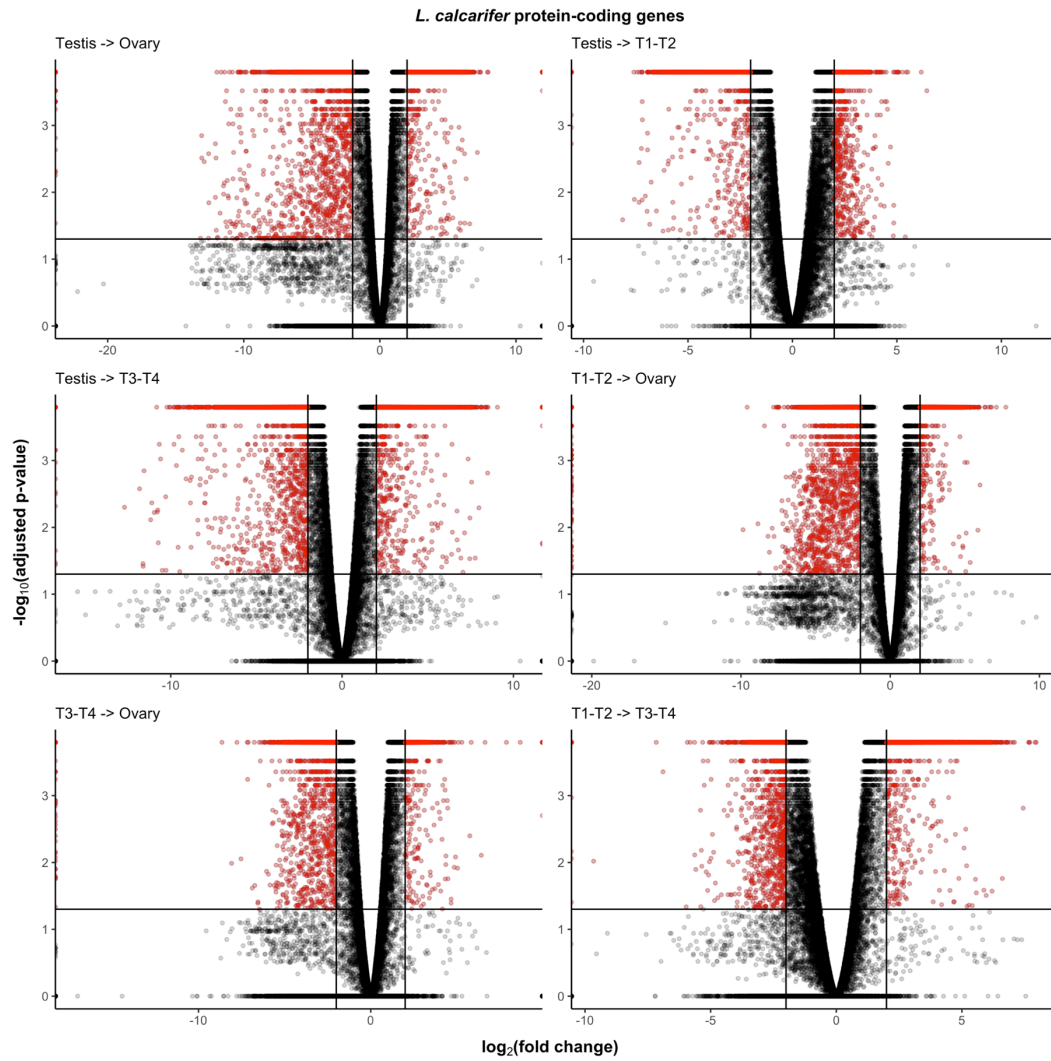


Figure C.2 – volcano plot for protein-coding genes for the differential expression analysis presented in chapter 6 of this thesis. Colour legend: Red – differentially expressed genes; black – non-differentially expressed genes.

Sequence analysis

FilTar: using RNA-Seq data to improve microRNA target prediction accuracy in animals

Thomas Bradley ^{1,2} and Simon Moxon ^{1,*}

¹School of Biological Sciences, University of East Anglia, Norwich NR4 7TJ, UK and ²Earlham Institute, Norwich Research Park, Norwich NR4 7UZ, UK

*To whom correspondence should be addressed.

Associate Editor: Yann Ponty

Received on June 17, 2019; revised on January 1, 2020; editorial decision on January 2, 2020; accepted on January 9, 2020

Abstract

Motivation: MicroRNA (miRNA) target prediction algorithms do not generally consider biological context and therefore generic target prediction based on seed binding can lead to a high level of false-positive predictions. Here, we present FilTar, a method that incorporates RNA-Seq data to make miRNA target prediction specific to a given cell type or tissue of interest.

Results: We demonstrate that FilTar can be used to: (i) provide sample specific 3'-UTR reannotation; extending or truncating default annotations based on RNA-Seq read evidence and (ii) filter putative miRNA target predictions by transcript expression level, thus removing putative interactions where the target transcript is not expressed in the tissue or cell line of interest. We test the method on a variety of miRNA transfection datasets and demonstrate increased accuracy versus generic miRNA target prediction methods.

Availability and implementation: FilTar is freely available and can be downloaded from <https://github.com/TBradley27/FilTar>. The tool is implemented using the Python and R programming languages, and is supported on GNU/Linux operating systems.

Contact: s.moxon@uea.ac.uk

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

MicroRNAs (miRNAs) exert widespread post-transcriptional control over mRNA expression in most animal lineages (Bartel, 2018), creating a need for the accurate identification of miRNA targets in order to better understand gene regulation. Traditional methods for providing experimental support for putative interactions include the use of reporter assays to test for a direct interaction between the miRNA and mRNA, or perturbation experiments to test for the effect of increased or decreased miRNA levels on target mRNA, or the corresponding proteins translated from these molecules (Kuhn *et al.*, 2008). More recent methods allow researchers to test for direct interactions between miRNA and putative targets using transcriptome-wide crosslinking and immunoprecipitation experiments. These methods usually test for binding between the putative miRNA target and argonaute (AGO) (Chi *et al.*, 2009; König *et al.*, 2010; Van Nostrand *et al.*, 2016), a key component of the miRNA-guided RISC (RNA-induced silencing complex), and in addition, some methods can be used to determine the identity of the miRNA which is guiding AGO to the target transcript (Helwak and Tollervey, 2014; Kudla *et al.*, 2011).

Currently available data for these types of experiments are generally limited in number and diversity of cell types and species.

Inspection of the TarBase resource (v8.0) (Karagkouni *et al.*, 2018), a database of published, experimentally supported miRNA interactions, reveals that, at the time of writing, even for a widely utilized model organism such as mouse, AGO immunoprecipitation datasets are available for only three cell lines and five tissues. The problem is exacerbated when examining records for other model organisms such as rat and zebrafish, in which no data from immunoprecipitation experiments are reported. This is likely because generating data of this type is usually prohibitively expensive in terms of skills, time and material resources needed to complete sophisticated transcriptome-wide, next-generation library preparation and sequencing protocols. The limited applicability of experimental approaches, therefore, underlies the continuing necessity of computational approaches for predicting miRNA targets.

There are a number of existing computational tools for predicting miRNA targets in animals. Algorithms such as TargetScan use complementarity between the seed sequence of the miRNA (Bartel, 2018; Lewis *et al.*, 2003) and a corresponding region of the 3'-UTR of its target as the basis of target prediction (Agarwal *et al.*, 2015; Friedman *et al.*, 2008; Garcia *et al.*, 2011; Grimson *et al.*, 2007; Lewis *et al.*, 2003, 2005). Alternatively, some miRNA target prediction algorithms do not require full complementarity in the miRNA seed region (Enright *et al.*, 2003; Gumienny and Zavolan, 2015;

John *et al.*, 2004; Khorshid *et al.*, 2013; Wang, 2016), or predict miRNA targeting to occur in the coding region of the transcript as well as the 3'-UTR (Reczko *et al.*, 2012). Most algorithms, in addition to considerations of seed complementarity, and the location of the target site within the transcript, also consider features such as the conservation of the miRNA target site in closely related species, the thermodynamic stability of the miRNA-mRNA duplex, and the structural accessibility of putative target sites to the miRNA-RISC complex, as variables which are also thought to influence miRNA targeting and subsequent transcript repression (Ritchie and Rasko, 2014).

Although intramolecular features are often considered, current miRNA target predictions currently do not account for the broader cellular context in which miRNA targeting occurs. The clearest indication of this is that current target prediction tools do not account for whether predicted targets are expressed within a given cell type or tissue. If the predicted target is not expressed, it cannot physically interact and be translationally inhibited or repressed by miRNA molecules. As expression profiles vary across different cell types and tissues, failing to consider whether a predicted target is expressed in a given cellular context may lead to false-positive results when making miRNA target predictions.

For the prediction of miRNA targets in the 3'-UTR, an additional complication is that the identity of an individual 3'-UTR may not be constant across different cell types or different biological conditions due to alternative cleavage and polyadenylation (APA) (Elkon *et al.*, 2013; Tian and Manley, 2017). APA is the process by which cellular polyadenylation machinery utilizes alternative polyadenylation sites located on precursor mRNA molecules to produce transcripts with alternative 3'-UTR sequences. Differential usage of polyadenylation sites in diverse tissues or biological conditions, can result in distinct 3'-UTR isoform abundance profiles existing between different cell types (Nam *et al.*, 2014). One consequence of the existence of 3'-UTR isoforms is that a miRNA target site may exist for some 3'-UTR isoforms of the same annotated mRNA but not others.

As a result, APA allows the differential usage of miRNA target sites by the cell, diversifying and modifying the effect of miRNAs in different cellular contexts. For example, in cancer cells, shortening of 3'-UTRs can activate oncogenes by increasing mRNA stability, partially through the reduction in the number of miRNA target sites in their 3'-UTRs, decreasing the extent to which they are repressed (Mayr and Bartel, 2009). In contrast, an extensive enrichment of longer 3'-UTRs and hence additional miRNA target sites have been discovered in mammalian brain tissue (Miura *et al.*, 2013), which has been hypothesized to serve as an extended platform for the regulation of gene expression (Wang and Yi, 2014). This evidence of context-specific miRNA action underlies the utility of methods which accounts for this information in order to increase the precision and sensitivity of miRNA target predictions.

Most databases of miRNA target predictions do not incorporate information relating to APA, and instead rely on default 3'-UTR annotations provided by public sequence databases such as Ensembl (Birney, 2004; Cunningham *et al.*, 2019) and RefSeq (Pruitt *et al.*, 2007, 2014), when identifying potential miRNA targets. Similarly, most prediction algorithms do not easily allow the user to generate predictions for multiple 3'-UTR isoforms of the same mRNA. An exception is TargetScan (v7) (Agarwal *et al.*, 2015). In this version, each mRNA transcript is associated with a distinct profile of relative 3'-UTR isoform abundances. From this profile, each scored target site is weighted by the abundance of the 3'-UTR segment containing the predicted target site relative to all 3'-UTRs of that transcript. The caveat of this analysis being that 3'-UTR profiles are generated from sequencing data obtained from only four human cell lines (Nam *et al.*, 2014), which is subsequently treated as being representative for all cell types. Although it was shown that this approach was superior to not incorporating 3'-UTR profile data at all, it was sub-optimal in comparison to using 3'-UTR profiles specific to each cellular context examined (Nam *et al.*, 2014). Crucially, a miRNA target prediction tool which enables the user to predict miRNA targets specific to a given tissue or cell line is currently lacking.

Presented in this article is FilTar, a tool which takes RNA-Seq data as input and generates miRNA target predictions tailored to specific cellular contexts. Specificity of target prediction is increased by utilising information from sequencing data both to filter out poorly or non-expressed targets and to refine 3'-UTR annotations. Analysis demonstrates that predicted miRNA targets gained and lost due to 3'-UTR reannotation behave like pre-existing predicted miRNA target and non-targets, respectively, in response to miRNA transfection. The cumulative effect of integrating these additional processing steps into conventional miRNA target prediction workflows is to increase prediction accuracy and to drastically alter the number of miRNA target predictions made between different cell types.

2 Materials and methods

All following steps were carried out using the FilTar tool. The workflow and parameters are described in detail below.

2.1 Implementation

FilTar is a command line tool for GNU/Linux operating systems written predominantly in the Python (v3.6.8) and R (v3.5.0) programming languages. Users can configure the tool to process available RNA-Seq datasets from public repositories such as the European Nucleotide Archive (ENA; <https://www.ebi.ac.uk/ena>) (Harrison *et al.*, 2019; Leinonen *et al.*, 2011) and the Sequence Read Archive (SRA; <https://ncbi.nlm.nih.gov/sra>) (Leinonen *et al.*, 2010), and also the user's own private sequencing data. All reported parameters are fully configurable within the FilTar tool. FilTar utilizes Snakemake (v5.4.0) (Köster and Rahmann, 2012) for workflow management. Most FilTar dependencies are managed using Conda (v4.6.6; <https://docs.conda.io/en/latest/>).

2.2 Data preprocessing

Reads were trimmed using Trim Galore (v0.5.0) (Krueger, 2015), a wrapper around Cutadapt (v1.16) (Martin, 2011), using default parameters with the exception of the 'length' and 'stringency' parameters which were set to 35 and 4, respectively.

2.3 3'-UTR reannotation

In order to build an index for the alignment of FASTQ reads to the genome, unmasked chromosomal reference genome assembly fasta files for human (GRCh38.p12) and mouse (GRCm38.p6) (Schneider *et al.*, 2017) were downloaded from release 94 of Ensembl (www.ensembl.org/index.html) (Cunningham *et al.*, 2019). All subsequent files obtained from the Ensembl resource were for this same release version. Splice-aware mapping of reads to the genome was achieved using HISAT2 (v2.1.0) (Kim *et al.*, 2015): The locations of exons and junction sites were determined by running the appropriate HISAT2 scripts on the relevant species-specific GTF (gene transfer format) annotation file also obtained from Ensembl. The 'hisat2-build' binary was executed using the 'ss' and 'exon' flags indicating splice site and exon co-ordinates built from the previous step.

The indexed genome was used for FASTQ read alignment using the 'hisat2' command. The 'rna-strandness' option was used for strand-aware alignment. The strandedness of RNA-seq datasets was determined using the 'quant' command of the Salmon (v0.11.3) (Patro *et al.*, 2017) RNA-seq quantification tool, by setting the 'library-type' option to 'A' for automatic inference of library type. The SAMtools (v1.8) (Li *et al.*, 2010) 'view' and 'sort' commands were used to sort data from sam to bam format, and to sort the resultant bam files, respectively.

Sorted bam files were converted to bedgraph format using the 'genomeCoverageBed' command of bedtools (v2.27.1) (Quinlan, 2014; Quinlan and Hall, 2010) using the 'bg', 'ibam' and 'split' options. Bedgraph files representing biological replicates of the same condition were merged using bedtools' 'unionbedg' command. FilTar then calculated the mean average coverage value for each record in the merged bedgraph file.

Existing transcript models were produced by converting Ensembl GTF annotations files (containing one or zero 3'-UTR annotations per protein-coding transcript) into genePred format using the UCSC 'gtfToGenePred' binary, and then from genePred format to bed12 format using the UCSC 'genePredToBed' binary (Kent *et al.*, 2002). APAtrop (Ye *et al.*, 2018), the 3'-UTR reannotation tool, was used to refine 3'-UTR annotations on a transcript-by-transcript basis by integrating information from the bed12 file and bedgraph files using the 'identifyDistal3UTR.pl' perl script with default parameters. FilTar then integrated existing transcript 3'-UTR models with the new models predicted by APAtrop—replacing existing 3'-UTR models for those transcripts in which APAtrop has made a reannotation. Only truncations or elongations of single exon 3'-UTR annotations were integrated into final 3'-UTR annotations; novel 3'-UTR predictions (*i.e.* prediction of 3'-UTRs for transcripts without a previous 3'-UTR annotation) were discarded and alterations of the 3'-UTR start site were also not permitted, due to the reannotation of 3'-UTR start sites by the APAtrop dependency as beginning at the start position of the final exon in standard Ensembl transcript models. No alterations to existing 3'-UTR annotations spanning multiple exons were permitted, as this is not intended functionality of the APAtrop tool.

2.4 miRNA target prediction

Target prediction for the analyses presented in this study was conducted using the TargetScan algorithm (v.7.01) (Agarwal *et al.*, 2015). Mature miRNA sequences were obtained from release 22 of miRBase (www.mirbase.org) (Griffiths-Jones, 2004; Kozomara *et al.*, 2019). The 3'-UTR sequence data required for target prediction can either be provided as multiple sequence alignments (MSAs) or single sequences, with the former option enabling the computation of 3'-UTR branch lengths and the probability of conserved targeting (Pct) for putative miRNA target sites.

Multiple sequence alignments are derived from 100-way (human reference) and 60-way (mouse reference) whole-genome alignments hosted at the UCSC genome browser (https://genome.ucsc.edu) (Kent *et al.*, 2002) generated using the threaded blockset-aligner (Blanchette, 2004) stored in MAF (multiple alignment format) format. MAF files are indexed, and the relevant alignment regions corresponding to 3'-UTR co-ordinates extracted using 'MafIO' functions contained within the Biopython (v1.72) library (Cock *et al.*, 2009). For human MSAs, during post-processing, distantly related species were removed, resulting in 84-way MSAs (Agarwal *et al.*, 2015).

If MSAs are not used, single sequences are extracted from DNA files using relevant 3'-UTR co-ordinates in bed format using the 'get-fasta' command of bedtools with the 's' option enabled. Individual exon sequences are then merged, creating a single contiguous 3'-UTR sequence. FilTar then converts miRNA and 3'-UTR sequence and identifier information to a format which can be parsed by TargetScan algorithms.

TargetScan is executed using both Ensembl 3'-UTR annotations, and updated annotations produced using FilTar for the purposes of the differential expression analyses reported in this study.

The FilTar tool is also fully compatible with the miRanda (v3.3a) (Enright *et al.*, 2003; John *et al.*, 2004) miRNA target prediction algorithm allowing users to identify non-canonical miRNA targets, that is predicted targets without a perfectly complementary seed match to the miRNA.

2.5 Transcript quantification

Human and mouse cDNA files were downloaded from Ensembl. Kallisto (v0.44.0) (Bray *et al.*, 2016) was used to index the cDNA data using the 'kallisto index' command with default parameters. Reads were pseudoaligned and relative transcript abundance quantified using the 'kallisto quant' executable, using the 'bias' option to correct for sequence-based biases. When kallisto was used with data derived from single-end RNA-sequencing experiments, 180 and 20 nt were used as required estimates of the mean average fragment length and SD, respectively.

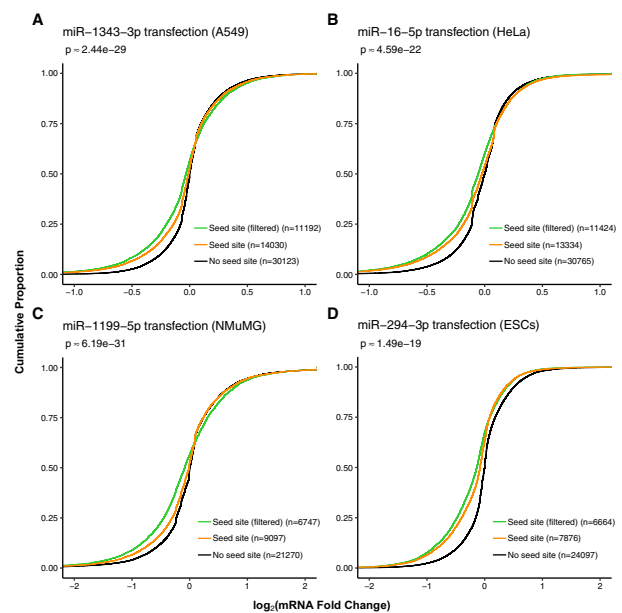


Fig. 1. Implementing an expression threshold on predicted miRNA targets improves miRNA target prediction accuracy. Results are derived from miRNA mimic and control transfection experiments. Curves show the cumulative \log_2 fold change distributions of: (i) protein-coding non-target transcripts (black), (ii) protein-coding seed target transcripts (orange) and (iii) expression filtered ($\text{TPM} \geq 0.1$) protein-coding seed target transcripts (green). Numbers in round brackets represent the number of mRNA transcripts contained in each distribution. Approximate P -values were computed using one-sided, two-sample, Kolmogorov–Smirnov tests between unfiltered and filtered target fold change distributions. Data presented for miRNA mimic transfection into (A) A549 and (B) HeLa cell lines, (C) normal murine mammary gland (NMuMG) cells and (D) mouse embryonic stem cells (ESCs)

2.6 Availability of data and materials

See **Supplementary Methods** for information regarding the selection and analysis of data used in this article. All data analysed in this study are publicly available and a table of relevant project accessions is given (**Supplementary Table S1**), along with relevant QC statistics (**Supplementary Table S2**). The FilTar tool is publicly and freely accessible for download (https://github.com/TBradley27/FilTar) with full supporting documentation (https://tbradley27.github.io/FilTar/).

3 Results

In order to benchmark the performance of the FilTar tool in a specific cellular context versus general miRNA target prediction we used RNA-Seq data from miRNA mimic transfection experiments in mouse and human cell lines. Fold change values represent changes in relative mRNA abundance in samples transfected with a miRNA mimic compared to samples transfected with a negative control.

3.1 Expression filtering

Predicted miRNA targets which were filtered according to their expression level, at a TPM (transcripts per million) (Li *et al.*, 2009) threshold of 0.1, as a whole, exhibited stronger repression after miRNA transfection than the full miRNA target set without expression filtering (**Fig. 1** and **Supplementary Fig. S1**). Predicted miRNA targets removed by FilTar generally exhibited low absolute fold change values suggesting that these are false-positive predictions in these specific cellular contexts (**Supplementary Fig. S2**). Implementing expression filters for a range of different TPM values reveals that increasing this threshold results in a stronger filtering effect on retained mRNAs (**Supplementary Fig. S3a**). However, increasing the expression threshold beyond a particular point (between 1 and 10 TPM for experiments analysed) leads to the removal of a considerable number of mRNA transcripts which are repressed by the transfection of a miRNA mimic (**Supplementary Fig. S3b**).

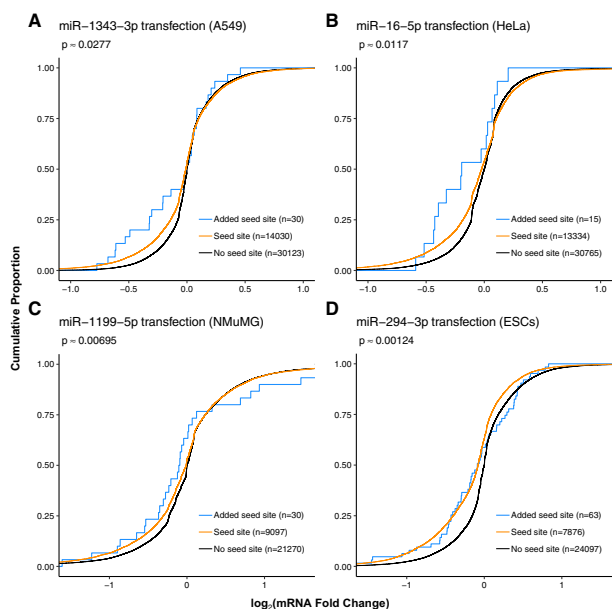


Fig. 2. 3'-UTR elongation by FilTar leads to the identification of additional valid miRNA targets. Curves show the cumulative \log_2 fold change distributions of: (i) protein-coding non-target transcripts (black), (ii) protein-coding seed target transcripts (orange) and (iii) predicted target transcripts deriving from FilTar 3'-UTR annotations but not Ensembl 3'-UTR annotations (blue). Approximate *P*-values were computed using one-sided, two-sample, Kolmogorov–Smirnov tests between pre-existing target and newly identified target fold change distributions. Otherwise as in [Figure 1](#)

The number and percentage of annotated protein-coding transcripts which are used in FilTar's 3'-UTR reannotation workflow, for each sample after expression filtering, are given in [Supplementary Table S3](#). Only those transcripts possessing a pre-existing 3'-UTR annotation spanning only a single exon are selected (see Materials and methods).

3.2 3'-UTR extension

Newly gained miRNA target predictions deriving from FilTar's refined 3'-UTR annotations of protein-coding transcripts (*i.e.* miRNA targets deriving from the elongation of existing 3'-UTR annotations), generally exhibited similar levels of repression to miRNA target predictions deriving from Ensembl 3'-UTR annotations ([Fig. 2](#) and [Supplementary Fig. S4](#)). Anomalies were results deriving from the transfection of miR-107 and miR-10a-5p miRNA mimics into HeLa cells in which newly identified miRNA target predictions did not exhibit a log fold change distribution commensurate with that exhibited by already existing miRNA target predictions ([Supplementary Fig. S4](#)).

3.3 3'-UTR truncation

Conversely, miRNA target transcripts that were removed as a result of FilTar truncating 3'-UTR annotations relative to standard Ensembl annotations, exhibited repression similar to that of annotated non-target transcripts ([Fig. 3](#) and [Supplementary Fig. S5](#)). In a minority of datasets analysed, removed target transcripts exhibited significantly less repression than target transcripts, but nonetheless exhibited greater repression than annotated non-target transcripts. In these datasets, the removed target log fold change distribution tended to align with the non-target distribution at the negative extremity, but not at small negative fold change value ranges—indicating that for a minority of datasets, labelled 'removed targets' may be mildly repressed by targeting miRNAs. Additional analysis demonstrated that for these datasets, such targets exhibited significantly weaker repression in response to miRNA transfection than 6-mer

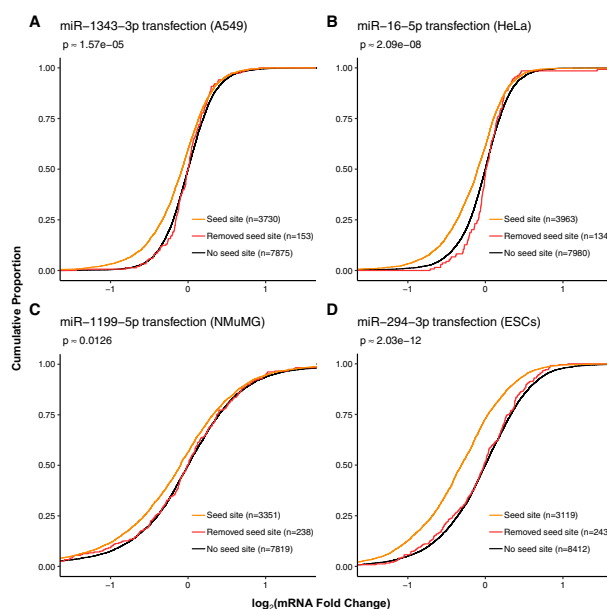


Fig. 3. 3'-UTR truncation by FilTar leads to the removal of false-positive miRNA target predictions. Curves are plotted of the cumulative \log_2 fold change distributions of expression filtered: (i) protein-coding non-target transcripts (black), (ii) protein-coding seed target transcripts (orange) and (iii) predicted target transcripts deriving from Ensembl 3'-UTR annotations but not FilTar 3'-UTR annotations (red). Approximate *P*-values were computed using one-sided, two-sample, Kolmogorov–Smirnov tests between non-target and discarded miRNA target fold change distributions. Otherwise as in [Figure 1](#)

targets, which are the weakest canonical miRNA target site type ([Bartel, 2018](#)) ([Supplementary Fig. S6](#)).

3.4 Cumulative effect of filtering and reannotation

When the FilTar reannotation and miRNA target prediction workflow was applied transcriptome-wide, to multiple organs and cell lines, using all annotated miRBase human miRNAs, there was a mean average gain and loss of miRNA target sites corresponding to 0.18% and 1.5% of the total original miRNA target sites predicted deriving from Ensembl 3'-UTR annotations ([Fig. 4](#)). This corresponds to a gain and loss of total miRNA seed sites in the tens and hundreds of thousands, respectively ([Supplementary Table S4](#)). Although a much larger proportion of miRNA seed sites (mean average of 26.3%) are lost through expression filtering ([Supplementary Fig. S7](#)), representing a loss of millions of miRNA seed sites ([Supplementary Table S4](#)). This is commensurate with the mean average of 34.0% of 3'-UTR bases lost when removing lowly expressed transcripts (<0.1 TPM) from target predictions ([Supplementary Table S5](#)), which is greater than the mean average of 2.0% of bases lost through 3'-UTR reannotation ([Supplementary Table S6](#)). When considering the combined effect of expression filtering and 3'-UTR reannotation, a mean average 36.1% of 3'-UTR bases are lost, affecting a mean average of 53.4% of protein-coding 3'-UTRs ([Supplementary Table S7](#)).

4 Discussion

Results show that FilTar is successfully able to utilize RNA-Seq data to reannotate protein-coding 3'-UTR sequences and filter based on expression data leading to a gain in specificity and sensitivity of target prediction evidenced through tests using experimental data.

Expression filtering target transcripts at even a modest expression threshold of 0.1 TPM leads to a loss of millions of seed sites in most datasets analysed ([Supplementary Table S4](#)), representing a radical reduction in the number of false-positive predictions associated with miRNA target prediction. This is indicative of the

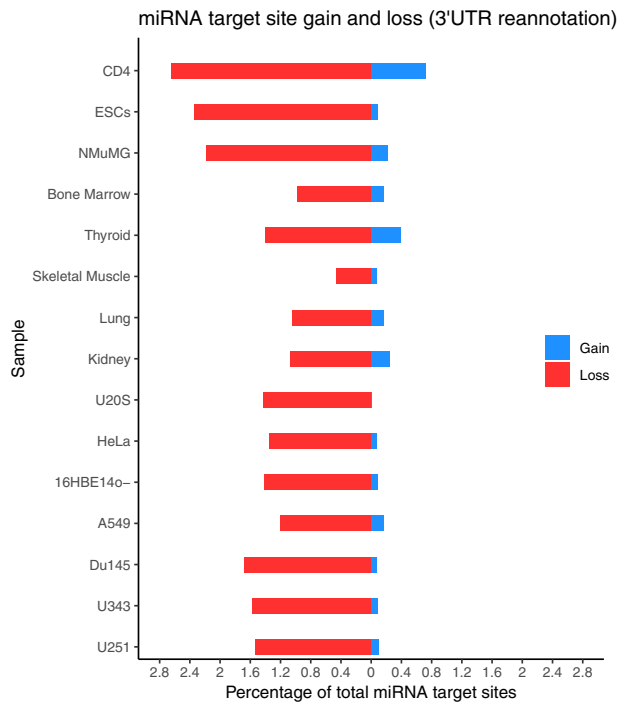


Fig. 4. Total miRNA target site gain and loss when applying FilTar to multiple sample types. FilTar is applied to the protein-coding transcriptome for all annotated human miRNAs for multiple tissues, organs and cell lines. Gained (blue) and lost (red) miRNA target sites are expressed as a percentage of the total number of target sites identified when deriving miRNA from Ensembl 3'-UTR annotations

importance of considering the biological plausibility of candidate miRNA interactions. The positive relationship between the expression threshold chosen and the extent of repression of retained mRNA transcripts is evidence for the robustness of this effect (Supplementary Fig. S3a). The increase in specificity conferred by expression filtering does, however, seem to be accompanied by a corresponding loss of sensitivity of miRNA target prediction when large expression threshold values are chosen (Supplementary Fig. S3b), indicating that sufficient caution ought to be exercised by the user when choosing expression threshold values. However, even for larger expression thresholds, the reduction in sensitivity is less than the increase in specificity conferred by expression filtering (Supplementary Fig. S3a).

The number of newly predicted miRNA target sites deriving from FilTar elongated 3'-UTR sequences is generally relatively low. For cell line datasets analysed, the maximum of number of newly predicted miRNA targets made for any single miRNA was 63, with the majority of datasets analysed yielding less than 15 newly predicted targets (Fig. 2 and Supplementary Fig. S4). The number of newly identified target transcripts is commensurate with the universally low proportion of 3'-UTRs extended, and the small proportion of bases added to the total of the 3'-UTR annotation (Supplementary Table S6), even though this still represents a substantial increase in the number of miRNA seed target sites identified. This is in contrast to 3'-UTR truncation in which the proportion of 3'-UTRs truncated and bases removed from the 3'-UTR annotation total are much greater. Analysis shows that there is a strong positive correlation between the number of 3'-UTR bases reannotated, and the number of predicted miRNA target sites gained or lost through reannotation (Supplementary Fig. S8a and b). The bias in 3'-UTR truncation as opposed to elongation can possibly be explained by either a pre-existing bias in standard Ensembl 3'-UTR annotations to generate long 3'-UTR models, or rather a bias in the FilTar reannotation workflow for 3'-UTR truncation rather than elongation. A potential bias in the standard Ensembl annotation workflow could potentially be explained by the method of transcript annotation, in which, although transcript models are built on a tissue-

specific basis, transcript models incorporated into the final Ensembl gene set typically only derive from the merging of RNA-sequencing reads from multiple different tissue samples (Aken *et al.*, 2016), therefore, creating a bias towards the annotation of longer 3'-UTRs. This effect may be exacerbated or supplemented by the existence of 3'-UTR isoforms within a given sample and transcript—creating relatively low abundance isoforms towards the distal end of the 3'-UTR, making annotation difficult and likely generating a large amount of uncertainty, biases and variability in different methods used to model 3'-UTRs.

Another possibility is that the shortening and extension of existing 3'-UTR annotations are qualitatively different problems requiring different respective sequencing depths. Within a given sample, a read sampling analysis demonstrates that there is a positive relationship, up to a point of saturation between sequencing depth and the number of bases used to elongate existing 3'-UTRs (Supplementary Fig. S9a). In addition, the saturation point for the addition of bases to 3'-UTRs is still substantially less than the proportion of bases removed at 3'-UTRs even at relatively low sequencing depths indicating that the discrepancy between proportion of 3'-UTR bases added or subtracted from the 3'-UTRs cannot be explained by insufficient sequencing depth. A similar positive relationship is observed between sequencing depth and the number of bases truncated from existing 3'-UTRs (Supplementary Fig. S9b), although far fewer reads seem to be required for saturation to occur, indicating a weaker reliance on sequencing depth for 3'-UTR truncation compared to 3'-UTR elongation.

Although as mentioned previously, the sequencing depth does seem to influence the extent of 3'-UTR reannotation, for a set of different biological samples, sequencing depth alone seems to have limited predictive value for this variable (Supplementary Fig. S10a and b). The likely explanation being that as well as sequencing depth, the extent of 3'-UTR reannotation is also determined by other key variables such as the cell type being analysed, read length used for sequencing, library preparation protocol, the use of single-end or paired-end sequencing, as well as additional researcher or lab-specific batch effects (Leek *et al.*, 2010). For example, as some cell types are biased towards shorter 3'-UTRs (Mayr and Bartel, 2009), while others towards longer 3'-UTRs (Miura *et al.*, 2013), generating radically different reannotation statistics irrespective of sequencing depth used.

As mentioned previously, there was generally a much larger number of miRNA target sites predicted to be removed than added during 3'-UTR reannotation. This is despite FilTar permitting 3'-UTR truncations only occurring on moderately-to-highly expressed transcripts, after discovery that the reannotation of the 3'-UTRs of lowly expressed transcripts generated a relatively large number of what seemed to be false-positive predictions (Supplementary Fig. S11). The likely cause being that low transcript expression leads to sporadic and inconsistent coverage across the 3'-UTR, in which there is insufficient information to correctly call 3'-UTR truncation. The default behaviour of the FilTar tool therefore is to only truncate the 3'-UTRs of transcripts which are not poorly expressed (*i.e.* TPM ≥ 5).

When examining 3'-UTR truncations further, for a minority of datasets analysed, some removed predicted miRNA targets seem to be marginally effective, with some transcripts exhibiting low levels of repression upon transfection of the miRNA mimic. Further analysis indicates that these marginally repressed transcripts exhibit even weaker repression than 6-mer targeted transcripts (Supplementary Fig. S6), one of the least effective canonical miRNA target types (Bartel, 2018), indicating that the efficacy of these site types is marginal. A possible explanation for the existence of these site types is that, for some transcript annotations for which the 3'-UTR was truncated, there may exist a small proportion of isoforms with longer 3'-UTRs, which are too low in abundance to be detected by APAtap, but nonetheless still confer a marginal level of repression to the transcript, and hence is detectable when analysing experimental data.

Investigations into the effect of utilising expression data when making transcriptome-wide miRNA target predictions can be

extended by closer examination of not only the refinement of 3'-UTR annotations across different biological contexts, and its effects on miRNA target prediction, but more precisely the definition of specific 3'-UTR profiles, incorporating information about 3'-UTR isoforms within a given cellular context (Agarwal *et al.*, 2015). This enables the weighting of miRNA target prediction scores on the basis of sequencing data applied by the user themselves, enabling even further and extended tailoring of miRNA target prediction to the specific biological context being researched. Previous analyses indicate that the most effective target predictions occur when those predictions are weighted on the basis of 3'-UTR isoform ratios (Nam *et al.*, 2014). In addition, the scope of FilTar's functionality can be increased by enabling the annotation of novel 3'-UTR sequences for transcripts without a current annotated 3'-UTR, and also for those 3'-UTRs which themselves span multiple exons. In addition, both the configurability and precision of FilTar can be improved in the future by, respectively, enabling use of additional tools for 3'-UTR reannotation (Gruber *et al.*, 2018a, b) and exploring the greater transcriptomic resolutions enabled by nascent single cell sequencing technologies.

5 Conclusion

FilTar utilizes RNA-Seq data to increase the accuracy of miRNA target predictions in animals by filtering for expressed mRNA transcripts and reannotating 3'-UTRs for greater specificity to a given cellular context of interest to the researcher. FilTar's compatibility with user-generated RNA-Seq data confers functionality across a wide range of potential biological contexts.

Acknowledgements

We would like to thank Daniel Mapleson, Robert Davey, Tamas Dalmay and members of the Dalmay Lab for helpful comments and discussion. We would like to thank Dagnė Daškevičiūtė for help with the identification of appropriate miRNA mimic transfection datasets. We would also like to thank Leighton Folkes for beta testing the tool.

Funding

This work has been supported by the BBSRC Norwich Research Park Biosciences Doctoral Training Partnership [grant number BB/J014524/1 to T.B.]. This research was supported in part by the University of East Anglia high-performance computing (HPC) team, NBI Computing infrastructure for Science (CiS) group and the Earlham Institute (EI) Scientific Computing group through use of HPC and data storage resources, and assistance provided for the use of these resources.

Conflict of Interest: none declared.

References





Agarwal,V. *et al.* (2015) Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, **4**, e05005.
 Aken,B.L. *et al.* (2016) The Ensembl gene annotation system. *Database*, **2016**, baw093.
 Bartel,D.P. (2018) Metazoan microRNAs. *Cell*, **173**, 20–51.
 Birney,E. (2004) An overview of Ensembl. *Genome Res.*, **14**, 925–928.
 Blanchette,M. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708–715.
 Bray,N.L. *et al.* (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**, 525–527.
 Chi,S.W. *et al.* (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, **460**, 479–486.
 Cock,P.J. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
 Cunningham,F. *et al.* (2019) Ensembl 2019. *Nucleic Acids Res.*, **47**, D745–D751.

Elkon,R. *et al.* (2013) Alternative cleavage and polyadenylation: extent, regulation and function. *Nat. Rev. Genet.*, **14**, 496–506.
 Enright,A.J. *et al.* (2003) MicroRNA targets in *Drosophila*. *Genome Biol.*, **5**, R1.
 Friedman,R.C. *et al.* (2008) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, **19**, 92–105.
 Garcia,D.M. *et al.* (2011) Weak seed-pairing stability and high target-site abundance decrease the proficiency of *lscy-6* and other microRNAs. *Nat. Struct. Mol. Biol.*, **18**, 1139–1146.
 Griffiths-Jones,S. (2004) The microRNA registry. *Nucleic Acids Res.*, **32**, D109–D111.
 Grimson,A. *et al.* (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, **27**, 91–105.
 Gruber,A.J. *et al.* (2018a) Terminal exon characterization with TECtool reveals an abundance of cell-specific isoforms. *Nat. Methods*, **15**, 832–836.
 Gruber,A.J. *et al.* (2018b) Discovery of physiological and cancer-related regulators of 3' UTR processing with KAPAC. *Genome Biol.*, **19**, 44.
 Gumienny,R. and Zavolan,M. (2015) Accurate transcriptome-wide prediction of microRNA targets and small interfering RNA off-targets with MIRZA-G. *Nucleic Acids Res.*, **43**, 1380–1391.
 Harrison,P.W. *et al.* (2019) The European Nucleotide Archive in 2018. *Nucleic Acids Res.*, **47**, D84–D88.
 Helwak,A. and Tollervey,D. (2014) Mapping the miRNA interactome by cross-linking ligation and sequencing of hybrids (CLASH). *Nat. Protoc.*, **9**, 711–728.
 John,B. *et al.* (2004) Human microRNA targets. *PLoS Biol.*, **2**, e363.
 Karagkouni,D. *et al.* (2018) DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA-gene interactions. *Nucleic Acids Res.*, **46**, D239–D245.
 Kent,W.J. *et al.* (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
 Khorshid,M. *et al.* (2013) A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. *Nat. Methods*, **10**, 253–255.
 Kim,D. *et al.* (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.
 König,J. *et al.* (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.*, **17**, 909–915.
 Köster,J. and Rahmann,S. (2012) Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520–2522.
 Kozomara,A. *et al.* (2019) miRBase: from microRNA sequences to function. *Nucleic Acids Res.*, **47**, D155–D162.
 Krueger,F. (2015) *Trim Galore. A Wrapper Tool around Cutadapt and FastQC to Consistently Apply Quality and Adapter Trimming to FastQ Files*. http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.
 Kudla,G. *et al.* (2011) Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast. *Proc. Natl. Acad. Sci. USA*, **108**, 10010–10015.
 Kuhn,D.E. *et al.* (2008) Experimental validation of miRNA targets. *Methods*, **44**, 47–54.
 Leek,J.T. *et al.* (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.*, **11**, 733–739.
 Leinonen,R. *et al.* (2010) The sequence read archive. *Nucleic Acids Res.*, **39**(suppl_1), D19–D21.
 Leinonen,R. *et al.* (2011) The European nucleotide archive. *Nucleic Acids Res.*, **39**, D28–D31.
 Lewis,B.P. *et al.* (2003) Prediction of mammalian microRNA targets. *Cell*, **115**, 787–798.
 Lewis,B.P. *et al.* (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.
 Li,B. *et al.* (2010) RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, **26**, 493–500.
 Li,H. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
 Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.*, **17**, 10–12.
 Mayr,C. and Bartel,D.P. (2009) Widespread shortening of 3' UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, **138**, 673–684.
 Miura,P. *et al.* (2013) Widespread and extensive lengthening of 3' UTRs in the mammalian brain. *Genome Res.*, **23**, 812–825.
 Nam,J.W. *et al.* (2014) Global analyses of the effect of different cellular contexts on microRNA targeting. *Mol. Cell*, **53**, 1031–1043.

- Patro,R. *et al.* (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, **14**, 417–419.
- Pruitt,K.D. *et al.* (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
- Pruitt,K.D. *et al.* (2014) RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.*, **42**, D756–D763.
- Quinlan,A.R. (2014) BEDTools: the Swiss-army tool for genome feature analysis. *Curr. Protoc. Bioinformatics*, **47**, 11.12.1–11.12.34.
- Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
- Reczko,M. *et al.* (2012) Functional microRNA targets in protein coding sequences. *Bioinformatics*, **28**, 771–776.
- Ritchie,W. and Rasko,J.E. (2014) Refining microRNA target predictions: sorting the wheat from the chaff. *Biochem. Biophys. Res. Commun.*, **445**, 780–784.
- Schneider,V.A. *et al.* (2017) Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.*, **27**, 849–864.
- Tian,B. and Manley,J.L. (2017) Alternative polyadenylation of mRNA precursors. *Nat. Rev. Mol. Cell Biol.*, **18**, 18–30.
- Van Nostrand,E.L. *et al.* (2016) Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods*, **13**, 508–514.
- Wang,L. and Yi,R. (2014) 3' UTRs take a long shot in the brain. *Bioessays*, **36**, 39–45.
- Wang,X. (2016) Improving microRNA target prediction by modeling with unambiguously identified microRNA-target pairs from CLIP-ligation studies. *Bioinformatics*, **32**, 1316–1322.
- Ye,C. *et al.* (2018) APATrap: identification and quantification of alternative polyadenylation sites from RNA-seq data. *Bioinformatics*, **34**, 1841–1849.

OPEN

Divergence in Transcriptional and Regulatory Responses to Mating in Male and Female Fruitflies

Emily K. Fowler ^{1*}, Thomas Bradley ^{1,2}, Simon Moxon ¹ & Tracey Chapman ¹

Mating induces extensive physiological, biochemical and behavioural changes in female animals of many taxa. In contrast, the overall phenotypic and transcriptomic consequences of mating for males, hence how they might differ from those of females, are poorly described. Post mating responses in each sex are rapidly initiated, predicting the existence of regulatory mechanisms in addition to transcriptional responses involving *de novo* gene expression. That post mating responses appear different for each sex also predicts that the genome-wide signatures of mating should show evidence of sex-specific specialisation. In this study, we used high resolution RNA sequencing to provide the first direct comparisons of the transcriptomic responses of male and female *Drosophila* to mating, and the first comparison of mating-responsive miRNAs in both sexes in any species. As predicted, the results revealed the existence of sex- and body part-specific mRNA and miRNA expression profiles. More genes were differentially expressed in the female head-thorax than the abdomen following mating, whereas the opposite was true in males. Indeed, the transcriptional profile of male head-thorax tissue was largely unaffected by mating, and no differentially expressed genes were detected at the most stringent significance threshold. A subset of ribosomal genes in females were differentially expressed in both body parts, but in opposite directions, consistent with the existence of body part-specific resource allocation switching. Novel, mating-responsive miRNAs in each sex were also identified, and a miRNA-mRNA interactions analysis revealed putative targets among mating-responsive genes. We show that the structure of genome-wide responses by each sex to mating is strongly divergent, and provide new insights into how shared genomes can achieve characteristic distinctiveness.

Mating is well-known to induce extensive behavioural and physiological changes in animals of many taxa. These include changes to fecundity, longevity, immunity, chemical signalling and sexual receptivity^{1–4}. Post mating responses (PMRs) can be initiated within seconds or minutes or may build up over several hours or days^{5,6}. They also vary in duration and may be sustained either temporarily, or permanently throughout life⁷. PMRs may act to optimise physiological and behavioural processes in mated individuals to facilitate subsequent reproductive effort or behaviour. However, the form of PMRs is expected to diverge significantly between the sexes. For example, in species in which both sexes mate multiply, mated females often show refractory responses associated with lowered willingness to mate and their removal from the mating arena in order to support and sustain the production of fertile eggs, at least until sperm supplies become depleted and sexual receptivity returns. In contrast, mated males are likely to be subject to selection pressures to minimise any refractory period, replenish ejaculates and rapidly re-enter reproductive competition in the mating arena.

PMRs are particularly well-studied in females of *Drosophila melanogaster* (reviewed in^{8,9}). During copulation, a male *D. melanogaster* transfers a suite of well over 150 seminal fluid proteins (Sfps) along with thousands of sperm to the female^{10,11}. Many of the changes which occur in females following mating are induced by Sfps, and the magnitude of female PMRs appears to be dependent upon the quantity and relative composition of Sfps received in the ejaculate^{12,13}. Sfp receipt increases oogenesis, ovulation and feeding¹⁴, reduces siesta sleep¹⁵, increases female aggression towards other females¹⁶ and reduces sexual receptivity toward males¹⁷. Other physiological effects of Sfps include facilitation of sperm storage and retention^{18,19}, changes to immune gene expression^{20,21} and to nutrient and water balance^{22,23}. The adaptive modulation of PMRs appears to be facilitated by the highly precise²⁴ and socially-flexible^{25,26} expression of Sfp-encoding genes.

¹School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, NR4 7TJ, UK. ²Earlham Institute, Norwich Research Park, Norwich, NR4 7UZ, UK. *email: e.fowler@uea.ac.uk

In contrast, in males surprisingly little is known about the physiological or behavioural changes associated with mating. Some type of refractory period is generally noted, as following mating, both sperm and Sfps must be replenished. Data are scarce, but it is suggested that in general seminal fluids are in more limited supply than are sperm themselves^{27–30}. The refuelling of a full complement of Sfps in particular is known in some cases to take time, e.g. over 24 h in *D. melanogaster* males^{29–31}. Potentially associated with this, physiological changes to the ejaculatory duct have also been noted in mated *D. melanogaster* males^{32,33}. Aside from these responses, there is some evidence from targeted gene expression studies that males invest in immune response molecules following mating, similarly to females³⁴. In contrast, while females increase and adapt their nutritional intake following mating, the same is not true for males³⁵.

Since males and females share the vast majority of their genome, sexually dimorphic traits such as PMRs are predicted to arise from the sex-specific regulation of gene expression³⁶. Hence measurements of genome-wide, mating-responsive gene expression profiles can provide insight into the underlying mechanisms involved in PMRs in both sexes. Transcriptomic profiles of PMRs have been generated for females of multiple insect species, e.g. the Mediterranean fruitfly *Ceratitis capitata*³⁷; the honey bee *Apis mellifera*³⁸; mosquitoes *Anopheles gambiae*³⁹ and *Aedes aegypti*⁴⁰ and the seed beetle *Callosobruchus maculatus*⁴¹. In transcriptomic studies of PMRs in female *D. melanogaster*, several have focussed on profiling responses in the whole female fly^{42–47}. Others have profiled individual body parts, such as the female reproductive tract^{48–50}, or heads⁵¹ which has helped to reveal additional complexity which can sometimes be obscured by whole body arrays and profiles⁵². Several other studies have also profiled the responses of females to the receipt of sperm or Sfps^{44,45,53,54}. Collectively, this work has revealed that PMRs can induce pervasive, genome-wide gene expression changes in reproductive, sensory and immune system genes with some similarities between signatures of mating and aging processes⁴⁶.

In contrast to females, transcriptomic studies of PMRs in males are scarce. Gene expression profiles of male and female *C. capitata* and *C. maculatus* revealed the presence of distinct, sex-specific transcriptional responses to mating^{37,41}. However, for *D. melanogaster*, no direct comparison of the transcriptomic mating response by males and females has previously been undertaken, and data on male responses are restricted to a single study using head tissue⁵⁵.

The timing of the different facets of PMRs in both sexes is also highly variable and distinct. This suggests that mechanisms in addition to expression changes in coding genes are likely to be important contributors to PMRs and should themselves show sex-specificity. Some PMRs are extremely rapid and may rely upon the release of neurotransmitters⁵⁶ or the actions of regulatory molecules. We have scant data so far of these aspects of PMRs, particularly in how changes in gene regulation versus gene expression are linked. Hence a significant part of our mechanistic understanding of the responses of both sexes to mating is still missing. Consistent with the idea that PMRs are achieved by a range of qualitatively different responses, recent data in female *D. melanogaster* show that regulatory molecules, such as miRNAs, can also change in response to mating^{46,57}. miRNAs are a group of small non-coding RNA molecules which play a key role in post-transcriptional gene regulation by binding complementary mRNA transcripts, inhibiting their translation into peptides. Well-known for their role in gene regulation during development, miRNAs are increasingly implicated in the expression of adult phenotypes, including the regulation of Sfps²⁴, male and female fertility and ovary morphology⁵⁸. These recent findings predict significant changes to the expression of coding and regulatory non-coding genes following mating in both sexes, but as yet there has been no genome-wide analysis of non-coding RNA responses to mating in males. Study of the expression profiles of miRNAs in tandem with mRNAs also offer the potential for new insights into the regulatory processes underlying the changes in transcript abundance.

In this study, we addressed the omissions noted above by testing two predictions: (i) that there are significant changes to the expression of both coding and regulatory non-coding genes between virgin and mated flies in both sexes, and (ii) that the mode and nature of PMR gene expression profiles of each sex are markedly different. The data supported both predictions. For the female head-thorax (HT) and male abdomen (Ab) tissues, >2000 genes were differentially expressed (DE) between virgin and mated status. Interestingly, for the female HT the majority of DE genes were downregulated following mating, while many of the same genes were upregulated in the mated male Ab. In contrast, only 125 genes were DE after mating in the female Ab, while mating did not significantly impact gene expression in the male HT. The magnitude of genome-wide change showed sex specificity and was much greater in females, with ~50% vs 15% of DE genes showing greater than two-fold change in females vs males, respectively. We identified novel mating-induced miRNAs in the abdomens of both sexes, with changes occurring in fewer miRNAs in males than in females.

Results

Sequencing QC. To determine the transcriptomic profiles of mated and virgin flies, we conducted high-throughput RNA sequencing (RNA-seq). We extracted the mRNA and small RNA (sRNA) fractions from a total of 16 samples, consisting of two treatments (virgin vs. three hours post-mated), two sexes, two body-parts (HT and Ab) and two biological replicates. FASTQ files generated from the sequencing reads were checked using FastQC (Babraham Bioinformatics) and no significant quality issues were discovered. RNA-seq reads had an average pseudo-alignment rate of 77.61% to the transcriptome (min 64.63%, max 85.1%, Supplementary Table S1a), and sRNA reads had an average of 84.97% alignment rate to the genome (min 82.49%, max 88.63%, Supplementary Table S1b). PCA plots showed clustering by sample for both mRNA-seq and sRNA-seq datasets (Fig. 1). Variation between replicates was generally very low. During differential expression analysis using the DESeq2 package⁵⁹, coding genes or miRNA exhibiting high variability in expression values between replicates are penalised, so genes are only called DE if the variation between treatments is over and above that attributable to the replicates. Size distribution profiles of sRNA read length showed peaks at 22 and 23 nt, corresponding to the expected length for miRNAs (Supplementary Table S1c). In previous *Drosophila* sequencing studies, we and others^{46,60} have noted a read length peak at 30 nt, accounting for >90% of the total reads. This fraction may contain

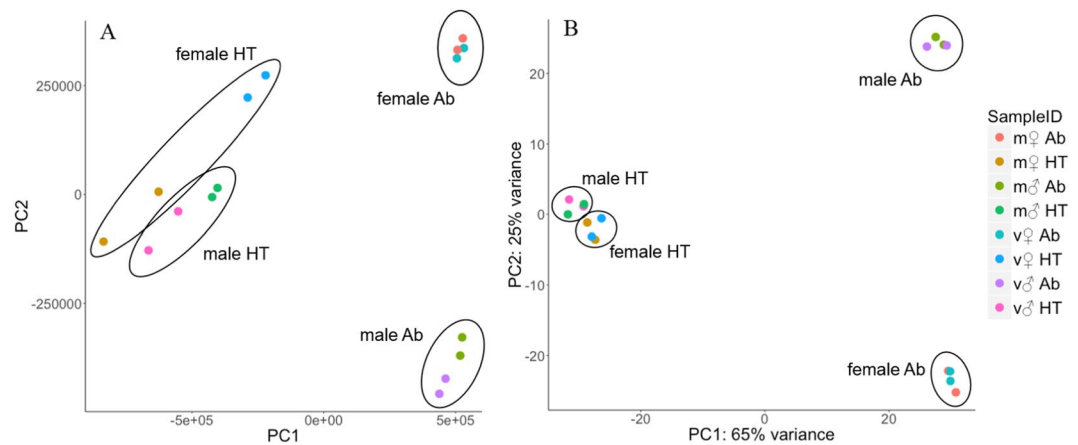


Figure 1. Principle component analyses of mRNA-sequencing (A) and miRNA sequencing data (B). Points are coloured by sample type: mated (m), virgin (v), male (δ), female (\varnothing), abdomen tissue (Ab) or head-thorax tissue (HT).

	Significance threshold	Total DE genes	Upregulated in mated flies (>2 fold)	Downregulated in mated flies (>2 fold)
mRNA				
♀ abdomen	$p < 0.05$	628	475 (110)	153 (32)
	$FDR q < 0.05$	125	106 (50)	19 (12)
♀ head-thorax	$p < 0.05$	3372	1206 (224)	2166 (1036)
	$FDR q < 0.05$	2040	628 (144)	1412 (804)
♂ abdomen	$p < 0.05$	3607	2273 (355)	1334 (38)
	$FDR q < 0.05$	2068	1507 (296)	561 (19)
♂ head-thorax	$p < 0.05$	329	265 (19)	64 (2)
	$FDR q < 0.05$	0	0	0
miRNA				
♀ abdomen	$p < 0.05$	16	12	4
	$FDR q < 0.05$	4	3	1
♀ head-thorax	$p < 0.05$	10	4	6
	$FDR q < 0.05$	0	0	0
♂ abdomen	$p < 0.05$	9	2	7
	$FDR q < 0.05$	2	0	2
♂ head-thorax	$p < 0.05$	2	2	0
	$FDR q < 0.05$	0	0	0

Table 1. Total numbers of mating-responsive mRNAs and miRNAs in male (δ) and female (\varnothing) head-thorax and abdomen tissues. Numbers of significant DE genes with a fold-change difference of >2 are shown in parentheses.

some longer sRNA species such as piRNAs, but is dominated by reads corresponding to the highly abundant 30 nt 2S rRNA fraction found in insect species. To remove this fraction, we incorporated a complimentary “blocking oligo” into the library construction⁶⁰, which successfully prevented adaptor ligation of 2S rRNA, thus increasing the proportion of reads derived from miRNAs.

Mating-Responsive mRNAs. The total numbers of genes showing DE in response to mating varied widely across sex and tissue type (Table 1; Supplementary Table S2a–d), as did the magnitude of the fold change in expression. The female HT and male Ab tissues had the greatest number of mating-responsive genes, with over 2000 in each case. In contrast, the number of genes affected by mating was much lower in the female Ab, in which only 125 genes were DE between the virgin and mated treatments. Strikingly, no genes were DE between the HT of virgin versus mated males with a q-value of < 0.05. Replicate to replicate variability of virgin and mated male head-thorax samples was comparable with other samples (Supplementary Fig. S1). Therefore, the absence of statistically significantly differentially expressed genes cannot be explained by replicate variability and appears to be a biologically relevant effect. The magnitude of change in gene expression was greater in females than males, with ~50% of DE genes showing greater than two-fold change regardless of body-part. Gene expression changes in the male abdomen, though involving numerous genes, were more subtle, with only 15% of DE genes showing over two-fold change. To detect signatures of enriched function amongst the DE genes we performed gene ontology

(GO) enrichment analyses on all sets of mating responsive genes (Table 2, Supplementary Table S3a–h). We also carried out GO analyses on the subset of genes in each sex/tissue type which exceeded a fold-change threshold of two (Supplementary Table S4a–f).

DE in the female abdomen. 125 protein coding genes were responsive to mating in female abdomens. Most of these (106) were upregulated in mated females. A GO enrichment analysis of all upregulated genes revealed a significant over-representation (FDR q -value < 0.05) of 30 biological process terms (Supplementary Table S3a). Many of the terms were related to translation and peptide synthesis due to the presence of 17 ribosomal protein encoding genes. More generally, “protein metabolism” was enriched which, aside from ribosomal protein genes, involved 16 peptidase-encoding genes, including the spermathecal endopeptidases *Send1* and *Send2*. At least six terms were related to “multi-organism process” and two terms involving response to heat were also enriched, generated by the presence of 13 genes encoding immune system, or stress response proteins. Additionally, the term “electron transport chain” was enriched, associated with six genes – *blw*, *CG3835*, *CG3731*, *CG4169*, *ND75* and *RFeSP*. Of the 106 significantly upregulated genes, 50 exceeded the two fold-change threshold. A GO analysis of these revealed four significant biological process terms, three of which were related to “response to bacterium”, and involved eight different genes. The other enriched term, “proteolysis”, consisted of 12 genes (Supplementary Table S4a). Of the 19 genes which were downregulated in mated female abdomens, there were no terms with an FDR q -value < 0.05 . However, five genes related to “carbohydrate metabolic process” were present in this subset: *CG32444*, *Mal-A1*, *Mal-A7* and *Mal-A8*, and *tobi* (Supplementary Table S3b).

DE in the female head-thorax. 2040 genes showed DE between virgin and mated females in the HT. Of these, 628 had higher expression in mated females. An enrichment analysis of the upregulated genes did not return any terms with an FDR q -value < 0.05 , although terms associated with ncRNA processing fell just below the significance threshold (Supplementary Table S3c). Excluding low fold-change (< 2 FC) DE genes from the GO analysis highlighted enrichment in two terms related to rRNA processing (Supplementary Table S4b). Of the 1412 downregulated genes, 141 biological process terms were significantly over-represented (Supplementary Table S3d). Almost all were linked to metabolic processes involving the generation of precursor metabolites and energy and the oxidation-reduction process, nucleoside phosphate metabolic process, and organonitrogen compound metabolic process. There were also terms associated with carbohydrate metabolism and oxoacid metabolism. Specific terms which were significantly enriched included “translation”, “ATP biosynthetic process”, “glycolytic process”, “muscle contraction” and “drug metabolic process”. More than half of the downregulated HT genes exceeded the two fold-change threshold. A GO analysis on this subset returned 106 enriched biological process terms, again mostly related to organonitrogen compound biosynthesis and energy metabolism, as well as carbohydrate metabolism and many other metabolic processes (Supplementary Table S4c).

DE in the male abdomen. 2068 protein-coding genes were DE between virgin and mated male abdomens, and 1507 of these were upregulated in the mated flies. A GO enrichment analysis of the upregulated genes returned 97 biological process terms with an FDR q -value < 0.05 (Supplementary Table S3e). At least 35 of the terms were related to the transport and localization of organic substances and proteins, and protein folding. Another two terms were related to proteolysis. The remaining terms were connected to metabolic processes. Similarly to the downregulated genes in the female HT, these included terms related to translation and energy generation. Additionally, the term “translational initiation” was enriched, driven by the presence of genes encoding eukaryotic translation initiation factor. Only 20% of the upregulated male abdomen genes had a fold change of over two, but this subset was enriched for 32 terms, mostly generated by the presence of ~50 ribosomal protein genes (Supplementary Table S4d). These terms included “translation” and “organonitrogen compound biosynthesis”. Terms involving energy generation also remained overrepresented. Among the higher fold-change subset, “defence response to Gram-positive bacterium” was enriched, whereas the terms connected to protein transport and folding were absent. The 561 downregulated genes were not significantly enriched for any biological process terms, although the term “response to nutrient levels” fell just below the significance threshold (Supplementary Table S3f). Only 19 genes were over two-fold DE in the downregulated set. A GO analysis of this subset also did not return any significantly enriched terms. However, six terms had a p -value of > 0.05 and were all connected to glutamate receptor signalling, involving four genes – *Rdl*, *Syt1*, *CG32447* and *mtt* (Supplementary Table S4e).

DE in the male head-thorax. Strikingly, no mating-responsive genes met the stringent q -value < 0.05 threshold in the comparison of virgin and mated male head-thorax tissue. However, a GO enrichment analysis on 329 DE genes with a p -value < 0.05 showed that, similarly to the male abdomen, terms involving ribosomal protein genes, such as “translation” were over-represented in the 264 upregulated genes (Supplementary Table S3g). When we analysed the 19 upregulated genes with greater than two-fold DE, 11 terms associated with defence response were enriched, represented by five genes – *AttB*, *CecA2*, *LysX*, *CecC* and *Drsl2* (Supplementary Table S4f). GO analysis of the 64 downregulated genes returned terms involving molybdopterin cofactor processing, caused by three genes – *cin*, *Mocs2* and *CG42503* (Supplementary Table S3h). All but two of the downregulated genes had a fold change of over two, so no further GO analyses were carried out on these genes.

Comparison of sex- and tissue-specific profiles of mRNAs. We compared the mating responsive genes in each sex and tissue type to one another, to examine differences in the transcriptomic profiles between males and females and between different body parts (Fig. 2). For comparisons of female HT and abdomen tissues, and female and male abdomens, we were able to use the set of DE genes with an FDR q -value of < 0.05 (Table 1). However, since there were no such genes falling below that cut-off in the male HT, we produced extended DE gene lists with a p -value of < 0.05 for the male HT, Ab and the female HT. This less stringent threshold for DE calling allowed us

Biological process term	♀ Ab (up) gene count	♀ HT (down) gene count	♂ Ab (up) gene count
electron transport chain	6	44	35
translation	17	79	91
organic substance biosynthetic process	25	227	231
organonitrogen compound metabolic process	45	395	436
primary metabolic process	56	543	589
organic substance metabolic process	60	599	643
metabolic process	67	694	720
mitochondrial protein processing		5	5
cell redox homeostasis		16	19
generation of precursor metabolites and energy		70	44
cofactor metabolic process		58	48
organophosphate metabolic process		73	63
oxidation-reduction process		136	93
cellular amide metabolic process		94	120
macromolecule biosynthetic process		117	125
organonitrogen compound biosynthetic process		153	148
small molecule metabolic process		168	149
cellular nitrogen compound biosynthetic process		154	160
cellular biosynthetic process		222	223
biosynthetic process		230	234
cellular nitrogen compound metabolic process		260	293
cellular macromolecule metabolic process		281	316
nitrogen compound metabolic process		481	535
cellular metabolic process		572	613
cellular process		843	899
cellular response to unfolded protein			8
translational initiation			15
cellular component biogenesis			16
Golgi organization			28
rRNA metabolic process			30
Golgi vesicle transport			31
protein folding			48
protein-containing complex subunit organization			90
macromolecule localization			109
organic substance transport			123
establishment of localization in cell			124
proteolysis			129
cellular localization			132
catabolic process			136
cellular protein metabolic process			226
protein metabolic process			329
macromolecule metabolic process			453
regulation of autophagy of mitochondrion		5	
muscle system process		9	
reactive oxygen species metabolic process		11	
cellular aldehyde metabolic process		12	
mitochondrial transport		22	
antibiotic metabolic process		25	
monovalent inorganic cation transport		29	
cellular homeostasis		43	
carbohydrate metabolic process		56	
carbohydrate derivative metabolic process		80	
drug metabolic process		81	
rRNA 3'-end processing	3		
response to heat	7		
multi-organism process	13		

Table 2. Representative GO biological process terms significantly enriched among the mating-responsive genes in the female abdomen (♀ Ab), male head-thorax (♂ HT) and male abdomen (♂ Ab). Term enrichment analysis was performed using GOrilla.

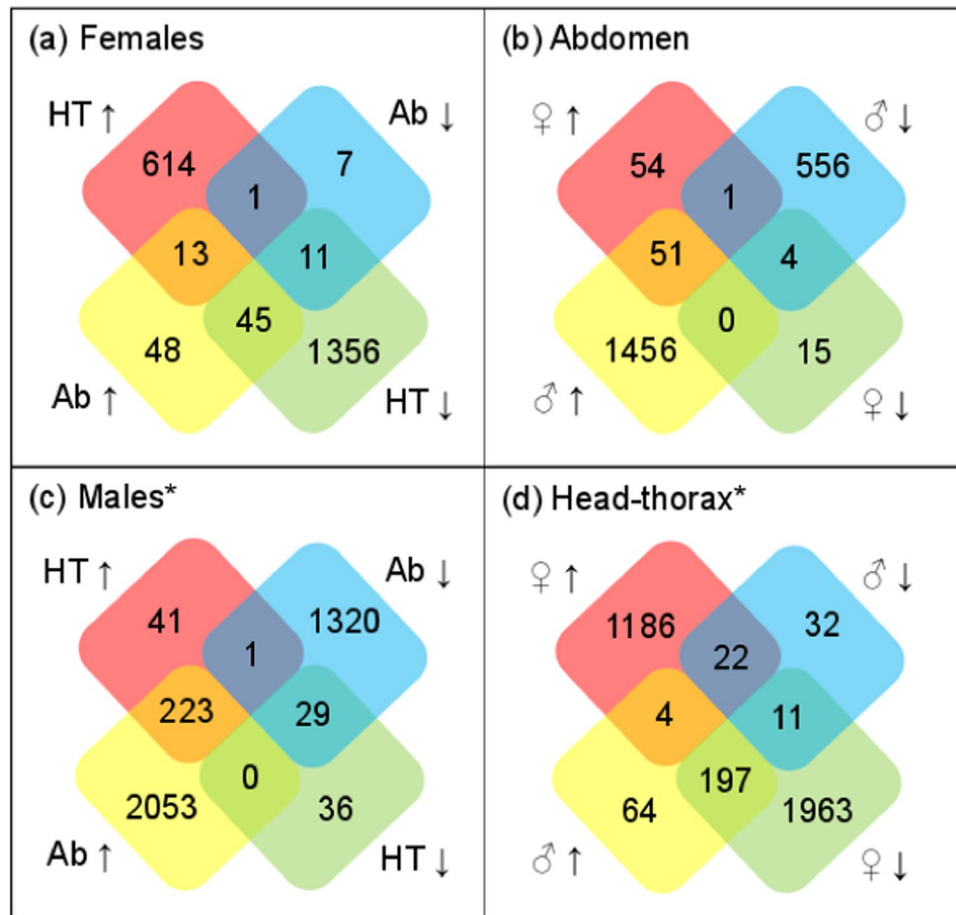


Figure 2. Overlap of the identities of up- and downregulated genes in response to mating in the different transcriptomes. (a) Numbers of DE genes upregulated (↑), or downregulated (↓) in response to mating in female abdomen (Ab) or head-thorax (HT). (b) Numbers of DE genes up- or downregulated in female or male Ab tissues. (c) Numbers of DE genes upregulated, or downregulated in response to mating in male Ab or HT. (d) Numbers of DE genes upregulated, or downregulated in response to mating in female (♀) or male (♂) HT tissue. *DE calling based on a significance threshold of $p < 0.05$.

to explore the most significant changes in gene expression in the mated male HT, and make fair comparisons with the male Ab and female HT. GO analyses were conducted on the overlapping genes (Supplementary Table S5) to detect shared signatures of functional enrichment.

We first compared the DE genes in the female HT and female Ab, and found a total overlap of 70 genes (Fig. 2a). Of these, 13 genes were upregulated in response to mating in both the Ab and HT tissues, including the turandot protein genes *TotA*, *TotC* and *TotX*, and the gene encoding juvenile hormone esterase (*Jhe*) (Supplementary Table S5a). There were also 11 genes downregulated in both tissues in females, including the five genes predicted to be involved in carbohydrate metabolism, mentioned earlier (Supplementary Table S5b). Another 46 genes were differentially expressed in both body-parts, but in opposing directions. All but one of the opposing genes were downregulated in response to mating in the HT, and upregulated in the Ab. Over-represented among these were 14 ribosomal protein genes, and six genes encoding proteins involved in oxidative phosphorylation (*ND-75*, *RFeSP*, *UQCR-C1*, *UQCR-C2*, *COX8*, *blw*) (Supplementary Table S5c).

We next considered whether the mating responsive genes of the female Ab were also DE in the male Ab (Fig. 2b). In this case almost all overlapping genes were expressed in the same direction, i.e. 51 genes were upregulated in both the female and male Ab in response to mating. Again, a GO analysis on the 51 genes revealed an enrichment in the terms “translation”, and “electron transport chain” (Supplementary Table S5d). Using the less stringent DE calling, we next compared all DE genes with a p -value < 0.05 in the male HT, to those from the male Ab (Fig. 2c). In contrast to the female, overlapping DE genes in the male tissues were always DE in the same direction (with the exception of one gene), i.e. 223 genes were upregulated in both HT and Ab, and 29 genes were downregulated in both tissues. Once again, GO terms associated with translation were enriched amongst the 223 overlapping upregulated genes, while there were no significantly enriched terms for the downregulated genes (Supplementary Table S5e,f).

Finally, we compared the male and female HT body parts (Fig. 2d). Unlike in the Ab samples, where overlapping DE genes tended to be expressed in the same direction in both sexes, in the HT they were more likely to

be expressed in opposite directions. For example, we found 197 genes that were upregulated in the male HT, but downregulated in the female HT. GO analysis of the overlapping genes again revealed an enrichment of terms related to translation, and organonitrogen compound metabolism (Supplementary Table S5g).

Comparison of mRNA profiles with existing studies. Several previous studies have investigated transcriptomic changes in response to mating in *D. melanogaster* females, with variation in the tissue type analysed, and the time-point captured following mating (Supplementary Table S6a). To compare the mating-responsive genes found in previous studies with our data here, we selected the studies most closely matching our experimental protocol. For example, we compared our female Ab samples to those studies that used either whole females, or reproductive tracts, and our female HT samples to those using heads or whole females. A summary of the studies and the numbers of mating-responsive genes identified in each one can be found in Supplementary Tables S6a,b.

First we examined the number of mating-responsive genes in our dataset which were also DE in other studies, regardless of the direction in expression change (Supplementary Table S6b–f). For our female abdomen samples, 65% of DE genes were also found to be mating-responsive in at least one other study. For the HT samples there was less, but still considerable, agreement with 29% of our DE genes also responding to mating in at least one other study. When we also considered the direction of DE, for the female Ab samples, 34.9% of our upregulated genes were also upregulated in at least one other study, and 63.2% of downregulated genes were downregulated in other studies. For the HT, 19.4% of genes increasing expression following mating in our study also did so in at least one previous study. For the downregulated genes in the head-thorax, agreement with other studies fell to 8.9% when direction of expression was taken into account. Interestingly, the female HT genes which showed opposing direction of DE in our study compared to previous studies were enriched for genes involved in translation, which we have shown to exhibit opposing expression in the head-thorax in comparison to abdomen samples. Studies measuring transcriptomic responses in whole flies would not have captured these body-part specific patterns.

To identify individual genes consistently found to be mating-responsive throughout the literature, we examined in more detail the DE genes found in our study and at least three previous studies. For upregulated genes in the abdomen, four genes were found to increase expression following mating in three other studies (*Uro*, *jhamt*, *CG17234*, *CG3290*), two were found in four other studies (*su(r)* and *CG17239*) and another two in five other studies (*CG31324* and *Send2*). For upregulated genes in the head-thorax, *CG6910*, *CG3036* and *Uro* were found in three other studies, *CG31324* was found in four other studies and *fit* was found in five other studies.

Sex- and tissue-specific miRNA profiles. Reads from the sRNA-Seq output aligned to 401 mature miRNAs. The principle component analysis revealed that the miRNA profiles of HT and Ab tissues were distinct (Fig. 1). Furthermore, the miRNA profile of male and female Ab body parts differed significantly. This was in contrast to the HT, in which both sexes had similar profiles (Fig. 1). To investigate the sex- or tissue-bias in miRNA expression, we conducted a differential expression analysis between sexes and tissues (Supplementary Table S7a–d). A comparison of HT and Ab body-parts revealed that 83 miRNAs and 71 miRNAs were HT or Ab biased, respectively, in both sexes (Table 3). For the Ab-biased miRNAs, 45 were biased in males only, and 29 in females. There were fewer HT-biased miRNAs which were specific to one sex - 21 miRNAs were HT-biased in males only, and 14 miRNAs were HT-biased in females only. When males were compared to females, as expected from the PCA, most sex-biased miRNAs were specific to the Ab (97 male-biased, 42 female biased), with fewer miRNAs specific to the head-thorax (9 male-biased, 12 female-biased). There were also some miRNAs that were male- or female-biased in both body parts (8 male-biased, 12 female-biased) (Table 4). We examined the identities of the sex- and tissue-biased miRNAs and found strong concordance with a meta-analysis on sex-bias using publicly available sRNA-Seq data⁶¹. Of the 37 female-biased miRNAs identified in the meta-analysis, 31 were also female-biased in this study. Similarly, 26 of the 28 male-biased miRNAs identified were also male-biased here. The fact that the majority of sex-biased miRNAs were also abdomen-biased in this study is consistent with the fact that male- or female-biased miRNAs tend to be expressed in the testes or ovaries, respectively⁶¹.

Mating-Responsive miRNAs. Significantly ($\text{padj} < 0.05$) differentially expressed miRNAs in response to mating were found in the Ab tissue of both males and females (Table 1, Supplementary Table S8a–d). In the female Ab, three miRNAs were significantly upregulated in response to mating (miR-14-3p, miR-997-5p, and miR-184-5p), and one miRNA was downregulated (miR-286-3p). In the male Ab, both strands of miR-927 were significantly downregulated in response to mating. No miRNAs were significantly DE in the HT of either sex.

mRNA-miRNA interactions. A number of mating-responsive mRNA targets of the differentially expressed miRNAs were identified, including genes which were dysregulated in the opposite direction of the targeting miRNA (*i.e.* an upregulated target of a downregulated miRNA, or a downregulated target of an upregulated miRNA) (Supplementary Table S9). Of the upregulated miRNAs, only miR-184-5p in the female abdomen was linked to any significantly downregulated targets, with one gene identified. For downregulated miRNAs, miR-286-3p of the female abdomen contains four significantly upregulated targets. For the male abdomen, there were 50 and 114 significantly upregulated targets for the downregulated miRNAs miR-927-3p and miR-927-5p, respectively. Six of these significantly upregulated genes are predicted to be targeted by both miR-927-3p and miR-927-5p, namely *Rpl37a*, *CG5707*, *Slh*, *twi*, *Su(w[*a*])* and *Nop60B*. Network visualisations of miR-927-3p and miR-927-5p interactions with all putative, mating-responsive mRNA targets (regardless of direction of differential expression) revealed the extensive targeting of both upregulated and downregulated mRNAs, including an additional 13 downregulated mRNAs targeted by both strands of miR-927 (Supplementary Fig. S3a,b). To test whether the predicted targets of miR-927 were functionally linked, we conducted a GO enrichment analysis on

Head-thorax biased miRNAs			Abdomen biased miRNAs		
<i>Both sexes (83)</i>			<i>Both sexes (71)</i>		
bantam-3p, 5p	miR-193-3p, 5p	miR-4952-5p	miR-1007-5p	miR-306-5p	miR-960-3p, 5p
let-7-5p	miR-210-3p, 5p	miR-4960-3p	miR-1012-3p	miR-308-3p, 5p	miR-961-3p, 5p
miR-1000-3p, 5p	miR-219-5p	miR-4968-5p	miR-1015-3p	miR-310-3p	miR-962-3p, 5p
miR-1001-3p, 5p	miR-2489-3p	miR-7-5p	miR-10-3p, 5p	miR-311-3p, 5p	miR-963-3p, 5p
miR-1004-3p, 5p	miR-252-3p, 5p	miR-87-3p, 5p	miR-12-3p, 5p	miR-312-3p, 5p	miR-964-3p, 5p
miR-1005-3p	miR-263b-5p	miR-927-5p	miR-184-5p	miR-313-3p, 5p	miR-982-5p
miR-1006-3p	miR-276a-3p, 5p	miR-929-3p	miR-2494-3p, 5p	miR-314-3p, 5p	miR-983-3p, 5p
miR-1009-3p	miR-276b-3p, 5p	miR-932-3p, 5p	miR-263a-3p, 5p	miR-316-3p, 5p	miR-984-3p, 5p
miR-1017-3p	miR-277-3p, 5p	miR-9383-3p	miR-275-3p	miR-31a-5p	miR-991-3p
miR-11-3p	miR-278-3p	miR-957-3p, 5p	miR-279-3p, 5p	miR-31b-5p	miR-997-5p
miR-124-3p, 5p	miR-284-3p, 5p	miR-969-3p	miR-281-1-5p	miR-33-5p	miR-9b-3p, 5p
miR-125-5p	miR-285-3p, 5p	miR-970-3p, 5p	miR-281-2-5p	miR-4919-3p, 5p	miR-9c-5p
miR-133-3p, 5p	miR-2c-3p, 5p	miR-971-3p, 5p	miR-281-3p	miR-92a-3p	miR-iab-4-3p
miR-137-5p	miR-307a-3p	miR-981-3p, 5p	miR-282-5p	miR-956-3p, 5p	miR-iab-8-5p
miR-13a-3p, 5p	miR-315-3p, 5p	miR-987-5p	miR-283-3p, 5p	miR-958-3p, 5p	
miR-13b-1-5p	miR-317-5p	miR-990-5p	miR-304-3p, 5p	miR-959-3p, 5p	
miR-1-3p, 5p	miR-34-3p, 5p	miR-993-3p, 5p	Females only (29)		
miR-14-3p	miR-4945-5p	miR-998-5p	miR-1008-5p	miR-2a-3p	miR-92b-3p
miR-190-5p	miR-4951-5p	miR-999-3p	miR-1014-5p	miR-2b-2-5p	miR-9372-5p
Females only (14)			miR-11-5p	miR-2b-3p	miR-989-3p, 5p
miR-100-3p	miR-307a-5p	miR-4956-5p	miR-13b-2-3p	miR-306-3p	miR-994-3p, 5p
miR-1007-3p	miR-317-3p	miR-4976-5p	miR-13b-3p	miR-318-3p, 5p	miR-995-3p
miR-1011-3p	miR-3-3p	miR-929-5p	miR-275-5p	miR-4917-3p	miR-996-3p, 5p
miR-274-5p	miR-375-3p, 5p	miR-9a-3p, 5p	miR-282-3p	miR-79-3p, 5p	miR-998-3p
Males only (21)			miR-2a-1-5p	miR-92a-5p	miR-9c-3p
miR-1003-3p	miR-263b-3p	miR-7-3p	Males only (45)		
miR-1010-3p	miR-2a-2-5p	miR-954-3p, 5p	miR-1014-3p	miR-4977-3p	miR-978-3p, 5p
miR-137-3p	miR-2a-3p	miR-969-5p	miR-125-3p	miR-4985-5p	miR-979-3p, 5p
miR-13b-3p	miR-2b-2-5p	miR-988-3p	miR-2498-3p, 5p	miR-8-3p, 5p	miR-980-5p
miR-14-5p	miR-2b-3p	miR-995-3p	miR-2499-3p, 5p	miR-929-5p	miR-982-3p
miR-219-3p	miR-307b-5p	miR-998-3p	miR-274-5p	miR-9369-3p, 5p	miR-985-3p
miR-2535b-3p	miR-4952-3p		miR-303-5p	miR-9370-5p	miR-986-3p
			miR-307b-3p	miR-972-3p	miR-992-3p
			miR-31a-3p	miR-973-3p, 5p	miR-997-3p
			miR-33-3p	miR-974-3p, 5p	miR-iab-4-5p
			miR-375-3p, 5p	miR-975-5p	miR-iab-8-3p
			miR-4966-3p, 5p	miR-976-3p	
			miR-4976-5p	miR-977-3p, 5p	

Table 3. miRNAs with tissue-biased expression in both sexes, or in males or females only. Numbers in parentheses are the total miRNAs in each category, inclusive of 3p and 5p strands where both are differentially expressed.

the upregulated predicted targets of each strand of miR-927. However, the analysis did not yield any enriched terms with an FDR q-value > 0.05.

When comparing predicted target and non-target transcripts of each differentially expressed miRNA for each comparison (e.g. mated versus virgin female Ab), the Kolmogorov-Smirnov analyses did not show differences between the fold change distributions of predicted targets and non-targets (Supplementary Fig. S6). Similarly we found no evidence for the overrepresentation of miR-184-5p, miR-286-3p or miR-927 targets amongst the corresponding set of mating-responsive mRNAs, when compared to all mRNAs (Fisher's Exact Test).

Discussion

We tested two major predictions (i) that there are significant changes to the expression of coding and regulatory non-coding genes following mating in both sexes, and (ii) that the mode and nature of PMR gene expression profiles of each sex are markedly different. The results supported both predictions and revealed significant insights into sex-specific functional variation in post-mating responses. Our data showed strong signatures of mating-responsive gene expression profiles that were unique, and spatially distinct, in each sex. In females, differential expression was generally of larger magnitude than in males. Gene expression in the female head-thorax was radically altered by mating, with 2040 genes showing differential expression of substantial magnitude, while 125

Female biased miRNAs		Male biased miRNAs		
<i>Whole fly (12)</i>		<i>Whole fly (8)</i>		
miR-286-3p	miR-956-5p	miR-1006-3p	miR-252-5p	miR-2c-3p
miR-2b-2-5p	miR-989-3p	miR-10-5p	miR-263a-5p	miR-993-3p
miR-308-3p, 5p	miR-994-5p	miR-133-3p	miR-263b-3p	
miR-318-3p	miR-9b-3p	<i>Abdomen only (97)</i>		
miR-92a-3p	miR-9c-5p	let-7-5p	miR-314-3p	miR-970-3p, 5p
miR-92b-3p		miR-1000-5p	miR-315-5p	miR-972-3p
<i>Abdomen only (42)</i>		miR-100-3p	miR-316-5p	miR-973-3p, 5p
bantam-3p	miR-311-3p, 5p	miR-1004-3p	miR-317-3p	miR-974-5p
miR-1003-3p	miR-312-3p, 5p	miR-1013-3p	miR-31a-3p, 5p	miR-975-5p
miR-1010-3p	miR-313-3p, 5p	miR-1015-3p	miR-31b-5p	miR-976-3p
miR-11-5p	miR-318-5p	miR-10-3p	miR-3-3p	miR-977-3p, 5p
miR-13b-2-5p	miR-7-5p	miR-12-3p	miR-34-3p, 5p	miR-978-3p, 5p
miR-13b-3p	miR-79-3p	miR-125-3p, 5p	miR-375-3p, 5p	miR-979-3p, 5p
miR-184-3p	miR-92a-5p	miR-12-5p	miR-4919-3p	miR-980-5p
miR-2489-3p	miR-9372-5p	miR-1-3p	miR-4966-3p, 5p	miR-981-3p
miR-275-3p	miR-988-3p	miR-2498-3p, 5p	miR-4976-5p	miR-982-3p, 5p
miR-279-3p	miR-989-5p	miR-2499-3p	miR-6-3p	miR-983-3p, 5p
miR-282-3p, 5p	miR-994-3p	miR-263a-3p	miR-8-3p, 5p	miR-984-3p, 5p
miR-284-3p, 5p	miR-995-3p	miR-274-5p	miR-87-3p	miR-985-3p
miR-2a-1-5p	miR-996-3p, 5p	miR-276b-3p	miR-929-5p	miR-986-3p
miR-2a-3p	miR-998-3p, 5p	miR-277-3p, 5p	miR-9369-3p, 5p	miR-987-5p
miR-2b-3p	miR-999-3p	miR-278-3p	miR-9370-5p	miR-991-3p
miR-306-3p, 5p	miR-9b-5p	miR-281-1-5p	miR-959-3p, 5p	miR-992-3p
miR-310-3p	miR-9c-3p	miR-281-2-5p	miR-960-3p, 5p	miR-997-5p
<i>Head-thorax only (12)</i>		miR-281-3p	miR-961-3p, 5p	miR-9a-3p, 5p
miR-281-1-5p	miR-375-3p	miR-303-5p	miR-962-3p, 5p	miR-iab-4-3p
miR-283-5p	miR-8-3p	miR-304-3p, 5p	miR-963-5p	miR-iab-8-5p
miR-314-3p, 5p	miR-956-3p	miR-307a-5p	miR-964-3p, 5p	
miR-316-5p	miR-958-3p, 5p	miR-307b-3p	miR-969-3p	
miR-33-5p	miR-980-3p	<i>Head-thorax only (9)</i>		
		miR-1017-3p	miR-210-5p	miR-990-5p
		miR-124-3p, 5p	miR-932-3p	miR-998-3p
		miR-190-3p	miR-957-3p	

Table 4. miRNAs with sex-biased expression in both body-parts (whole fly), or in the head-thorax or abdomen only. Numbers in parentheses are the total miRNAs in each category, inclusive of 3p and 5p strands where both are differentially expressed.

genes responded to mating in female abdomens. In contrast, in males there were no mating-responsive expression changes in the head-thorax at all under the same significance criteria, whereas male abdomens showed differential expression in 2068 genes.

The large number of DE genes in the female head-thorax and male abdomen is consistent with known PMR activity and phenotypes in those different body-parts. For example, the receipt of the ‘sex peptide’ seminal fluid protein from males during mating causes neurological changes in the female brain that affect feeding behaviour, sleep patterns, sexual receptivity and aggression levels^{14–16}. Hence even if the primary site of Sfp receipt is within the female reproductive tract in the abdomen, Sfps can cause many changes in other parts of the body by binding to receptors located in the nervous system including the brain⁶². Our knowledge of PMR phenotypes in males is limited, but biological processes known to be affected by mating are located within the abdomen, namely the replenishment of ejaculate components and morphological changes to the ejaculatory duct^{31,33}. The low numbers of DE genes seen in the female abdomen and male head-thorax suggests there is lower activity of biological processes in those body-parts following mating, or a low requirement for active *de novo* gene transcription. However, it is also possible that gene expression changes in different tissues and cell types within the major body parts tested are occurring, but counteracting one another. The gene expression patterns we describe are supported by specific validated genes reported from other studies. Previous transcriptomic studies on females varied considerably in the numbers of DE genes detected in response to mating, from just 38⁴³ to over 2000⁴⁴ in whole females assayed a few hours after mating. One study⁴⁹ also found a ten-fold difference in the numbers of mating-responsive genes in the spermatheca compared to the seminal receptacle, indicating that tissues with related functions can also show distinctive responses.

We also observed sex-specific functional enrichment amongst mating-responsive genes. For example, in male abdomens, the predominant response was an upregulation in genes associated with protein folding, localization and processing through the Golgi apparatus and endoplasmic reticulum. Genes encoding signal recognition particles (SRP), SRP receptors, translocation channel proteins and p24 family proteins⁶³ were upregulated, as well as genes encoding coatomer-proteins that form COPI and COPII vesicles^{64,65}. This sex-specific response implies an increase in the production of secreted or transmembrane proteins, and is consistent with male replenishment of Sfps that become depleted following mating³¹. Most of the secretory pathway transcripts were upregulated less than two-fold in mated versus virgin males, which suggests a complex coordinated and on-demand regulation machinery for the production of Sfp proteins. In female abdomens, the transcription of genes encoding immune effectors was elevated following mating, consistent with previous observations^{43,46,47,50}. The significance of this robust response is not yet clear but may stem from either the transfer of pathogens through mating, damage to the female genital tract, or induction by ejaculate proteins^{20,66}.

As well as sex-specific responses, a core set of shared genes were differentially expressed in both male and female abdomens. Among these was an over-representation of ribosomal protein (RP) genes and genes involved in the electron transport chain. This implies that both sexes have an increased requirement for translation and energy generation following mating. In males, increased translation in the abdomen is consistent with the replenishment of Sfps. Indeed, a previous study reported a burst of ribosome synthesis in the accessory glands (the main site of Sfp production) of males between 30 minutes and 6 hours following mating⁶⁷. Increased translation in mated female abdomens may be required for egg activation and the progression of vitellogenic oocytes, requiring the translation of maternal mRNAs and enhanced yolk protein synthesis, respectively^{68,69}.

Interestingly, some mating-responsive processes that were upregulated in the abdomen of both sexes were downregulated in the female head-thorax. Indeed, downregulated genes constituted the majority of DE genes in the female head-thorax, in direct contrast to all other comparisons, in which DE was generally due to the upregulation of genes in response to mating. Reduced expression of RP genes and genes associated with energy generation in the head-thorax, and elevation of those same genes in the abdomen, is suggestive of a mating-induced 'switch' in tissue-specific resource allocation. This could reflect a compensatory mechanism to counterbalance the increased demand for energy and translation in the abdomen, an idea that would be interesting to test.

It can be somewhat difficult to directly compare transcriptomic studies across different laboratories, given the variance in experimental design, diet, fly strain, tissue, transcriptomic methods and analysis. Nevertheless, there are some interesting contrasts to explore with existing studies of the transcriptomic responses of female *D. melanogaster* to mating^{42–49,51}. To minimise confounding variation, we compared our data with studies that had used similar time points and tissues and in general, there was good overlap. Specific genes were robustly differentially expressed in response to mating across multiple studies. These included *fit*, *CG31324* and *Send2* which were upregulated in response to mating in this and in five other studies. *Send2* encodes the serine protease spermathecal endopeptidase 2 which, along with *Send1*, is exclusively expressed in secretory cells of the female spermatheca⁷⁰. Although the exact function of *Send2* itself is unknown, products of the spermathecal secretory cells are required for the recruitment of sperm for storage, and sustained egg laying⁷⁰. The gene product of *CG31324*, which was also consistently upregulated in the mated female abdomen, is currently unknown. The product of *fit* is associated with feeding behaviour. In females, it is downregulated in starvation conditions⁷¹ and acts as a negative feedback regulator to suppress the intake of protein-rich food⁷². Mated females have an increased appetite¹⁴, and a preference for protein-rich food when compared to virgins⁷³. Therefore the upregulation of *fit* following mating may be triggered by the protein-component of an increased food intake.

Our data revealed that post-mating changes in both sexes have the potential to be regulated by small RNA molecules, as has been reported previously for females^{46,57}. The pluripotentiality of miRNA targeting allows multiple genes to be regulated simultaneously under the influence of miRNA 'hubs'²⁴. This facilitates the coordinated expression of genes with related functions in response to an appropriate single stimulus, such as mating. At the most stringent significance threshold, we identified four miRNAs that differed in expression between virgin and mated female abdomens, and two in male abdomens. Both the 3p and 5p strands of miR-927 were downregulated in males following mating. In support of a role for this miRNA in regulating reproductive processes, deletion of miR-927 in male *D. melanogaster* is reported to reduce adult fertility⁵⁸. Interestingly, both strands of miR-927 were also among the most significantly downregulated miRNAs in the female abdomen (albeit below the padj threshold of 0.05), suggesting that this miRNA may play a role in the regulation of post-mating responses common to both sexes.

Of the four miRNAs that were significantly differently expressed in mated females, the two with the greatest fold change were miR-184-5p and miR-997-5p. Increased expression of miR-184 is consistent with the essential role of this miRNA in the regulation of oogenesis. Females lacking miR-184 show an age-progressive failure to produce eggs⁷⁴ and their fecundity is unaffected by the presence of sex peptide⁵⁷. Interestingly, overexpression of miR-184 in both sexes causes a severe reduction in lifespan⁷⁵ and this may offer clues to the mechanisms underlying reduced lifespan in mated females. miR-997 was completely absent in virgin females and in female head-thorax tissue, but was detectable in female abdomens following mating. Notably, in males, miR-997 expression was also restricted to the abdomen, but is stably expressed regardless of mating status. One possibility is that miR-997 is expressed solely by males and transferred to females during mating. Extracellular miRNAs can be transported stably within microvesicles, which are released into the ejaculate by secondary cells of the male accessory gland^{76,77}. Once in the female, miRNAs contained within the microvesicles have the potential to target female mRNA molecules, and thus alter female post-mating responses⁷⁶.

The miRNA-mRNA interaction analysis identified a number of genes that were differentially expressed in the opposite direction to significantly upregulated or downregulated miRNAs, indicating a potential response of the coding transcriptome to miRNA differential expression after mating. However, global differences in expression between all predicted targets and non-targets of differentially expressed miRNAs were not observed. A potential

explanation is that miRNA differential expression influences physiological change by mediating the repression of a restricted set of predicted targets. The signal for this type of repression would be obscured in a global analysis of all predicted targets and nontargets. The global correlation analysis of mating-responsive miRNAs and mRNAs also revealed that a number of differentially expressed mRNAs in the male abdomen had the potential to be targeted by miR-927-3p or 5p strands. Of the mRNA targets that were expressed in the opposite direction to miR-927, at least 38 are described as having a role in developmental processes, although there was no overall significant enrichment of functional terms, suggesting that putative targets of miR-927 have diverse functions. Interestingly, a number of DE mRNAs were predicted to be targets of both strands of miR-927, opening up the possibility that the two mature miRNAs are acting cooperatively to mediate repression.

Conclusion

Our results provide the first direct comparison of the transcriptomic responses of male and female *Drosophila* to mating, and the first comparison of mating-responsive miRNAs in both sexes in any species. Our data reveal that there were marked sex- and body part-specific responses to mating, in profiles of mRNAs and miRNAs in *D. melanogaster*. However, some transcriptional responses were also shared by males and females. There were also sex-specific differences in the magnitude of gene expression changes, with females generally showing a greater magnitude of DE. In addition, while many of the same genes were differentially expressed between body-parts and sexes, the direction of DE of these genes was sex- or tissue-specific. Taken together, our results show that while both males and females invest in enhanced protein and energy production in the abdomen, males have a much broader response than females, and additionally invest in the production of secretory protein pathway components. In contrast, in the head-thorax, females showed the greatest transcriptional response through the downregulation of both small- and macro-molecule metabolism and energy production, while the transcriptional profile of males remained largely unchanged. Our results reveal the extent of quantitative and qualitative variation in sex-specific responses to mating and highlight novel potential roles for regulatory molecules in shaping the expression of sex differences.

Materials and Methods

Sample preparation. Wildtype *D. melanogaster* flies were from a large laboratory population originally collected in the 1970s in Dahomey (Benin). Flies were reared on standard sugar yeast (SY) medium (100 g brewer's yeast powder, 50 g sugar, 15 g agar, 30 ml Nipagin (10% w/v solution), and 3 ml propionic acid, per litre of medium) in a controlled environment (25 °C, 50% humidity, 12:12 hour light:dark cycle). Larvae were raised at a standard density of 100 per vial (glass, 75 × 25 mm, each containing 7 ml SY medium). Male and female adults were separated within 6 hours of eclosion using ice anaesthesia and stored in single sex vials at a density of 10/vial for 6 days. For the mated treatment, a single male was placed with a female and the time of mating was recorded. Immediately after mating the male was removed to a separate vial to prevent further matings. All mated flies were then flash frozen at 3 hours after start of mating in liquid N₂. For the virgin treatment, males and females were housed individually in vials for ~3–4 hours before flash freezing. Frozen flies were stored at –80 °C until use. The sample size for each treatment was 50 males and 50 females. The entire experiment was repeated exactly, using fresh egg collections to generate two biological replicates. Therefore, in total 16 samples were generated: 2 sexes × 2 treatments (mated/virgin) × 2 body parts (HT and Ab) × 2 replicates.

RNA extraction. To prepare tissue for RNA extraction, 50 flies from each sex, treatment and biological replicate were separated into HT and Ab tissues on dry ice, and the body parts were then pooled for RNA extraction (note that both body parts were intact, and thus the Ab contained the germline). Tissues were disrupted by grinding under liquid nitrogen, then total RNA was extracted using the miRvana miRNA isolation kit (Ambion, AM1561), according to the kit protocol. RNA was eluted in RNA storage solution (1 mM sodium citrate, pH 6.4 ±/– 0.2, Ambion). Samples were DNase treated to remove residual genomic DNA (Ambion Turbo DNA-free kit, AM1907). RNA was assessed for quantity and quality using a NanoDrop 8000 spectrophotometer.

Library construction and sequencing. The 16 samples were sent to the Earlham Institute provider (Norwich Research Park, UK) for mRNA and sRNA library construction, and sequencing. Libraries were constructed using the Illumina TruSeq kit. For the sRNA libraries, a modified 'blocking oligo' was also used to preclude adapter ligation to the highly abundant 30 nt 2S rRNA⁶⁰. Non-directional, single end RNA-seq was conducted using the Illumina HiSeq. 2500 platform with 50nt read length.

Sequencing analysis. Kallisto version 0.46.0⁷⁸ was used to pseudoalign reads to the Berkeley *Drosophila* Genome Project 6 (BDGP6) cDNA sequences downloaded from Ensembl (release 89⁷⁹). A kallisto index was created using the "kallisto index" command (k-mer size 31). Kallisto quant was used to obtain transcript count estimates and parameters were set to include 100 bootstrap samples and to perform sequence bias correction. Transcript to gene mappings were obtained using biomaRt⁸⁰ and transcript counts were aggregated in Sleuth (version 0.28.1)⁸¹ before calling pairwise differential expression between mated and virgin samples of the same body part and sex. Small RNA reads were converted from FASTQ to FASTA format and then processed to trim sequencing adaptors using a custom Perl script (available in the Supplementary Material) recognising the first 8 bases of the adapter sequence ("TGGAATTC"). Trimmed reads were then aligned to miRBase (v22.0) *D. melanogaster* mature miRNA sequences using PatMaN⁸² (parameters -e 0 -g 0). A custom Perl script (see Supplementary Material) was used to parse the alignment files and generate an aligned read count table across all samples. DESeq2 (version 1.14.1)⁵⁹ was used for normalisation of counts between samples and calling differentially expressed miRNAs.

miRNA target prediction. Prediction of miRNA target sites were conducted using the TargetScan algorithm (version 4.1 – ‘TargetScanS’)⁸³, which predicted targets on the basis of complementarity of the 3′ untranslated region (3′UTR) to the mature miRNA seed sequence^{83,84}. In order to run TargetScan, using custom shell scripts, mature miRNA sequences were downloaded from miRBase, filtered for *D. melanogaster* miRNAs, and processed to produce a tab-delimited three column file, with columns sequentially referring to miRNA name, miRNA seed sequence (i.e. nucleotides 2–8 of the miRNA from the 5′ end), and NCBI taxonomic ID (i.e. ‘7227’ for *D. melanogaster*). The R (v3.5.1)⁸⁵ biomaRt package (v2.38.0)^{80,86} was used for the download of *D. melanogaster* 3′UTR sequences necessary for the running of TargetScan, along with transcript–gene mappings: The *useMart* function was used to select the Ensembl mart. For the selected mart, the *useDataset* function was used to select *D. melanogaster* ensembl gene models for release 89 of Ensembl. Afterwards the *getBM* function was used to extract stable gene and transcript ids, external gene names and 3′UTR sequences for all *D. melanogaster* transcript models. Otherwise default parameter values were used when calling biomaRt functions. For each gene model, the transcript splice-isoform (denoted by the ensembl transcript ID) with the longest annotated 3′UTR was designated as being representative for that gene. In cases where a gene model possessed multiple transcript isoforms corresponding to the maximum 3′UTR length for that gene, one transcript isoform was selected at random. Gene models in which none of the corresponding transcript models possessed an annotated 3′UTR sequence was not used for miRNA target prediction with TargetScan. For use with TargetScan, 3′UTR data were deposited in a three-column tab-delimited text file sequentially containing an identifier column containing the ensembl gene ID, ensembl transcript ID, and the external gene name; a column containing the *D. melanogaster* NCBI taxonomic ID (i.e. ‘7227’), and a final column containing the 3′UTR sequence. TargetScan was then subsequently executed with the 3′UTR data file and the previously described miRNA data file.

Comparison of miRNA predicted targets and nontargets. Transcript expression data was pre-processed before statistical testing: The mean average relative abundance for each mRNA, in units of normalised transcripts per million (TPM)⁸⁷, was computed from both replicates for each sample type (e.g. female abdomen). mRNA with average normalised TPM values equal to zero were discarded for each sample type and not used for further analysis. A pseudocount (AKA offset) of 1 was added to all remaining average normalised TPM values. Log₂ fold change values were computed from offset average normalised TPM values for each individual differential expression analysis (e.g. virgin male abdomen vs. mated male abdomen). The log₂ fold change was subsequently used as a metric of the magnitude of mRNA differential expression between conditions. Exploratory Data Analysis, in which cumulative plots of upregulated, downregulated and not differentially expressed genes were constructed with respect to 3′UTR length and 3′UTR predicted target site frequency, indicated that 3′UTR length was a potential confounding variable when examining mRNA dysregulation between the virgin and mated conditions (Supplementary Figs S2ab and S4ab). A simulation, in which TargetScan was ran as described previously, but using 401 randomly generated miRNA seed sequences instead of 401 *D. melanogaster* miRNA seed sequences, and with subsequent cumulative plot construction with respect to 3′UTR target site frequency, provided further evidence that 3′UTR length was a confounding variable (Supplementary Fig. S5a,b). In subsequent analyses, a sampling method was used to normalise 3′UTR length for mRNA expression data when comparing predicted target 3′UTRs to predicted nontarget 3′UTRs for any given comparison: A histogram of 3′UTR sequence lengths was constructed separately for predicted target and nontarget datasets, starting from 0, in increments of 200nt, and to a maximum representing the maximum sequence length from both target and non-target datasets. Each break of the two histograms are iterated through, and for each iteration, log fold change values for predicted target and nontarget datasets are subsetted to fall within the 3′UTR sequence length range given by the individual histogram breaks. Within this range, of the target and nontarget log fold change vectors, if vector sizes are unequal, the vector with the largest number of records is sampled to match the number of observations contained within the smaller vector. Log fold change values for both predicted target and nontarget datasets are concatenated for each iteration, in order to create log fold change distributions which are normalised for 3′UTR length. In the case of the Fisher Exact test, an identical sampling procedure is implemented with the exception that transcript identifiers are sampled instead of log fold change values. We investigated whether changes in mRNA expression could be influenced by the changes in expression of differentially expressed miRNAs, by a process of miRNA targeting. Two-sample, one-sided Kolmogorov-Smirnov (KS) tests (using the *ks.test* function of the R *stats* package) were implemented to test for the inequality between miRNA target and miRNA nontarget fold change distributions for a given miRNA, and the one-sided Fisher Exact Test (using the *fisher.test* function of the R *stats* package) to test for an enrichment of the target sites of differentially expressed miRNAs in mRNAs differentially expressed in the opposite direction. For the KS test, the value of the ‘alternative’ parameters was set to ‘greater’ when testing the effect of upregulated miRNAs, and set to value of ‘less’ when testing the effect of downregulated miRNAs. For the Fisher test, the value of the ‘alternative’ parameter was always set to a value of ‘greater’. Otherwise, default parameters were used for both statistical test functions. Fisher Exact and KS tests were also similarly conducted to test for potential combinatorial effects of different pairwise combinations of miRNAs which were differentially expressed in the same direction with predicted target sets designated as those mRNAs with predicted targets for both differentially expressed miRNAs. The false discovery rate (FDR) was set at 0.05, with the Benjamini-Hochberg method used to correct for multiple comparisons⁸⁸. To counteract the stochasticity introduced by the sampling described previously, for each test, p-values were calculated 100 times, and the mean average p-value was taken as being representative.

Network visualisations. Network visualisations of predicted interactions between differentially expressed coding genes and differentially expressed miRNA for the mated male abdomen were completed using Cytoscape (v3.4.0)⁸⁹. Network visualisations were not completed for other comparisons, which either did not possess any

differentially expressed miRNAs, or the number of differentially expressed coding genes predicted to be targeted by differentially expressed miRNAs was judged to be too low for network visualisations to be informative.

Gene ontology enrichment analysis. GO analyses were conducted using the GOrilla enrichment analysis and visualisation tool using the default parameter settings^{90,91}. Unranked target lists of genes were compared to a background of genes for which reads were obtained in our sequencing analysis. Background reference lists were tailored to each sex and body part, to minimise sampling bias⁹². The cut-off for statistical significance was an FDR q-value < 0.05.

Data archiving. Raw sequencing data for this study is stored at the Sequence Read Archive (SRA) using the BioProject accession: PRJNA521155.

Received: 28 June 2019; Accepted: 24 September 2019;

Published online: 06 November 2019

References

1. Oku, K., Price, T. A. R. & Wedell, N. Does mating negatively affect female immune defences in insects? *Animal Biol.* **69**, 117 (2019).
2. South, A. & Lewis, S. M. The influence of male ejaculate quantity on female fitness: a meta-analysis. *Biol. Rev. Camb. Philos. Soc.* **86**, 299–309 (2011).
3. Thomas, M. L. Detection of female mating status using chemical signals and cues. *Biological Reviews. Biol. Rev. Camb. Philos. Soc.* **86**, 1–14 (2011).
4. Johnstone, R. A. & Keller, L. How males can gain by harming their mates: Sexual conflict, seminal toxins, and the cost of mating. *Am. Nat.* **156**, 368–377 (2000).
5. Kalb, J. M., Dibeneditto, A. J. & Wolfner, M. F. Probing the function of *Drosophila melanogaster* accessory-glands by directed cell ablation. *Proc. Natl. Acad. Sci. USA* **90**, 8093–8097 (1993).
6. Peng, J. *et al.* Gradual release of sperm bound sex-peptide controls female postmating behavior in *Drosophila*. *Curr. Biol.* **15**, 207–213 (2005).
7. Ram, K. R. & Wolfner, M. F. Sustained post-mating response in *Drosophila melanogaster* requires multiple seminal fluid proteins. *PLoS Genet.* **3**, e238 (2007).
8. Ravi Ram, K. & Wolfner, M. F. Seminal influences: *Drosophila* Acp5 and the molecular interplay between males and females during reproduction. *Integr. Comp. Biol.* **47**, 427–45 (2007).
9. Sirot, L. K. *et al.* Sexual conflict and seminal fluid proteins: a dynamic landscape of sexual interactions. *CSH Perspect. Biol.* **7** (2015).
10. Findlay, G. D. *et al.* Proteomics reveals novel *Drosophila* seminal fluid proteins transferred at mating. *PLOS Biol.* **6**, 1417–1426 (2008).
11. Sepil, I. *et al.* Quantitative proteomics identification of seminal fluid proteins in male *Drosophila melanogaster*. *Mol. Cell. Proteomics.* **18**, S46–S58 (2018).
12. Wigby, S. *et al.* Seminal fluid protein allocation and male reproductive success. *Curr. Biol.* **19**, 751–757 (2009).
13. Sirot, L. K., Wolfner, M. F. & Wigby, S. Protein-specific manipulation of ejaculate composition in response to female mating status in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* **108**, 9922–9926 (2011).
14. Carvalho, G. B. *et al.* Allosteric modulation of feeding behavior by the sex peptide of *Drosophila*. *Curr. Biol.* **16**, 692–696 (2006).
15. Isaac, R. E. *et al.* *Drosophila* male sex peptide inhibits siesta sleep and promotes locomotor activity in the post-mated female. *Proc. Royal Soc. Lond. B.* **277**, 65–70 (2010).
16. Bath, E. *et al.* Sperm and sex peptide stimulate aggression in female *Drosophila*. *Nat. Ecol. Evol.* **1**, 0154 (2017).
17. Chapman, T. *et al.* The sex peptide of *Drosophila melanogaster*: female post-mating responses analyzed by using RNA interference. *Proc. Natl. Acad. Sci. USA* **100**, 9923–8 (2003).
18. Avila, F. W. *et al.* Sex peptide is required for the efficient release of stored sperm in mated *Drosophila* females. *Genetics.* **186**, 595–600 (2010).
19. Neubaum, D. M. & Wolfner, M. F. Mated *Drosophila melanogaster* females require a seminal fluid protein, Acp36DE, to store sperm efficiently. *Genetics.* **153**, 845–57 (1999).
20. Peng, J., Zipperlen, P. & Kubli, E. *Drosophila* sex-peptide stimulates female innate immune system after mating via the Toll and Imd pathways. *Curr. Biol.* **15**, 1690–1694 (2005).
21. Fedorka, K. M. *et al.* Post-mating disparity between potential and realized immune response in *Drosophila melanogaster*. *Proc. Royal Soc. Lond. B.* **274**, 1211–1217 (2007).
22. Ribeiro, C. & Dickson, B. J. Sex Peptide Receptor and Neuronal TOR/S6K Signaling Modulate Nutrient Balancing in *Drosophila*. *Curr. Biol.* **20**, 1000–1005 (2010).
23. Cognigni, P., Bailey, A. P. & Miguel-Aliaga, I. Enteric neurons and systemic signals couple nutritional and reproductive status with intestinal homeostasis. *Cell Metab.* **13**, 92–104 (2011).
24. Mohorianu, I. *et al.* Control of seminal fluid protein expression via regulatory hubs in *Drosophila melanogaster*. *Proc. Royal Soc. Lond. B.* **285** (2018).
25. Bretman, A., Fricke, C. & Chapman, T. Plastic responses of male *Drosophila melanogaster* to the level of sperm competition increase male reproductive fitness. *Proc. Royal Soc. Lond. B.* **276**, 1705–11 (2009).
26. Mohorianu, I. *et al.* Genomic responses to the socio-sexual environment in male *Drosophila melanogaster* exposed to conspecific rivals. *RNA.* **23**, 1048–1059 (2017).
27. Bingham, J., Chapman, T. & Partridge, L. Effects of body size, accessory gland and testis size on pre- and postcopulatory success in *Drosophila melanogaster*. *Animal Behav.* **64**, 915–921 (2002).
28. Lefevre, G. & Jonsson, U. B. Sperm transfer, storage, displacement, and utilization in *Drosophila Melanogaster*. *Genetics.* **42**, 1719–1736 (1962).
29. Linklater, J. R. *et al.* Ejaculate depletion patterns evolve in response to experimental manipulation of sex ratio in *Drosophila melanogaster*. *Evolution.* **61**, 2027–2034 (2007).
30. Hihara, F. Effects of the male accessory gland secretion on oviposition and remating in females of *Drosophila melanogaster*. *Zoological magazine.* **90**, 307–316 (1981).
31. Sirot, L. K. *et al.* Seminal fluid protein depletion and replenishment in the fruit fly, *Drosophila melanogaster*: an ELISA-based method for tracking individual ejaculates. *Behav. Ecol. Sociobiol.* **63**, 1505–1513 (2009).
32. Norville, K., Sweeney, S. T. & Elliott, C. J. H. Postmating change in physiology of male *Drosophila* mediated by serotonin (5-HT). *J. Neurogenet.* **24**, 27–32 (2010).
33. Cohen, A. B. & Wolfner, M. F. Dynamic changes in ejaculatory bulb size during *Drosophila melanogaster* aging and mating. *J. Insect Physiol.* **107**, 152–156 (2018).

34. Winterhalter, W. E. & Fedorka, K. M. Sex-specific variation in the emphasis, inducibility and timing of the post-mating immune response in *Drosophila melanogaster*. *Proc. Royal Soc. Lond. B.* **276**, 1109–1117 (2009).
35. Camus, M. F. *et al.* Dietary choices are influenced by genotype, mating status, and sex in *Drosophila melanogaster*. *Ecol. Evol.* **8**, 5385–5393 (2018).
36. Williams, T. M. & Carroll, S. B. Genetic and molecular insights into the development and evolution of sexual dimorphism. *Nat. Rev. Genet.* **10**, 883–883 (2009).
37. Gomulski, L. M. *et al.* Transcriptome profiling of sexual maturation and mating in the mediterranean fruit fly, *Ceratitis capitata*. *Plos One.* **7** (2012).
38. Kocher, S. D. *et al.* Genomic analysis of post-mating changes in the honey bee queen (*Apis mellifera*). *BMC Genom.* **9**, 232 (2008).
39. Rogers, D. W. *et al.* Molecular and cellular components of the mating machinery in *Anopheles gambiae* females. *Proc. Natl. Acad. Sci. USA* **105**, 19390–19395 (2008).
40. Alfonso-Parra, C. *et al.* Mating-Induced Transcriptome Changes in the Reproductive Tract of Female *Aedes aegypti*. *PLOS Negl. Trop. Dis.* **10** (2016).
41. Immonen, E. *et al.* Mating Changes Sexually Dimorphic Gene Expression in the Seed Beetle *Callosobruchus maculatus*. *Genome Biol. Evol.* **9**, 677–699 (2017).
42. Innocenti, P. & Morrow, E. H. Immunogenic males: a genome-wide analysis of reproduction and the cost of mating in *Drosophila melanogaster* females. *J. Evol. Biol.* **22**, 964–73 (2009).
43. Lawniczak, M. K. N. & Begun, D. J. A genome-wide analysis of courting and mating responses in *Drosophila melanogaster* females. *Genome.* **47**, 900–910 (2004).
44. McGraw, L. A., Clark, A. G. & Wolfner, M. F. Post-mating gene expression profiles of female *Drosophila melanogaster* in response to time and to four male accessory gland proteins. *Genetics* **179**, 1395–408 (2008).
45. McGraw, L. A. *et al.* Genes regulated by mating, sperm, or seminal proteins in mated female *Drosophila melanogaster*. *Curr. Biol.* **14**, 1509–14 (2004).
46. Zhou, S., Mackay, T. & Anholt, R. R. Transcriptional and epigenetic responses to mating and aging in *Drosophila melanogaster*. *BMC Genom.* **15**, 927 (2014).
47. Delbare, S. Y. N. *et al.* Roles of female and male genotype in post-mating responses in *Drosophila melanogaster*. *J. Hered.* **108**, 740–753 (2017).
48. Mack, P. D. *et al.* Mating-responsive genes in reproductive tissues of female *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* **103**, 10358–10363 (2006).
49. Prokupek, A. M. *et al.* Transcriptional profiling of the sperm storage organs of *Drosophila melanogaster*. *Insect Mol. Biol.* **18**, 465–75 (2009).
50. Kapelnikov, A. *et al.* Mating induces an immune response and developmental switch in the *Drosophila* oviduct. *Proc. Natl. Acad. Sci. USA* **105**, 13912–7 (2008).
51. Dalton, J. E. *et al.* Dynamic, mating-induced gene expression changes in female head and brain tissues of *Drosophila melanogaster*. *BMC Genomics.* **11**, 541 (2010).
52. Chintapalli, V. R., Wang, J. & Dow, J. A. T. Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. *Nat. Genet.* **39**, 715–720 (2007).
53. Domanitskaya, E. V. *et al.* The hydroxyproline motif of male sex peptide elicits the innate immune response in *Drosophila* females. *FEBS J.* **274**, 5659–5668 (2007).
54. Gioti, A. *et al.* Sex peptide of *Drosophila melanogaster* males is a global regulator of reproductive processes in females. *Proc. Royal Soc. Lond. B.* **279**, 4423–32 (2012).
55. Ellis, L. L. & Carney, G. E. Mating alters gene expression patterns in *Drosophila melanogaster* male heads. *BMC Genom.* **11** (2010).
56. Heifetz, Y. *et al.* Mating regulates neuromodulator ensembles at nerve termini innervating the *Drosophila* reproductive tract. *Curr. Biol.* **24**, 731–7 (2014).
57. Fricke, C. *et al.* MicroRNAs influence reproductive responses by females to male sex peptide in *Drosophila melanogaster*. *Genetics.* **198**, 1603–19 (2014).
58. Chen, Y. W. *et al.* Systematic study of *Drosophila* microRNA functions using a collection of targeted knockout mutations. *Dev. Cell.* **31**, 784–800 (2014).
59. Love, M. I., Huber, W. & Anders, S. J. G. B. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
60. Fowler, E. K. *et al.* Small RNA populations revealed by blocking rRNA fragments in *Drosophila melanogaster* reproductive tissues. *Plos One.* **13**, e0191966 (2018).
61. Marco, A. Sex-biased expression of microRNAs in *Drosophila melanogaster*. *Open Biol.* **4** (2014).
62. Haussmann, I. U. *et al.* Multiple pathways mediate the sex-peptide-regulated switch in female *Drosophila* reproductive behaviours. *Proc. Royal Soc. Lond. B.* **280** (2013).
63. Saleem, S. *et al.* *Drosophila melanogaster* p24 trafficking proteins have vital roles in development and reproduction. *Mech. Dev.* **129**, 177–191 (2012).
64. Jensen, D. & Schekman, R. COPII-mediated vesicle formation at a glance. *J. Cell Sci.* **124**, 1–4 (2011).
65. Jayaram, S. A. *et al.* COPI Vesicle Transport Is a Common Requirement for Tube Expansion in *Drosophila*. *Plos One.* **3** (2008).
66. Morrow, E. H. & Innocenti, P. Female postmating immune responses, immune system evolution and immunogenic males. *Biol. Rev.* **87**, 631–638 (2012).
67. Schmidt, T., Stumm-Zollinger, E. & Chen, P. S. Protein metabolism of *Drosophila melanogaster* male accessory glands—III: Stimulation of protein synthesis following copulation. *Insect Biochemistry.* **15**, 391–401 (1985).
68. Soller, M., Bownes, M. & Kubli, E. Mating and sex peptide stimulate the accumulation of yolk in oocytes of *Drosophila melanogaster*. *Eur. J. Biochem.* **243**, 732–738 (1997).
69. Krauchunas, A. R., Sackton, K. L. & Wolfner, M. F. Phospho-regulation pathways during egg activation in *Drosophila melanogaster*. *Genetics.* **195**, 171–80 (2013).
70. Schnakenberg, S. L., Matias, W. R. & Siegal, M. L. Sperm-Storage Defects and Live Birth in *Drosophila* Females Lacking Spermathecal Secretory Cells. *PLOS Biol.* **9** (2011).
71. Fujikawa, K. *et al.* Characteristics of genes up-regulated and down-regulated after 24 h starvation in the head of *Drosophila*. *Gene.* **446**, 11–17 (2009).
72. Sun, J. H. *et al.* *Drosophila* FIT is a protein-specific satiety hormone essential for feeding control. *Nat. Commun.* **8** (2017).
73. Kubli, E. Sexual Behavior: Dietary Food Switch Induced by Sex. *Curr. Biol.* **20**, R474–R476 (2010).
74. Iovino, N., Pane, A. & Gaul, U. MiR-184 has multiple roles in *Drosophila* female germline development. *Dev. Cell.* **17**, 123–133 (2009).
75. Gendron, C. M. & Pletcher, S. D. MicroRNAs mir-184 and let-7 alter *Drosophila* metabolism and longevity. *Aging Cell.* **16**, 1434–1438 (2017).
76. Green, D., Dalmay, T. & Chapman, T. Microguards and micromessengers of the genome. *Heredity.* **116**, 125–34 (2016).
77. Corrigan, L. *et al.* BMP-regulated exosomes from *Drosophila* male reproductive glands reprogram female behavior. *J. Cell Biol.* **206**, 671–688 (2014).
78. Bray, N. L. *et al.* Near-optimal probabilistic RNA-seq quantification. *Nat. Biotech.* **34**, 525 (2016).

79. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
80. Durinck, S. *et al.* Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–91 (2009).
81. Pimentel, H. *et al.* Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat. Methods.* **14**, 687 (2017).
82. Prüfer, K. *et al.* PatMaN: rapid alignment of short sequences to large databases. *Bioinformatics.* **24**, 1530–1 (2008).
83. Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell.* **120**, 15–20 (2005).
84. Ruby, J. G. *et al.* Evolution, biogenesis, expression, and target predictions of a substantially expanded set of *Drosophila* microRNAs. *Genome Res.* **17**, 1850–1864 (2007).
85. R Core Team. R: a language and environment for statistical computing. Vienna, Austria (2015).
86. Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics.* **21**, 3439–3440 (2005).
87. Li, B. *et al.* RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics.* **26**, 493–500 (2010).
88. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J. Royal Stat. Soc. B.* **57**, 289–300 (1995).
89. Smoot, M. E. *et al.* Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics.* **27**, 431–432 (2011).
90. Eden, E. *et al.* GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC bioinform.* **10**, 48 (2009).
91. Eden, E. *et al.* Discovering motifs in ranked lists of DNA sequences. *PLoS Comp. Biol.* **3** (2007).
92. Timmons, J. A., Szkop, K. J. & Gallagher, I. J. Multiple sources of bias confound functional enrichment analysis of global -omics data. *Genome Biol.* **16**, 186–186 (2015).

Acknowledgements

We thank the Norwich Research Park Science Links Seed Corn fund and NERC (NE/R000891/1) and the BBSRC (BB/L003139/1) for funding. TB was supported by the BBSRC Norwich Research Park Biosciences Doctoral Training Partnership (BB/J014524/1).

Author contributions

T.C. and S.M. conceived the study, E.K.F. conducted the practical work, E.K.F., T.B. and S.M. analysed the data, E.K.F. and T.B. wrote the paper, E.K.F., T.B., S.M. and T.C. revised and commented on the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41598-019-51141-9>.

Correspondence and requests for materials should be addressed to E.K.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019