

Dissecting the RNA-binding activity and function of HOW(S)  
during *Drosophila melanogaster* spermatogenesis

Michaela Agapiou

Submitted in accordance with the requirements for the degree of  
Doctor of Philosophy

The University of Leeds  
School of Molecular and Cellular Biology

April 2021



The candidate confirms that the work submitted is their own and that appropriate credit has been given where reference has been made to the work of others.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

The right of Michaela Agapiou to be identified as Author of this work has been asserted by Michaela Agapiou in accordance with the Copyright, Designs and Patents Act 1988.



# Acknowledgements

First, I would like to thank my supervisor Julie Aspden. I want to thank her for her guidance, support and kindness, as well as her contagious optimism and passion that always kept this project moving forward. Choosing to accept a PhD with Julie was one of the easiest decisions I've ever made and not once over the last four and half years did I regret it. Thanks also to my two co-supervisors Amanda Bretman and Thomas Edwards; without their expertise and valuable input this thesis would not have covered such a wide spectrum of biology, which is what made this project so fun to take on.

Thank you to the past and present members of Aspden, Bretman and Edwards labs, particularly: Isabel Birds, Katerina Douka, Mark Handley, Tayah Hopes, Karl Norris, Ioannis Tsagakis and Eleanor Walton. They've been the most incredible colleagues and friends, and I've learnt so much from them. They've kept me company while dissecting hundreds of flies, given me so much thoughtful feedback, always kept me thinking, and frequently provided me with sweet treats, hugs and shoulders to cry on.

Thank you to the many fantastic facilities at the University of Leeds. The Bioimaging facility allowed me take beautiful photos of testes, the Next Generation Sequencing facility looked after my previous RNA from over 6,000 hand-dissected fly testes, the Electron Microscopy facility helped with the best side project, and the Mass Spectrometry facility helped to figure out a few of the protein puzzles that came along the way. And a big thanks to Brian Jackson and Laura Wilkinson Hewitt from the Protein Production facility who helped me express a fly protein (and once, rather unexpectedly, a human protein) in bacteria. The LeedsOmics research community was also fantastic in providing lots of training and help, but most importantly it connected me to another bunch of amazing colleagues, including: Veronica DeJesus, Euan McDonnell, Peter Mulhair, Laura Najera Cortazar and David Orr.

I'm very grateful to The Story Collider team, especially Erin Barker, Paula Croxson and Nisse Greenberg, who took me on as an intern for three months as part of my PhD. It was, and continues to be, a delight to work with such a lovely and thoughtful group of people.

Thank you to my parents, Siu Yin and Agapios, my sisters, Melpo, Ariana and Olympia, and my partner's family. I have been incredibly lucky to be supported by so many friends and family members outside of the lab, and I am constantly inspired by you all. Finally, thank you to my partner, Robin. You've been the best teammate through this whole PhD journey. You've taught me how great following your gut can be, and even when that has led me to being an ocean away you always cheer me on.



# Abstract

Held out wings (HOW) is an RNA-binding protein essential for spermatogenesis in *Drosophila melanogaster*. HOW is a signal transduction and activation of RNA (STAR) protein family member. Its orthologues include Quaking (mammals) and GLD-1 (*Caenorhabditis elegans*); both are important post-transcriptional regulators of RNAs in processes including gametogenesis and myelination. Loss of HOW in *D. melanogaster* testes results in male infertility. Similar to Quaking, multiple protein isoforms of HOW exist, including the longer HOW(L) and the shorter HOW(S). HOW(L) is nuclear and regulates *bam* mRNA to control mitotic divisions prior to meiosis (Monk et al., 2010). The HOW(S) isoform is cytoplasmic and its function is not well characterised.

To determine which RNAs are bound by HOW(S) in germ cells, *in vivo* RIP-seq was performed. This identified 343 genes and 121 transcripts bound by HOW(S), which were enriched for the GO terms related to signal transduction, consistent with HOW being a STAR protein. Hipk, a signalling kinase, was identified in the HOW(S) RIP-seq and knockdown of Hipk in the testis revealed a potential role for it in spermatogenesis, as these flies displayed a range of defects in testis morphology and fertility.

Motif analysis found the (A/U/G)CUAAC motif enriched in 3'-UTRs of the HOW(S) bound mRNAs. This sequence is similar to the consensus sequences of other STAR proteins and is found in the 3'-UTR of Hipk transcripts. Fluorescence anisotropy assays revealed that HOW's STAR domain has strong nanomolar affinity for RNA oligos containing this motif. A novel motif was identified within 5'-UTRs, GCG(A/U)G, which HOW's STAR domain bound with micromolar affinity.

Together, this work has identified many RNAs bound by HOW(S) in the cytoplasm of male germ cells. These RNAs have helped to expand our understanding of HOW(S)-RNA interactions and may contribute to understanding the importance of HOW(S) RNA-binding activity during spermatogenesis.



# Table of contents

Acknowledgements . . . . .	iii
Abstract . . . . .	v
List of tables . . . . .	xiv
List of figures . . . . .	xvii
List of abbreviations . . . . .	xix
1. Introduction . . . . .	1
1.1. Spermatogenesis in <i>D. melanogaster</i> . . . . .	1
1.1.1. The stem cell niche of the <i>D. melanogaster</i> testis . . . . .	3
1.1.2. Mitosis and the switch to meiosis . . . . .	5
1.1.3. Meiosis and spermiogenesis . . . . .	5
1.2. RNA-binding proteins and spermatogenesis . . . . .	7
1.2.1. Formation of the germline . . . . .	8
1.2.2. GSC homeostasis and mitosis . . . . .	9
1.2.3. Meiosis . . . . .	11
1.2.4. Spermiogenesis . . . . .	12
1.3. The STAR protein family . . . . .	13
1.3.1. The three STAR subfamilies . . . . .	13
1.3.2. Structure of the STAR domain . . . . .	14
1.3.2.1. The maxi-KH domain of STAR proteins . . . . .	14
1.3.2.2. The QUA1 and QUA2 regions of the STAR domain . . . . .	17
1.3.3. RNA sequence features of the STAR-RNA interactions . . . . .	19
1.3.4. STAR proteins in spermatogenesis . . . . .	23
1.3.4.1. Sam68 subfamily . . . . .	23
1.3.4.2. Quaking subfamily . . . . .	24
1.3.4.3. SF1 subfamily . . . . .	25
1.4. Held out wings . . . . .	26
1.4.1. HOW is an essential protein with multiple isoforms . . . . .	26
1.4.2. HOW(L) and HOW(S) can regulate the same RNAs . . . . .	29
1.4.3. Independent functions of HOW(L) and HOW(S) . . . . .	30
1.4.4. Regulation of HOW expression and activity . . . . .	31
1.4.5. HOW in germ cells . . . . .	32
1.5. Project Objectives . . . . .	33
2. Materials and Methods . . . . .	35
2.1. Fly husband and stocks . . . . .	35
2.1.1. Fly husbandry . . . . .	35
2.1.2. Fly stocks . . . . .	35

Table of contents

2.2.	Genotyping <i>D. melanogaster</i> . . . . .	39
2.2.1.	Genomic DNA extraction . . . . .	39
2.2.2.	Polymerase chain reaction . . . . .	39
2.2.3.	Agarose gel electrophoresis . . . . .	40
2.3.	Quantification of transcript expression . . . . .	40
2.3.1.	RNA extraction from <i>D. melanogaster</i> testes . . . . .	40
2.3.2.	cDNA synthesis . . . . .	40
2.3.3.	qRT-PCR . . . . .	41
2.3.4.	Relative quantification of gene expression . . . . .	42
2.4.	Fertility assays . . . . .	42
2.4.1.	Sperm competition assay . . . . .	42
2.4.1.1.	Experimental design . . . . .	42
2.4.1.2.	Analysis . . . . .	43
2.4.2.	3-day mating assay . . . . .	43
2.4.2.1.	Experimental design . . . . .	43
2.4.2.2.	Analysis . . . . .	44
2.5.	Immunofluorescence . . . . .	44
2.5.1.	Dissections for immunofluorescence . . . . .	44
2.5.2.	Staining and microscopy . . . . .	44
2.5.3.	Image analysis . . . . .	47
2.5.3.1.	Counting hub cells . . . . .	47
2.5.3.2.	Morphology assessment . . . . .	47
2.6.	RIP-seq . . . . .	48
2.6.1.	Ribonucleoprotein immunoprecipitation . . . . .	48
2.6.1.1.	Dissection and lysis . . . . .	48
2.6.1.2.	HA pull-down . . . . .	48
2.6.1.3.	RNA elution and extraction . . . . .	48
2.6.1.4.	Protein elution . . . . .	49
2.6.1.5.	Sodium dodecyl sulphate polyacrylamide gel electrophoresis . . . . .	49
2.6.1.6.	Western blotting . . . . .	49
2.6.2.	Sequencing . . . . .	52
2.6.3.	Computational analysis . . . . .	52
2.6.3.1.	Filtering . . . . .	52
2.6.3.2.	Genome alignment and read counting . . . . .	52
2.6.3.3.	Transcriptome mapping . . . . .	53
2.6.3.4.	Differential gene and transcript enrichment . . . . .	53
2.6.3.5.	Principal Component Analysis . . . . .	54
2.6.3.6.	Gene ontology . . . . .	54
2.6.3.7.	Motif enrichment . . . . .	54
2.6.3.8.	Circular RNA alignment and annotation . . . . .	54
2.7.	Protein sequence alignment and structure prediction . . . . .	55
2.7.1.	Protein sequence alignment . . . . .	55

2.7.2.	Structure prediction . . . . .	55
2.8.	Cloning the STAR domain for protein purification . . . . .	55
2.8.1.	Gene synthesis and cloning reactions . . . . .	55
2.8.2.	Transformation into DH5 $\alpha$ cells and sequence analysis . . . . .	56
2.9.	Protein expression and purification . . . . .	57
2.9.1.	Small-scale protein expression . . . . .	57
2.9.2.	Small-scale protein purification . . . . .	58
2.9.3.	Large-scale protein expression . . . . .	58
2.9.4.	Large-scale His-STAR purification . . . . .	58
2.9.5.	Large-scale His-GST-STAR purification . . . . .	59
2.9.6.	Large-scale His-GST-STAR purification by Protein Production Facility . . . . .	60
2.10.	Mass spectrometry . . . . .	60
2.11.	Fluorescence anisotropy . . . . .	60
2.11.1.	Oligonucleotides synthesis and structure prediction . . . . .	60
2.11.2.	Experimental setup . . . . .	61
2.11.3.	Analysis . . . . .	62
2.12.	Schematics and figures . . . . .	62
3.	Identifying the targets of HOW(S) in <i>Drosophila melanogaster</i> testis . . . . .	63
3.1.	Introduction . . . . .	63
3.2.	Results . . . . .	64
3.2.1.	HOW(S)-HA expression in the male germline . . . . .	64
3.2.2.	Assessing the impact of HOW(S)-HA overexpression on male fertility in <i>D. melanogaster</i> . . . . .	70
3.2.3.	Enrichment and purification of RNAs bound to HOW(S) in the testes . . . . .	74
3.2.4.	Genome aligning and transcriptome mapping of RIP-seq data . . . . .	79
3.2.5.	Verifying the applicability of edgeR for RIP-seq differential enrichment analysis . . . . .	84
3.2.6.	Identifying the differentially enriched genes and transcripts from the HOW(S)-HA RIP-seq . . . . .	88
3.2.7.	‘Signal transduction’ genes and transcripts are enriched in the HOW(S)-HA RIP-seq . . . . .	92
3.2.8.	Finding binding motifs in the 3’-UTR of RNAs enriched in the HOW(S)-HA pull-down . . . . .	95
3.2.9.	Finding binding motifs in the 5’-UTR of RNAs enriched in the HOW(S)-HA pull-down . . . . .	98
3.2.10.	Finding the nearest half-sites to the 3’-UTR core motif . . . . .	101
3.2.11.	Identifying circRNAs pulled-down by HOW(S)-HA . . . . .	102
3.3.	Discussion . . . . .	105
3.3.1.	Overexpression of HOW(S)-HA in GSCs and early spermatogonia does not alter testis morphology or male fertility . . . . .	105

Table of contents

- 3.3.2. Identification of RNAs that HOW(S) binds in GSCs and spermatogonia . . . . . 105
- 3.3.3. Sequence drivers of HOW(S)-RNA binding . . . . . 106
- 3.3.4. Conclusions . . . . . 107
- 4. Understanding the role of HOW(S) regulated events in germ cells . . . . . 109
  - 4.1. Introduction . . . . . 109
  - 4.2. Results . . . . . 110
    - 4.2.1. Conditional knockdown of HOW in the testis results in large morphological changes . . . . . 110
    - 4.2.2. Selecting transcripts enriched in the HOW(S)-HA RIP-seq for phenotypic analysis . . . . . 115
    - 4.2.3. Hipk transcript expression in the testes of Hipk RNAi knockdowns and parental stocks . . . . . 119
    - 4.2.4. Assessing testis morphology phenotypes of the conditional Hipk knockdowns . . . . . 121
    - 4.2.5. The impact of conditional knockdown of Hipk on male fertility . . . 129
  - 4.3. Discussion . . . . . 135
    - 4.3.1. Screening RNAs bound by HOW(S) for function in spermatogenesis 135
    - 4.3.2. Hipk function in spermatogenesis . . . . . 135
    - 4.3.3. Conclusions . . . . . 137
- 5. Characterising the molecular mechanism of RNA binding by HOW . . . . . 139
  - 5.1. Introduction . . . . . 139
  - 5.2. Results . . . . . 141
    - 5.2.1. Optimisation of expression and purification of HOW's STAR domain 141
      - 5.2.1.1. Defining the STAR domain for protein purification . . . . . 141
      - 5.2.1.2. Small-scale screening of STAR domain constructs and growth conditions . . . . . 144
      - 5.2.1.3. Large-scale expression and purification of the STAR domain. 146
    - 5.2.2. Assessing the RNA-binding capacity of the HOW STAR domain . . 157
      - 5.2.2.1. Designing RNA oligos for RNA-binding experiments . . . . 157
      - 5.2.2.2. Measuring the RNA-binding affinity of the HOW STAR domain . . . . . 160
  - 5.3. Discussion . . . . . 167
    - 5.3.1. HOW's STAR domain binds the 3'-UTR motif, but not the 5'-UTR motif, with high affinity . . . . . 167
    - 5.3.2. Guanosine in the first position of the HOW binding motif may be preferred over adenosine . . . . . 167
    - 5.3.3. The effect of half-sites and RNA structure on HOW-RNA interactions 169
    - 5.3.4. Conclusions . . . . . 169
- 6. Discussion . . . . . 171
  - 6.1. HOW(S) binds RNAs associated with signal transduction functions . . . . 172
  - 6.2. *Hipk* mRNA could be an important target of HOW(S) for spermatogenesis 173

6.3.	Expanding the consensus sequence that drives HOW(S)–RNA interactions	174
6.4.	Additional features may contribute to HOW(S)–RNA interactions . . . . .	175
6.5.	HOW(S) function in the testis . . . . .	175
6.6.	Conclusions . . . . .	176
A.	Appendix I . . . . .	179
A.1.	DNA and amino acid sequences . . . . .	179
A.2.	Test statistic tables . . . . .	182
A.3.	RIP-seq data tables . . . . .	186
A.4.	qRT-PCR primer efficiencies . . . . .	202
A.5.	SIMV NP fluorescence anisotropy . . . . .	204
	References . . . . .	205



# List of tables

1.1.	Consensus sequences for STAR proteins. . . . .	20
1.2.	HOW's RNA binding targets. . . . .	28
2.1.	<i>D. melanogaster</i> stocks used. . . . .	36
2.2.	Protocol for PCR thermocycler. . . . .	39
2.3.	Primers for genotyping the UAS-HOW-S-HA fly stock. . . . .	39
2.4.	Fast cycle qRT-PCR thermocycler protocol. . . . .	41
2.5.	Primers used in qRT-PCR. . . . .	41
2.6.	Buffers and solutions used for immunofluorescence. . . . .	45
2.7.	Antibodies used for immunofluorescence. . . . .	46
2.8.	Buffers and solutions used for RIP-seq. . . . .	50
2.9.	Antibodies used for western blotting. . . . .	51
2.10.	Primers for cloning STAR domain into pOPIN vectors. . . . .	56
2.11.	Thermocycler protocol for PCR with Q5 enzyme. . . . .	56
2.12.	Buffers and solutions used for cloning. . . . .	57
2.13.	Buffers and solutions used in protein expression and purification. . . . .	61
3.1.	Mating rates of focal males in sperm competition assay. . . . .	72
3.2.	Breakdown of read counts through the pipeline by percentage. . . . .	81
3.3.	circRNAs from HOW(S)-HA RIP-seq with 10 reads or more mapped. . . . .	104
4.1.	Expression and sequence features of 12 transcripts enriched in the HOW(S)- HA RIP-seq selected for phenotypic experiments . . . . .	116
4.2.	Morphology and fertility defects observed from <i>Hipk</i> and <i>lola</i> conditional knockdowns. . . . .	118
4.3.	Statistics for the number of F <sub>2</sub> progeny produced. . . . .	134
5.1.	RNA oligos designed for RNA-binding experiments with HOW's STAR domain. . . . .	159
5.2.	Apparent K <sub>D</sub> of the STAR domain for RNAs containing the 5'- and 3'- UTR binding motifs. . . . .	165
A.1.	Statistics for the number of progeny produced after first mating in sperm competition assay. . . . .	182
A.2.	Statistics for the sperm competition assay. . . . .	183
A.3.	Statistics for transcript feature assignments. . . . .	183
A.4.	Statistics for <i>Hipk</i> expression determined via qRT-PCR. . . . .	183
A.5.	Statistics for the number of hub cells. . . . .	184
A.6.	Statistics for the number of vials with F <sub>2</sub> progeny in the 3-day mating assay. . . . .	185
A.7.	Breakdown of read counts through the pipeline. . . . .	187
A.8.	343 significantly enriched genes from the HOW(S)-HA RIP-seq. . . . .	188

*List of tables*

- A.9. 121 significantly enriched transcripts from the HOW(S)-HA RIP-seq. . . . 197
- A.10. GO terms from the enriched genes identified in the HOW(S)-HA RIP-seq. 201

# List of figures

1.1.	Schematic of cell types in <i>D. melanogaster</i> spermatogenesis. . . . .	2
1.2.	Schematic of the <i>D. melanogaster</i> testis stem cell niche. . . . .	4
1.3.	Stages of nuclear elongation during spermiogenesis. . . . .	7
1.4.	Formation of pole cells in <i>D. melanogaster</i> embryos. . . . .	9
1.5.	Schematic of the distal region of the adult <i>C. elegans</i> gonad. . . . .	10
1.6.	The domain architecture of the three STAR subfamilies. . . . .	14
1.7.	The variable loop in the maxi-KH domain of STAR proteins is extended. . .	16
1.8.	RNA binds in the hydrophobic cleft of STAR domains. . . . .	16
1.9.	GLD-1 and T-STAR dimerise via different mechanisms. . . . .	18
1.10.	Structural RNA features can increase STAR proteins' affinity for RNA. . .	21
1.11.	Schematic of the six HOW RNA transcripts. . . . .	27
1.12.	Schematic of HOW and Quaking protein isoforms. . . . .	27
1.13.	Schematic of HOW-dependent <i>neurexin-IV</i> splicing events. . . . .	30
3.1.	Schematic representing the PCR genotyping of the UAS-HOW-S-HA trans- genic fly line. . . . .	65
3.2.	Genotype of UAS-HOW-S-HA via PCR and agarose gel electrophoresis. . .	65
3.3.	Schematic of crosses performed to express HOW(S)-HA in germ cells. . . .	67
3.4.	<i>nanos</i> -GAL4:VP16-64277 and -4937 can drive HOW(S)-HA expression. . .	68
3.5.	HOW(S)-HA is predominately cytoplasmic. . . . .	69
3.6.	Schematic of sperm competition experimental design. . . . .	70
3.7.	HOW(S)-HA expressing flies do not produce fewer progeny than the parental fly lines. . . . .	71
3.8.	HOW(S)-HA overexpression flies and its two parental lines perform poorly compared to WT in sperm competition experiments. . . . .	73
3.9.	Schematic of RIP-seq protocol. . . . .	74
3.10.	Mouse anti-HA cannot verify the success of the HOW(S)-HA pull-down in western blots. . . . .	75
3.11.	HOW(S)-HA pull-down confirmed with rabbit anti-HA. . . . .	76
3.12.	Successful enrichment of HOW(S)-HA in pull-down replicates 2 and 3. . . .	77
3.13.	UAS-HOW-S-HA line has leaky expression of HOW(S)-HA. . . . .	78
3.14.	Schematic of RIP-seq differential enrichment analysis pipeline. . . . .	80
3.15.	HOW(S)-HA pull-down samples are more similar to each other than the total RNA samples. . . . .	82
3.16.	No difference in the percentage of reads mapped to transcript features between total RNA and pull-down RNA samples. . . . .	83
3.17.	Most genes and transcripts are neither enriched or depleted. . . . .	85
3.18.	BCV for the <i>nanos</i> -GAL4 samples should be set to the HOW(S)-HA BCV. . .	87
3.19.	343 genes are enriched in HOW(S)-HA RIP-seq. . . . .	89

List of figures

3.20.	121 transcripts are enriched in HOW(S)-HA RIP-seq. . . . .	90
3.21.	Small overlap in significantly enriched genes and transcripts. . . . .	91
3.22.	Signal transduction is the most enriched GO term from the list of HOW(S)- HA pull-down genes. . . . .	92
3.23.	GO terms related to signalling and communication are enriched among the HOW(S)-HA pull-down transcripts. . . . .	93
3.24.	78 significant enriched genes and 22 transcripts are signal transduction related. . . . .	94
3.25.	Motif enriched in the 3'-UTRs of HOW(S)-HA bound transcripts is very similar to the HRE defined within stripe mRNA. . . . .	96
3.26.	HOW(S)-enriched transcripts are more likely to contain multiple (A/U/G) CUAAC motifs. . . . .	97
3.27.	Motif enriched in the 5'-UTR of HOW(S)-HA RIP-seq transcripts is very different to the 3'-UTR motifs. . . . .	99
3.28.	HOW(S)-enriched transcripts are more likely to contain multiple GCG(A/U)G motifs. . . . .	100
3.29.	The majority of 3'-UTR core motif sites have a UAAY half-site less than 20 nucleotides away. . . . .	101
3.30.	Schematic of circRNA analysis pipeline. . . . .	103
4.1.	<i>nanos</i> -GAL4>HOW-RNAi testes are small and have few germ cells. . . . .	112
4.2.	No meiotic cells in <i>nanos</i> -GAL4>HOW-RNAi testes. . . . .	113
4.3.	Loss of elongating spermatids in <i>nanos</i> -GAL4>HOW-RNAi testes. . . . .	114
4.4.	Schematic of the RNAi crosses. . . . .	117
4.5.	Quantification of Hipk knockdown in the testes via qRT-PCR. . . . .	120
4.6.	Parental lines of the Hipk conditional knockdowns show all three phases of spermatogenesis. . . . .	122
4.7.	Testes from conditional Hipk RNAi 1 knockdowns have populations of germ cells from all three phases of spermatogenesis. . . . .	123
4.8.	Testes from conditional Hipk RNAi 2 maternal knockdowns are small and do not have late spermatids. . . . .	125
4.9.	Maternal Hipk RNAi 2 knockdown produce strong mutant phenotypes in the testes of their progeny. . . . .	126
4.10.	Testes from Hipk knockdowns do not have more hub cells than the parental lines. . . . .	128
4.11.	Schematic of 3-day fertility assay. . . . .	129
4.12.	The percentage of pairings that produced F <sub>2</sub> progeny was lowest between WT females and the F <sub>1</sub> males from the Hipk 2 maternal cross. . . . .	131
4.13.	Fertile Hipk 2 maternal cross knockdown males sire very few progeny. . . . .	133
5.1.	RNA-interacting residues are conserved between HOW and GLD-1 STAR domains. . . . .	142

5.2.	HOW(S)'s unique C-terminal amino acids are not close to the RNA-binding region. . . . .	143
5.3.	His- and His-GST-tags express STAR domain best in small scale expression screens. . . . .	145
5.4.	HisTrap purification and PreScission cleavage of His-STAR. . . . .	147
5.5.	Gel filtration of His-STAR suggests protein degradation. . . . .	148
5.6.	Purification of His-GST-STAR also suggests degradation. . . . .	150
5.7.	MS/MS analysis suggests the presence of full length STAR domain and a C-terminal degradation. . . . .	152
5.8.	Truncated STAR is 10 or 11 amino acids shorter than the full-length STAR protein. . . . .	153
5.9.	Ni <sup>2+</sup> affinity chromatography of His-GST-STAR. . . . .	155
5.10.	Purification of His-GST-STAR produces predominantly full-length STAR domain. . . . .	156
5.11.	Schematic of the 3'-UTRs of eight <i>lola</i> transcripts and the location of (A/U/G) CUAAC motifs. . . . .	158
5.12.	Schematic of the 3'-UTRs of the four <i>hipk</i> transcripts and the location of (A/U/G)CUAAC motifs. . . . .	158
5.13.	Schematic of the 5'-UTRs of the seven <i>jvl</i> transcripts and the location of GCG(A/U)G motifs. . . . .	158
5.14.	Schematic of unbound vs protein-bound RNA in fluorescence anisotropy assays. . . . .	160
5.15.	The positive protein control, SIMV NP, binds poly(A) <sub>10</sub> and poly(U) <sub>10</sub> . . .	161
5.16.	The STAR domain binds the <i>lola</i> core oligo selectively. . . . .	162
5.17.	STAR domain binds the 3'-UTR motif derived oligos. . . . .	163
5.18.	The predicted fold of the 'Hipk-RA core + half-site' oligo has a stem-loop. . . . .	164
5.19.	STAR domain binds the 'Hipk-RA core' oligo with highest affinity. . . . .	166
A.1.	Primer efficiencies for primers used to evaluate Hipk RNAi 1 knockdown. . .	202
A.2.	Primer efficiencies for primers used to evaluate Hipk RNAi 2 knockdown. . .	203
A.3.	SIMV NP binds all seven oligos derived from <i>lola</i> , <i>Hipk</i> and <i>jvl</i> transcripts. . .	204



# List of abbreviations

Abbreviation	Description
ANOVA	analysis of variance
Bam	bag of marbles
BBP	branch point binding protein
BCV	biological coefficient of variation
BDSC	Bloomington Drosophila Stock Center
Bgcn	benign gonial cell neoplasm
bp	base pair
CDS	coding sequence
circRNA	circular RNA
CySC	cyst stem cell
DAZ	deleted in azoospermia
DAZL	deleted in azoospermia-like
DREME	Discriminative regular expression motif elicitation
DSHB	Developmental Studies Hybridoma Bank
DTT	dithiothreitol
E	embryonic day (when E is followed by a number, e.g. E11)
EMSA	electrophoretic mobility shift assay
FA	fluorescence anisotropy
FBF	<i>fem-3</i> mRNA binding factor
fwd	forward
gDNA	genomic DNA
GLD-1	defective in germline development 1
GO	gene ontology
GORila	GO enrichment analysis and visualization tool
GSC	germline stem cell
GST	glutathione-S-transferase
GTF	gene transfer format
HA	haemagglutinin
Hipk	homeodomain-interacting protein kinase
HOW	held out wings
HOW(L)	held out wings long
HOW(M)	held out wings medium
HOW(S)	held out wings short
HRE	HOW response element
jvl	javelin-like
KH	K homology
LB	Luria–Bertani
lola	longitudinals lacking

## List of figures

MBP	maltose binding protein
MS	mass spectrometry
msl-2	male-specific lethal 2
NP	nucleoprotein
Nrx-IV	neurexin IV
Nsr	novel spermatogenesis regulator
nt	nucleotide
P	position (when P is followed by a number, e.g. P1)
PAGE	polyacrylamide gel electrophoresis
PBS	phosphate buffered saline
PCA	principal component analysis
PCR	polymerase chain reaction
PGC	primordial germ cell
PSI	P-element somatic inhibitor
qRT-PCR	quantitative reverse transcription PCR
RBD	RNA-binding domain
RBMXL2	RBMX-like 2
RBM5	RNA-binding motif 5
RBP	RNA-binding protein
rev	reverse
RIP	ribonucleoprotein immunoprecipitation
RNAi	RNA interference
RNP	ribonucleoprotein particle
Rok	Rho kinase
Rp49	ribosomal protein 49
RRM	RNA recognition motif
SDS	sodium dodecyl sulphate
SEC	size exclusion chromatography
SEM	standard error of the mean
SLM-1	Sam68-like mammalian protein 1
SF1	splicing factor 1
shRNA	short hairpin RNA
SIMV NP	Simbu orthobunyavirus nucleoprotein
SLIRP1	SRA stem-loop interacting RNA binding protein 1
STAR	signal transduction and activation of RNA
SUMO	small ubiquitin-like modifier
SXL	sex lethal
T-STAR	testis-STAR
TB	terrific broth
TER94	transitional endoplasmic reticulum 94
TGCT	testicular germ cell tumour
TIAR	TIA-1 related protein

UAS	upstream activation sequence
UTR	untranslated region
VDRC	Vienna Drosophila Resource Center
WT	wild type
Zfh1	Zinc finger homeodomain 1
ZnF	zinc finger
2YT	2X yeast-tryptone

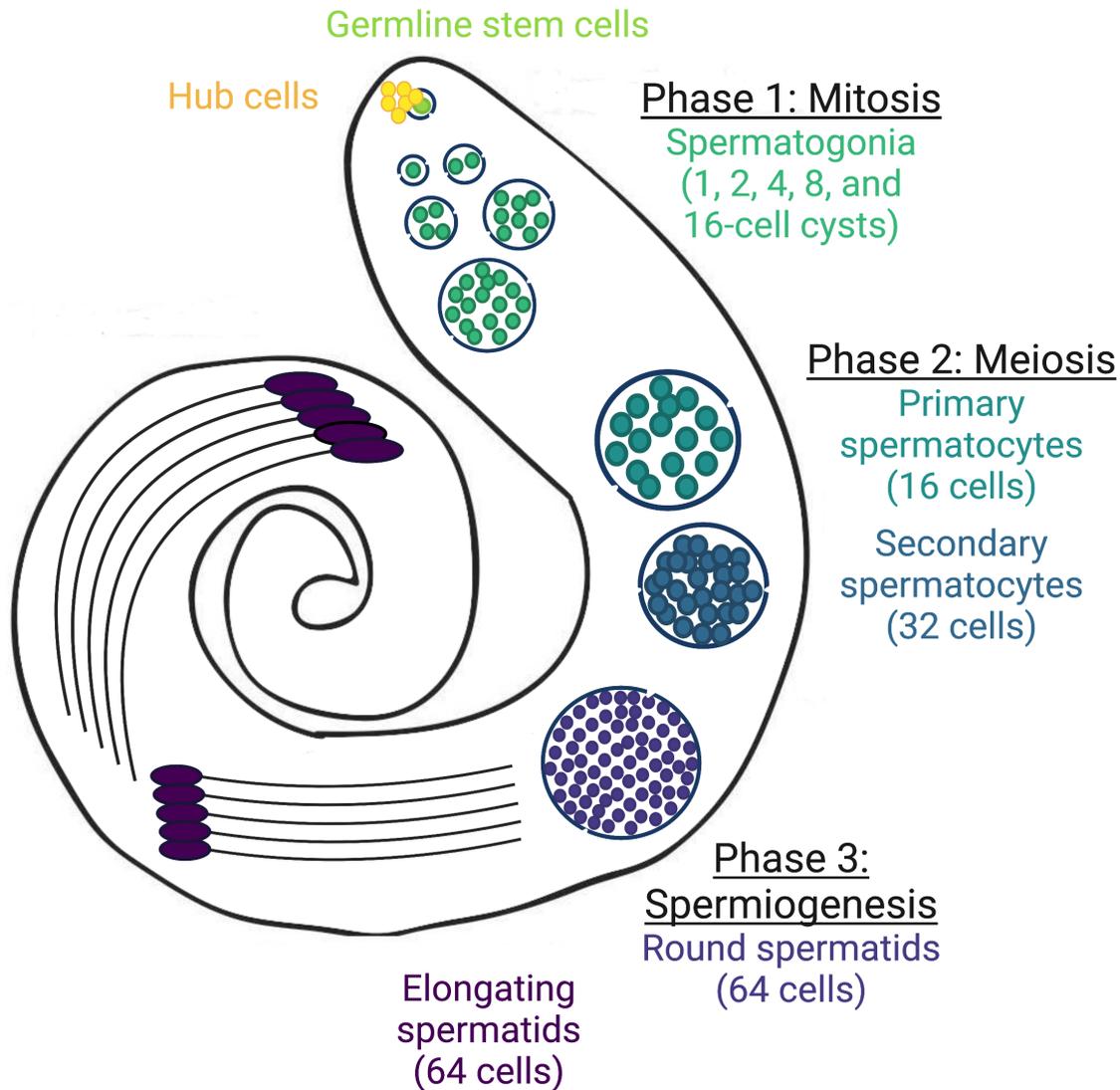


# 1. Introduction

## 1.1. Spermatogenesis in *D. melanogaster*

Spermatogenesis is the process where diploid germline stem cells (GSCs) differentiate into haploid spermatozoa. This process not only includes the meiotic divisions that are unique to gametogenesis but also substantial physical changes as the cells develop from round GSCs into mature sperm cells that contain specific structures and organelles, such as the flagellar axoneme and the acrosome. Additionally, the testis is an incredibly productive organ, and human testes generate  $\sim 100$  million sperm each day in adult males (Sharpe, 1994). Thus, maintaining the stem cell population is key to the continuation of successful spermatogenesis.

*D. melanogaster* has been used as a model organism to study many processes and spermatogenesis is no exception. For example, much has been learnt about how the stem cell niche, the microenvironment in which stem cells reside, support and maintain the stem cell population in the *D. melanogaster* testis. Similar to mammalian testes, the *D. melanogaster* testis also expresses the largest number of tissue-specific genes (Chintapalli et al., 2007; Soumillon et al., 2013; Uhlén et al., 2015). While there are some differences between spermatogenesis in flies and mammals, such as recombination not occurring during meiosis in *D. melanogaster* males (McKee et al., 2012), many of the underlying features and mechanisms of spermatogenesis are conserved between humans and flies. For example, the three main phases of differentiation in spermatogenesis follow the same broad structure in flies and mammals, making *D. melanogaster* an appropriate model. First is the mitotic phase, when the transit-amplifying mitotic divisions occur, during this time the cells are referred to as spermatogonia (Fig 1.1). The second phase is meiosis, where the spermatogonia mature into spermatocytes and undergo the two rounds of meiosis to become haploid spermatids (Fig 1.1). The third and final phase is spermiogenesis, where the major morphological changes occur with the round spermatids differentiating into mature spermatozoa (Fig 1.1).



**Figure 1.1. Schematic of cell types in *D. melanogaster* spermatogenesis.** The *D. melanogaster* testis is a coiled blind ended tube (outlined in black). At the apical tip is a cluster of somatic hub cells. The germline stem cells (GSCs) surround the hub (one shown here for clarity), and each GSC is surrounded by two cyst stem cells. When the GSC divides asymmetrically one of the child cells becomes a spermatogonium that undergoes 4 rounds of mitosis during the first phase of spermatogenesis. The cyst is surrounded by two mature cyst cells throughout spermatogenesis. The 16-cell spermatogonial cyst then grows in volume and matures into primary spermatocytes. After meiosis I the secondary spermatocytes form a 32-cell cyst, and after meiosis II a 64-cell cyst of round spermatids. The spermatids (5 are shown and without the surrounding cyst cells for clarity) then elongate and mature into spermatozoa during spermiogenesis. Adapted from Witt et al. (2019).

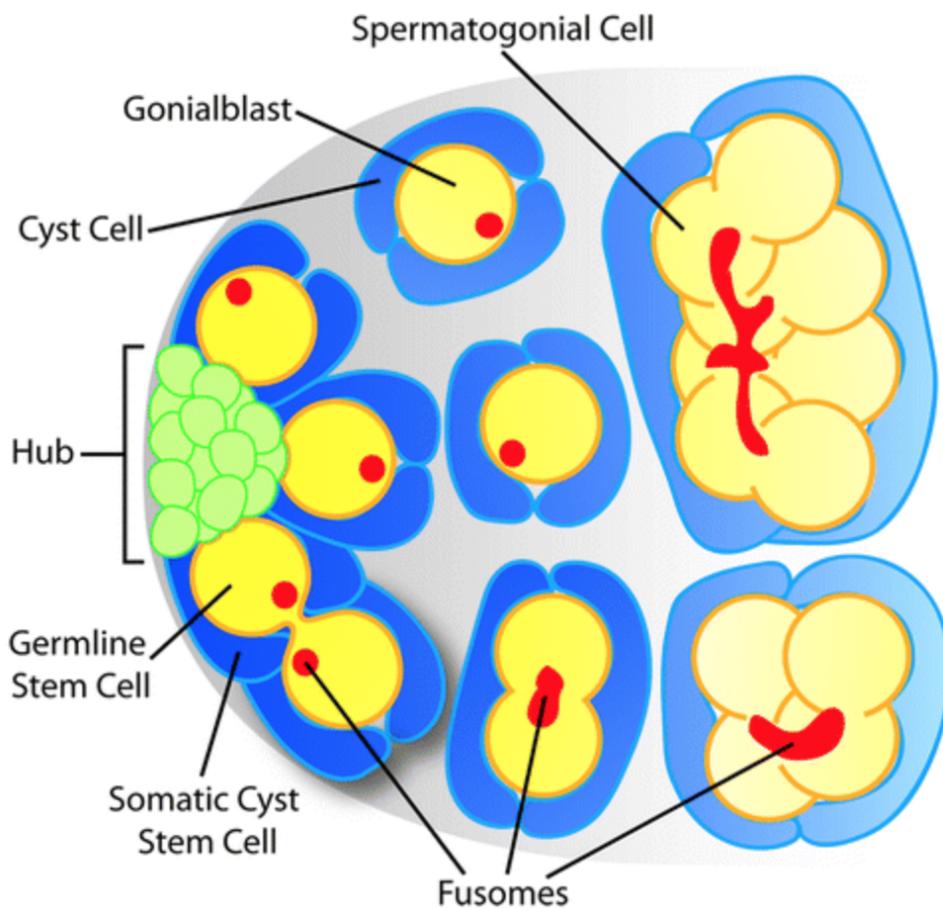
### 1.1.1. The stem cell niche of the *D. melanogaster* testis

Spermatogenesis in *D. melanogaster* occurs in the two testes, which are coiled blind ended tubes each connected to a seminal vesicle that join to a single ejaculatory duct. Each testis has a single stem cell niche at the apical tip, and as the germ cells develop they move away from the tip and move through the testis (Fig 1.1). This clear cellular architecture of the *D. melanogaster* testis is one of its advantages as a model for spermatogenesis.

There are two stem cell populations that are supported by the somatic hub cells at the apical tip, the GSCs and the cyst stem cells (CySCs). The cluster of hub cells are surrounded by 6–12 GSCs and each GSC is surrounded by two CySCs (Fig 1.2; Spradling et al., 2011). The contact between the GSCs and the hub cells provide polarity to the GSCs and ensure the asymmetric division of the GSC into a new GSC that remains attached to the hub and a new gonialblast (or spermatogonium) that is displaced from the hub (Yamashita et al., 2003; Yamashita et al., 2007). This asymmetry helps to maintain the balance between self-renewal and differentiation. The two CySCs also divide asymmetrically when a GSC does, the cells surrounding the newly divided GSC maintain their stem identity while the two cells around the new gonialblast are now cyst cells (Fig 1.2; Cheng et al., 2011). These mature cyst cells do not divide again, instead they grow to accommodate the germ cyst as the gonialblast divides and differentiates (Fig 1.2; Zoller and Schulz, 2012).

The intercellular signalling between these three cell types — hub cells, CySCs and GSCs — is essential for maintaining the GSC population. While multiple pathways are involved (reviewed in Matunis et al., 2012), one of the most prominent and best characterised pathways is the JAK-STAT signalling pathway. Without either JAK or STAT proteins (Hopscotch and Stat92E in *D. melanogaster*, respectively), GSCs are not able to self-renew, instead they differentiate and the stem cell population is lost (Kiger et al., 2001; Tulina and Matunis, 2001). The signalling begins with the hub cells that secrete the protein Unpaired, this activates JAK-STAT signalling in both the CySCs and the GSCs (Kiger et al., 2001; Tulina and Matunis, 2001). The activation of JAK-STAT directly promotes the cell adhesion to the hub cells in GSCs and self-renewal in CySCs (Leatherman and Dinardo, 2010). GSC self-renewal is maintained by signalling from the CySCs via Zinc-finger homeodomain 1 (*Zfh1*). *Zfh1* is a target of Stat92E, and when expressed by CySCs ectopically it can maintain the self-renewal of GSCs even when removed from the hub cells (Leatherman and Di Nardo, 2008). Typically, Unpaired does not reach the gonialblasts or its surrounding cyst cells, which are not in direct contact with the hub (Fig 1.2; Kiger et al., 2001; Tulina and Matunis, 2001), and thus these cells do not self-renew and the gonialblast proceeds with the mitotic phase of spermatogenesis.

1. Introduction



**Figure 1.2. Schematic of the *D. melanogaster* testis stem cell niche.** The somatic cells (green) reside at the apical tip of the *D. melanogaster* testis. The germ cells (yellow) directly next to the hub cells are the germline stem cells (GSCs) and each is surrounded by two somatic cyst stem cells (CySCs; blue). When a GSC divides one cell remains in contact with the hub, the other becomes a gonialblast surrounded by two mature cyst cells (also in blue). As the spermatogonia divide from this gonialblast the fusomes (red) keep the cells connected and the two cyst cells grow to keep the cyst encapsulated. Image from Matunis et al. (2012).

### 1.1.2. Mitosis and the switch to meiosis

During the mitotic phase, a gonialblast undergoes four rounds of cell division to result in a cyst of 16 spermatogonia. These divisions, often referred to as transit-amplifying divisions, are employed by many types of adult stem cells before they enter terminal differentiation to decrease the burden of cell division on the stem cells directly (Nelson et al., 2019). The sibling spermatogonia within a cyst are connected because they do not complete cytokinesis when they divide. Instead intercellular bridges, also called ring canals, form and the branched vesicular fusome links all the cells of a cyst together (Fig 1.2; Hime et al., 1996). These bridges, which are also seen in mammals, allows the cells within the cyst to stay coordinated through their differentiation from spermatogonia to spermatozoa (reviewed in Greenbaum et al., 2011). However, spermatogonia are able to breakaway from the cysts, dedifferentiate back into GSCs and regain contact with the hub cells (Brawley and Matunis, 2004). This dedifferentiation increases when the GSCs are killed with irradiation, implying that that this process can be a repair mechanism to replenish the stem cell population (Cheng et al., 2008). Dedifferentiation of spermatogonia has since been in the mice testis too (Nakagawa et al., 2010).

The mitotic divisions help to maintain the GSC population, however, the number of these divisions must be tightly regulated so as not to result in an overproliferation of undifferentiated cells. One study counted the number of cells inside 112 primary spermatocyte cysts and found that 99% of them contained 16 cells, implying that they had undergone exactly four rounds of mitosis (Insko et al., 2009). One of the main factors that regulates the number of mitotic divisions is the protein Bag of marbles (Bam). Bam had previously been identified, along with Benign gonial cell neoplasm (Bgcn), to regulate the switch from mitosis to meiosis, as null mutants of either of these genes results in no spermatocytes and an accumulation of undifferentiated spermatogonia in the testis (Gönczy et al., 1997). When the level of Bam protein was increased (by deleting the PEST motif that targets Bam for rapid turnover) several of the primary spermatocyte cysts had only 8 cells (rather than 16), while flies with just one copy of the *bam* gene had cysts that entered meiosis after five or more rounds of mitosis (Insko et al., 2009). However, flies with one copy of *bgcn* did not undergo extra rounds of mitosis (Insko et al., 2009). Thus, it appears that levels of Bam, but not Bgcn, influences the number of mitotic divisions prior to meiosis, but both are needed to complete the transition from mitosis to meiosis.

### 1.1.3. Meiosis and spermiogenesis

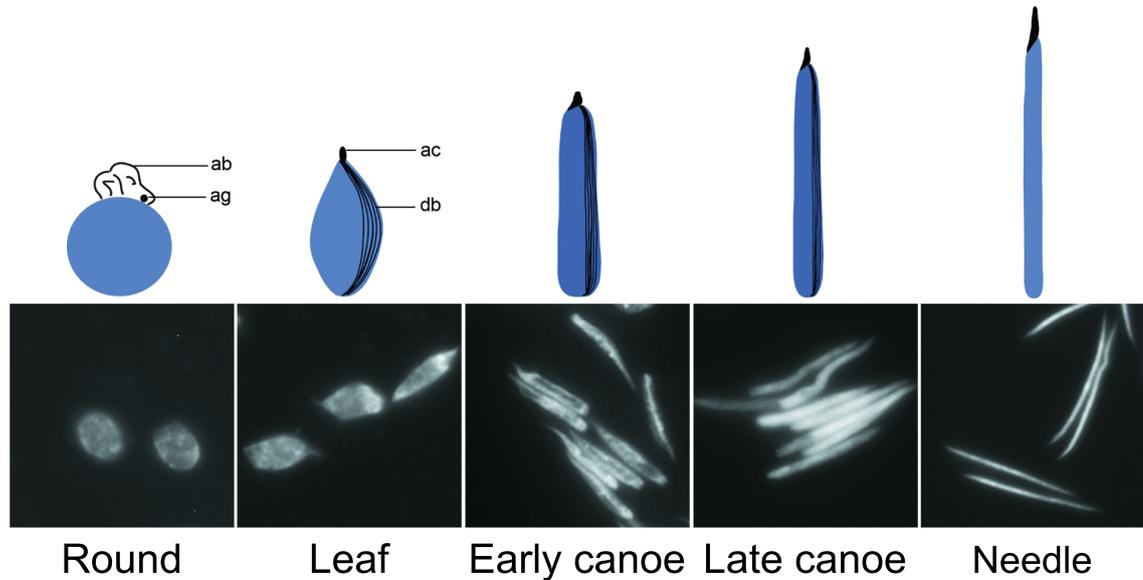
After the completion of the fourth round of mitosis, DNA is replicated in the premeiotic S phase. Following this there is an extended G<sub>2</sub> phase where the spermatocytes grow substantially in volume, increasing around 25-fold (McKee et al., 2012). During this time the cells become very transcriptionally active (reviewed in White-Cooper, 2010 and Lim et al., 2012). As there is little transcription during spermiogenesis, many of the proteins that are required for spermatid differentiation are transcribed in primary spermatocytes

## 1. Introduction

and these RNAs are translationally repressed until after meiosis (Schäfer et al., 1995). One of the key events for meiosis I to proceed, is the derepression of *twine* mRNA translation by Boule and eIF4G2 (Maines and Wasserman, 1999; Baker and Fuller, 2007). Twine is a homologue of the yeast phosphatase Cdc25, and dephosphorylates Cdc2, which triggers the transition into the first meiotic division after the G<sub>2</sub> phase (Sigrist et al., 1995). eIF4G2 is also required for the increased protein expression of Cyclin B during the G<sub>2</sub>/M transition (Baker and Fuller, 2007).

Intriguingly, in flies with null or loss of function mutations in genes that have a key role in meiosis (e.g. *twine* and *boule*) the spermatocytes can still begin spermatid differentiation without having completed meiosis, though crucially they will not produce haploid cells (Alphey et al., 1992; Eberhart et al., 1996). However, there are a set of genes, referred to as ‘meiotic arrest genes’, whereby if they are mutated there is no differentiation of the primary spermatocytes beyond the G<sub>2</sub> phase, i.e. there are no meiotic divisions and no signs of spermatid differentiation (White-Cooper et al., 1998). Many of these meiotic arrest genes are exclusively expressed in spermatocytes and contribute to either the testis-specific meiotic arrest complex (tMAC) or the group of testis-specific TBP-associated factors (tTAFs) which are likely paralogues of the somatic TFIID complex (reviewed in White-Cooper, 2010). These two complexes are responsible for stimulating a large amount of the transcriptional activity that takes place in the primary spermatocytes, resulting in the expression of hundreds of genes (White-Cooper, 2010; Laktionov et al., 2018).

In spermiogenesis the round spermatids differentiate into mature sperm cells, many specialised organelles develop and there are a host of physical changes (Fig 1.1; reviewed in Fabian and Brill, 2012). The round spermatids become polarised so that the flagellar axoneme elongates towards the apical tip and the nuclei bundle at the basal side of the cyst. The round nuclei also elongate in several stages named after the shapes they adopt: leaf, early canoe, late canoe, and the final needle-shaped nuclei (Fabian and Brill, 2012; Fig 1.3). The chromatin in the nucleus changes from being histone based to protamine based. As described earlier, many of the genes required for spermiogenesis are transcribed in the primary spermatocytes. However, during the elongation of the spermatids and before the switch from histones to protamines, a small number of genes are post-meiotically transcribed (Barreau et al., 2008). The transcribed RNAs are localised to the distal end of the spermatid tails, i.e. towards the apical tip of the testis, and are categorised as either ‘comet’ genes or ‘cup’ genes based on the RNA localisation patterns in the tails (Barreau et al., 2008). Once elongation and nuclear shaping are complete the individualisation complex moves through the cyst and the intercellular bridges between the cells are removed, along with any redundant organelles. Finally, the mature sperm cells are coiled into the base of the testis and can then be stored in the seminal vesicles.



**Figure 1.3. Stages of nuclear elongation during spermiogenesis.** During spermiogenesis the spermatid nucleus elongates through several stages: round, leaf, early canoe, late canoe, and the final needle-shaped nucleus. The top row shows schematics of each of these stages. Ab is acroblast, ag is acrosomal granule, ac is acrosome, db is dense body (adapted from Fabian and Brill, 2012). The bottom row is microscopy images from squashed preparations of testes with DNA stained using Hoechst dye (adapted from Hundertmark et al., 2018).

## 1.2. RNA-binding proteins and spermatogenesis

Gene expression is incredibly dynamic during spermatogenesis. The testis is a very transcriptionally active tissue, with many tissue-specific genes and proteins (Chintapalli et al., 2007; Soumillon et al., 2013; Uhlén et al., 2015). At the same time, the transcriptome and the proteome are poorly correlated in the testis, more so than in other organs in humans (Cagney et al., 2005). This suggests that a substantial amount of post-transcriptional gene regulation is occurring. Some modes of gene regulation are well known, for example, as described in the previous section, there are many genes that are transcribed early in spermatogenesis but are translationally repressed and stored for later stages.

RNA-binding proteins (RBPs) are central to post-transcriptional gene regulation and there are many RBPs that are required for normal spermatogenesis. The importance of RNA processing in spermatogenesis in different species has been recently reviewed (Legrand and Hobbs, 2018), and there are additional reviews on RBPs involved in mouse spermatogenesis (Qi, 2016; Idler and Yan, 2012). Many of these RBPs are conserved in *D. melanogaster* or were even first found to have roles in germ cells in flies. In this section, I have selected examples for each phase of spermatogenesis to highlight how RBPs are important throughout this process. RBPs are immensely diverse with many modes of binding and action, and this selection of RBPs have a range of functions from splicing, to RNA localisation, to activating and repressing mRNA translation. Several of these examples also highlight how

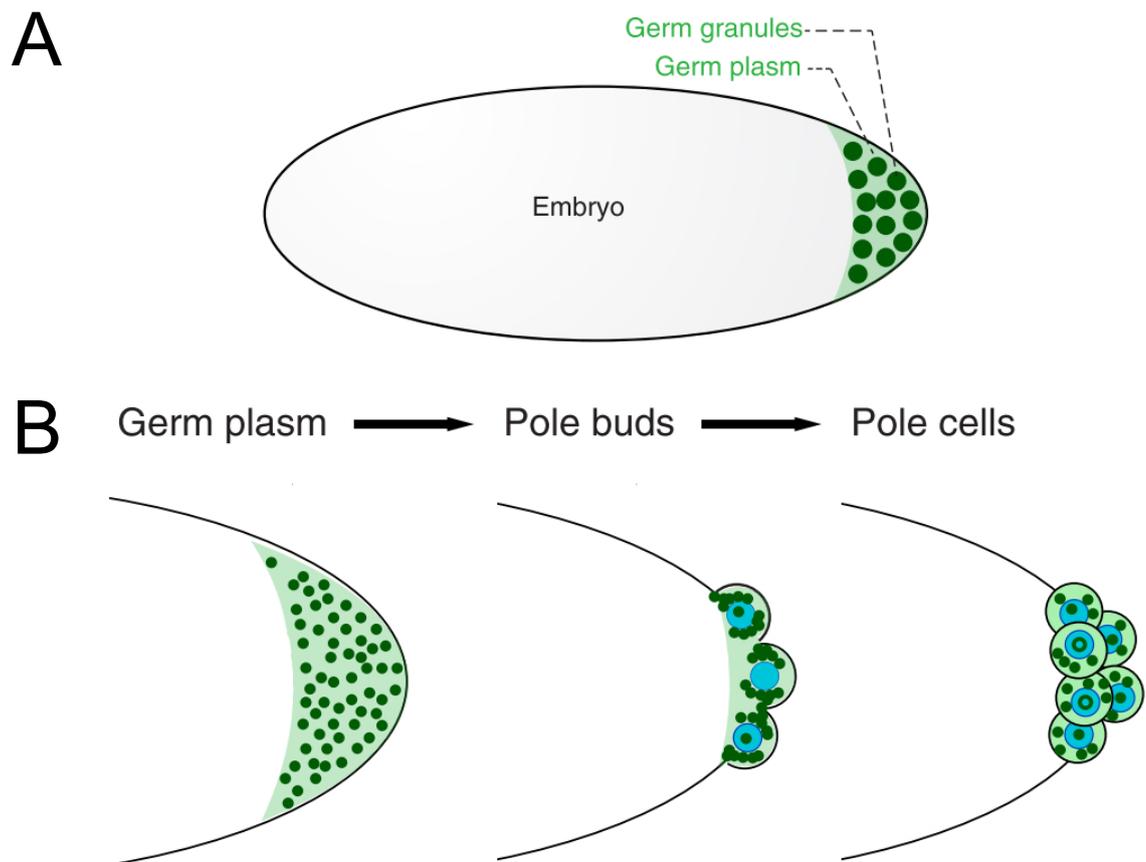
## 1. Introduction

omics technologies have been used to understand the function of these RBPs (reviewed further in Hentze et al., 2018 and Hafner et al., 2021).

### 1.2.1. Formation of the germline

While the development of the primordial germ stem cells (PGCs) during embryogenesis is not specific to the development of sperm cells, as it is necessary for oogenesis too, it would be remiss not to discuss PGCs as RBPs are immensely important for their development. Research in *D. melanogaster* embryos has been foundational in understanding how the germline forms. The early *D. melanogaster* embryo is a multinucleate syncytium, and the first cells to form are the PGCs, also known as pole cells in flies (Fig 1.4). Without the specialised cytoplasm, the germ plasm, at the posterior pole of the embryo the PGCs would not form. Germ granules, comprised of RNAs and proteins, are a crucial component of the germ plasm (reviewed in Trcek and Lehmann, 2019). Most of the core proteins in the germ granules are RBPs, such as Vasa (an ATP-dependent RNA helicase), Aubergine (a member of the PIWI family), Nanos (a translational repressor with a zinc finger domain protein), and Oskar (an RBP with a lipase-like OSK domain), and several of these are functionally conserved in mammals (Trcek and Lehmann, 2019). For example, in addition to *D. melanogaster* embryos Vasa has been detected in ribonucleoprotein (RNP) granules in the germ cells of developing embryos in species ranging from humans to oysters (*Crassostrea gigas*) to zebrafish (reviewed in Gustafson and Wessel, 2010). Mutations in Vasa in *D. melanogaster* result in mutant embryos with no germ granules or pole cells (Schüpbach and Wieschaus, 1989).

Many other animal embryos form a specialised germ plasm, though mammals are an exception to this. However, RBPs are still important to mammalian PGC formation and maintenance. For example, TIA-1 related protein (TIAR, also known as TIAL1) is a translational repressor (Mazan-Mamczarz et al., 2006) and mice deficient in TIAR are sterile and do not generate either spermatogonia or oogonia (Beck et al., 1998). PGCs are present in both wild type (WT) and *tiar*<sup>-/-</sup> at embryonic day 11.5 (E11.5), though there are fewer in the *tiar*<sup>-/-</sup> embryos. By E13.5 PGCs are totally absent in the *tiar*<sup>-/-</sup> mice (Beck et al., 1998). The exact mechanism of TIAR function for the survival of PGCs has not been elucidated. However, TIAR has been shown to colocalise with another RBP, Nanos3, in mice (Yamaji et al., 2010). And similar to TIAR, when *nanos3* is knocked out there are a reduced number of PGCs at E12.5 and none present at E15.5 (Tsuda et al., 2003).



**Figure 1.4. Formation of pole cells in *D. melanogaster* embryos.** A) The early *D. melanogaster* embryo is a single multinucleated cell. At the posterior end is a type of specialised cytoplasm, the germ plasm (green), which contains germ granules (dark green) formed of proteins and RNAs. B) When the nuclei (blue) migrate to the edge of the embryo the pole cells are the first cells to form. The germ plasm is encapsulated as the cells bud to form the cytoplasm of the new pole cells, and some germ granules become localised to the nucleus. Images from Trcek and Lehmann (2019).

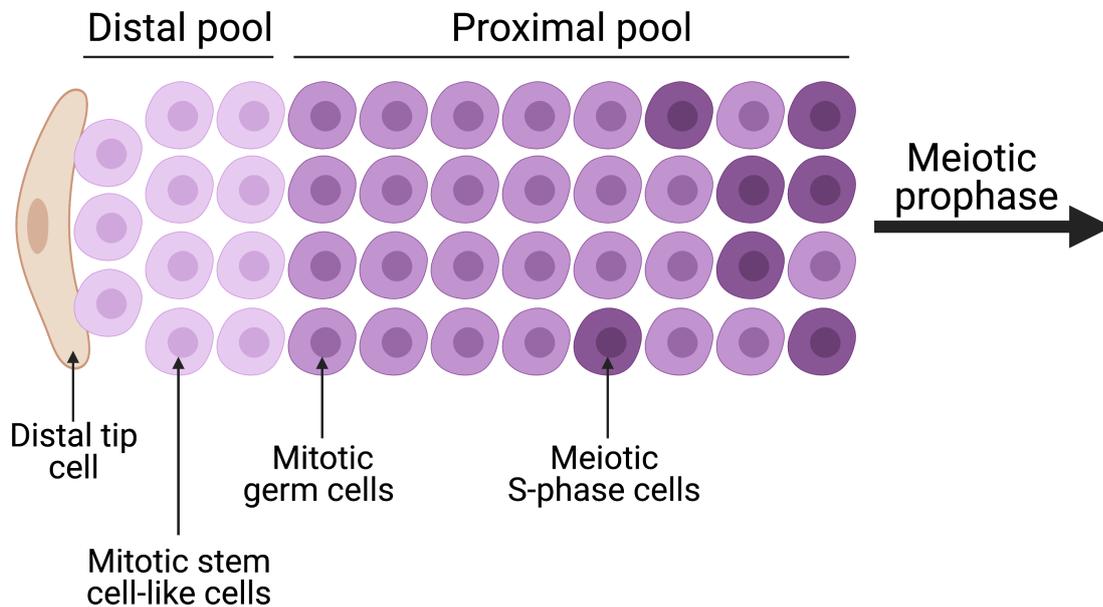
### 1.2.2. GSC homeostasis and mitosis

In *D. melanogaster*, there is a single *nanos* gene whose protein product acts as a translational repressor. It is required for PGC maintenance and GSC homeostasis in the female germline but not the male germline (Gilboa and Lehmann, 2004; Wang and Lin, 2004; Spradling et al., 2011). However, in mice and humans there are three Nanos proteins, and Nanos2 is specifically important for spermatogenesis (Tsuda et al., 2003). Nanos2 is highly expressed in the mouse testis and knockout of Nanos2 results in a loss of spermatogonia (Tsuda et al., 2003). Additional work found that Nanos2 is essential for GSC self-renewal, as a disruption of Nanos2 expression results in a loss of these stem cells and overexpression sees an overproliferation of undifferentiated spermatogonia (Sada et al., 2009). Nanos2 suppresses differentiation by promoting the formation of RNPs, which capture mRNAs such as *Dazl* and *Sohlh2*, and inhibits them from being translated into proteins that promote mitotic proliferation and differentiation (Reynolds et al., 2005; Zhou

## 1. Introduction

et al., 2015). Along with mRNAs, mTOR protein is found in these RNPs, which further inhibits translation in the GSCs (Zhou et al., 2015).

In *C. elegans* there is a pool of GSCs in the distal tip of the gonad arm, these cells are mitotically active but maintain a stem cell-like state (Cinquin et al., 2010). The germ cells in the proximal pool are also mitotically active but transition to a more differentiated state to enter meiosis as they move away from the distal tip (Fig 1.5; Cinquin et al., 2010). The two *fem-3* mRNA binding factor (FBF) proteins, FBF-1 and FBF-2, maintain the self-renewal of the mitotically active GSCs by repressing many of the RNAs needed for differentiation and meiosis. Hundreds of FBF mRNA targets have been identified through RIP-chip (RNP immunoprecipitation followed by microarray analysis) and the majority of the targets appear to be downregulated by FBF (Kershner and Kimble, 2010). One of the important promoters of meiosis that FBF represses is *GLD-1*, an RBP that will be discussed in more detail in section 1.3, by directly binding to the 3'-UTR of *gld-1* mRNA (Crittenden et al., 2002). Unsurprisingly, given that the FBF proteins repress a number of RNAs necessary for meiosis, germ cells in double FBF mutant worms all enter meiosis and no stem cell-like or undifferentiated mitotic cells remain in the distal tip (Crittenden et al., 2002).



**Figure 1.5. Schematic of the distal region of the adult *C. elegans* gonad.** At the distal tip of each *C. elegans* gonad is a somatic distal tip cell (beige), which helps to maintain the distal pool of mitotic germ cells (light purple) in a stem cell-like state. As the cells divide and move proximally through the gonad the mitotic germ cells (purple) begin to differentiate and enter meiosis (dark purple).

### 1.2.3. Meiosis

The Y-linked gene *deleted in azoospermia* (*DAZ*) encodes an RBP with one RNA recognition motif (RRM) domain and was found to be deleted in 12–15% of a cohort of azoospermic men (Reijo et al., 1995). The DAZ protein family also includes DAZ-like (DAZL) and Boule in humans (reviewed in VanGompel and Xu, 2011). All three proteins are expressed almost exclusively in germ cells and each have been shown to be important for spermatogenesis in several species (VanGompel and Xu, 2011). DAZL and Boule both have roles in meiosis, the latter is most likely the founding member of this protein family and it is the only DAZ family protein found in *D. melanogaster* (Xu et al., 2001). Boule is also functionally conserved, for example, human Boule is able to rescue the meiosis defect phenotype in fly *boule* mutants (Xu et al., 2003), and the loss of Boule has been associated with meiotic arrest in human spermatogenesis too (Luetjens et al., 2004). However, in mice Boule might be more important for spermiogenesis than meiosis (VanGompel and Xu, 2010). As described in section 1.1.3, Boule in *D. melanogaster* activates the translation of *twine* mRNA to trigger the entry of spermatocytes into meiosis (Maines and Wasserman, 1999). Translation activation activity has since been demonstrated for DAZL from *Xenopus laevis*, human and mouse, as well as human DAZ and Boule proteins via MS2 tethering assays carried out in *X. laevis* oocytes (Collier et al., 2005). This activity is dependent on the DAZ family proteins interacting with poly(A) binding protein 1 (PABP1) with or without the poly(A) tails on the target mRNAs (Collier et al., 2005). Additional non-translational roles for the DAZ protein family have also been explored but these have not been directly tied to their function in meiosis (VanGompel and Xu, 2011).

Point mutations in the *RBMX-like 2* (*RBMXL2*) gene, which encodes a nuclear RBP, have been detected in infertile men (Westerveld et al., 2004) but its function was only recently characterised via an RBMXL2 mouse knockout model (Ehrmann et al., 2019). These mice had much smaller testes than WT and produced no mature sperm, while no impact on female fertility was observed. Histology and immunofluorescence experiments revealed a block in meiosis at the diplotene substage of meiotic prophase (Ehrmann et al., 2019). A combination of *in vivo* RNA-seq and CLIP-seq experiments identified that a loss of RBMXL2 results in an increase in the use of cryptic splicing sites, altering the expression and/or amino acid sequences of several proteins including Meioc (Meiosis specific gene with coiled-coil domain) and Kdm4d (Lysine-specific demethylase 4D; Ehrmann et al., 2019). Though the exact mechanism of how RBMXL2 regulates splicing during meiosis is unclear, there is evidence it works with the splicing factor Tra2 $\beta$  to repress cryptic splicing (Ehrmann et al., 2019).

## 1. Introduction

### 1.2.4. Spermiogenesis

RNA-binding motif 5 (RBM5) was identified as an RBP necessary for spermiogenesis in a mouse mutagenesis screen. Male mice with a point mutation in the *Rbm5* gene (referred to as *Rbm5<sup>sda/sda</sup>* mice) were sterile, while females were unaffected (O’Bryan et al., 2013). This missense mutation results in a conversion of an arginine residue to a proline (R263P) in the second RRM domain of RBM5, a highly conserved residue in the RNA-binding interface (Song et al., 2012). The *Rbm5<sup>sda/sda</sup>* mice showed a defect in spermatid differentiation, as round spermatids were present in the testis but there were no elongating spermatids or mature sperm cells (O’Bryan et al., 2013). RIP-chip was performed on round spermatids isolated from WT testes, this identified 11 putative RNA targets of RBM5, four of which had previously been linked to spermatogenesis (O’Bryan et al., 2013). 8 of the 11 putative targets had aberrant splicing patterns in round spermatids from *Rbm5<sup>sda/sda</sup>* mice compared with those isolated from WT. Further, it was demonstrated that for one of the targets, *St5*, its aberrant splicing resulted in a decrease in ST5 protein and an increase in phosphorylation of ERK1 and ERK2 (the downstream functional targets of ST5) in round spermatids (O’Bryan et al., 2013). Though, this defect in *St5* splicing is unlikely to be the only cause of the spermatid differentiation arrest seen in the *Rbm5<sup>sda/sda</sup>* mice.

The final example of an RBP important in spermatogenesis is Orb2, which is one of the two *Drosophila* cytoplasmic and polyadenylation element binding (CPEB) proteins, and has multiple roles in spermatogenesis (Xu et al., 2012; Xu et al., 2014). Male flies with mutant alleles that affect the expression of the longer 75 kDa Orb2 protein isoform are sterile and show defects in both meiosis and spermiogenesis (Xu et al., 2012). While the function of Orb2 in spermiogenesis is difficult to separate from its essential role in meiosis, some observations have been made with an Orb2 mutant which has the N-terminal poly-Q domain deleted (Orb2<sup>ΔQ</sup>). In Orb2<sup>ΔQ</sup> males meiosis proceeds as normal but many of the testes exhibit abnormal spermatid differentiation (Xu et al., 2012). For example, 70% of the Orb<sup>ΔQ</sup> mutants had over-elongated flagellar axonemes (Xu et al., 2012), indicating that the spermatid defects observed in the Orb2 knockout males were not just due to the failure of meiosis. A subsequent study found that Orb2 directly binds *apkc-RA* mRNA and affects its localisation in spermatids, where it is distributed in a comet pattern (Xu et al., 2014). The localisation of aPKC protein at the apical side of spermatid cysts is important for determining the correct orientation of the elongating spermatids (Xu et al., 2014), i.e. flagellar axonemes should elongate towards the apical end of the testis and spermatid nuclei should bundle towards the basal end (Fig 1.1). It is likely that Orb2 regulates other RNAs as well, as a defect in *apkc-RA* mRNA localisation alone does not explain other defects observed in Orb2 mutants, such as over-elongation.

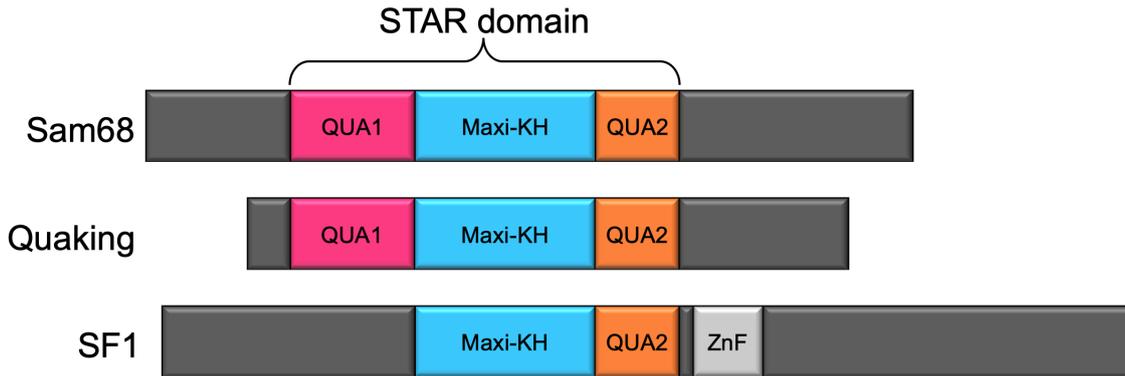
## 1.3. The STAR protein family

### 1.3.1. The three STAR subfamilies

In the last section, I presented that across the animal kingdom RBPs have roles in all stages of spermatogenesis. In *D. melanogaster* one of the RBPs that is essential in the earlier stages of spermatogenesis is held out wings (HOW; Monk et al., 2010). HOW belongs to the signal transduction and activation of RNA (STAR) protein family, and several of the STAR proteins are important in gametogenesis (Paronetto et al., 2006; Monk et al., 2010; Priti and Subramaniam, 2015). The STAR proteins contain a STAR domain (sometimes referred to as the STAR/GSG domain), and these RBPs have functions relating to many different RNA processes, including alternative splicing, nuclear export of RNAs, RNA stability, and mRNA translation (Lukong and Richard, 2003; Chénard and Richard, 2008; Volk et al., 2008). Post-translational modifications by kinases or acetyltransferases can alter their affinity for RNA, making their RNA functions responsive to various signalling pathways, hence the protein family name (Lukong and Richard, 2003; Babic et al., 2004; Israeli et al., 2007).

There are three paralogous STAR subfamilies, each named after the best known mammalian proteins from the subfamily. The oldest is the splicing factor 1 (SF1) subfamily, which can be found across eukaryotes. The SF1 proteins recognise the branch point sequence in introns and is crucial for the splicing of a subset of mRNAs in both *Saccharomyces cerevisiae* and humans (Rutz and Séraphin, 2000; Tanackovic and Krämer, 2005). The other two subfamilies, Quaking and Sam68, are found in animals (Beadell and Haag, 2014; Vernet and Artzt, 1997). Proteins in these younger two subfamilies can also regulate splicing, but some of their functions have diversified to the regulation of mRNA stability and translation (Lukong and Richard, 2003; Volk et al., 2008). The STAR proteins have the somewhat unusual feature of containing just a single RNA-binding domain (RBD). This STAR domain is made up of three main regions: QUA1, a K homology (KH) domain, and QUA2, though the SF1 proteins do not have the QUA1 region (Fig 1.6). How these regions each impact the RNA-binding properties of the STAR proteins will be explored in this section along with their functions in spermatogenesis.

## 1. Introduction



**Figure 1.6. The domain architecture of the three STAR subfamilies.** The domain architecture of human STAR proteins each representing one of the three STAR subfamilies. Sam68 and Quaking both have the full STAR domain with QUA1 (pink), maxi-KH (blue) and QUA2 (orange) regions. SF1 does not have a QUA1 region but does have a zinc finger (ZnF; light grey) domain.

---

### 1.3.2. Structure of the STAR domain

The structures of RBDs, with and without RNA, can give insights into the mechanisms and functions of the RBPs they originate from. High resolution crystal structures of the STAR domain have been solved for GLD-1 (*C. elegans*), Quaking (*H. sapiens*), and one of the Sam68 subfamily proteins — testis-STAR (T-STAR; *H. sapiens*; Teplova et al., 2013; Feracci et al., 2016). Additionally, the NMR solution structure of KH-QUA2 domain of SF1 (*H. sapiens*), which does not have the QUA1 region, has been solved (Liu et al., 2001). These structures have shown that not only are the subfamilies distinct from one another in their primary amino acid sequences but in their structures too.

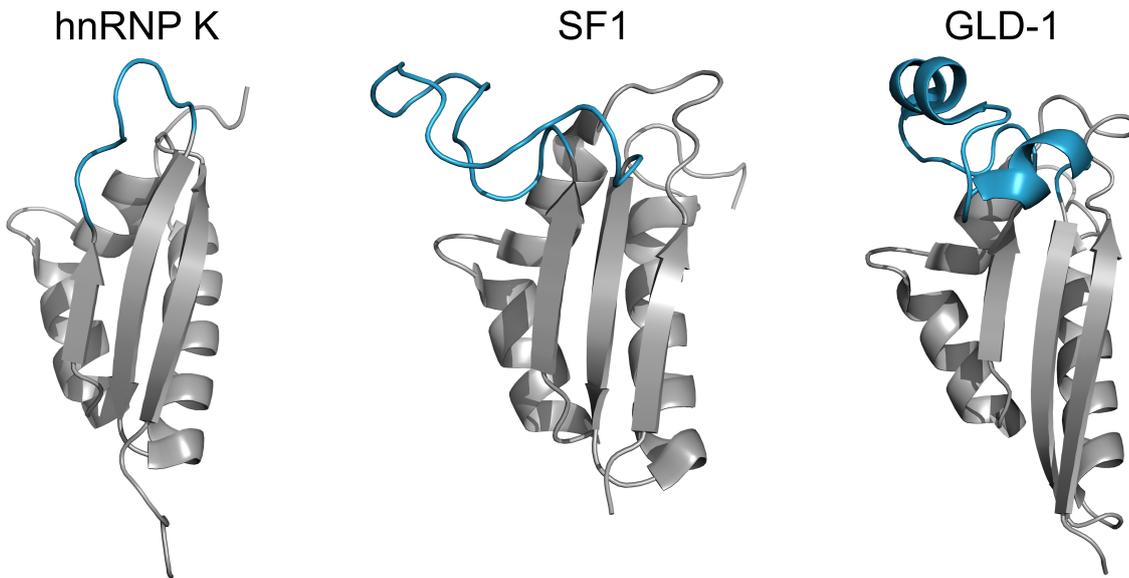
#### 1.3.2.1. The maxi-KH domain of STAR proteins

At the core of the STAR domain is the KH domain, first identified from the protein hnRNP K (Siomi et al., 1993). The KH domain is one of the most common RBDs, alongside the RRM and zinc finger (ZnF) domains (Corley et al., 2020). The typical binding surface of a KH domain can recognise just 4 nucleotides, these are often adenosines, which KH domains bind more frequently than the other 7 most common RBDs (Valverde et al., 2008; Corley et al., 2020). This is reflected in the consensus binding sequences of STAR proteins (see section 1.3.3, Table 1.1). KH domains do not form many, or any, stacking interactions with RNAs, this combined with their small binding surface means that a KH domain tends to only have micromolar affinity for RNA (Valverde et al., 2008; Corley et al., 2020). Subsequently, RBPs frequently have tandem domain repeats to increase their specificity for their cognate RNAs (reviewed for KH domains in Nicastro et al., 2015). An extreme example of KH domain repeats is vigilin, which has at least 14 KH repeats and in some organisms 15 (Cheng and Jansen, 2017). This use of multiple repeated domains to

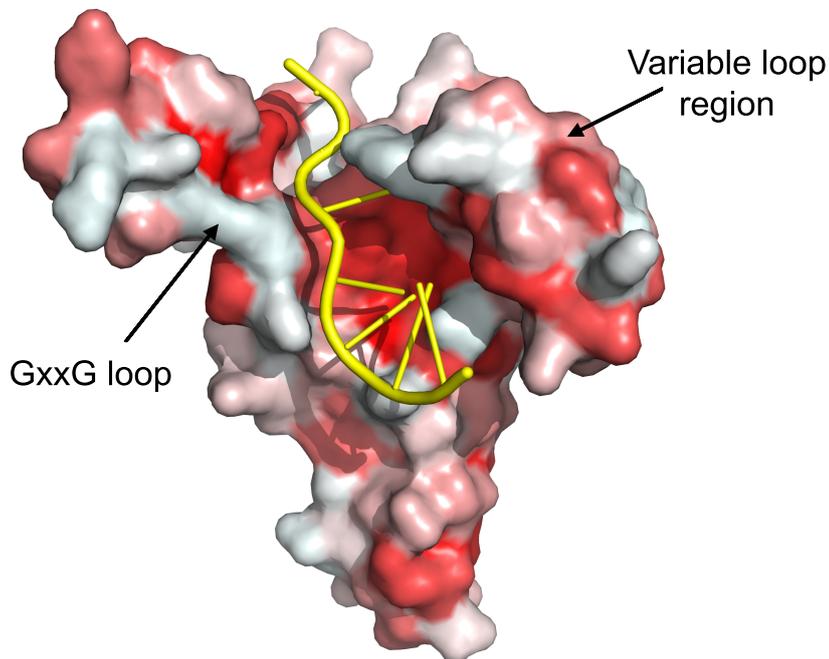
increase specificity and affinity is also seen in RRM and ZnF containing proteins (Lunde et al., 2007).

By contrast, the STAR proteins are unusual in having a single KH domain. Yet, several of the protein members have low nanomolar affinity for their cognate RNAs and some are able to recognise 6 or 7 nucleotides (Garrey et al., 2006; Carmel et al., 2010; Feracci et al., 2016). One of the ways that STAR proteins increase their binding surface is by having an expanded KH domain, also known as a maxi-KH domain. Most KH domains are around 70 amino acids (Corley et al., 2020), while the maxi-KH domain is around 100 amino acids. This expanded domain follows the eukaryotic type I topology of KH domains:  $\beta_1\alpha_1\alpha_2\beta_2\beta_3\alpha_3$ , but includes expansions in the loops between these features (Grishin, 2001; Cheng and Jansen, 2017). The variable loop region between  $\beta_2$  and  $\beta_3$  is expanded across all STAR proteins. GLD-1 and T-STAR both have two additional short  $\alpha$  helices within this variable loop region (Fig 1.7; Teplova et al., 2013; Feracci et al., 2016).

The structures of STAR proteins bound to RNA show that, similar to most KH domains, RNA binds to the maxi-KH domain in the hydrophobic cleft that is created, in part, by the GxxG loop and the variable loop region (Fig 1.8; Liu et al., 2001; Teplova et al., 2013; Feracci et al., 2016). The GxxG loop is one of the hallmarks of KH-RNA interactions (Valverde et al., 2008), and without it RNA binding is either impaired or abolished (Nakel et al., 2010; Silva et al., 2011). The GxxG loop, often GP(R/Q)G in STAR proteins, is situated between the first two  $\alpha$ -helices. RNA interactions with the maxi-KH domain are also stabilised by amino acids in  $\alpha_1$ ,  $\alpha_2$  and  $\beta_2$  (Liu et al., 2001; Teplova et al., 2013; Feracci et al., 2016).



**Figure 1.7. The variable loop in the maxi-KH domain of STAR proteins is extended.** Three KH domains with the variable loop regions between  $\beta_2$  and  $\beta_3$  highlighted in blue. Left: hnRNPK's third KH domain, the variable loop is 8 residues long (PDB: 1zzi; Backe et al., 2005). Middle: SF1's maxi-KH domain, the variable loop is 23 residues long (PDB: 1k1g; Liu et al., 2001). Right: GLD-1's maxi-KH domain, the variable loop region is 27 residues long and forms two short  $\alpha$ -helices (PDB: 4jvy; Teplova et al., 2013).



**Figure 1.8. RNA binds in the hydrophobic cleft of STAR domains.** Molecular surface representation of GLD-1's KH and QUA2 regions with hydrophobic residues coloured in red (based on hydrophobicity scale from Eisenberg et al., 1984). RNA (CUAA-CAA; yellow) binds to GLD-1 in the hydrophobic cleft partly formed by the GxxG loop and variable loop region (PDB: 4jvy; Teplova et al., 2013).

### 1.3.2.2. The QUA1 and QUA2 regions of the STAR domain

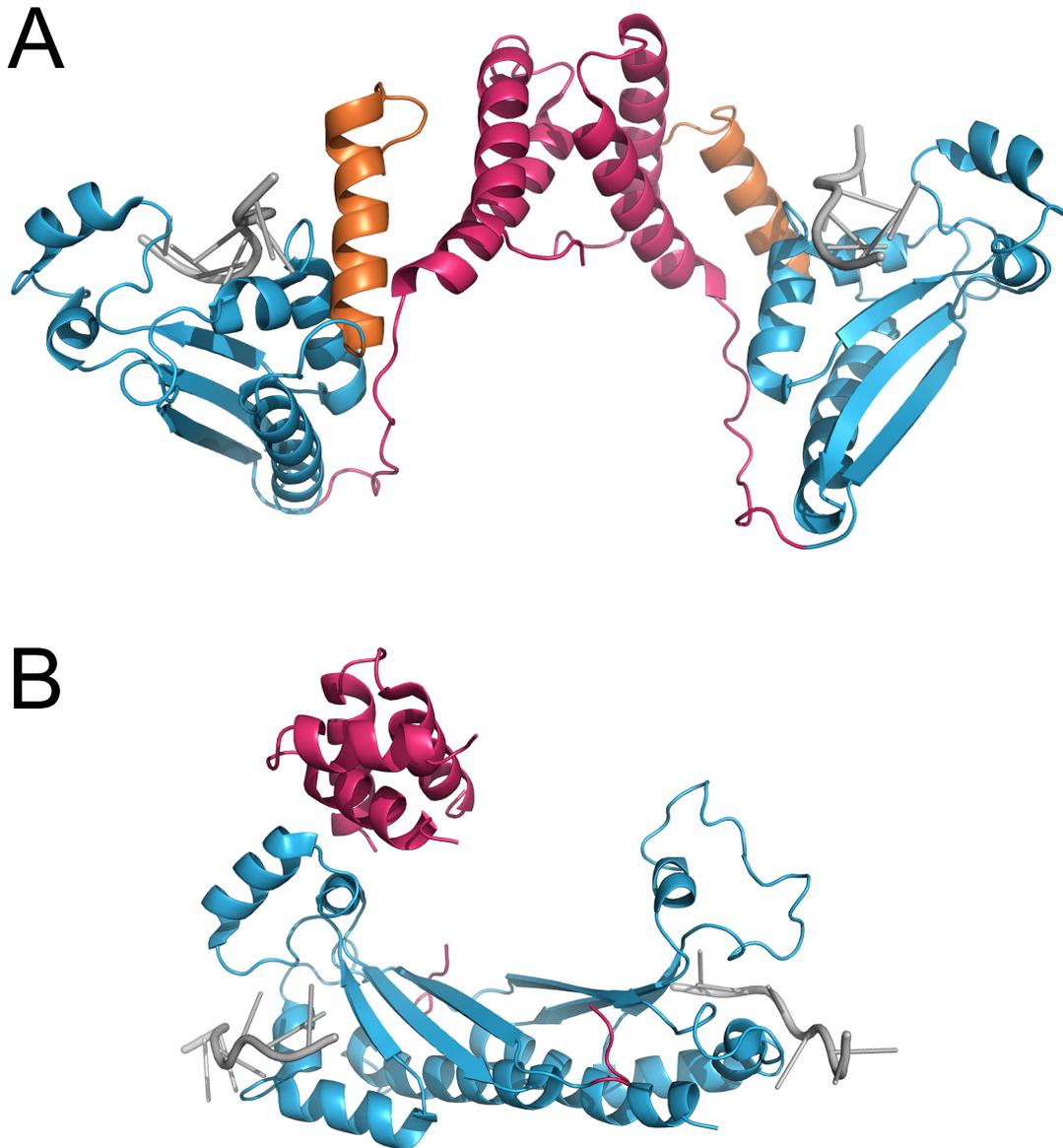
Additional increases in RNA binding specificity by the STAR proteins come from the two flanking regions of the KH domain: QUA1 and QUA2. Broadly, QUA1 is important for dimerisation and QUA2 for extending the RNA-binding surface. However, how each STAR protein utilises these varies. For example, the SF1 proteins do not have the QUA1 region (Liu et al., 2001; Nicastro et al., 2015), and while the Sam68 proteins have a QUA2 region, it does not interact with RNA (Feracci et al., 2016). STAR proteins with the QUA1 region, i.e. those in the Quaking and Sam68 subfamilies, form symmetric dimers. Solution and crystal structures of this region alone have shown that dimerisation is independent of RNA-binding (Beuck et al., 2010; Beuck et al., 2012; Meyer et al., 2010). However, the structures of the full STAR domains unexpectedly revealed that Sam68 and T-STAR dimerise differently to the GLD-1 and Quaking proteins (Fig 1.9; Teplova et al., 2013; Feracci et al., 2016). Sam68 and T-STAR form compact dimers where two KH domains also interact (Feracci et al., 2016), while GLD-1 and Quaking structures are extended and only QUA1 regions have intermolecular protein-protein interactions (Teplova et al., 2013). This could contribute to reducing the formation of heterodimers between different STAR subfamily proteins.

Though the QUA1 region does not interact with RNA, it has been shown to improve the affinity STAR proteins have for RNA. For example, mutations in the QUA1 region of GLD-1 that disrupt the homodimerisation interface can decrease GLD-1's affinity by up to two orders of magnitude (Beuck et al., 2010). Deletion of the entire QUA1 region in Sam68, as well as mutating specific residues in the region, abrogated Sam68's ability to stimulate inclusion of exon 5 in CD44 transcript (Meyer et al., 2010). This relationship between dimerisation and RNA binding can be utilised to regulate the activity of the STAR proteins, as there is evidence that phosphorylation alters the stability of STAR dimers. For example, one of the many post-translational modification sites in Sam68 is tyrosine 103 in the QUA1 region, which can be phosphorylated (Meyer et al., 2010). In NMR experiments mutating this residue has only a small negative effect on dimerisation stability but *in vitro* splicing assays show that mutating this tyrosine has severe impacts on Sam68's alternative splicing capabilities (Meyer et al., 2010). This demonstrates how the STAR proteins can be regulated in signalling pathways to directly affect RNA metabolism.

The other C-terminal flanking region of the STAR domain is QUA2, and unlike QUA1, it directly interacts with RNA and extends the RNA-binding surface of the STAR domain. The QUA2 region is situated close to the crucial GxxG loop of the KH domain and is thus able to continue the RNA binding surface (Liu et al., 2001; Teplova et al., 2013). The SF1 proteins, which lack QUA1, are still able to recognise six or seven nucleotides as a result of the QUA2 region (Liu et al., 2001; Garrey et al., 2006). In GLD-1, protein-protein interactions between QUA1 and QUA2 region help to orient QUA2 relative to the KH domain, aiding these QUA2-RNA interactions (Teplova et al., 2013). However, NMR data and direct binding assays show that the QUA2 region of Sam68 and T-STAR do not interact with RNA or affect the proteins' affinity for RNA (Feracci et al., 2016).

## 1. Introduction

Consequently, these proteins recognise much shorter trinucleotide sequences (Feracci et al., 2016).



**Figure 1.9. GLD-1 and T-STAR dimerise via different mechanisms.** A) Structure of GLD-1's STAR domain with CUAACAA, which dimerises only via the QUA1 region (PDB: 4jvy; Teplova et al., 2013). B) T-STAR's STAR domain with AUUAAA, which dimerises via the QUA1 region and the KH domain. Note that though the QUA1 region appears asymmetric, NMR experiments indicate that it is symmetric in solution (PDB: 5emo; Feracci et al., 2016). The QUA1 region is coloured in pink, KH domain in blue, QUA2 in orange, and RNA molecules in grey.

### 1.3.3. RNA sequence features of the STAR-RNA interactions

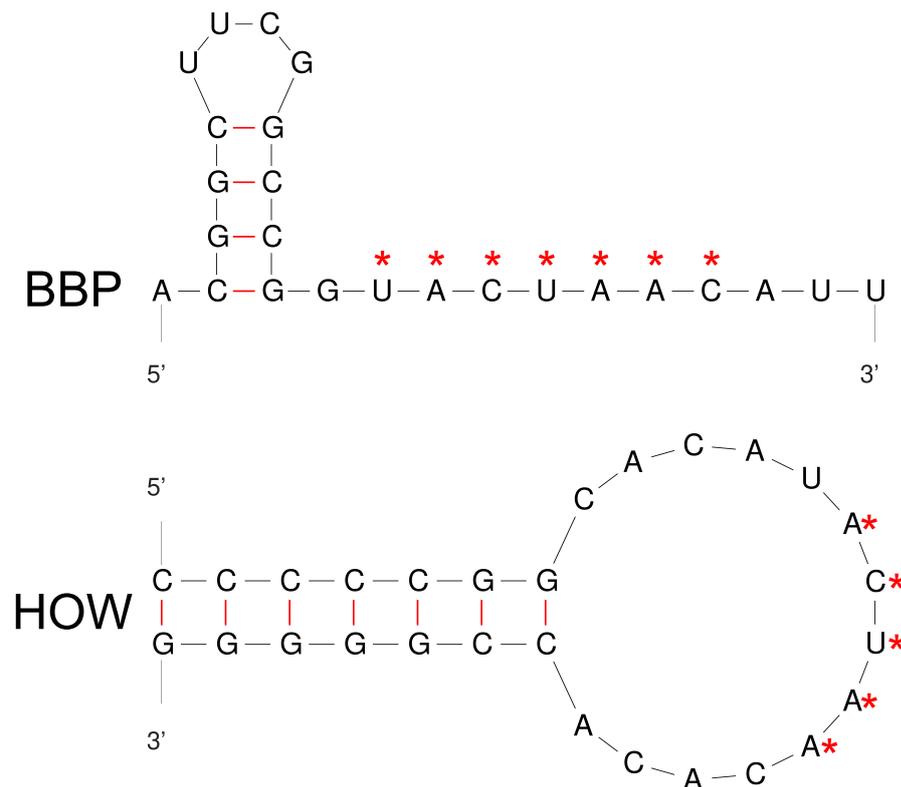
The consensus binding sequences have been defined for proteins from each of the STAR subfamilies, through direct binding experiments, such as electrophoretic mobility shift assays (EMSAs) and fluorescence anisotropy (FA), as well as *in vitro* SELEX and *in vivo* CLIP-seq (summarised in Table 1.1). Several of these studies have also investigated how protein dimerisation and RNA secondary structures impact binding.

SF1 is the ancestral STAR protein, and, accordingly, the other two subfamilies have consensus sequences derived from the SF1 sequence. In *S. cerevisiae* the branch point sequence is strictly conserved and thus the SF1 orthologue, branch point binding protein (BBP), has an optimal binding sequence to match: UACUAAC (Garrey et al., 2006). The mammalian branch point sequence is less conserved, so while the optimal binding sequence for SF1 in humans is similar (ACUNAC), SF1 displays greater flexibility than BBP and only the uridine and second adenosine are critical for binding (Table 1.1; Berglund et al., 1997; Corioni et al., 2011). Studies investigating SF1–RNA interactions also revealed that RNA secondary structure can affect the RNA-binding activity of the STAR proteins. For example, BBP has a greater affinity for RNA when the UACUAAC sequence is downstream of a stem-loop structure (Fig 1.10, top), though this activity is dependent on BBP’s two ZnF domains not its STAR domain (Garrey et al., 2006). However, a stem-loop structure adjacent to the consensus sequence makes no difference to the human SF1 protein’s affinity for RNA, which only has one ZnF domain (Fig 1.6; Garrey et al., 2006).

## 1. Introduction

**Table 1.1. Consensus sequences for STAR proteins.** The current consensus binding sequences for eight different STAR proteins, divided into the three subfamilies. The majority of these experiments used just the STAR domains from each protein. The consensus sequences for Sam68 and T-STAR were aligned based on structural data from T-STAR and GLD-1, which show the KH domain interacting with the 3'-end of the RNAs in both structures (Teplova et al., 2013; Feracci et al., 2016).

Protein (Organism)	Consensus sequence (5' → 3')						Reference
BBP <i>S. cerevisiae</i>	U	A	C	U	A	A C	Garrey et al. (2006)
SF1 <i>H. sapiens</i>		A	C	U	N	A C	Corioni et al. (2011)
GLD-1 <i>C. elegans</i>	(U>G>C/A)	A	(C>A)	U	(C/A>U)	A	Ryder et al. (2004)
STAR-2 <i>C. elegans</i>	U	A	(A>C)	U	(A>>C)	A	Carmel et al. (2010)
Quaking <i>M. musculus</i>	N	A	(A>C)	U	(A>>C)	A	Carmel et al. (2010)
HOW <i>D. melanogaster</i>		A	(C>A)	U	A	A	Israeli et al. (2007)
Sam68 <i>H. sapiens</i>				(A/U)	A	A	Feracci et al. (2016)
T-STAR <i>H. sapiens</i>				(A/U)	A	A	Feracci et al. (2016)



**Figure 1.10. Structural RNA features can increase STAR proteins' affinity for RNA.** Top: BBP has a higher affinity for RNA when its consensus binding sequence (marked with asterisks) is adjacent to a stem-loop structure, as opposed to an entirely linear structure (Garrey et al., 2006). Bottom: HOW has a higher affinity for RNA when its consensus binding sequence (marked with asterisks) is in a loop of 12 nucleotides or more, as opposed to an entirely linear structure (Israeli et al., 2007).

The Quaking subfamily bind very similar sequences to the SF1 subfamily. A combination of FA direct binding assays and competition assays were carried out to identify the optimal binding sequences of the mouse Quaking protein and two of the Quaking proteins in *C. elegans*, GLD-1 and STAR-2 (Ryder et al., 2004; Carmel et al., 2010). STAR-2 and Quaking have almost identical consensus sequences with UA(A>C)U(A>>C)A and NA(A>C)U(A>>C)A, respectively (Carmel et al., 2010). Additionally, all three proteins share strong preferences for adenosines in the second and sixth positions and uridine in the third position (Table 1.1; Ryder et al., 2004; Carmel et al., 2010). The U<sub>3</sub> and A<sub>6</sub> are equivalent to the uridine and adenosine that are critical for SF1 and BBP binding.

An optimal consensus binding sequence for *D. melanogaster* orthologue to Quaking, held out wings (HOW), has not been determined. However, several binding sites have been characterised on individual mRNAs. This was first done with *stripe* mRNA, a previously identified target of HOW (Nabel-Rosen et al., 1999). In these experiments biotinylated RNA was coupled to streptavidin beads, then different concentrations (200, 20 or 2 nM) of the long isoform of HOW, HOW(L), was incubated with the beads, and binding was then verified by western blotting (Israeli et al., 2007). This identified a pentamer sequence,

## 1. Introduction

ACUAA, in the 3'-UTR of *stripe* that HOW specifically bound (Israeli et al., 2007), which is also in the 3'-UTR of the HOW targets *dpp* and *miple1* (Israeli et al., 2007; Toledano-Katchalski et al., 2007). While an optimal binding sequence was not characterised it was demonstrated that, similar to GLD-1, HOW has a preference for a cytidine in the second position of the pentamer (Table 1.1; Israeli et al., 2007). Additionally, HOW had a higher affinity for RNA that had the pentamer within a hairpin loop, where the loop is at least 12 nucleotides long, compared to an unstructured oligo (Fig 1.10, bottom; Israeli et al., 2007).

It has also been proposed that the Quaking subfamily bind bipartite sequences, composed of a core binding site and a neighbouring UAAY half-site (Galarneau and Richard, 2005; Galarneau and Richard, 2009). This was first identified for Quaking where the Quaking response element was defined as NACUAAY-N<sub>1-20</sub>-UAAY, using a combination of SELEX and EMSAs (Galarneau and Richard, 2005). These experiments showed that if the half-site was mutated to GAGC Quaking no longer bound to the RNA (Galarneau and Richard, 2005), and similar results were reported for GLD-1 (Galarneau and Richard, 2009). However, a later study found that these mutations to the half-site may have introduced secondary structures to the RNAs making the core site inaccessible to the proteins, and that this was a more likely cause for the decrease in affinity rather than the loss of the half-site (Carmel et al., 2010). Additional FA experiments showed that one core site is sufficient for nanomolar binding but that the presence of two core sites increased the affinity by at least one order of magnitude for both Quaking and GLD-1 (Carmel et al., 2010). Finally, FA experiments with GLD-1 tested different spacer lengths (0, 2, 4, 6 and 12 nucleotides) between two core sites, and GLD-1 had the highest affinity for the oligos with 6 or 12 nucleotide spacers (Carmel et al., 2010).

In contrast to the five to seven nucleotide long sequences that the SF1 and Quaking subfamilies bind, the Sam68 subfamily have much shorter consensus sequences. SELEX experiments defined the Sam68 binding site as UAAA and the T-STAR binding site as two U(U/A)AA repeats 3–25 nucleotides apart (Lin et al., 1997; Galarneau and Richard, 2009). Follow up EMSA experiments showed that Sam68 could bind to the same aptamers identified in the SELEX experiments as T-STAR, and binding for both T-STAR and Sam68 required both U(U/A)AA repeats to be present (Galarneau and Richard, 2009). However, when the structure of T-STAR was solved it showed that it only bound three nucleotides directly (Feracci et al., 2016). Subsequent *in vivo* T-STAR CLIP-seq using mouse testis and FA assays with T-STAR and Sam68 redefined their optimal binding sites as two (A/U)AA repeats 15 or more nucleotides apart (Feracci et al., 2016). Unlike the EMSA experiments that indicate two U(U/A)AA repeats were needed for binding (Galarneau and Richard, 2009), the FA experiments showed that one (A/U)AA is sufficient, though two enhanced the affinity for both Sam68 and T-STAR (Feracci et al., 2016). Additionally, this improved affinity for two repeats versus one was abrogated when mutations were introduced that disrupted the dimerisation interfaces of both T-STAR and Sam68 (Feracci et al., 2016). It is somewhat surprising that the spacer needed for T-STAR and Sam68 is

more than double the one required for GLD-1 when GLD-1's dimer is much less compact than T-STAR's (Fig 1.9), it would be interesting to have a structure or model of these STAR dimers bound to a single RNA molecule.

To summarise, the SF1 and Quaking proteins bind broadly similar five to seven nucleotide long sequences, with subtle differences in their preferences in certain nucleotide positions, while the Sam68 proteins bind much shorter trimers (Table 1.1). Other factors beside the nucleotide sequence also contribute to the RNA-binding activity of STAR proteins, including the RNA secondary structure, the dimerisation state of the proteins, and post-translational modifications.

### 1.3.4. STAR proteins in spermatogenesis

#### 1.3.4.1. Sam68 subfamily

Three *KHDRBS* genes make up the Sam68 subfamily in vertebrates. *KHDRBS1* encodes Sam68 which is ubiquitously expressed, including the testis (Stoss et al., 2004). *KHDRBS2* is known as Sam68-like mammalian protein 1 (SLM-1), and is predominately expressed in the brain (Di Fruscio et al., 1999; Stoss et al., 2004). *KHDRBS3* encodes T-STAR (also known as SLM-2), and is highly expressed in the testis and brain (Venables et al., 1999; Stoss et al., 2004).

Mouse Sam68 is the most well-studied of the Sam68 subfamily proteins, including for its role in spermatogenesis. Male *Sam68* null mice are infertile and they have smaller testes than their littermate controls (Paronetto et al., 2009). Histological images revealed that there were few round spermatids or elongated spermatids in the *Sam68*<sup>-/-</sup> testes. Spermatozoa were produced by some of the *Sam68*<sup>-/-</sup> mice but these cells were immotile, abnormally shaped and were infertile in *in vitro* fertilisation conditions (Paronetto et al., 2009).

In WT mice, Sam68 is expressed in the nuclei of spermatogonia, primary spermatocytes and round spermatids (Paronetto et al., 2006). In the nucleus of primary spermatocytes, Sam68 has been shown to interact with RNA polymerase II (RNAPII) and the splicing factors SRSF1 and hnRNP A2 (Paronetto et al., 2011). The exclusion of exon 8 in *Sgce* mRNA is one of the well-characterised Sam68 splicing events, in the testis and other tissues. Instead, in the *Sam68*<sup>-/-</sup> germ cells exon 8 was almost always included (Paronetto et al., 2011). This showed that Sam68 can regulate alternative splicing in mouse male germ cells.

In the cytoplasm of germ cells, Sam68 has been shown to be a translational regulator. Sam68, upon phosphorylation by ERK1/2 and the maturation promoting factor complex, is cytoplasmically localised in secondary spermatocytes (Paronetto et al., 2006). There it is associated with the translation initiation factors eIF4E and eIF4G, as well as with polysomes, which are multiple ribosomes bound to a single RNA that they are likely to

## 1. Introduction

be actively translating (Paronetto et al., 2006; Paronetto et al., 2009). In *Sam68*<sup>-/-</sup> mice it was found that there was a specific decrease in *Spag16*, *Nedd1*, and *Spdya* mRNAs in polysome fractions, compared to *Sam68*<sup>+/-</sup> mice (Paronetto et al., 2009). There was also a corresponding decrease in the protein levels of SPAG16, NEDD1 and SPDYA (Paronetto et al., 2009). SPAG16 is required for sperm motility as it forms part of the flagellar axoneme (Zhang et al., 2006). SPDYA is cell cycle regulator and *Spdya*<sup>-/-</sup> mice are infertile and exhibit meiotic arrest in male and female mice (Tu et al., 2017). NEDD1 is important for microtubule organisation and has a role in meiosis in oocytes (Ma et al., 2010), but a role in spermatogenesis has not been established. It therefore seems likely that the phenotype resulting from Sam68 null mice is due to the requirement of Sam68 activity to upregulate the translation of its specific mRNA targets during meiosis. Sam68 may also regulate translation indirectly via microRNAs, as several are upregulated when Sam68 is knocked out in male germ cells (Messina et al., 2012).

Despite T-STAR's specific and high expression in the testis, it is not essential for spermatogenesis (Ehrmann et al., 2013). However, T-STAR's homologue in *D. melanogaster* is novel spermatogenesis regulator (Nsr), which is also highly expressed in the testis according to modENCODE data (Brown et al., 2014). Nsr is necessary for spermatid individualisation and coiling (Ding et al., 2010). In WT flies, the investment cones of the 64 spermatids in a cyst move synchronously in the cyst, but in *nsr* mutants these cones are scattered and the spermatids do not complete individualisation into mature spermatozoa (Ding et al., 2010). RNA-seq was carried out on WT and *nsr* mutant testes, and among the genes downregulated in the mutant testes were *kl-2*, *kl-3* and *kl-5* (Ding et al., 2010). Spermatids in *kl-3* and *kl-5* mutants also show defects in individualisation and coiling, mirroring the *nsr* mutants (Timakov and Zhang, 2000; Ding et al., 2010). In the *nsr* mutant testes the levels of the primary transcripts of *kl-2*, *kl-3* and *kl-5* are similar to that of WT levels, suggesting that their downregulation is occurring post-transcriptionally (Ding et al., 2010), however, direct binding to these RNAs by Nsr has not been demonstrated.

### 1.3.4.2. Quaking subfamily

Mammalian Quaking is expressed in many tissues, including the testis (Kondo et al., 1999), but it has mostly been studied for its function in glial cells due to its dysregulation in glioblastomas and people with schizophrenia (Chénard and Richard, 2008). However, the *C. elegans* orthologue, defective in germline development 1 (GLD-1), is well-studied for its many role in gametogenesis (reviewed in Lee and Schedi, 2010).

The first role found for GLD-1 in germ cells was as a tumour suppressor during oogenesis (Francis et al., 1995b). *C. elegans* hermaphrodites produce sperm cells during the larval stages before later switching to producing oocytes. In *gld-1* null worms, hermaphrodites are able to complete spermatogenesis but during oogenesis, specifically during meiosis, tumours form and no oocytes develop (Francis et al., 1995b). Two decades after these

observations were made, it was found that GLD-1 does have a role in suppressing tumours during spermatogenesis, but here it functions redundantly with PUF-8 (Priti and Subramaniam, 2015). Double *puf-8;gld-1* mutants result in germline tumours in male and hermaphrodite worms, while single mutants of either of these genes does not result in germline tumours in males (Priti and Subramaniam, 2015). GLD-1 also functions redundantly with GLD-2 and GLD-3, neither of which are STAR proteins, to regulate the mitosis to meiosis switch in males and hermaphrodites (Francis et al., 1995a; Kadyk and Kimble, 1998). In both *gld-1;gld-2* and *gld-1;gld-3* null mutants germline tumours form before the entry to meiosis (Kadyk and Kimble, 1998).

GLD-1 primarily acts as a translational repressor, and has yet to be shown to positively regulate any mRNAs to which it binds (Wright et al., 2011). Unlike Sam68, GLD-1 and its targets are associated with subpolysomal fractions, i.e. fractions with mRNAs bound to only one ribosome, ribosomal subunits or smaller RNP complexes. In *gld-1* mutants there is a shift of its RNA targets towards the polysome fractions (Scheckel et al., 2012). One of the ways GLD-1 promotes meiosis, is by directly binding the 3'-UTR of *glp-1*, a Notch receptor, and repressing its translation (Marin and Evans, 2003). Thus, GLP-1 levels are restricted to the distal ends of the *C. elegans* gonad arms where Notch signalling promotes mitotic proliferation and maintains the stem cell population, for example, by promoting FBF expression (Lamont et al., 2004; see section 1.2.2 for more on FBF).

#### 1.3.4.3. SF1 subfamily

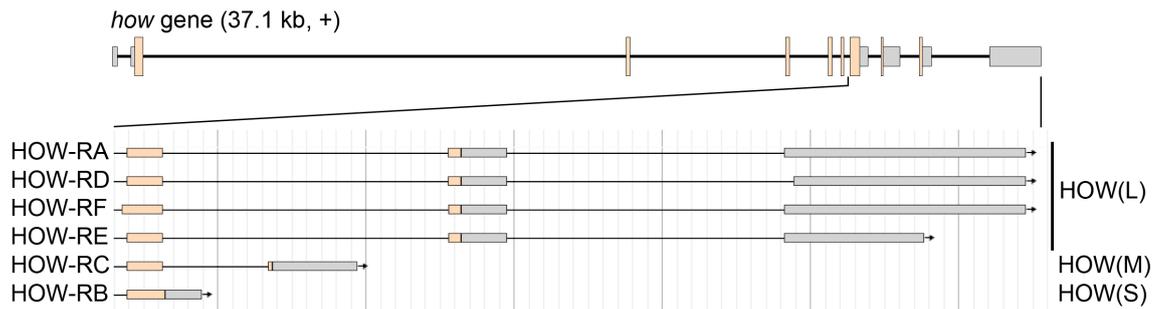
SF1 proteins are the least studied STAR proteins in relation to spermatogenesis. However, one study has found that lower levels of SF1 in mice can lead to lower incidences of testicular germ cell tumours (TGCTs; Zhu et al., 2010). *Sf1*<sup>-/-</sup> mice are embryonically lethal, but mice with one copy of *Sf1* (*Sf1*<sup>+/-</sup>) are viable, fertile and have lower SF1 protein levels (Shitashige et al., 2007; Zhu et al., 2010). When the *Sf1*<sup>+/-</sup> mice were mated to 129-*Ter* or M19 mice, which typically have very high incidences of TGCTs, the frequency of TGCTs were significantly reduced (Zhu et al., 2010). However, it is unclear what causes this reduction in TGCTs or how this affects the process of spermatogenesis in these mice.

## 1.4. Held out wings

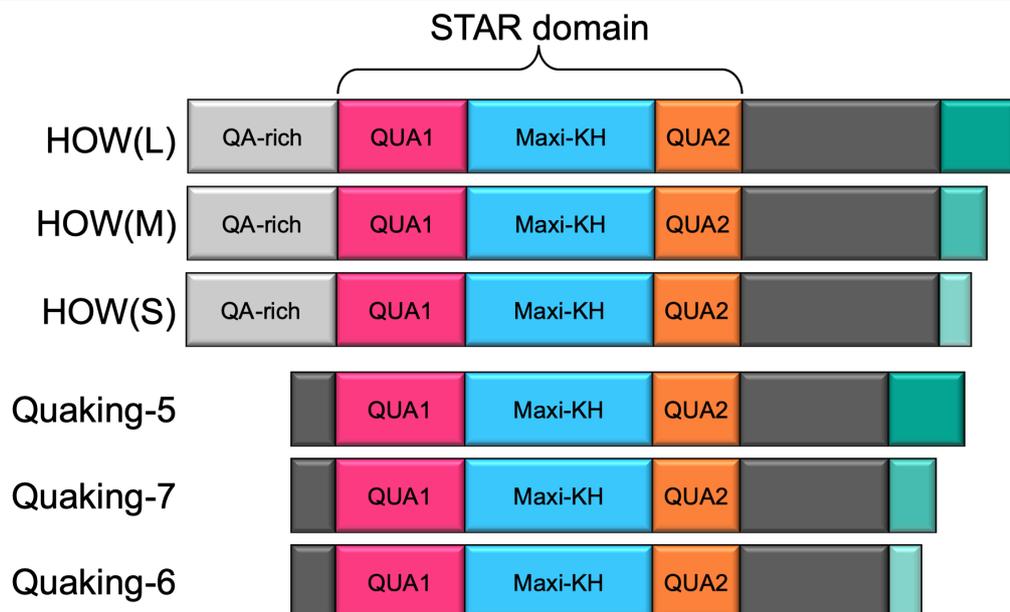
### 1.4.1. HOW is an essential protein with multiple isoforms

The *D. melanogaster* orthologue to Quaking is held out wings (HOW), which is an essential protein for flies. Total loss of function *how* mutants, generated in mutagenesis screens, are embryonically lethal (Baehrecke, 1997; Zaffran et al., 1997). This includes the *how<sup>e44</sup>* mutant, which has a missense mutation causing a change from the highly conserved arginine 185 residue to a cysteine, which abrogates the RNA binding activity of HOW (Baehrecke, 1997; Nabel-Rosen et al., 1999). Partial loss of function mutants were also generated, many of these can reach the pupal stage of development but the majority of these flies die during metamorphosis (Baehrecke, 1997; Zaffran et al., 1997). Some of these partial loss of function mutants, for example the *how<sup>r17</sup>* mutant, are able to survive to adulthood but these flies consistently have wings that do not fold properly over their back, i.e. they have held out wings, and they also have blisters on their wings due to issues with cell adhesion (Baehrecke, 1997; Lo and Frasch, 1997; Zaffran et al., 1997). As well as wing development, HOW has been shown to be important for leg development (Fortier et al., 2006), the maturation of tendon and glial cells (Nabel-Rosen et al., 1999; Edenfeld et al., 2006), mesoderm spreading (Toledano-Katchalski et al., 2007), and spermatogenesis (Monk et al., 2010).

Currently, there are six annotated RNA isoforms for HOW on FlyBase (Larkin et al., 2021), which vary from one another at their 3' ends (Fig 1.11). These transcripts result in the translation of three protein isoforms, HOW long (L), medium (M) and short (S). These three protein isoforms mirror mammalian Quaking which also has three distinct isoforms (Fig 1.12). Four of the six RNA isoforms translate to nearly identical protein sequences, which represent the long isoform of HOW(L). HOW(L), unlike the other protein isoforms, has a nuclear localisation signal in its C-terminus (HPYQR), which is conserved in the long isoform of Quaking (Quaking-5; Volk et al., 2008). HOW-RC is annotated as translating to the medium length isoform of HOW, HOW(M); however, there is no published evidence of this protein isoform yet. Finally, HOW-RB is translated to the shortest HOW protein isoform, HOW(S), which has six amino acids at its C-terminus that are unique to this isoform. Overall, HOW(L) has been studied more than HOW(S), this includes in spermatogenesis where HOW(S)'s role is poorly understood. Subsequently more RNAs that HOW(L) binds have been identified than for HOW(S) (Table 1.2).



**Figure 1.11. Schematic of the six HOW RNA transcripts.** Top: the *how* gene is 37.1 kb and on the sense strand. UTRs are in grey, coding regions in beige. Bottom: Zoom in of the 3'-end of the HOW gene showing where the six HOW transcripts vary. They are all identical upstream of the region depicted here. HOW-RA, -RD, -RF and -RE are translated into HOW(L) protein, HOW-RC to HOW(M), and HOW-RB to HOW(S). Bold vertical lines are every 1,250 nt. Adapted from JBrowse (Buels et al., 2016).



**Figure 1.12. Schematic of HOW and Quaking protein isoforms.** *D. melanogaster* HOW and mammalian Quaking both have three main protein isoforms that have unique C-terminal ends (green). HOW proteins are longer with a QA-rich region at the N-terminus.

## 1. Introduction

**Table 1.2. HOW's RNA binding targets.** All of these RNAs have been identified as HOW-bound RNA targets from pull-downs either of HOW (using antibodies) or biotinylated RNA (using streptavidin beads) followed by PCR to detect RNAs or western blotting to detect the presence of HOW protein. The RNAs are classified based on whether they have been demonstrated to bind HOW(L), HOW(S) or both. Unspecified refers to experiments that used a general HOW antibody and with lysates that contain both isoforms of HOW.

HOW isoform	RNA target	Reference
HOW(L)	<i>string</i>	Nabel-Rosen et al. (2005)
	<i>dpp</i>	Israeli et al. (2007)
	<i>miple1</i>	Toledano-Katchalski et al. (2007)
	<i>falten</i>	Toledano-Katchalski et al. (2007)
	<i>CG31638</i>	Toledano-Katchalski et al. (2007)
	<i>like-AP180</i>	Toledano-Katchalski et al. (2007)
	<i>Death-associated inhibitor of apoptosis 1</i>	Reuveny et al. (2009)
HOW(S)	<i>male-specific lethal 2</i>	Graindorge et al. (2013)
	<i>dgrasp</i>	Giuliani et al. (2014)
Both	<i>stripe</i>	Nabel-Rosen et al. (1999)
	<i>neurexin IV</i>	Edenfeld et al. (2006) Rodrigues et al. (2012)
Unspecified	<i>bag of marbles</i>	Monk et al. (2010)

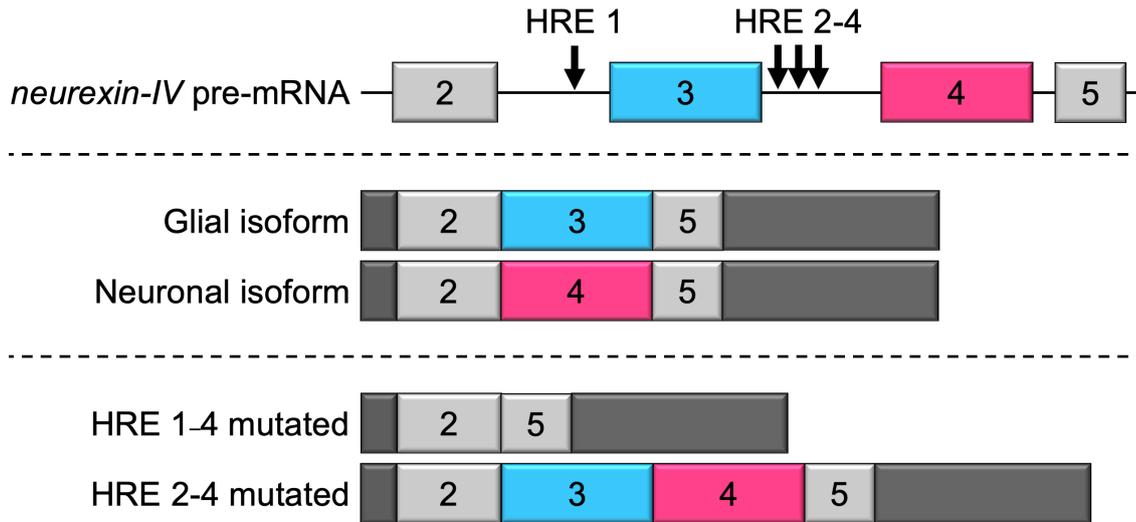
### 1.4.2. HOW(L) and HOW(S) can regulate the same RNAs

The first mRNA that was identified as bound by HOW was the *stripe* mRNA in tendon precursor cells during embryogenesis (Nabel-Rosen et al., 1999). Stripe is a transcription factor and key regulator of tendon cell differentiation. In the *how* loss of function mutant embryos tendon precursor cells undergo premature differentiation, and there are elevated levels of Stripe protein (Nabel-Rosen et al., 1999). In WT embryos HOW is first seen in the nucleus of precursor tendon cells, and in mature tendon cells HOW is located in both the nucleus and the cytoplasm (Nabel-Rosen et al., 1999). *In vitro* assays demonstrated that both HOW(L) and HOW(S) could bind the 3'-UTR of *stripe* (Nabel-Rosen et al., 1999), and GFP reporter assays revealed that they have differing effects on the mRNAs they bind — HOW(L) appears to have a destabilising effect, which is dependent on its nuclear localisation, while HOW(S) stabilises (Nabel-Rosen et al., 2002). These and additional experiments led to the model that in early tendon cell differentiation HOW(L) negatively regulates Stripe protein by retaining *stripe* mRNA in the nucleus, while the expression of HOW(S) later in embryogenesis promotes the release and stabilisation of *stripe* mRNA in the cytoplasm (Nabel-Rosen et al., 2002). This allows Stripe protein to be expressed at the appropriate time for tendon cell maturation during embryogenesis.

Glial cells are another cell type where HOW(L) and HOW(S) have both been shown to function (Edenfeld et al., 2006; Rodrigues et al., 2012). However, unlike in tendon cells where they work in opposing ways, in glia they both regulate the alternative splicing of *neurexin IV* (*nrx-IV*). While HOW(S) does not have a nuclear localisation signal, in glial cells it forms a complex with Crooked neck in the cytoplasm which then translocates to the nucleus where it can regulate alternative splicing (Edenfeld et al., 2006). In both *how* and *nrx-IV* mutants the blood–brain barrier is leaky, though more so in *nrx-IV* flies. And overexpression of either HOW(L) or HOW(S) can partially rescue this phenotype in *how* mutants (Rodrigues et al., 2012).

There are two RNA isoforms of *nrx-IV*, the main isoform in glial cells includes exon 3, while in neurons the predominant RNA isoform includes exon 4 (Fig 1.13; Stork et al., 2009). These different isoforms result in proteins with different N-terminal discoidin domains with different adhesive properties (Stork et al., 2009), thus the alternative splicing of *nrx-IV* must be regulated for proper blood–brain barrier formation. There are four HOW binding sites surrounding exon 3, one at the 5' end and three at 3' end (Fig 1.13). When all of these sites are mutated to stop HOW binding there is an increase in the novel *nrx-IV* transcript that includes neither exons 3 or 4, while mutating the binding sites downstream of exon 3 results in *nrx-IV* transcripts that contain both exons 3 and 4 (Fig 1.13; Rodrigues et al., 2012). Thus, the model for normal *nrx-IV* splicing in glial cells is that HOW is required to bind to the 5' binding site to ensure inclusion of either exon 3 and 4, and it needs to bind to the 3' binding sites to suppress the inclusion of exon 4 (Rodrigues et al., 2012).

## 1. Introduction



**Figure 1.13. Schematic of HOW-dependent *neurexin-IV* splicing events.** Top: A segment of the *neurexin-IV* pre-mRNA that contains both exons 3 (blue) and exon 4 (pink) and four HOW response elements (HRE; arrows). Middle: In glial cells the predominate *neurexin-IV* mRNA isoform contains exon 3, while in neurons the predominate isoform contains exon 4 (mutually exclusive cassette exons). Bottom: If HRE 1–4 are mutated there is an increase in a novel *neurexin-IV* mRNA isoform with neither exon 3 or 4. If HRE 2–4 are mutated both exons 3 and 4 included.

### 1.4.3. Independent functions of HOW(L) and HOW(S)

In *how* mutant embryos mesoderm invagination is delayed and the subsequent mesoderm spreading occurs abnormally (Nabel-Rosen et al., 2005; Toledano-Katchalski et al., 2007). During this period of embryogenesis only HOW(L) is expressed, and thus functions independently from HOW(S). Similar to its role in tendon precursor cells, HOW(L) represses the expression of mRNAs to control the progression of development (Nabel-Rosen et al., 2005; Toledano-Katchalski et al., 2007). Without HOW(L) to negatively regulate *string*, another homologue of yeast Cdc25, there is an excess of cell division that occurs prematurely (Nabel-Rosen et al., 2005). Though, in most *how* mutant embryos mesoderm invagination does eventually take place but it does so asynchronously (Nabel-Rosen et al., 2005). In mesoderm spreading, HOW(L) has been shown to bind directly to several RNAs (Table 1.2). Of these overexpression of *miple1* had the most similar defects in mesoderm spreading as *how* mutants, which included uneven expression and scattered activation of MAPK. Suggesting that HOW(L) is required to downregulate *miple1* expression for correct mesoderm development (Toledano-Katchalski et al., 2007).

HOW(L) also functions to retain transcripts in the nucleus. Previously this had been paired with its destabilising and repressor function, however, in the case of HOW(L)'s interaction with *male-specific lethal 2* (*misl-2*) it does not affect the stability or splicing of the RNA but promotes its nuclear localisation (Graindorge et al., 2013). *Misl-2* is required in male flies for hypertranscription of the X chromosome, but it must be repressed in females. This female-specific repression is coordinated by the RBP Sex lethal (SXL),

which binds to both the 5'-UTR and 3'-UTR of *msl-2* to affect its alternative splicing and repress its translation (reviewed in Moschall et al., 2017). HOW(L) binds to SXL and directly binds the 5'-UTR of *msl-2*, in a separate site to SXL, to retain *msl-2* in the nucleus. Most of the experiments were carried out in S2 cells but it was also demonstrated that when HOW is knocked down in the salivary glands in female flies, *msl-2* was derepressed (Graindorge et al., 2013).

The HOW(S) isoform is specifically important for the stability of *dgrasp* mRNA (also known as *grasp65*) in the follicular epithelium that covers oocytes (Giuliani et al., 2014). During stage 10 of oogenesis *dgrasp* mRNA is required at the plasma membrane, where it is also presumed to be locally translated (Schotman et al., 2008). Mutants with reduced levels of *dgrasp* mRNA and protein results in the disorganisation and loss of integrity of the follicular epithelium (Schotman et al., 2008). In *how* mutant cells there is a dramatic decrease in *dgrasp* mRNA compared to neighbouring cells with normal HOW expression. Additionally, in the *how* mutant cells *dgrasp* mRNA is not seen in punctae, as it is in WT cells (Giuliani et al., 2014). These punctae are presumed to be RNPs, and it has been suggested that HOW is required for *dgrasp* RNP formation, which in turn enhance the stability of *dgrasp*. While a precise causal link here has not been established, it was demonstrated in *in vitro* degradation assays that HOW(S) can protect *dgrasp* from RNase degradation. This protection was dependent on a binding site that, unusually for an RBP, was within the coding sequence (CDS) of *dgrasp*, rather than a UTR (Giuliani et al., 2014). Together, these studies show that the function of HOW is very dependent on the cell type it is expressed in as well as its subcellular localisation.

#### 1.4.4. Regulation of HOW expression and activity

In embryos the expression of HOW(S) is post-transcriptionally regulated, even though both *how(S)* and *how(L)* mRNA isoforms are present in the early embryo only HOW(L) protein is present (Lo and Frasch, 1997; Nabel-Rosen et al., 1999). This regulation is likely to be via its unique 3'-UTR, as when this is deleted in transgenic flies the overexpression of HOW(S) using a GAL4 driver is reportedly more successful than constructs that contain its 3'-UTR (Nabel-Rosen et al., 1999; Nabel-Rosen et al., 2002).

Similar to Sam68, post-translational modifications of HOW, namely phosphorylation, have been shown to modify the dimerisation and RNA binding activity of HOW in certain contexts. HOW has two MAPK/ERK consensus sites, threonine 59 and 64, which are phosphorylated in HOW(L) but not HOW(S), in S2 cells, embryonic somatic muscle cells and cardioblasts (Nir et al., 2012). This difference between isoforms is likely due to the nucleus, where HOW(L) is localised, containing more active MAPK/ERK than the cytoplasm, where HOW(S) is localised (Nir et al., 2012). This phosphorylation appears to stabilise the dimerisation of HOW, as there was about a 50% decrease in HOW dimers detected in cells transfected with HOW(L)<sup>TTAA</sup> mutants (where the threonine 59 and 64 phosphorylation sites are mutated to alanines) compared to cells transfected with WT

## 1. Introduction

HOW(L). Additionally, HOW(L)<sup>EG</sup> mutants, where glutamic acid 106 is mutated to a glycine disrupting the dimerisation interface of HOW, show very little phosphorylation compared to WT HOW(L) (Nir et al., 2012). This dimerisation and phosphorylation of HOW can affect its RNA binding activity, HOW(L)<sup>EG</sup> showed virtually no RNA binding activity, while the HOW(L)<sup>TTAA</sup> variant showed less binding than WT HOW(L) (Nir et al., 2012). However, this effect of dimerisation on RNA binding activity appears to be dependent on the RNA target as HOW(L)<sup>EG</sup> is able to repress *msl-2* in the same fashion as WT HOW(L), while HOW<sup>R185C</sup>, which cannot bind RNA, cannot repress *msl-2* (Graindorge et al., 2013).

### 1.4.5. HOW in germ cells

HOW is expressed in the earliest stages of gametogenesis, it is found in both the nucleus and cytoplasm, though it appears to be more concentrated in the nucleus (Monk et al., 2010; Monk et al., 2011). HOW is important for maintaining GSCs in male and female flies but it is only needed for the transit-amplifying divisions in spermatogonia and not in the cystocytes (the female germline equivalent to spermatogonia; Monk et al., 2010; Monk et al., 2011). Loss of function *how* mutants can be induced in the GSCs using the FLP-FRT recombination system. When *how* mutant clones are induced in the testes and ovaries, the mutant GSC clones are lost at a much greater rate than WT GSC clones (Monk et al., 2010; Monk et al., 2011). When *how* is knocked down in the testes, using RNAi, many of the flies are agametic, but when there are germ cells present there is significantly more cells positive for acridine orange, a marker of apoptosis (Monk et al., 2010). While in the ovaries, there is not an increase in acridine orange positive cells, and the mutant GSC clones are able to survive for a longer time after the induction than the male GSCs (Monk et al., 2011). From these differences, it has been suggested that the female GSCs lacking in HOW are lost due to premature differentiation rather than cell death, while male GSCs are not able to survive without HOW. Overexpression of either HOW(L) or HOW(S) in the testis can partially rescue the loss of GSC phenotype in *how* mutant flies, though HOW(L) can do this to a greater extent than HOW(S) (Monk et al., 2010). However, whether overexpression of either isoform of HOW can rescue the loss of GSC *how* mutant phenotype in ovaries was not tested.

Altering the expression of *how* in the testis, but not the ovaries, disrupts the transit-amplifying mitotic divisions. Spermatogonia with loss of function *how* rarely progressed beyond the 2-cell stage. While the overexpression of HOW(L), but not HOW(S), frequently resulted in at least one extra round of spermatogonial division prior to meiosis (Monk et al., 2010), mirroring the phenotype seen in *bam* mutant testes (Insko et al., 2009). In ovaries, cystocytes with mutant *how* can progress to the 16-cell stage, and an overexpression of HOW(L) does not alter the number of cell divisions but does result in an increase in ‘GSC-like’ cells, these cells cannot be confirmed as GSCs as it is difficult to distinguish between GSCs and their child cystoblast cells. This phenotype in ovaries also mirrors *bam*

mutants; loss of *bam* in the ovaries cause an excess of GSCs and GSC tumours develop (McKearin and Ohlstein, 1995).

Given the overlap in the phenotypes it was investigated and confirmed that HOW can bind *bam* mRNA, probably via a HOW binding site in its 3'-UTR. HOW most likely represses *bam* mRNA, as in both tissues HOW has complimentary expression to Bam. In the male germ cells, HOW is expressed in GSCs and up to the 2-cell spermatogonia, while Bam is expressed from the 4-cell cysts (Monk et al., 2010). In the female germline, HOW is expressed in GSCs and the immediate cystoblast child cell but not in 2-cell cysts, while Bam can be detected from the 2-cell cyst stage (Monk et al., 2011). As described in section 1.1.2, Bam is required in spermatogenesis to control the number of mitotic divisions and to promote entry into meiosis. In the female germline, Bam is required earlier for the child cells from GSC divisions to differentiate into cystoblasts, hence loss of *bam* in the ovaries results in GSC tumours (McKearin and Ohlstein, 1995). Thus, the difference in HOW affecting mitosis in spermatogenesis and not oogenesis could be due to the different roles of Bam in these two processes. However, the dysregulation of *bam* in *how* mutant testes does not explain the loss of GSCs in these testes and there is currently little understanding as to why this is.

An additional difference between the role of HOW in the testes and ovaries is HOW's interaction with Cyclin B. The 2-cell spermatogonial cysts with mutant *how* express Cyclin A but lack Cyclin B, suggesting a delay in the G2 phase. Overexpression of Cyclin B in testes where *how* is knocked down with RNAi, can partially rescue of the loss of GSC phenotype, while removing a copy of *cyclin B* in flies with *how* knocked down in the testis results in an even greater loss of GSCs (Monk et al., 2010). However, *cyclin B* mRNA is not pulled down with HOW in immunoprecipitation assays from embryo lysates, unlike *bam* mRNA, suggesting that HOW does not directly regulate Cyclin B (Monk et al., 2010). Mutant *how* cells in the female germline had no irregularities with Cyclin B expression and the cystocytes were able to progress through mitosis and reach the 16-cell cyst stage (Monk et al., 2011).

## 1.5. Project Objectives

Multiple members of the STAR protein family, including HOW, have important functions in the nucleus and cytoplasm of germ cells (Paronetto et al., 2006; Lee and Schedi, 2010; Monk et al., 2010). The STAR family can connect signalling pathways to RNA regulation and their functions are diverse (Lukong and Richard, 2003; Volk et al., 2008). Demonstrated functions of HOW include: nuclear retention of RNAs (Graindorge et al., 2013), RNA stability (Nabel-Rosen et al., 2002), and alternative splicing (Edenfeld et al., 2006). These functions are dependent on the cell type and the cellular localisation of HOW. The longer isoform of HOW is exclusively localised to the nucleus, while the shorter isoform is present in either the cytoplasm or both the nucleus and cytoplasm depending on the

## 1. Introduction

cell type (Volk et al., 2008), and in general the function of HOW in the nucleus is well characterised whereas the role of cytoplasmic HOW is less well understood.

In germ cells, HOW is present in both the nucleus and the cytoplasm (Monk et al., 2010; Monk et al., 2011), and overexpression of HOW(S) indicates that this isoform is restricted to the cytoplasm in male germ cells (Monk et al., 2010). HOW is an essential protein for *D. melanogaster* spermatogenesis, as knock down of *how* using RNAi results in mostly agametic males. In partial loss of function *how<sup>r17/r4</sup>* mutant flies, the testes are not totally agametic but do show a decrease in the number of GSCs surrounding the hub, this loss can be partially rescued by overexpression of HOW(L) or HOW(S) (Monk et al., 2010). Thus, it is reasonable to assume that HOW(S) has a function in the cytoplasm of germ cells that is yet to be elucidated. To date, only 3 RNAs have been identified as bound by HOW(S) specifically, *stripe*, *nrx-IV* and *dgrasp* (Table 1.2), and none of these targets were identified from germ cells. Additionally, no *in vitro* or *in vivo* transcriptome wide studies have been carried out on either HOW(S) or HOW(L), and HOW's affinity for RNA is not as well characterised as other STAR proteins, such as GLD-1 and Quaking.

This project aims to better understand the role of HOW(S)'s RNA-binding activity in spermatogenesis. To this end, the first objective is to identify the RNAs that HOW(S) binds in *D. melanogaster* male germ cells. To achieve this, I intend to use RIP-seq to identify the RNAs bound by HOW(S) *in vivo* in the GSCs and early spermatogonia, where HOW(S) is endogenously expressed. The second objective is to enhance our understanding of the function of HOW(S) by investigating the RNAs HOW(S) binds in spermatogenesis. The third and final objective is to better characterise and quantify HOW(S)-RNA interactions, using the transcriptome-wide data generated from the first objective.

## 2. Materials and Methods

### 2.1. Fly husband and stocks

#### 2.1.1. Fly husbandry

Flies were kept in a 25 °C humidified room with a 12:12 hour light:dark cycle and raised on 10 ml standard sugar-yeast-agar medium (50 g/L sugar, 100 g/L yeast, 10 g/L agar, 0.3% (v/v) propionic acid, 1.97 mM Nipagin M; Bass et al., 2007).

Crosses, unless stated otherwise, were generated by collecting unmated male and female flies. 5–10 flies of each sex were placed into a vial together with grains of active baker’s yeast (day 1). On day 3 and 5 adult flies were placed into fresh vials and on day 8 flies were removed from the vials.

#### 2.1.2. Fly stocks

*D. melanogaster* lines used are listed in Table 2.1. Fly stocks used were mostly from Vienna Drosophila Resource Center (VDRC) and Bloomington Drosophila Stock Center (BDSC), with others generously given by T. Volk and A. Bretman (Table 2.1).

The RNA interference (RNAi) lines from VDRC come from three different libraries: the GD, KK and shRNA libraries (Dietzl et al., 2007). The GD library is for long hairpin RNAi and is P-element based, the chromosome for each insertion is mapped but the precise location is unknown, the transformation vector used for this library is pMF3. The KK library is also for long hairpin RNAi but instead is  $\Phi$ C31 based, insertions were targeted to the VIE-260b landing site on chromosome 2, the transformation vector used for this library is pKC26. The shRNA library is for short hairpin microRNAs and is  $\Phi$ C31 based, insertions were targeted to the attP40 landing site at 25C6 on chromosome, the transformation vector used for this library is pWALIU20.

The RNAi lines from BDSC were generated by the Transgenic RNAi Project (TRiP). These lines have shRNAs inserted using the  $\Phi$ C31 method, which targeted either the attP40 (25C6 on chromosome 2) or attP2 (68A4 on chromosome 3) landing sites. Fly stocks from different generations of TRiP used different vectors, some of these are described as being better for the ‘soma’ and/or the ‘germline’, however, the ‘germline’ in the context of TRiP refers to the female germline, and the lines that are suitable for the ‘soma’ are suited for use in the male germline. The first generation of lines used the pVALIU10 vector, which works well in the ‘soma’. The pVALIU20 vector works better in the ‘soma’ than the pVALIU10, and also works in the ‘germline’. pVALIU21 and 22 vectors are suited for ‘germline only’ (Perkins et al., 2015).

**Table 2.1. *D. melanogaster* stocks used.** Each fly line is listed with its name, the source of the fly stock, and a description of the stock which includes as much genotypic information as possible as well as any pertinent phenotypic information.

Fly line name	Source	Description
bif RNAi 1	VDRC – #109722	y, w <sup>1118</sup> ; P{KK105557}VIE-260B ; From the KK library
bif RNAi 2	VDRC – #330390	w <sup>1118</sup> ; P{VSH330390}attP40 ; From the shRNA library
CLIP-190 RNAi 1	VDRC – #330453	w <sup>1118</sup> ; P{VSH330453}attP40/CyO ; From the shRNA library
CLIP-190 RNAi 2	VDRC – #107176	y, w <sup>1118</sup> ; P{KK107824}VIE-260B ; From the KK library
CycG RNAi 1	VDRC – #106846	y, w <sup>1118</sup> ; P{KK109269}VIE-260B ; From the KK library
CycG RNAi 2	BDSC – #29315	y <sup>1</sup> v <sup>1</sup> ; ; P{y[+t7.7] v[+t1.8]=TRiP.JF02474}attP2 Generated with the pVALIUM10 vector
Df31 RNAi 1	BDSC – #43185	y <sup>1</sup> sc* v <sup>1</sup> sev <sup>21</sup> ; P{y[+t7.7] v[+t1.8]=TRiP.GL01529}attP40 ; Generated with the pVALIUM22 vector
Df31 RNAi 2	BDSC – #35718	y <sup>1</sup> sc* v <sup>1</sup> sev <sup>21</sup> ; ; P{y[+t7.7] v[+t1.8]=TRiP.GLV21083}attP2 Generated with the pVALIUM21 vector
Hipk RNAi 1	BDSC – #35363	y <sup>1</sup> sc* v <sup>1</sup> sev <sup>21</sup> ; ; P{y[+t7.7] v[+t1.8]=TRiP.GL00276}attP2 Generated with the pVALIUM22 vector
Hipk RNAi 2	BDSC – #60084	y <sup>1</sup> sc* v <sup>1</sup> sev <sup>21</sup> ; P{y[+t7.7] v[+t1.8]=TRiP.HMC05078}attP40 ; Generated with the pVALIUM20 vector
HOW RNAi	BDSC – #55665	y <sup>1</sup> sc* v <sup>1</sup> sev <sup>21</sup> ; P{y[+t7.7] v[+t1.8]=TRiP.HMC03820}attP40 ; Generated with the pVALIUM20 vector

Fly line name	Source	Description
jvl RNAi 1	VDRC – #108229	y, w <sup>1118</sup> ; P{KK104198}VIE-260B ; From the KK library
jvl RNAi 2	BDSC – #43177	y <sup>1</sup> v <sup>1</sup> ; ; P{y[+t7.7] v[+t1.8]=TRiP.GL01520}attP2 Generated with the pVALIUM22 vector
lola RNAi 1	VDRC – #101925	y, w <sup>1118</sup> ; P{KK110256}VIE-260B ; From the KK library
lola RNAi 2	BDSC – #35721	y <sup>1</sup> sc* v <sup>1</sup> sev <sup>21</sup> ; ; P{y[+t7.7] v[+t1.8]=TRiP.GLV21086}attP2 Generated with the pVALIUM21 vector
<i>nanos</i> -GAL4:VP16-4937	BDSC – #4937	w <sup>1118</sup> ; ; P{w[+mC]=GAL4::VP16-nos.UTR}CG6325[MVD1]
<i>nanos</i> -GAL4:VP16-7303	BDSC – #7303	P{w[+mC]=GAL4::VP16-nos.UTR}MVD2, w <sup>1118</sup> ; ;
<i>nanos</i> -GAL4:VP16-64277	BDSC – #64277	y <sup>1</sup> w*/Dp(1;Y)B[S]Yy[+] ; ; P{w[+mC]=GAL4::VP16-nos.UTR}1C. Males have bar eyes and normal body colour Females have WT eyes and yellow body colour
Scarlet	A. Bretman	Homozygous <i>scarlet</i> mutant
smash RNAi 1	VDRC – #100636	y, w <sup>1118</sup> ; P{KK104543}VIE-260B/CyO ; From the KK library
smash RNAi 2	BDSC – #52982	y <sup>1</sup> v <sup>1</sup> ; P{y[+t7.7] v[+t1.8]=TRiP.HMJ21669}attP40/CyO ; Generated with the pVALIUM20 vector
Syx1A RNAi 1	VDRC – #33112	w <sup>1118</sup> ; ; P{GD564}v33112 From the GD library
Syx1A RNAi 2	BDSC – #25811	y <sup>1</sup> v <sup>1</sup> ; ; P{y[+t7.7] v[+t1.8]=TRiP.JF01829}attP2 Generated with the pVALIUM10 vector
Syx6 RNAi 1	VDRC – #104795	y, w <sup>1118</sup> ; P{KK109340}VIE-260B ; From the KK library

<b>Fly line name</b>	<b>Source</b>	<b>Description</b>
Syx6 RNAi 2	BDSC – #28505	$y^1 v^1 ; ; P\{y[+t7.7] v[+t1.8]=TRiP.JF03125\}attP2/TM3, Sb^1$ Generated with the pVALIUM20 vector
TER94 RNAi 1	VDRC – #24354	$w^{1118} ; P\{GD9777\}v24354 ;$ From the GD library
TER94 RNAi 2	BDSC – #35608	$y^1 sc^* v^1 sev^{21} ; ; P\{y[+t7.7] v[+t1.8]=TRiP.GL00448\}attP2$ Generated with the pVALIUM22 vector
toc RNAi	VDRC – #24084	$w^{1118} ; ; P\{GD14282\}v24084$ From the GD library
UAS-HOW-S-HA	T. Volk	3xHA-tagged HOW(S) transgene, likely homozygous on the X chromosome
WT	A. Bretman	Dahomey wild type strain

## 2.2. Genotyping *D. melanogaster*

### 2.2.1. Genomic DNA extraction

Genomic DNA (gDNA) was extracted from 50 mg of flies using DNeasy Blood & Tissue Kit (Qiagen; #69504) and following the supplementary protocol ‘Purification of total DNA from insects using the DNeasy Blood & Tissue Kit’. DNA was eluted with a 150  $\mu$ L of the ‘Buffer AE’ provided with the kit. DNA was quantified using a NanoDrop spectrophotometer.

### 2.2.2. Polymerase chain reaction

Polymerase chain reaction (PCR) was performed with each reaction containing 3 mM  $MgCl_2$  (Thermo Fisher), 10 mM deoxynucleotides, 0.2  $\mu$ M primers, 0.625 U AmpliTaq 360 (Thermo Fisher), 1X AmpliTaq 360 Buffer (Thermo Fisher), 1  $\mu$ L of gDNA, and  $H_2O$  made up to total reaction volume of 25  $\mu$ L. Annealing temperature used for all primers was 55  $^{\circ}C$  and 40 cycles were carried out (Table 2.2). Primer pair combinations used: ‘HOW fwd’ and ‘HOW rev’, ‘UAS fwd’ and ‘HOW rev’, ‘HOW fwd’ and ‘HA rev’, ‘HA fwd’ and ‘SV40 rev’. The Rp49 primer pair was used as a positive control and as a negative control ‘HA fwd’ and ‘HOW rev’ were paired (Table 2.3).

**Table 2.2. Protocol for PCR thermocycler.**

Step	Temperature	Time	Cycles
Initial denaturation	95 $^{\circ}C$	3 mins	
Denaturation	95 $^{\circ}C$	30 secs	
Annealing	55 $^{\circ}C$	30 secs	40
Extension	72 $^{\circ}C$	1 min/kb	
Final extension	72 $^{\circ}C$	5–15 mins	

**Table 2.3. Primers for genotyping the UAS-HOW-S-HA fly stock.**

Primer name	Primer sequence (5' $\rightarrow$ 3')
Rp49 fwd	CCAGTCGGATCGATATGCTAA
Rp49 rev	TCTGCATGAGCAGGACCTC
HOW fwd	TTTGTCGGTCGCATTTTGGG
HOW rev	CTTCTTGTCGCGCATGGAAC
HA fwd	TACCCATACGATGTTCCCTGAC
HA rev	GTTCCAGATTACGCTGCT
UAS fwd	ACGGAGCGACAATTCAATTCA
SV40 rev	GGCATTCCACCACTGCTCCC

## 2. Materials and Methods

### 2.2.3. Agarose gel electrophoresis

1% (w/v) agarose gels were made with 1X TAE buffer (40 mM Tris base, 20 mM acetic acid, 1 mM EDTA) and SYBR Safe (0.01% v/v; Thermo Fisher). Prior to loading, DNA samples were mixed with 6X DNA loading dye (0.25% (w/v) bromophenol blue, 0.25% (w/v) xylene cyanol). 5–10  $\mu\text{L}$  of GeneRuler 1 kb Plus DNA Ladder (Thermo Fisher) was used for each gel. Gels were run at  $\sim 100$  V until bands were sufficiently separated. Gels were visualised with UV light using a ChemiDoc (Bio-Rad Laboratories).

## 2.3. Quantification of transcript expression

### 2.3.1. RNA extraction from *D. melanogaster* testes

0-3 day old male flies were anaesthetised on ice and testes were dissected in PBS with 1 U/ $\mu\text{L}$  RNasin Plus RNase inhibitor (Promega) and 1 mM dithiothreitol (DTT). 15–30 pairs of testes were dissected per sample and were stored at  $-80$  °C.

RNA was extracted using either the Quick-RNA Minirep or Microprep kits (Zymo Research). Tissue was homogenised with a micropestle in lysis buffer and the kit instructions were followed for RNA extraction with on-column DNase I treatment. RNA was eluted with 17–50  $\mu\text{L}$  with nuclease-free water. The 260/280 and 260/230 ratios were evaluated with a NanoDrop spectrophotometer (Thermo Scientific). If 260/280 was below 1.90 and/or 260/230 was below 1.40, then RNA samples were processed again using the RNA Clean and Concentrator kit (Zymo Research). After this samples were measured with a Qubit Fluorometer (Invitrogen) to determine the concentration of the RNA samples.

Some samples were treated with TURBO DNase (Invitrogen) if gDNA was detected in the RNA samples from the no reverse transcription control reactions used in quantitative reverse transcription PCR (qRT-PCR). TURBO DNase treatment was used following manufacturer's instructions: 1  $\mu\text{L}$  TURBO DNase for up to 10  $\mu\text{g}$  RNA in a 50  $\mu\text{L}$  reaction with 1X TURBO DNase buffer, and incubated at 37 °C for 30 minutes. RNA was then isolated using the RNA Clean and Concentrator kit, and evaluated with a NanoDrop and Qubit Fluorometer as before.

### 2.3.2. cDNA synthesis

RNA samples were denatured by placing them at 70 °C for 5 minutes before being placed on ice. Equal amounts of RNA (up to a maximum of 1  $\mu\text{g}$ ) were used with the qScript cDNA synthesis kit (Quantabio). Following the kit instructions, 1  $\mu\text{L}$  of qScript and 4  $\mu\text{L}$  of the 5X reaction mix were added to each reaction. Nuclease-free water was added to make a final reaction volume of 20  $\mu\text{L}$ . cDNA synthesis reactions were incubated at 22 °C for 5 minutes, followed by 42 °C for 30 min and then 85 °C for 5 min.

### 2.3.3. qRT-PCR

qRT-PCR was set up so that each well contained: 10  $\mu$ L of PowerUp SYBER Green Master Mix (Applied Biosystems), 0.3  $\mu$ M forward primer, 0.3  $\mu$ M reverse primer, 12.5 ng of cDNA, and nuclease-free water to make a final reaction volume of 20  $\mu$ L. All reactions were performed in triplicate in 96-well plates. No template controls were run for each primer pair and no reverse transcription reactions were run with each RNA sample used to make the cDNA. The thermocycler used to perform qRT-PCR was the CFX Connect Real-Time PCR Detection System (Bio-Rad Laboratories) in fast cycling mode, as detailed in Table 2.4.

All primers for qRT-PCR (Table 2.5) were designed to anneal at 60  $^{\circ}$ C. Standard curves were run on each primer set using a cDNA pool from all samples, except for Hipk primer standard curve where a cDNA pool was made from only the parental fly lines and not the Hipk knockdown progeny. Primer efficiencies were calculated from the standard curves and all were within 95–105% (Appendix I, Figs A.1 and A.2).

**Table 2.4. Fast cycle qRT-PCR thermocycler protocol.**

Step	Temperature	Time	Cycles
Uracil-DNA glycosylase activation	50 $^{\circ}$ C	2 mins	
Dual-Lock DNA polymerase	95 $^{\circ}$ C	2 mins	
Denaturation	95 $^{\circ}$ C	3 secs	40
Annealing/extension	60 $^{\circ}$ C	30 secs	
Melt curve	65–95 $^{\circ}$ C, increasing by 0.5 $^{\circ}$ C	5 secs	

**Table 2.5. Primers used in qRT-PCR.**

Primer name	Primer sequence (5' $\rightarrow$ 3')	Exon-junction spanning
Actin 5C fwd	TACTCTTTCACCACCACCGC	No
Actin 5C rev	GGCCATCTCCTGCTCAAAGT	No
Hipk fwd	CGGCTGCCTCCTCTAGCAACATT	Yes
Hipk rev	CCGTTGGCGTTGCAGTTGCTTA	Yes
FMRP fwd	CGCGGATAGATTACAGGGCCA	Yes
FMRP rev	GGCCACCTCAACGGTTTCCT	Yes
RpL22 fwd	TGGATGTGGCCGACTTCGAGA	Yes
RpL22 rev	AGCGCTCGAAGGTGACGTTGT	Yes

## 2. Materials and Methods

### 2.3.4. Relative quantification of gene expression

For relative quantification the  $\Delta\Delta C_q$  method was applied (Taylor et al., 2019). For the samples used in quantifying *Hipk* expression in the *Hipk* RNAi 1 knockdowns *RpL22* and *actin 5C* were used as reference genes. For the samples used in quantifying *Hipk* expression in the *Hipk* RNAi 2 knockdown *Rpl22* and *FMRP* were used as reference genes. In both analyses expression was relative to the *nanos*-GAL4 parental levels.

Statistical analyses were carried out in IBM SPSS Statistics 26. The data for each biological triplicate was tested for normality using the Shapiro–Wilk test. The maternal cross knockdowns for both *Hipk* 1 and 2 were not normal ( $p < 0.05$ ), so non-parametric Kruskal–Wallis  $H$  tests were performed on two sets of data. The first set included results from the *nanos*-GAL4 samples, *Hipk* RNAi line 1, *Hipk* 1 maternal cross and the *Hipk* 1 paternal cross. The second set included results from the *nanos*-GAL4 samples, *Hipk* RNAi line 2, *Hipk* 2 maternal cross and the *Hipk* 2 paternal cross. The *nanos*-GAL4 samples used in each set were independent of one another.

## 2.4. Fertility assays

### 2.4.1. Sperm competition assay

#### 2.4.1.1. Experimental design

All flies used for mating in the sperm competition assay were progeny from vials that had no more than 5 males and 5 females, to ensure all larvae had enough access to food and development was not hindered. For this assay, unmated 2–4 day old male and female flies were used. All female flies were Scarlet, and the males used were Scarlet, WT, UAS-HOW-S-HA, *nanos*-GAL4:VP16-64277, and *nanos*-GAL4>HOW(S)-HA (see Table 2.1 for descriptions of stocks).

Experiments were conducted in a 25 °C humidified room. Individual scarlet-eyed females were first mated with either a focal male or a scarlet-eyed male in 7.5 mL vials supplemented with active baker’s yeast. No more than 500 vials were set up per assay so that all vials could be monitored for mating. After mating, the male flies were removed. Any pairs that had not mated after 4 hours were marked as having not mated.

The next day scarlet-eyed males were added to the vials that previously had a focal male in and vice versa. After mating the females were moved into a new 7.5 mL vial with fresh active baker’s yeast and left to lay embryos for 24 hours. For the second mating, any pairs that not mated after 8 hours were marked as having not mated. 2 weeks later the progeny from both the first and second mating vials were counted and phenotype based on eye colour and/or shape.

For the sperm competition assay, ‘Position 1’ (P1) refers to when the focal male is mated first to the female and the scarlet-eyed male second. ‘Position 2’ (P2) is when the scarlet-eyed male is the first male and the focal male is second.

### 2.4.1.2. Analysis

From the sperm competition assay three measurements were made. 1) The number of progeny produced from the first mating. 2) The mating rates of focal males in both P1 and P2. 3) The ‘offence’ and ‘defence’ sperm competition abilities of the focal males, which is determined by the percentage of progeny they sired that eclose from the second vial in the experiment.

Statistical analyses were carried out in IBM SPSS Statistics 26. For the number of progeny produced from the first mating, the data from each genotype was tested for normality using the Shapiro–Wilk test. Only the *nanos*-GAL4 data were normal ( $p > 0.05$ ), so the non-parametric Kruskal–Wallis  $H$  test was performed. The null hypothesis was rejected ( $p < 0.05$ ) so Dunn’s test followed by a Bonferroni correction was carried out for the post-hoc pairwise comparisons.

To compare the sperm competition abilities of the different genotypes Kruskal–Wallis  $H$  tests were performed on the ‘offence’ and ‘defence’ data as none of the data were normal, as tested by the Shapiro–Wilk test ( $p < 0.05$ ). The null hypothesis was not rejected for the offence data (i.e. when the focal males were in P2) but it was for the defence results (i.e. when the focal males were in P1). Thus, post-hoc pairwise comparisons were performed between the different focal male genotypes with the P1 data. The post-hoc test was a Dunn test followed by a Bonferroni correction.

### 2.4.2. 3-day mating assay

#### 2.4.2.1. Experimental design

The 3-day mating assay was designed to assess the fertility of male flies with conditional knockdowns of different genes in the testes. For each RNAi line the cross was performed in both directions. To differentiate the progeny from each cross they are referred to as coming from a maternal or paternal cross. For example, when using the *Hipk* RNAi 1 line, if the RNAi line is female and the GAL4 driver male in the cross the progeny are referred to as ‘*Hipk* 1 maternal’ males, and ‘*Hipk* 1 paternal’ males when the cross is carried out in the opposite direction.

All flies used for mating in the 3-day mating assay were progeny from vials that had no more than 5 males and 5 females, to ensure all larvae had enough access to food and development was not hindered. For this assay, unmated 0–6 day old males and unmated 3–5 day old female flies were used. All female flies used in the assay were WT flies, the

## 2. Materials and Methods

male flies were the three parental lines (*nanos*-GAL4:VP16-64277, Hipk RNAi 1, Hipk RNAi 2) and the four types of knockdown males (Hipk 1 maternal, Hipk 1 paternal, Hipk 2 maternal, and Hipk 2 paternal; see Table 2.1 for stock descriptions).

### 2.4.2.2. Analysis

Statistical analyses were carried out in IBM SPSS Statistics 26. For the number of vials which produced F<sub>2</sub> progeny, Pearson's chi-squared tests were carried out on the two sets of Hipk knockdown data. The first set included the results from the *nanos*-GAL4 males, Hipk RNAi line 1, Hipk 1 maternal cross and the Hipk 2 paternal cross. The second set included the results from the *nanos*-GAL4 males, Hipk RNAi line 2, Hipk 2 maternal cross and the Hipk 2 paternal cross. The data for the *nanos*-GAL4 males were the same in both sets. The *p*-value was below 0.05 for both chi-squared tests so pairwise chi-squared tests were carried out within the sets with Yates's correction for continuity applied, as the degrees of freedom was 1 in these tests.

The number of progeny produced from the 3-day mating assays were compared within each Hipk 1 and Hipk 2 sets, as described above. First, the data from each type of F<sub>1</sub> male was tested for normality using the Shapiro–Wilk test. The data were all normal ( $p < 0.05$ ) except for the Hipk 1 paternal cross. Thus, the non-parametric Kruskal–Wallis *H* test was performed on the Hipk 1 set and a one-way ANOVA was performed on the Hipk 2 set. In both tests the null hypothesis was rejected ( $p < 0.05$ ). Dunn's test followed by a Bonferroni correction was carried out for the post-hoc pairwise comparisons of the Hipk 1 set, and Tukey's honestly significant difference test was carried out for the post-hoc pairwise comparisons of the Hipk 2 set.

## 2.5. Immunofluorescence

### 2.5.1. Dissections for immunofluorescence

5–10 male flies were anaesthetised on ice or with CO<sub>2</sub> and dissected in 1X PBS on a silicone dissecting dish with number 5-SA entomological forceps. Dissected testes then placed into 1.5 mL tubes with 100–500  $\mu$ L PBS.

### 2.5.2. Staining and microscopy

Testes were fixed by 4% paraformaldehyde in PBS for 2 hours at room temperature or overnight at 4 °C. They were then washed in 1X PBX (Table 2.6) 3 times for 2 minutes at room temperature. Next, they were blocked in blocking buffer (Table 2.6) for 1 hour at room temperature or overnight at 4 °C. Primary antibodies were used in 1X PBX and testes were incubated for two hours at room temperature or overnight at 4 °C. Testes were

then washed in 1X PBX 3 times for 15 minutes each at room temperature. They were then incubated with secondary antibodies in PBX for either 2 hours at room temperature or overnight at 4 °C. This method was then repeated sequentially if testes were being stained with more than one primary antibody. Samples were then mounted onto slides with Vectashield antifade mounting medium with DAPI (Vector Laboratories) and left for at least 12 hours before imaging. All antibodies used are described and listed in Table 2.7.

Slides were imaged using a Zeiss LSM880 Upright Confocal Microscope with the Zen imaging software. 10X dry and 40X oil-immersion objectives were used for locating samples and imaging samples, respectively. When using the oil objective, a layer of mineral oil was applied on the slide. The lasers used were Argon 458, 488 and 514 nm, and the Diode 405 nm.

**Table 2.6. Buffers and solutions used for immunofluorescence.** Names and component concentrations of buffers and solutions. All descriptions are for 1X solutions unless stated otherwise.

<b>Solution name</b>	<b>Component concentrations</b>
Blocking buffer	0.1% Triton X-100, 2% (v/v) normal goat serum (Thermo Fisher), in 1X phosphate buffered saline (PBS)
PBX	0.1% (v/v) Triton X-100, 0.5% (v/v) normal goat serum, in PBS

## 2. Materials and Methods

**Table 2.7. Antibodies used for immunofluorescence.** Primary antibodies are listed in the top half of the table and secondary antibodies in the bottom half.

<b>Antibody</b>	<b>Additional information</b>	<b>Stock concentration (<math>\mu\text{g}/\text{mL}</math>) and dilution used</b>	<b>Supplier (catalogue reference)</b>
Armadillo	Mouse IgG2a Monoclonal Hybridoma supernatant	27 1:100	DSHB (N2 7A1)
Boule	Rabbit IgG Polyclonal Affinity purified	133 1:100	Proteintech (13720-1-AP)
FMRP	Mouse IgG1 Monoclonal Hybridoma supernatant	59 1:100	DSHB (5A11)
HA	Mouse IgG2b Monoclonal Ascites fluid	400 1:100	Roche (12CA5)
Vasa	Rat IgM Monoclonal Hybridoma supernatant	44 1:200	DSHB (anti-vasa)
Goat anti-rat IgM	Alexa Fluor 488 Polyclonal Affinity purified	2000 1:400	Thermo Fisher (A-21212)
Goat anti-mouse IgG2a	Alexa Fluor 555 Polyclonal Affinity purified	2000 1:400	Thermo Fisher (A-21137)
Goat anti-mouse IgG2b	Alexa Fluor 594 Polyclonal Affinity purified	2000 1:400	Thermo Fisher (A-21145)
Goat anti-mouse IgG1	Alexa Fluor 633 Polyclonal Affinity purified	2000 1:400	Thermo Fisher (A-21126)
Goat anti-rabbit IgG (H+L)	Alexa Fluor 633 Polyclonal Affinity purified	2000 1:400	Thermo Fisher (A-21070)

### 2.5.3. Image analysis

#### 2.5.3.1. Counting hub cells

Positive staining for Armadillo and negative staining for Vasa was used to identify hub cells at the apical tip of testes. Z-stacks, with at least 10% overlap at 40X were then taken to image across the depth of the hub. Hub cells were then counted from these stacked images in Fiji (Schindelin et al., 2012).

Statistical analyses were carried out in IBM SPSS Statistics 26. The results from each sample were tested for normality using the Shapiro–Wilk test, all were normal ( $p > 0.05$ ). The number of hub cells were compared using one-way ANOVAs on the two sets of Hipk knockdown data. The first set included the results from the *nanos*-GAL4 males, Hipk RNAi line 1, Hipk 1 maternal cross and the Hipk 2 paternal cross. The second set included the results from the *nanos*-GAL4 males, Hipk RNAi line 2, Hipk 2 maternal cross and the Hipk 2 paternal cross. The data for the *nanos*-GAL4 males were the same in both sets. The  $p$ -value was below 0.05 for both tests so Tukey’s honestly significant difference test was carried out for post-hoc pairwise comparisons within each set.

#### 2.5.3.2. Morphology assessment

Tiled images of testes were taken when the area of the hub (marked by positive Armadillo staining) was at its largest, scale bars were provided on the images. With no information given about the genotype, five volunteers were asked to independently assess the tiled images into one of three categories: 1) no phenotype, 2) weak phenotype, 3) strong phenotype. One or two images were given as examples for each of the categories with some features highlighted. Features of the no phenotype category highlighted were: a rounded apical tip and the presence of long spermatids. Features for the weak phenotype: a pointed apical tip combined with the presence of long spermatids. Features for the strong phenotype: a pointed apical tip and short or no spermatids and a shorter testis length.

After receiving the survey results, the classification for each image was decided on the ‘mode’ of the results as a way to extract a consensus. There was not a consensus for 2 of the 69 images, these were polarised with each of the image having received two votes for no phenotype, two votes for strong phenotype and one vote for weak phenotype. These 2 images were classified as weak phenotype.

## 2.6. RIP-seq

### 2.6.1. Ribonucleoprotein immunoprecipitation

#### 2.6.1.1. Dissection and lysis

Approximately 50 1–3 day old male flies were anaesthetised on ice at a time. The reproductive systems of the male flies were dissected in PBS with 30 U/mL RNasin Plus RNase inhibitor. After 10 pairs of testes were dissected they were gathered together and snap frozen in liquid nitrogen. Samples were stored at -80 °C. At least 1000 pairs of testes were collected per sample.

An equal volume of RIP lysis buffer (Table 2.8) was added to each sample. The tissue was homogenised with a micropestle first then passed through a 23 gauge needle at least 5 times. The lysates were incubated on ice for 30 minutes, and inverted half way through. After this the lysates were snap frozen.

#### 2.6.1.2. HA pull-down

180 µL of anti-HA beads (Pierce, #88837) per sample were used and were washed 5 times in ice-cold NT2 buffer. After the final wash the beads were resuspended in NT2 buffer (Table 2.8) at 5.5 times the volume of the lysate, if necessary samples were divided into multiple 1.5 mL tubes. To each tube of beads 0.2 U/µL of RNase inhibitor and 20 mM EDTA pH 8 were added.

The lysates were thawed and centrifuged at 15,000 xg for 15 minutes at 4 °C. 20 µL or 10% (whichever was less) of the supernatant was set aside and snap frozen as the undiluted lysate. The remaining supernatant was added the anti-HA beads in NT2 buffer. 150 µL per sample was taken as diluted lysate and snap frozen. Beads and lysate were incubated while tumbling for 1 hour at room temperature.

The supernatant from all samples were saved as flow through sample and snap frozen. 500 µL of NT2 buffer was added to all tubes and any samples split over multiple 1.5 mL tubes were now merged. Beads were washed 4 more times in 1 mL of NT2 buffer.

#### 2.6.1.3. RNA elution and extraction

For RNA elution, 750 µL of sample and beads were resuspended in 100 µL NT2 buffer with 20 mg/mL proteinase K to release the RNP components. Samples were incubated for 30 minutes at 65 °C. At the same time 100 µL of diluted lysate and 300 µL of flow through samples were also treated with proteinase K in the same manner.

After treatment the sample was separated from the beads. 300 µL of Trizol was added per 100 µL of sample and incubated for 5 minutes at room temperature. 0.15% (v/v)

chloroform was added and the tubes were shaken vigorously for 15 seconds, followed by incubation at room temperature for 3 minutes. Samples were centrifuged at 12,000  $\times g$  for 15 minutes at 4 °C. The aqueous phase was collected and 1 volume of isopropanol was added along with 1  $\mu\text{L}$  glycoblue and 0.3 M NaCl. The RNA was precipitated at -80 °C for at least 3 hours.

After precipitation, samples were centrifuged at the maximum speed (13,300  $\times g$ ) at 4 °C for 20 minutes. The pellet was washed twice with 70% ethanol with 5 minute spins in between. After the final spin the supernatant was removed and the pellet was left to air-dry for 10–15 minutes. The pellet was resuspended in 18  $\mu\text{L}$  nuclease-free water.

#### 2.6.1.4. Protein elution

For protein elution the remaining 250  $\mu\text{L}$  of the sample and beads were taken. The NT2 buffer was replaced with 25  $\mu\text{L}$  of 2X protein sample buffer (Table 2.8), vortexed and boiled at 95 °C for 10 minutes. This was repeated a second time and elution 1 and elution 2 samples were stored at -20 °C.

#### 2.6.1.5. Sodium dodecyl sulphate polyacrylamide gel electrophoresis

Sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) was used to separate proteins. Gels were handcast with a final acrylamide concentration of 10 or 15% (depending on the protein size) for the resolving portion of the gels, and 5% acrylamide for the stacking gels. Samples were diluted in 2X protein sample buffer or 4X Laemmli Buffer (BioRad).  $\beta$ -mercaptoethanol was added to all samples at a final concentration of 5% (v/v) before being heated at 95 °C for at least 5 minutes and loaded onto acrylamide gels. 5  $\mu\text{L}$  of PageRuler Prestained Protein Ladder (Thermo Fisher) was loaded onto every gel. Gels were run in running buffer (Table 2.8) at 150 V until the protein bands of the ladder were sufficiently separated.

#### 2.6.1.6. Western blotting

Protein samples were run on SDS-PAGE gels (see section 2.6.1.5). Proteins were transferred to Amersham Protran 0.2  $\mu\text{m}$  nitrocellulose membranes (GE Healthcare) in transfer buffer (Table 2.8) at 200 mA for 1.5 hours. After transfer membranes were stained with Ponceau S (VWR) to confirm even transfer of proteins across the membranes. Membranes were then rinsed before being blocked in 5% (w/v) milk in 1X PBS-T for at least 1 hour at room temperature or overnight at 4 °C. Membranes were incubated with primary antibodies diluted in 1X PBS-T for either 4 hours at room temperature or overnight at 4 °C. Membranes were washed 3 times for 5 minutes in PBS-T before being incubated with the appropriate HRP-linked secondary antibody (diluted in 1X PBS-T) for 1 hour at room temperature. After this, blots were washed 3 times for 5 minutes in 1X PBS-T. EZ-ECL

## 2. Materials and Methods

(Biological Industries) was used for chemiluminescent, and signal was detected using X-ray film. Antibodies are listed with dilutions used and suppliers in Table 2.9.

**Table 2.8. Buffers and solutions used for RIP-seq.** Names and component concentrations of buffers and solutions. All descriptions are for 1X solutions unless stated otherwise.

<b>Solution name</b>	<b>Component concentrations</b>
NT2 buffer	50 mM Tris-HCl pH 7.4, 150 mM NaCl, 1 mM MgCl <sub>2</sub> , 0.05% (v/v) IGEPAL
PBS-T	0.1% Tween 20, in 1X PBS
2X protein sample buffer	125 mM Tris-HCl pH 6.8, 5% (w/v) SDS, 25% (v/v) glycerol, 10% (v/v) $\beta$ -mercaptoethanol, 0.004% (w/v) bromphenol blue
RIP lysis buffer	50 mM Tris-HCl pH 8, 150 mM NaCl, 10 mM MgCl <sub>2</sub> , cOmplete mini protease inhibitor cocktail (Roche), 1% (v/v) IGEPAL, 24 U/mL Turbo DNase (Thermo Fisher), 30 U/mL RNasin Plus RNase inhibitor (Promega)
Running buffer	25 mM Tris base, 250 mM glycine, 0.1% (w/v) SDS
Transfer buffer	25 mM Tris base, 192 mM glycine, 20% (w/v) SDS

**Table 2.9. Antibodies used for western blotting.** Primary antibodies are listed in the top part of the table and secondary antibodies in the bottom part.

<b>Antibody</b>	<b>Additional information</b>	<b>Stock concentration (µg/mL)</b>	<b>Supplier (catalogue reference)</b>
Armadillo	Mouse IgG2a Monoclonal Hybridoma supernatant	27	DSHB (N2 7A1)
HA	Mouse IgG2b Monoclonal Ascites fluid	400	Roche (12CA5)
HA	Rabbit IgG Polyclonal Affinity purified	1000	Abcam (ab9110)
PSI	Rabbit IgG Polyclonal Affinity purified	Unknown	Custom (kind gift from Don Rio, UC Berkeley)
β-tubulin	Mouse IgG1 Monoclonal Hybridoma supernatant	42	DSHB (E7)
Horse anti-mouse IgG	HRP-linked	153	Cell Signalling Technology (7076S)
Goat anti-rabbit IgG	HRP-linked	65.7	Cell Signalling Technology (7074S)

## 2. Materials and Methods

### 2.6.2. Sequencing

The RNA from the lysates and elutions of three HOW(S)-HA samples and the two parental controls were prepared first using the Ribo-Zero rRNA Removal Kit (Illumina) followed by TruSeq Stranded Total RNA Library Prep (Illumina). 100 ng of each sample was pooled and sequenced on the same lane on the NextSeq 500 Illumina sequencer using the High Output Kit v2.5 (75 Cycles), i.e. 75 bp single-end sequencing.

### 2.6.3. Computational analysis

#### 2.6.3.1. Filtering

The adapter sequence (AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC) was trimmed from reads with Cutadapt (Martin, 2011). Reads were filtered out if the quality score was below than 20 for 10% or more of the read. Filtering was done using the ‘Filter by quality’ tool in Galaxy (version 1.0.2; Gordon, 2010).

Subread (version 2.0.0) was used for rRNA and tRNA removal (Liao et al., 2013). The tRNA fasta file was from release 6.22 of the *D. melanogaster* genome. The rRNA fasta file was from RiboGalaxy’s Shared Data library (UUID: 0f0983aa-3afb-4b5f-a417-23593a3df1ef).

#### 2.6.3.2. Genome alignment and read counting

The remaining reads after filtering were aligned to the *D. melanogaster* genome release 6.22 using Subread. The type parameter (`-t`) was set to 0 to indicate RNA-seq reads and the remaining parameters were left at their default settings.

featureCounts was used to count reads to the following genomic features: ‘gene’, ‘5UTR’, ‘CDS’ and ‘3UTR’ (Liao et al., 2014). The gtf file used for this was from *D. melanogaster* genome release 6.22. parameter `-s` was set to 1 to indicated the library preparation results in a stranded library, `-g` was used to group features into the gene ID meta-feature. The remaining parameters were left at their default settings. Count tables from the ‘gene’ feature was used in the differential enrichment analyses.

Statistical analyses to compare the percentage of reads mapped to a feature between the HOW(S)-HA total RNA and pull-down RNA samples were carried out using SciPy’s statistical functions module (Virtanen et al., 2020). The data for each sample was tested for normality using the Shapiro–Wilk test. The data for the total RNA reads assigned to the 5UTR feature were not normal ( $p < 0.05$ ) so a Wilcoxon signed-rank test was carried out to compare the percentage of reads assigned to this feature from the total RNA and pull-down RNA. The rest of the data were normal and so paired *t*-tests were carried out for the ‘CDS’ and ‘3UTR’ features.

### 2.6.3.3. Transcriptome mapping

Transcriptome indexing and quasi-mapping was carried out using Salmon (Patro et al., 2017). A decoy-aware transcriptome was indexed with a pre-computed decoy sequence file provided by the Salmon developers. The transcriptome and decoy files used were based on the *D. melanogaster* genome from Ensembl release 97, which corresponds to FlyBase's release 6.22. The  $k$ -mer size selected for indexing the transcriptome was 31.

The remaining reads after filtering were quasi-mapped to the transcriptome with library parameter (-l) set to SR to indicate that the library preparation results in a reverse stranded library type. The  $k$ -mer parameter (-k) was set to 31.

### 2.6.3.4. Differential gene and transcript enrichment

The following analysis was executed in RStudio, version 0.99.486, with R version 3.6.2 (RStudio Team, 2020). Gene counts were directly imported into R. Transcript counts from Salmon were imported using the tximport package (version 1.14.0; Soneson et al., 2015) using txOut=TRUE and countsFromAbundance="scaledTPM" while importing.

Differential enrichment for both genes and transcripts was done with the edgeR package (version 3.28.0; Robinson et al., 2010; McCarthy et al., 2012). Low expression genes or transcripts were filtered using the filterByExpr function, libraries were normalised using calcNormFactors.

Two factors were defined for the HOW(S)-HA samples: condition and pairing. Condition referred to whether a sample was a total RNA or a pull-down RNA sample (i.e. from the lysate or elution). Pairing referred to the three pairs of total and pull-down RNA samples. Thus, the design matrix was submitted as follows: `model.matrix(~sample$condition + sample$pair)`. One factor was defined for the *nanos*-GAL4 samples to differentiate between the total RNA and pull-down RNA. The design matrix was submitted as: `model.matrix(~group)`.

For the HOW(S)-HA samples, dispersion was estimated with the design matrix taken into account. Then testing for differential genes and transcripts was carried out with the quasi-likelihood F-test, as per the edgeR manual instructions. For the *nanos*-GAL4 samples, the exact test was used with the dispersion set as the square of the biological coefficient of variation (BCV). This BCV was either 0.1 or the square root of the common dispersion from the HOW(S) data, depending on the analysis.

Finally, genes and transcripts from the HOW(S)-HA RIP-seq data were classed as significantly enriched or depleted if they had an adjusted  $p$ -value  $< 0.05$ , had a  $\log_2$ (Fold Change) above 1 or below -1, and if it did not meet these same thresholds in the *nanos*-GAL4 data.

## 2. Materials and Methods

### 2.6.3.5. Principal Component Analysis

Principal component analysis (PCA) was performed using the PCAtools (version 1.2.0; Blighe, 2019) package in R (version 3.6.2). PCA was carried out on the triplicate HOW(S)-HA pull-down samples with the  $\log_2$ (Counts Per Million) from the transcript-level data, as calculated by the edgeR package in section 2.6.3.4. The `pca` function was used with the `removeVar` parameter set to 0.1, this removes the lower 10% of variables based on variance. The `biplot` function was used to generate a graph of PC1 against PC2.

### 2.6.3.6. Gene ontology

Gene ontology analysis was carried out in Gene Ontology enRIchment anaLysis and visu-aLizAtion tool (GORilla; Eden et al., 2007; Eden et al., 2009). *D. melanogaster* genes under the GO term ‘signal transduction’ (GO:0007165) were accessed via FlyBase’s controlled vocabulary tool.

### 2.6.3.7. Motif enrichment

Discriminative Regular Expression Motif Elicitation (DREME) from MEME Suite (MEME version 5.1.0, with Python version 2.7.15) was used to carry out the motif enrichment analysis on 5’- and 3’-UTRs (Bailey, 2011). Control sequences were generated from the list of transcripts that were not filtered out by edgeR’s `filterByExpr` function. For the differentially enriched gene list, the UTR sequences were generated from the highest expressed transcripts for that enriched gene. DREME was implemented using the `-norc` flag so only the strand given was searched and not the complimentary sequences, `-rna` flag was used to indicate the sequences were of RNA not DNA, and `-m` was set to 25 to stop searching after 25 motifs had been found.

### 2.6.3.8. Circular RNA alignment and annotation

An annotation file was retrieved from Ensembl’s release 97 for *D. melanogaster* in the gene transfer format (GTF; Zerbino et al., 2018), which corresponds to FlyBase’s release 6.22. All rows with ‘gene’ in the third column were removed with a custom script, referred to as the modified GTF file. The file was then converted to the GenePred table format using the `gtfToGenePred` function from the `ucsc-gtftogenepred` package (version 366). Finally, a new first column was added with `ens97` added to every row with a custom script, referred to as the modified GenePred file.

Spliced Transcripts Alignment to a Reference (STAR; version 2.7.3a) was used to align the RIP-seq reads to release 6.22 of the *D. melanogaster* genome (Dobin et al., 2013). First, the genome was indexed using by setting the `--runmode` parameter to `genomeGenerate`. The `--sjdbOverhang` parameter was set to 75, `--genomeSAindexNbases` was set to 13, the

## 2.7. Protein sequence alignment and structure prediction

modified GTF file described above was used as the input annotation file with FlyBase’s all chromosome fasta file from release 6.22 of the *D. melanogaster* genome. All other parameters were left at their default settings. RIP-seq reads were used after the same filtering steps described in section 2.6.3.1. Reads were aligned to the genome by setting the `--runmode` parameter to `alignReads`. The `--chimSegmentMin` parameter was set to 15 and `--quantMode` set to `GeneCounts`, all other parameters were left at their default settings.

CIRCexplorer2 (version 2.3.8) was used to identify circRNAs from the STAR aligned files (Zhang et al., 2016). First, the `Chimeric.out.junction` output files from STAR were parsed using CIRCexplorer2’s `parse` function, with the `-t` parameter set to `STAR`. The `back_spliced_junction.bed` output files were then annotated using STAR’s `annotate` function with the modified GenePred file described earlier and FlyBase’s all chromosome fasta file from release 6.22 of the *D. melanogaster* genome. The values in the `circularRNA_known.txt` output files were used as the final counts for all circRNAs identified.

## 2.7. Protein sequence alignment and structure prediction

### 2.7.1. Protein sequence alignment

Clustal Omega was used to align the protein sequences of STAR proteins in Figure 5.1B, the default settings were used (Madeira et al., 2019). Asterisks indicate positions which have a single, fully conserved residue. Colons indicate conservation between groups of strongly similar properties. Full stops indicate conservation between groups of weakly similar properties. Protein sequences were retrieved from FlyBase (Larkin et al., 2021) or UniProt (Consortium, 2021).

### 2.7.2. Structure prediction

The 3D structure of HOW(S), with the HOW-PB amino acid sequence (FBpp0083576), was predicted using the I-TASSER online server with default settings (Zhang, 2008; Roy et al., 2010; Yang et al., 2014).

## 2.8. Cloning the STAR domain for protein purification

### 2.8.1. Gene synthesis and cloning reactions

A codon optimised STAR domain, amino acids 72–266 (HOW-PB; FBpp0083576), in the pGEX-6P1 vector was ordered from Genewiz (Sequence A.3 in Appendix I).

## 2. Materials and Methods

PCR was performed to create a product of the codon optimised STAR domain with extensions for cloning into pOPIN vectors. Primers used are listed in Table 2.10. Gradient PCR, with annealing temperatures ranging from 50 °C to 70 °C at 4 °C increments, was performed using Q5 high fidelity polymerase (New England Biolabs). Reactions contained 5 µL of 5X Q5 reaction buffer, 0.2 µM dNTPs, 0.5 µM of each primer, 1 µL of STAR-pGEX-6P1 template DNA, 0.02 U/µl Q5 DNA polymerase, and nuclease-free water for a final reaction volume of 25 µL. PCR cycles were carried out as described in Table 2.11

**Table 2.10. Primers for cloning STAR domain into pOPIN vectors.** Extension sequence for pOPIN vectors in capital letters, STAR domain sequence in lower case.

Primer name	Primer sequence (5' → 3')
pOPIN-STAR fwd	AAGTTCTGTTTCAGGGTCCCacccagagcattgcccgatta
pOPIN-STAR rev	CTGGTCTAGAAAGCTTtttagcttttggcgggtgtatcac

**Table 2.11. Thermocycler protocol for PCR with Q5 enzyme.**

Step	Temperature	Time	Cycles
Initial denaturation	98 °C	30 secs	
Denaturation	95 °C	5 secs	
Annealing	Variable	20 secs	30
Extension	72 °C	20 secs	
Final extension	72 °C	2 mins	

PCR products were run on a 1% (w/v) agarose gel, as described in section 2.2.3. Agarose gel electrophoresis. Product from the reaction with an annealing temperature of 70 °C was DpnI digested. DpnI digestion reaction contained: 1 µL DpnI enzyme (New England Biolabs), 2 µL 10X CutSmart Buffer (New England Biolabs), 17 µL STAR domain PCR product. Reaction was incubated at 37 °C for 30 minutes. The DpnI treated PCR product was purified with the QIAquick PCR Purification Kit (Qiagen) and eluted with 30 µL of nuclease-free water.

Cloning of the STAR domain into four different pOPIN (F, J, M, or S3C) vectors was carried out using the NEBuilder HiFi DNA Assembly Master Mix (New England Biolabs). Uncut vector DNA was cut and gel purified by the Protein Production Facility (University of Leeds). Each reaction contained a 1:2 molar ratio of vector to STAR domain PCR product, 10 µL HiFi DNA Assembly Master Mix, and nuclease-free water for a final reaction volume of 20 µL. Reactions were incubated at 50 °C for 1 hour. Samples were then stored at -20 °C or used immediately for transformations into bacterial cells.

### 2.8.2. Transformation into DH5α cells and sequence analysis

2 µL of each of the 4 pOPIN vectors containing the STAR domain was transformed into 50 µL of DH5α cells. These were placed on ice for 30 minutes, heat shocked at 42 °C

for 45 seconds, then returned to ice for further 2 minutes. Following this, 250  $\mu\text{L}$  of SOC media (Table 2.12) was added and samples were incubated for 60 minutes at 37 °C and shaking at 1000 rpm.

250  $\mu\text{L}$  of the transformation mix was spread across LB-agar (Table 2.12) plates containing 100  $\mu\text{g}/\text{mL}$  ampicillin and 20  $\mu\text{g}/\text{ml}$  X-gal. Plates were incubated at 37 °C overnight or 21 °C over 3 nights. White colonies were picked and placed into 10 mL of LB media with ampicillin (100  $\mu\text{g}/\text{mL}$ ). Cultures were incubated overnight at 37 °C shaking at 250 rpm. Plasmid DNA was then isolated using Wizard Plus SV Minipreps DNA Purification System (Promega).

Diagnostic digests were performed using HindIII HF and NcoI HF restriction enzymes (New England Biolabs). Reactions were set up as follows: 5  $\mu\text{L}$  of purified plasmid DNA, 2  $\mu\text{L}$  10X CutSmart buffer, 0.5  $\mu\text{L}$  of each restriction enzyme, and 12  $\mu\text{L}$  nuclease-free water. Reactions were incubated for 37 °C for 60 minutes.

Samples were run on an agarose gel (as in section 2.2.3) to see the results of the restriction digest. Plasmids with a product of the correct size from the restriction digest were sent off for Sanger sequencing to confirm the correct STAR domain sequence had been inserted into the four different vectors.

**Table 2.12. Buffers and solutions used for cloning.** Names and component concentrations of buffers and solutions used for cloning. All descriptions are for 1X solutions and made up with water unless stated otherwise.

Solution name	Component concentrations
Luria-Bertani (LB) medium	10 g/L tryptone, 10 g/L NaCl, 5 g/L yeast extract
LB-agar	LB with additional 15 g/L agar
SOC medium	20 g/L tryptone, 5 g/L yeast extract, 8.56 mM NaCl, 2.5 mM KCl, 10 mM $\text{MgCl}_2$ , 20 mM glucose, and adjust to pH 7

## 2.9. Protein expression and purification

### 2.9.1. Small-scale protein expression

Plasmid DNA was transformed into BL21(DE3) cells using the transformation protocol used for DH5 $\alpha$  cells (see 2.8.2). Cells were plated onto LB-agar plates with ampicillin (100  $\mu\text{g}/\text{mL}$ ) and plates were incubated either overnight at 37 °C or for 72 hours at 21°C. Two colonies were picked from each plate to grow in 10 mL cultures of either LB or TB media (Table 2.13), with the appropriate antibiotics, shaking overnight at 37 °C.

Two identical 96 deep well plates were set up to compare different growth and incubation conditions for all four STAR-pOPIN constructs. Cells were grown in either LB or TB

## 2. Materials and Methods

media and each well had 1.6 mL of media (with 100 µg/mL ampicillin) and 400 µL of the overnight culture. Both 96 well plates were incubated at 37 °C and the OD<sub>600</sub> was checked periodically. Once the reading reached 0.5, 400 µM IPTG was added to each well. One plate was placed overnight in an 18 °C incubator, the other was incubated at 37 °C for 4 hours. When plates were removed from the incubators they were centrifuged at 4000 rpm for 15 minutes. The supernatant was removed and the plates were stored at -20 °C.

### 2.9.2. Small-scale protein purification

A Hamilton STAR Liquid Handling Robot was set up by Protein Production Facility staff (University of Leeds) for protein purification with the MagneHis Protein Purification System (Promega). The programmed protocol resuspended the bacterial pellets, lysed the cells, added the MagneHis-Ni particles for a 2 minute incubation, removed the supernatant, washed the beads and eluted the protein in 50 µL into a 96 well plate.

2X protein loading buffer (Table 2.13) was added to each well and samples were run on 10 or 12% SDS-PAGE gels at 150 V (see 2.6.1.5 for more details). The gels were then placed in Coomassie stain, followed by Coomassie destain (Table 2.13). Once sufficiently destained the gels were then imaged using a G:BOX XX9.

### 2.9.3. Large-scale protein expression

BL21(DE3) cells were transformed with either the His-STAR construct (from the pOPINF vector) or the His-GST-STAR construct (from the pOPINJ vector), and were grown in six 10 mL cultures of 2YT media (Table 2.13) with ampicillin (100 µg/mL). 10 mL cultures were incubated while shaking overnight at 37 °C. Each 10 mL starter culture was poured into 1 L of the appropriate media (2YT or LB) with ampicillin (100 µg/mL) and placed on a shaking incubator. Once OD<sub>600</sub> was between 0.6–0.8, the 1 L cultures were induced with 400 µM IPTG overnight at 18 °C.

After overnight induction the 1 L cultures were poured into 1 L centrifuge bottles and spun at 4500 rpm for 15 minutes. Pellets from two 1 L cultures were combined into 50 mL tubes and stored at -80 °C.

### 2.9.4. Large-scale His-STAR purification

His-STAR was expressed in BL21(DE3) cells as described in section 2.9.3. Bacterial pellets from 2 L of culture were lysed in 20 mL of bacterial lysis buffer (Table 2.13). Once the pellet had thawed the following were added while keeping the sample on ice: an EDTA-free protease inhibitor cocktail tablet, 200 µg DNase I, 200 µg RNase I, 500 µL B-PER (Thermo Scientific) and 100 mg of lysozyme. Once these all dissolved the lysate was then

passed through an Avestin C3 high pressure cell disrupter before being centrifuged for 45 minutes at 16,000 rpm at 4 °C.

After centrifugation a portion of the pellet was taken as the ‘insoluble fraction’ and 20 µL of supernatant was taken as the ‘soluble fraction’ for SDS-PAGE analysis. The remaining supernatant was filtered through a 0.45 µm filter, and passed over the HisTrap column that had been equilibrated with bacterial lysis buffer. This was followed by a 50 mL high salt wash (Table 2.13), which was collected in two 25 mL fractions. Increasing concentrations of imidazole buffers, all at pH 8, were passed through the column in 25 mL fractions to elute the protein: 75 mM, 150 mM, 200 mM, 250 mM, 300 mM, 400 mM and 500 mM. 20 µL of every fraction was collected for protein quantification and SDS-PAGE.

The His-tag was cleaved from His-STAR protein in the imidazole elution fractions, except for the 75 mM elution, using 120 µg/mL of PreScission Protease with samples rotating overnight at 4 °C. Based on SDS-PAGE analysis the appropriate samples were pooled together and concentrated to 10 mL using a 10 kDa molecular weight cut-off filter. Concentrated protein was passed through a 0.22 µm filter prior to size exclusion chromatography (SEC).

SEC was carried out using a HiLoad 26/600 Superdex 75 pg column (GE Healthcare), which was equilibrated overnight with degassed size exclusion buffer (Table 2.13) at 4 °C with an AKTA prime pump liquid chromatography system measuring absorbance at 280 nm. Filtered protein was loaded onto and flowed through the column at 2 mL/min collecting 3 mL fractions. Samples from the fractions containing protein, based on the absorbance readout were collected and run on an SDS-PAGE gel (see section 2.6.1.5) to assess purity of fractions.

### 2.9.5. Large-scale His-GST-STAR purification

His-GST-STAR was expressed in BL21(DE3) cells (section 2.9.3). Bacterial cell pellets were lysed as in section 2.9.4. After centrifugation a portion of the pellet was taken as the ‘insoluble fraction’ and 20 µL of supernatant was taken as the ‘soluble fraction’ for SDS-PAGE analysis. The remaining supernatant was filtered through a 0.45 µm filter.

50 mL of the filtered supernatant was incubated in a column Glutathione Sepharose 4B resin (GE Healthcare), which had been equilibrated with bacterial lysis buffer. The column was rotated at 4 °C for 1.5 hours. Following this, the 50 mL of supernatant, and any remaining filtered supernatant that had not been incubated with the resin, was flowed through the column. The column was washed with 500 mL of GST wash buffer (Table 2.13), 20 µL was collected for every 100–200 mL that was passed through. Protein was eluted from the column using 40 mL GST elution buffer, followed by a 50 mL of GST wash buffer. The elution and elution wash samples were each treated with PreScission protease to cleave the His-GST-tag from the STAR protein, samples were rotated overnight at 4 °C.

## 2. Materials and Methods

After SDS-PAGE analysis of the samples from the purification, the cleaved elution and elution wash samples were passed through a HisTrap column and eluted with 25 mM and 500 mM imidazole buffers (both at pH 8). Samples were collected for SDS-PAGE analysis and the 500 mM elution fraction was concentrated. Samples from the gel and from the concentrated fraction were sent to the Biomolecular Mass Spectrometry Facility at the University of Leeds (see section 2.10).

### 2.9.6. Large-scale His-GST-STAR purification by Protein Production Facility

His-GST-STAR was expressed in BL21(DE3) cells (section 2.9.3). Bacterial pellets were lysed and protein was purified by Dr Brian Jackson and Laura Wilkinson Hewitt from the Protein Production Facility (University of Leeds). The purification process involved a HisTrap purification with an AKTA to isolate the His-GST-STAR protein from the bacterial lysis, followed by cleavage with 3C protease overnight to removed the His-GST-tag. HisTrap purification was performed again to isolate the STAR protein from the tag and protease prior to SEC with a 26/600 Superdex 75 column. The final buffer used that the purified protein was stored in is listed under SEC PPF buffer in Table 2.13.

Samples from the fractions containing protein, based on the chromatogram were collected and run on an SDS-PAGE gel (see section 2.6.1.5) to assess purity of fractions. Fractions that contained pure STAR protein were pooled and used for subsequent fluorescence anisotropy experiments (see section 2.11.2) or snap frozen with 5% (v/v) glycerol added.

## 2.10. Mass spectrometry

Mass spectrometry (MS) experimentation and analysis was carried out by Dr James Ault and Rachel George from the Biomolecular Mass Spectrometry Facility at the University of Leeds. Protein bands were excised from Coomassie stained SDS-PAGE gels and trypsin digested for peptide identification via MS/MS. Mass determination of purified protein in solution was performed using liquid chromatography MS.

## 2.11. Fluorescence anisotropy

### 2.11.1. Oligonucleotides synthesis and structure prediction

RNA oligonucleotides (oligos) were synthesised by Integrated DNA Technologies. Oligos were 3' labelled with 6-carboxyfluorescein and all oligos were desalted after synthesis.

Mfold was used to predict RNA secondary structures from the oligos, and also to create the images of RNAs in Figures 1.10 and 5.18 (Zuker, 2003). Mfold was used through

**Table 2.13. Buffers and solutions used in protein expression and purification.**

Names and component concentrations of buffers and solutions used in protein expression and purification. All descriptions are for 1X solutions and made up with water unless stated otherwise.

Solution name	Component concentrations
Bacterial lysis buffer	200 mM NaCl, 20 mM Tris-HCl pH 7.5, 10% Triton-X
Coomassie destain	50% (v/v) methanol, 10% (v/v) acetic acid
Coomassie stain	0.1% (w/v) Coomassie Blue R-250, 50% (v/v) methanol, 10% (v/v) acetic acid
GST elution buffer	150 mM NaCl, 25 mM Tris-HCl pH 7.5, 80 mM L-glutathione
GST wash buffer	150 mM NaCl, 25 mM Tris-HCl pH 7.5
High salt wash buffer	1 M NaCl, 20 mM Tris-HCl pH 8
Luria–Bertani (LB) medium	10 g/L tryptone, 10 g/L NaCl, 5 g/L yeast extract
LB-agar	LB with 15 g/L agar
2X protein loading buffer	125 mM Tris-HCl pH 6.8, 4% (w/v) SDS, 40% (v/v) glycerol, 0.2% (w/v) bromophenol blue, 10 mM DTT
Size exclusion buffer	100 mM NaCl, 20 mM Tris-HCl pH 8, 5% (v/v) glycerol, 1 mM DTT
Terrific broth (TB) medium	20 g/L tryptone, 24 g/L yeast extract, 4 mL/L glycerol, 17 mM $\text{KH}_2\text{PO}_4$ , 72 mM $\text{K}_2\text{HPO}_4$
SEC PPF buffer	25mM HEPES pH 7.6, 150mM NaCl, 1mM TCEP
2X Yeast-tryptone (2YT) media	16 g/L tryptone, 10 g/L NaCl, 10 g/L yeast extract

their web server (<http://www.unafold.org/mfold/applications/rna-folding-form.php>) with default settings.

### 2.11.2. Experimental setup

Binding assays between 3' fluorescein labelled RNA oligos and purified protein were carried out in triplicate in black 384-well OptiPlates (Perkin Elmer), control wells (with no RNA) were carried out once. HOW STAR domain was purified as described in section 2.9.6. Simbu orthobunyavirus nucleoprotein (SIMV NP) was used as a positive control and was purified by Georgia Pangratiou (as described in Pangratiou, 2020).

20  $\mu\text{L}$  of RNA binding buffer (20 mM Tris-HCl pH 7.5, 150 mM NaCl, 0.01% Triton X-100) was first added to every well. 20  $\mu\text{L}$  of protein was added to the first column and tritrated across the 24 rows. Finally, 20  $\mu\text{L}$  of 5 nM fluorescein labelled RNA oligo was added to the appropriate rows, and for the control rows 20  $\mu\text{L}$  of RNA binding buffer was added instead. The plate was left to equilibrate for at least 45 minutes prior to data collection on a Spark 10M Multimode Microplate Reader (Tecan) with a 485 nm (20 nm, bandwidth)

## 2. Materials and Methods

excitation filter, and parallel (S) and perpendicular (P) channel emission filters at 535 nm (25 nm bandwidth).

### 2.11.3. Analysis

Anisotropy values were calculated using the emission values for S and P signals (corrected with the control values from protein only wells) with the following equation:

$$Anisotropy = \frac{S - P}{S + 2P}$$

The logistic function in OriginPro 2020 V2 was first used to determine the theoretical minimum and maximum anisotropy values ( $A_1$  and  $A_2$ , respectively), using the equation below:

$$y = A_2 + \frac{(A_1 - A_2)}{1 + \left(\frac{x}{x_0}\right)^p}$$

Where  $y$  is either anisotropy or fraction bound,  $x$  is protein concentration,  $x_0$  is the dissociation constant and  $p$  is the Hill coefficient.

$A_1$  and  $A_2$  were used to calculate the fraction of RNA bound. The logistic function was used again to fit a curve for the fraction bound data. The dissociation constant ( $x_0$ ) from these curves was reported as the apparent  $K_D$ .

## 2.12. Schematics and figures

The following schematics were created with [BioRender.com](https://www.biorender.com): Figures 1.5, 3.1, 3.3, 3.6, 3.9, 4.4, 4.11 and 5.14. Additionally, Figure 1.1 was adapted from Witt et al. (2019) by Karl Norris and myself with [BioRender.com](https://www.biorender.com), and is used under [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/). Figure 1.2 is from Matunis et al. (2012), which is licensed under [CC BY-NC 3.0](https://creativecommons.org/licenses/by-nc/3.0/). Figure 1.3 is adapted from Fabian and Brill (2012), which is used under [CC BY-NC 3.0](https://creativecommons.org/licenses/by-nc/3.0/), and Hundertmark et al. (2018), which is used under [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/). Protein structures in Figures 1.7, 1.8, 1.9 and 5.2 were produced using MacPyMol (version 1.7.2.3).

# 3. Identifying the targets of HOW(S) in *Drosophila melanogaster* testis

## 3.1. Introduction

The testis is a very transcriptionally active tissue, from humans to flies the testis expresses more unique genes than other organs (Chintapalli et al., 2007; Soumillon et al., 2013). In addition, the testis is the tissue in which the transcriptome and the proteome are the most poorly correlated. In a study of eight different human tissues the testis had the lowest correlation, 0.138, by contrast, liver tissue had the highest correlation, 0.432 (Cagney et al., 2005). This means that as a tissue the testis is the site of a substantial amount of post-transcriptional gene regulation. This emphasises the testis, and the process of spermatogenesis, as a particularly interesting tissue to study post-transcriptional gene regulation. RNA binding proteins (RBPs) are important throughout the life cycle of an RNA, and there are numerous examples of their importance in regulating gene expression in spermatogenesis e.g. the DAZ and STAR protein families (Paronetto et al., 2009; VanGompel and Xu, 2011; Priti and Subramaniam, 2015).

*D. melanogaster* has extensively been used as a model for spermatogenesis, as it has for many other biological processes. The *D. melanogaster* testis in particular has a clearly defined cellular architecture suitable for studying the different cell populations through the process of spermatogenesis. The three main phases of spermatogenesis occur in flies as it does in mammals: 1) mitosis — where the spermatogonia amplify in number, 2) meiosis — where the spermatocytes become haploid, and 3) spermiogenesis — where the spermatids differentiate into mature spermatozoa. Held out wings (HOW) is a member of the STAR protein family, and is the *D. melanogaster* orthologue of the human Quaking protein. HOW has previously been shown to be an RBP essential for spermatogenesis (Monk et al., 2010). It is expressed in the germ cells at the earliest stages of spermatogenesis, in the germline stem cells (GSCs) and during the first two (out of the typical four) mitotic divisions of the spermatogonia (Monk et al., 2010). Additionally, it is expressed in some of the somatic cells of the testis, the hub cells and the mature cyst cells (Monk et al., 2010).

In the testis, HOW(L), the longer and nuclear HOW isoform, regulates the number of the divisions during the mitotic phase, through direct interaction with *bag of marbles* and an indirect interaction with Cyclin B, though the exact mechanism behind this has not been determined (Monk et al., 2010). Yet the role of the cytoplasmic, shorter isoform, HOW(S), remains poorly understood. None of the RNAs with which HOW(S) interacts in the testis have been identified, and subsequently there is no understanding of HOW(S) function in this tissue. In other tissues a small number of HOW(S)–RNA interactions have

### 3. Identifying the targets of HOW(S) in *Drosophila melanogaster* testis

been characterised. HOW(S) improves the stability of *stripe* mRNA in tendons (Nabel-Rosen et al., 1999), and *dgrasp* mRNA in ovaries (Giuliani et al., 2014). In glial cells HOW(S) is located in both the cytoplasm and nucleus, and it has been shown to regulate the alternative splicing of *Neurexin IV* (Edenfeld et al., 2006; Rodrigues et al., 2012). However, an *in vivo*, transcriptome-wide assessment of which mRNAs HOW(S) binds to has yet to be performed. Here, this chapter aims to determine which mRNAs HOW(S) binds in the cytoplasm of germ cells, for the first time.

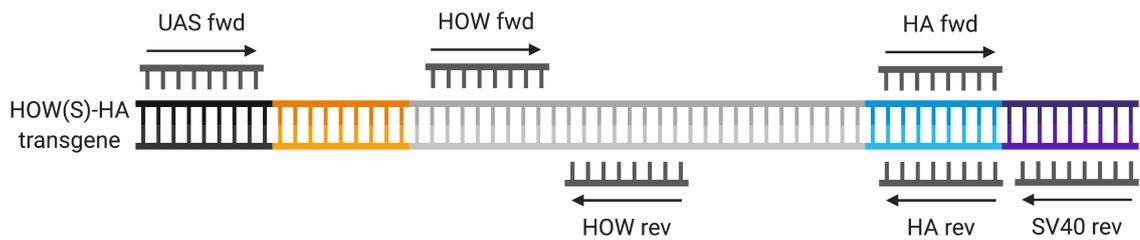
RIP-chip, which was later developed into RIP-seq, was one of the earliest high throughput methods for identifying RNAs associated with RBPs of interest (Tenenbaum et al., 2000; Zhao et al., 2010a). This method is high throughput while still being well suited for studying protein–RNA interactions in a tissue as it does not require any cross-linking steps, which is needed for CLIP-seq (Licatalosi et al., 2008). While RIP-seq cannot identify specific binding sites within the RNAs, it is suited for this study where HOW(S) is being expressed in just a few cells in a tissue. Thus, given that the low starting material is a limiting factor, it is an advantage here that RIP-seq does not involve any digestion of the RNA (Keene et al., 2006; Zhao et al., 2010a). To better understand the previously unexplored role of HOW(S) in spermatogenesis, the first aim of this thesis is to identify the RNAs HOW(S) binds in germ cells using RIP-seq.

## 3.2. Results

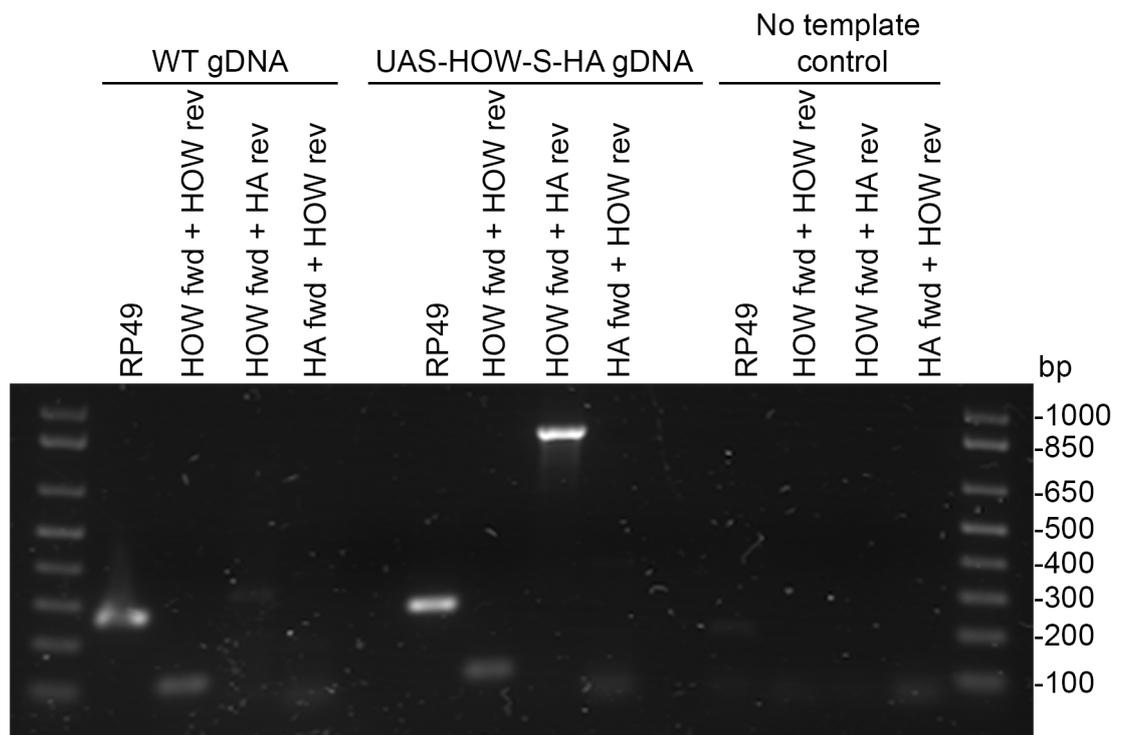
### 3.2.1. HOW(S)-HA expression in the male germline

In the *D. melanogaster* testis HOW is expressed in the hub cells, cyst cells, germline stem cells (GSCs) and in spermatogonia up to the 4-cell stage (i.e. for the first two mitotic divisions; Monk et al., 2010). To identify the RNAs which HOW(S) binds in the male germ cells, a strategy was designed to specifically probe HOW(S)–RNA interactions *in vivo*. To do this I made use of UAS-HOW-S-HA fly line, (provided by Professor Talila Volk), in combination with a *nanos*-GAL4 driver. In adult flies, the *nanos*-GAL4 fly line expresses the GAL4 transcription factor in these early differentiating germ cells (White-Cooper, 2012). Thus, mating these two fly lines would enable the limited expression of HA-tagged HOW(S) in GSCs and spermatogonia.

To confirm the presence of the HA-tag adjacent to the HOW(S) coding sequence (CDS), genomic DNA (gDNA) from UAS-HOW-S-HA flies was extracted and PCRs were carried out using HOW forward and HA reverse primers (Figure 3.1). The HOW primer pair and ribosomal protein 49 (Rp49) primer pair were used as a positive controls. HA forward primers combined with HOW reverse primers were used as negative controls. No signal was seen in the no template control, and, as expected, the ‘HOW fwd + HA rev’ PCR resulted in a product at the predicted size of ~829 base pairs (bp) for the UAS-HOW-S-HA gDNA but not in the gDNA from wild type (WT) flies (Fig 3.2). This confirmed the presence of the HOW(S)-HA transgene in the UAS-HOW-S-HA flies.



**Figure 3.1. Schematic representing the PCR genotyping of the UAS-HOW-S-HA transgenic fly line.** The UAS-HOW-S-HA fly line contains a HOW(S)-HA CDS with an upstream activation sequence (UAS) promoter region from the UAS<sub>t</sub> construct (see FlyBase tool FBto0000342; Brand and Perrimon, 1993). The transgene (from left to right in the schematic, regions not to scale) is comprised of 5 UAS repeats (black), partial promoter and 5'-UTR sequences from Hsp70 (orange), the HOW(S) CDS (grey), a 3x HA-tag (blue) and the SV40 3'-UTR and terminator sequence (purple). HOW primers were designed to detect both the WT HOW gene and the HOW(S)-HA transgene. When the HOW forward primer is combined with a reverse primer for the HA-tag (HA rev) a product will be amplified in UAS-HOW-S-HA gDNA but not in WT samples.

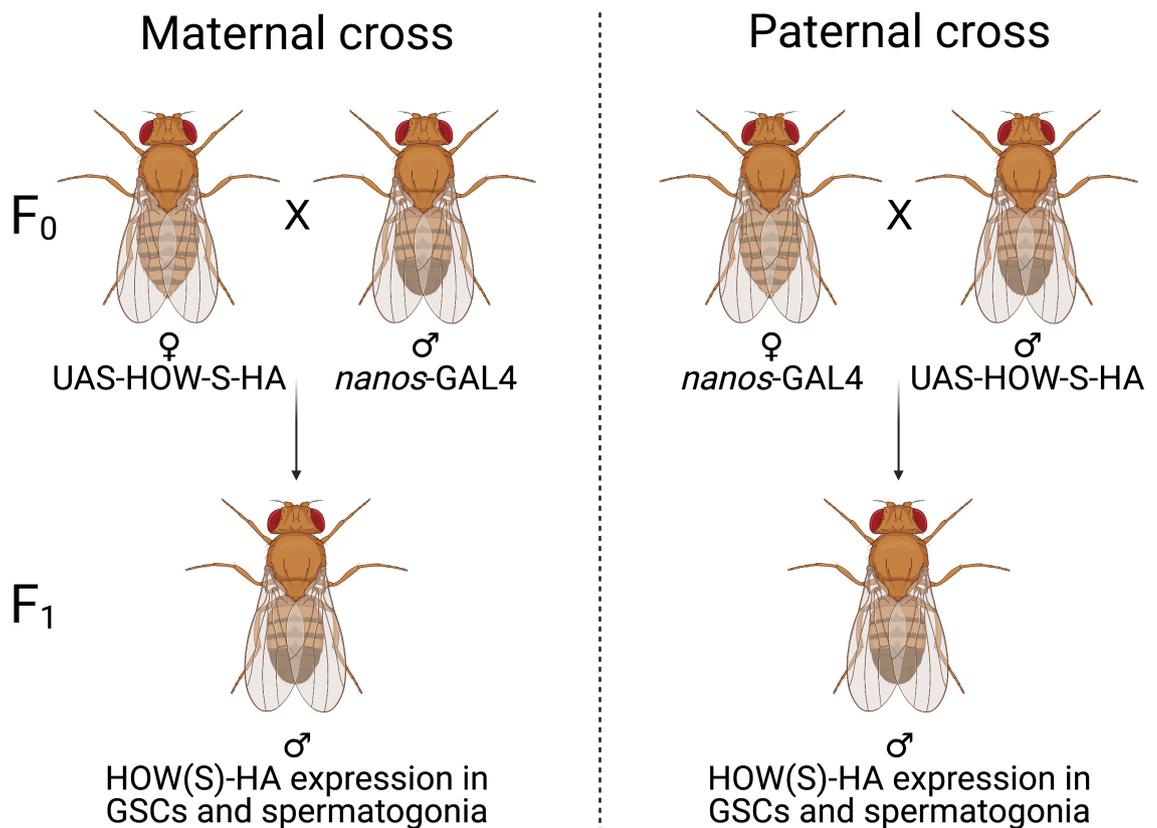


**Figure 3.2. Genotype of UAS-HOW-S-HA via PCR and agarose gel electrophoresis.** gDNA from WT and UAS-HOW-S-HA flies were extracted and PCR was performed with Rp49, HOW and HA primers. Expected product sizes: 207 bp for 'Rp49', 109 bp for 'HOW fwd + HOW rev', 829 bp for 'HOW fwd + HA rev', no amplification expected for 'HA fwd + HOW rev'. Rp49 and HOW are present in both fly lines. For the combined HOW fwd and HA rev primers only the UAS-HOW-S-HA gDNA sample has a product of the correct size.

### 3. Identifying the targets of *HOW(S)* in *Drosophila melanogaster testis*

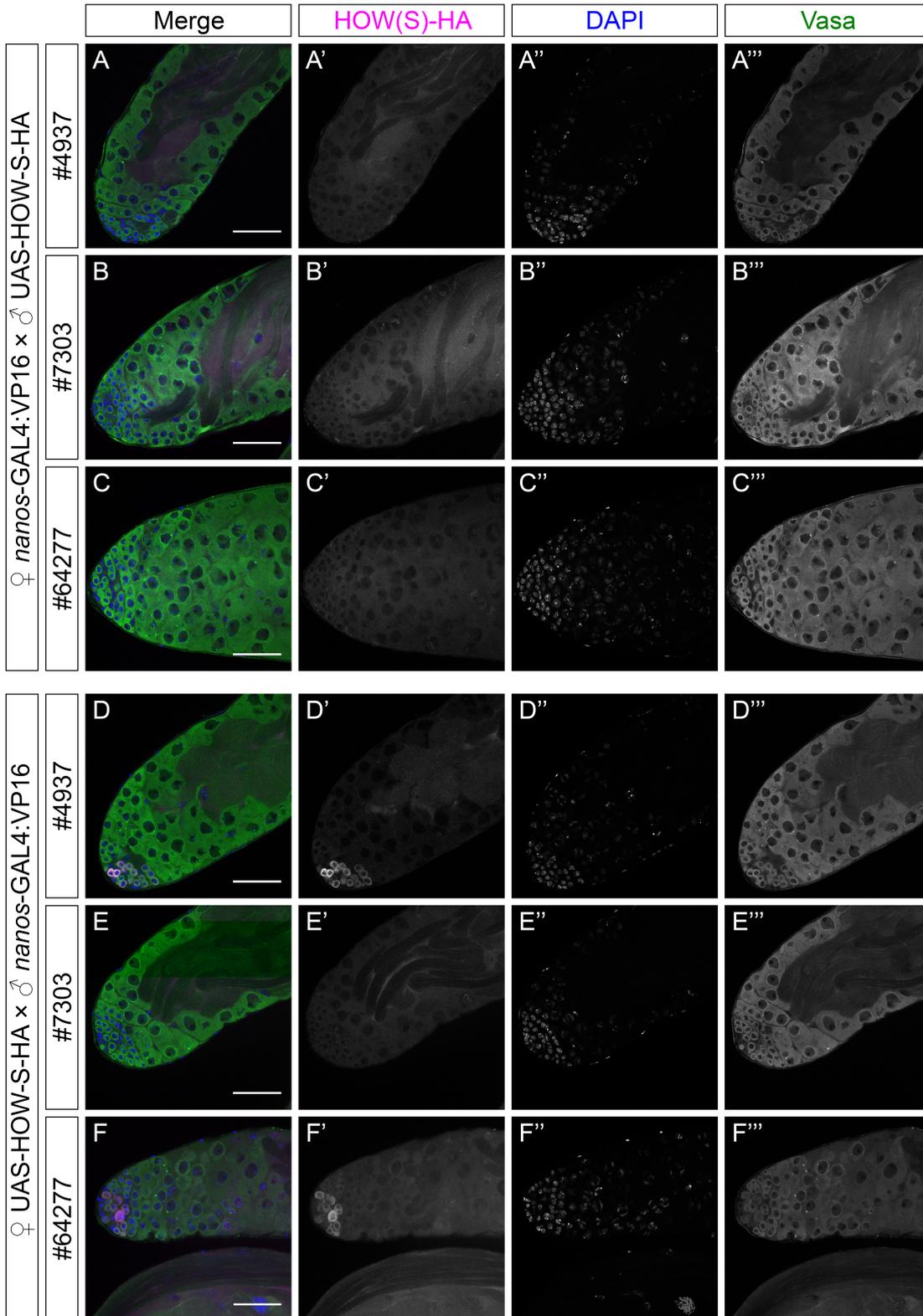
To ensure the entire transgene was precisely as expected, the whole CDS from the *HOW(S)*-HA transgene needed to be sequenced. Therefore, additional PCRs were carried out on the UAS-*HOW-S-HA* gDNA, which could then be sequenced. The UAS forward and *HOW* reverse primers were used for the start of the CDS, the *HOW* forward and HA reverse for the middle, and the HA forward and SV40 reverse for the 3' end (Fig 3.1). The resulting sequencing of the genomic sequence (Sequence A.1, Appendix I) indicated there were a few minor differences to the FlyBase sequence for *HOW(S)* (*HOW-RB*; FBtr0084178). Importantly, the translated sequence (Sequence A.2, Appendix I) does not contain the nuclear localisation signal found in the *HOW(L)* protein isoform, while the final 6 amino acids are **GGLFAR**, which are unique to the *HOW(S)* isoform (*HOW-PB*; FBpp0083576).

With the UAS-*HOW-S-HA* line established, an appropriate *nanos*-*GAL4* driver was needed to drive the expression of *HOW(S)*-HA in the biologically relevant stages of spermatogenesis. Three different *nanos*-*GAL4:VP16* lines (BDSC: #4937, #7303 and #64277) were tested. Each of these lines were crossed with the UAS-*HOW-S-HA* line in both directions (Fig 3.3). 5–10 testes from the progeny of each cross were dissected and immunostained with anti-Vasa (a germ cell marker) and anti-HA to determine which *nanos*-*GAL4* line was the most effective. Confocal microscopy images of these testes revealed that both #4937 and #64277 lines were able to drive expression of *HOW(S)*-HA in the spermatogonia but only with in the maternal crosses, i.e. when the UAS-*HOW-S-HA* flies were the females in the  $F_0$  cross (Figs 3.3 and 3.4). Line #7303 was unable to drive *HOW* expression (Fig 3.4). The importance of the direction of the cross and that line #7303 has the *GAL4* driver on its X chromosome (see Table 2.1.2 for genotype), strongly suggests that the HA-tagged *HOW(S)* transgene is on the X chromosome. One could confirm this by crossing the UAS-*HOW-S-HA* flies with white-eyed flies and assessing the eye colours of the progeny.



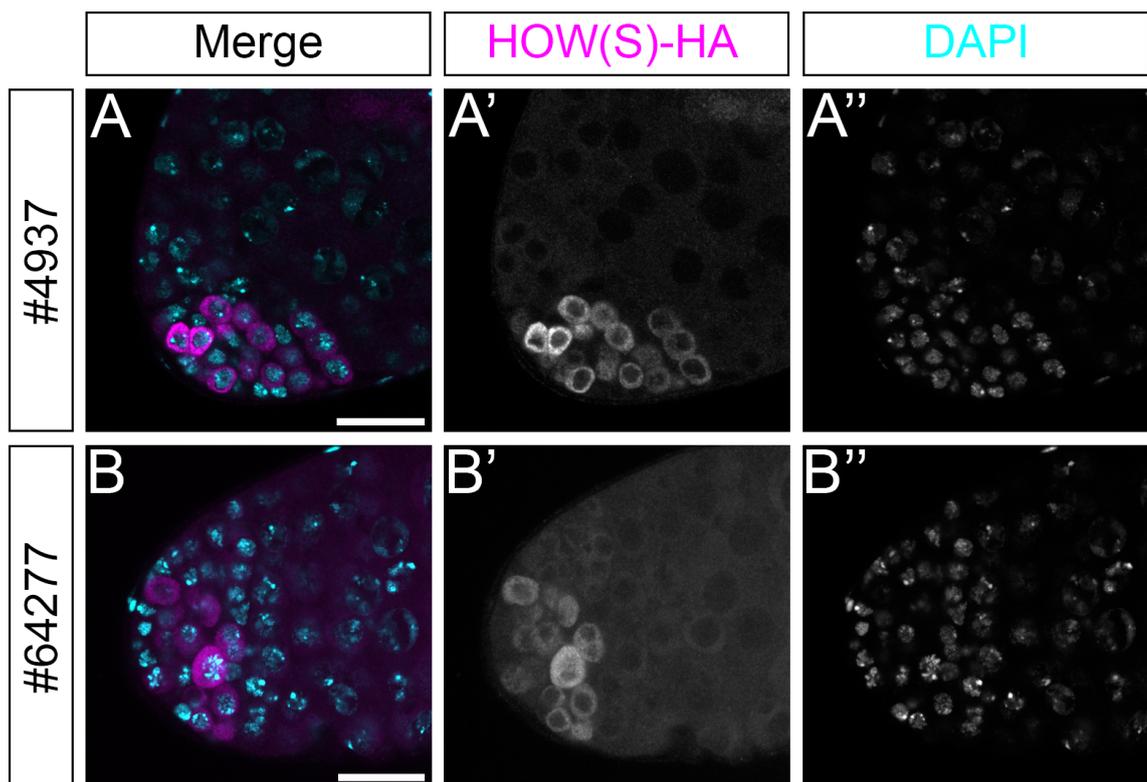
**Figure 3.3. Schematic of crosses performed to express HOW(S)-HA in germ cells.** The UAS-HOW-S-HA line was crossed to three different *nanos*-GAL4 lines to test which was best for driving expression in the GSCs and spermatogonia in the F<sub>1</sub> progeny. The crosses were performed in both directions, when the UAS-HOW-S-HA line was the female in the cross (maternal cross, left) and when the UAS-HOW-S-HA line was the male in the cross (paternal cross, right).

3. Identifying the targets of *HOW(S)* in *Drosophila melanogaster* testis



**Figure 3.4.** *nanos-GAL4:VP16-64277* and *-4937* can drive *HOW(S)-HA* expression. Confocal microscopy images of testes stained with anti-HA (magenta) and anti-Vasa (green) and counter-stained with DAPI (blue). The testes were dissected from progeny of crosses between UAS-HOW-S-HA flies and three different *nanos-GAL4:VP16* lines (#4937 — A, D; #7303 — B, E; #64277 — C, F). The direction of the cross is indicated by the far-left panels. Bar = 50  $\mu$ m.

From the confocal images of the testes with positive HA staining, it is possible to see that HOW(S)-HA is predominantly cytoplasmic because the anti-HA staining, for the most part, does not overlap with the DAPI staining (Fig 3.5). This is consistent with HOW(S) being localised to the cytoplasm (Monk et al., 2010; Rodrigues et al., 2012). From the confocal images it could be determined if *nanos*-driven expression of HOW(S)-HA resembled endogenous HOW expression in germ cells. HOW is expressed from GSCs up to 2-cell spermatogonia, with reduced expression in 4-cell spermatogonia (Monk et al., 2010). The *nanos*-GAL4 line is described and demonstrated as driving expression in GSCs and all spermatogonia (Kiger et al., 2001; White-Cooper, 2012). Here, we only see HOW(S)-HA expression up to the 4-cell germ cysts (Fig 3.5A) with the #4937 line and 2-cell germ cysts with the #64277 line (Fig 3.5B), when the largest spermatogonial cysts typically have 16 cells after the customary 4 rounds of mitosis. This suggests that the HOW(S)-HA protein might be less stable than other proteins and is being rapidly degraded after expression in the GSCs and early spermatogonia. Moving forward the #64277 line was used and is hereby referred to as the *nanos*-GAL4 line.

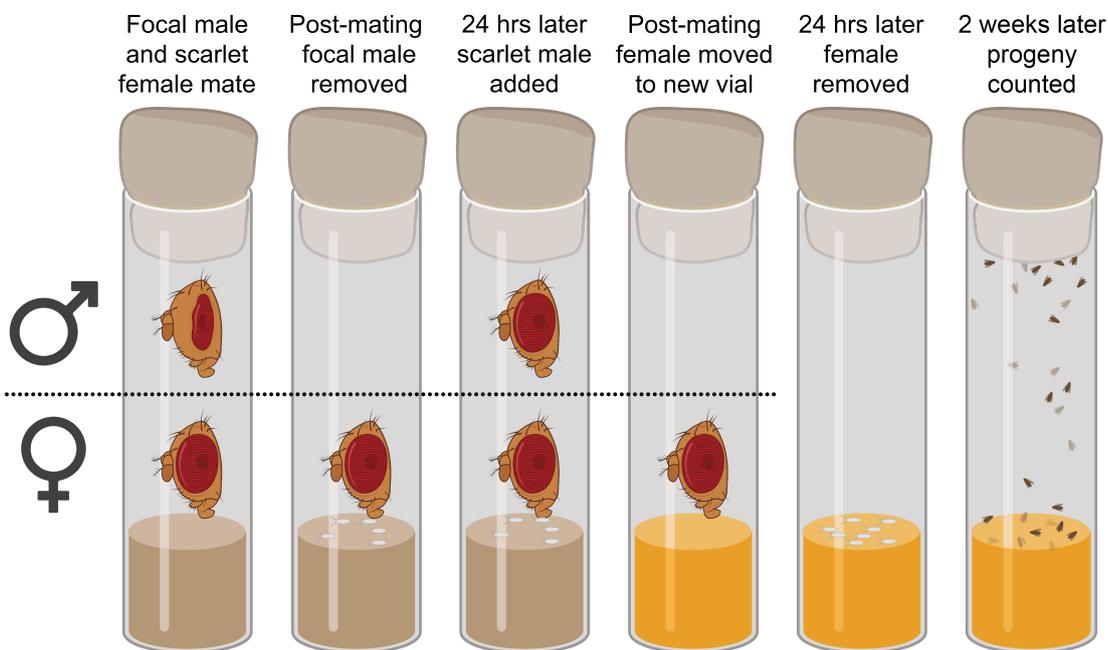


**Figure 3.5. HOW(S)-HA is predominantly cytoplasmic.** Enlarged confocal microscopy images from Figure 3.4. Testes stained with anti-HA (magenta) and anti-Vasa (not shown) and counter-stained with DAPI (blue). The testes were dissected from progeny of crosses between UAS-HOW-S-HA flies and two different *nanos*-GAL4:VP16 lines: A) #4937, B) #64277. Bar = 25  $\mu$ m.

### 3. Identifying the targets of *HOW(S)* in *Drosophila melanogaster testis*

#### 3.2.2. Assessing the impact of *HOW(S)*-HA overexpression on male fertility in *D. melanogaster*

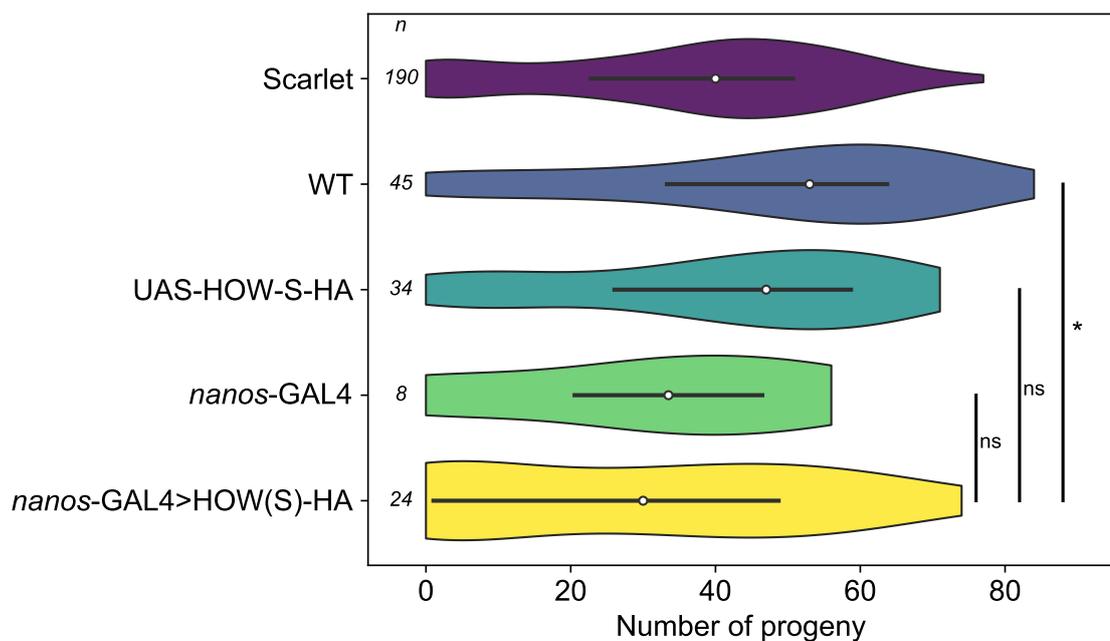
To determine if there is any effect on male fertility as a result of over-expressing *HOW(S)*-HA in the germ cells, sperm competition assays were performed. In *D. melanogaster*, as in many insects, when a female mates twice there will be more progeny from the second mating than the first (Parker, 1970). Sperm competition assays can be designed to compare the ‘offence’ and ‘defence’ ability of flies with different genotypes. The sperm competition assay carried out here (Fig 3.6) involved the mating of a scarlet-eyed female (homozygous *st/st* flies) with two males 24 hours apart, this scarlet eye colour allows us to track the offspring through the experiment. The males from the four genotypes of interest, also referred to as ‘focal males’, are tested in both position 1 (P1), where they are the first male to mate with the female and the second is a scarlet-eyed male, and vice versa (position 2; P2). The female is then left to lay embryos for a further 24 hours (Fig 3.6). Based on the eye phenotype of the progeny (either scarlet or matching the focal male’s eye phenotype) the sire can be attributed to either of the males. This sperm competition assay allows us to observe and analyse three different phenotypes: 1) the number of progeny after the first mating, 2) the mating rates of focal males, and 3) the sperm competition rates.



**Figure 3.6. Schematic of sperm competition experimental design.** A single scarlet-eyed female is mated first with a focal male with bar eyes or WT eyes. After mating the first male is removed. The next day a scarlet-eyed male is added to the vial, after mating the female is moved into a new vial and allowed to lay embryos for 24 hours before the female is removed. 2 weeks later the progeny are counted and phenotyped based on eye colour and/or shape.

To examine the reproductive success of the HOW(S)-HA expressing males (driven by *nanos*-GAL4), four different genotypes were compared: 1) WT flies, with WT eyes, 2) the UAS-HOW-S-HA parental line (homozygous *UAS-HOW-S-HA/UAS-HOW-S-HA*) with orange eyes that should not express HOW(S)-HA, 3) the *nanos*-GAL4 driver parental line (homozygous *nanos*-GAL4/*nanos*-GAL4) with bar eyes that should not express HOW(S)-HA, and 4) the HOW(S)-HA overexpressing flies (*UAS-HOW-S-HA/nanos*-GAL4) with bar eyes.

Firstly, the number of progeny after the first mating (i.e. those from the first vial in Fig 3.6) were counted for the four focal male genotypes and the scarlet-eyed males. The median number of progeny sired by the HOW(S)-HA overexpressing males (30) was fewer than the other genotypes but only significantly so when compared to WT (post-hoc Dunn-Bonferroni test:  $p = 0.018$ ) and not to either of the parental lines (Fig 3.7; see Table A.1 in Appendix I for test statistics). Thus, male fertility does not seem to be impaired by overexpression of HOW(S)-HA any moreso than its parental lines.



**Figure 3.7. HOW(S)-HA expressing flies do not produce fewer progeny than the parental fly lines.** Violin plot of the number of progeny sired by the five male genotypes after the first mating in the sperm competition assay. White circles represent the median, and the grey lines represent the interquartile ranges. The  $n$  number is on the left-hand side of each of the distributions, and is the number of individual males that were mated in total after the three experimental batches were performed. The overexpressing HOW(S)-HA flies (dark teal) had the lowest mean and was significantly different to the WT and UAS-HOW-S-HA flies. Kruskal–Wallis one-way analysis of variance test had a  $p$ -value of 0.004. Post-hoc Dunn-Bonferroni pairwise comparisons found *nanos*-GAL4>HOW(S)-HA was only significantly different to the WT results.  $p < 0.05$  (\*). For the genotypes of the transgenic flies see Table 2.1, for test statistics see Table A.1.

### 3. Identifying the targets of *HOW(S)* in *Drosophila melanogaster testis*

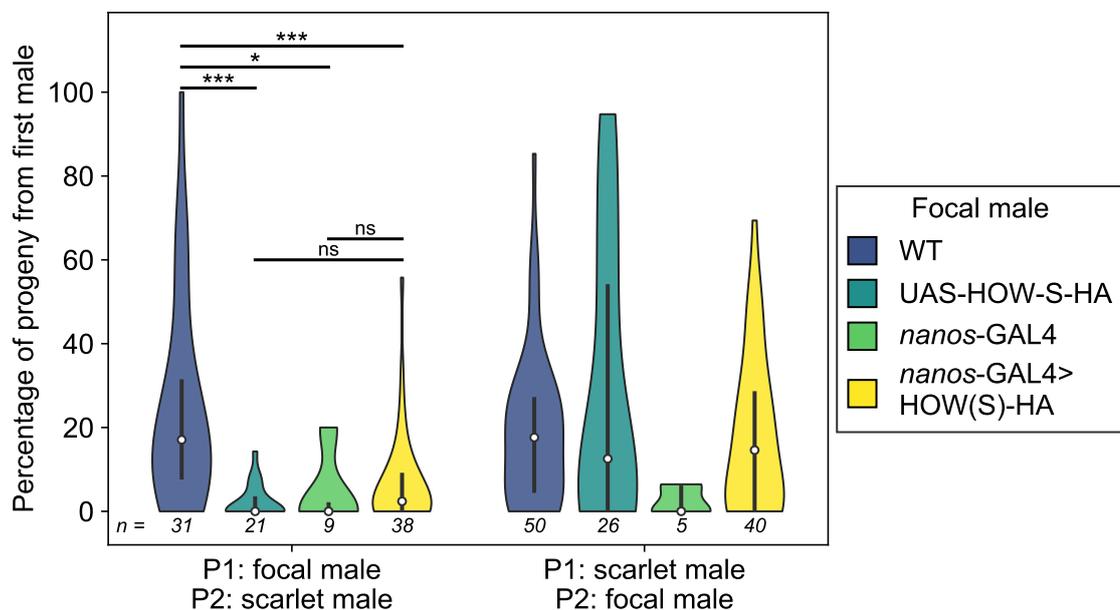
The second assessment from the sperm competition assay is whether overexpression of *HOW(S)* affects the mating rates of the focal males. The P2 mating rates for all focal males are much lower than the P1 rates, as would be expected (Table 3.1). However, the transgenic fly lines perform worse than WT in P1 and especially in P2. The *nanos-GAL4* line has the lowest mating rates of any of the lines, and is particularly low in P2. This poor mating rate results in low *n* numbers for the remainder of the assay. The mating rates of the males overexpressing *HOW(S)* was not particularly different to either of its parental lines (Table 3.1).

**Table 3.1. Mating rates of focal males in sperm competition assay.** The percentage of males that mated with a female when the focal male was in either P1 or P2, with the number of pairings below in brackets. All P1 pairings had a much higher mating rate than P2 pairings.

Position	Focal male	Mating rate (Starting number of pairs)
P1	WT	91.4% (232)
	UAS-HOW-S-HA	88.8% (143)
	<i>nanos-GAL4</i>	71.9% (89)
	<i>nanos-GAL4</i> > <i>HOW(S)</i> -HA	88.3% (273)
P2	WT	48.3% (140)
	UAS-HOW-S-HA	25.7% (117)
	<i>nanos-GAL4</i>	5.9% (85)
	<i>nanos-GAL4</i> > <i>HOW(S)</i> -HA	19.2% (273)

In this assay, P1 measures the focal male's 'defence' ability and P2 the 'offence' ability. In P1 the percentage of offspring the three transgenic focal males sired was significantly less than that of WT, i.e. a weaker sperm defence ability (Fig 3.8; see Table A.2 in Appendix I for test statistics). Both of the parental lines perform very poorly; the median for both the parental lines is 0%, as opposed to the WT median of 17.1%. The HOW(S)-HA flies have a higher median (2.4%) though this is not significantly different from either of its parental lines.

In P2 none of the focal males were significantly different from one another (Fig 3.8). Additionally, the two parental lines have very different medians, though the accuracy of the *nanos*-GAL4 median is limited because low *n* number for these males. Thus, it is difficult to draw any conclusions about the sperm offence abilities of the HOW(S)-HA males.

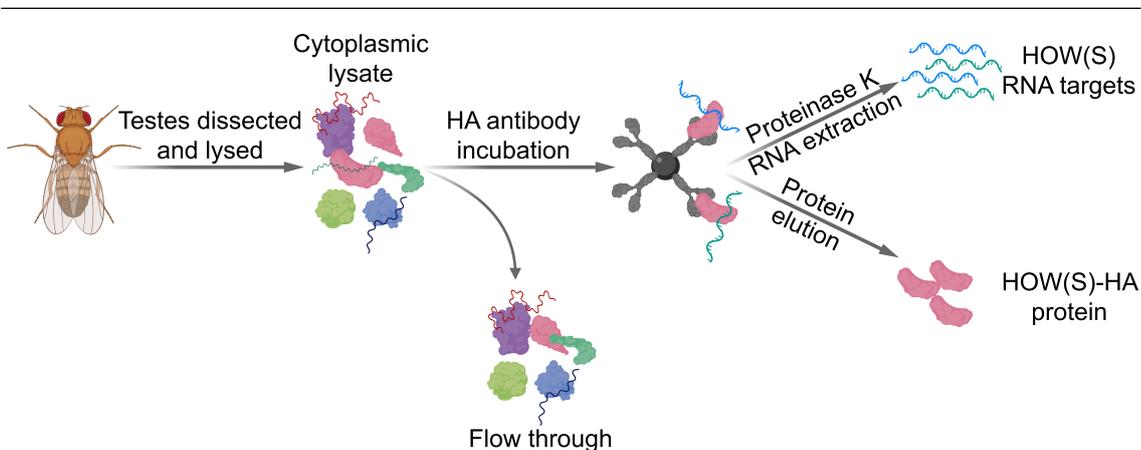


**Figure 3.8. HOW(S)-HA overexpression flies and its two parental lines perform poorly compared to WT in sperm competition experiments.** Violin plot of the percentage of progeny sired from the first male that the female mated. White circles represent the median, and the grey lines represent the interquartile ranges. Kruskal–Wallis one-way analysis of variance test had a  $p$ -value  $< 0.001$ . Post-hoc Dunn-Bonferroni pairwise comparisons showed for P1 that the only significant differences between the categories were with wild type and each of the other three focal male genotypes. There were no significant differences between any of the different focal males for P2.  $p < 0.05$  (\*),  $p < 0.001$  (\*\*\*)). For the genotypes of the transgenic flies see Table 2.1, for test statistics see Table A.2.

### 3. Identifying the targets of HOW(S) in *Drosophila melanogaster* testis

#### 3.2.3. Enrichment and purification of RNAs bound to HOW(S) in the testes

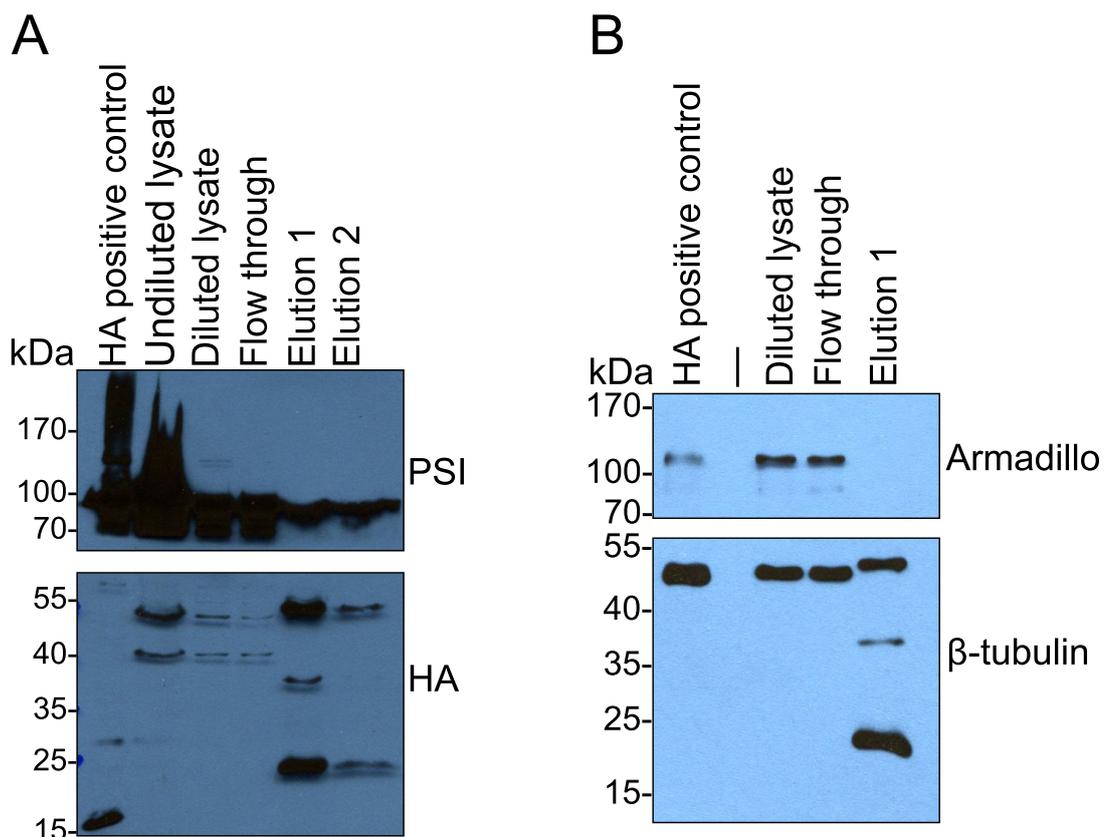
Having successfully shown the expression of HOW(S)-HA in the cytoplasm of GSCs and early spermatogonia, the next step was to determine whether this HA-tagged HOW(S) could be pulled down. The RIP-seq approach involves extracting the cytoplasmic lysate from the tissue of interest, in this case the *D. melanogaster* testis, and incubating it with anti-HA magnetic beads. This allows the HOW(S) HA-tagged protein to be isolated from the rest of the lysate. From a portion of these beads protein is eluted to confirm successful pull-down via western blotting, and from the remaining beads the RNA is extracted for sequencing, making it possible to identify the RNAs bound to HOW(S)-HA *in vivo* (Figure 3.9).



**Figure 3.9. Schematic of RIP-seq protocol.** Testes are dissected and snap frozen. They are then lysed in polysome lysis buffer and the cytoplasmic lysate is incubated with anti-HA magnetic beads. Following incubation some of the beads are boiled in protein sample buffer to elute the protein. The remaining beads undergo proteinase K elution followed by RNA extraction before sequencing to identify HOW(S) targets. For further details see section 2.6. RIP-seq in Materials and Methods.

To determine the ability to pull-down HOW(S)-HA, a small-scale pull-down experiment was performed. The expression of HA-tagged HOW(S) was driven by *nanos*-GAL4 and 250 pairs of testes from these flies were dissected. The cytoplasmic lysate from these testes was incubated with the anti-HA beads. After incubation the flow through (i.e. the unbound lysate) was removed and the beads washed, protein bound to the anti-HA beads was eluted (see Materials and Methods 2.6). Protein samples were collected from the lysate, flow through and wash steps, which were then assessed alongside the eluted protein by western blot (Fig 3.10). HA antibody was used to detect HOW(S)-HA and an antibody against P-element somatic inhibitor (PSI), another RBP essential for male fertility was used to determine the level of non-specific pull-down. The predicted molecular weight of HOW(S)-HA (Sequence A.2) is ~49 kDa. Protein bands of ~49 kDa were detected in all the samples from the pull-down, and enriched in elution 1 (Fig 3.10A). The presence of two smaller bands in the elution 1 lane suggested that there may have been some

degradation. However, after probing the samples with anti- $\beta$ -tubulin (Fig 3.10B) the same three-banding pattern appeared in the elution 1. The size of two of the bands in the elution match the size of heavy and light chains of mouse IgG (50 and 23.5 kDa, respectively). This suggested that the anti-mouse secondary antibody used for anti-HA (and anti- $\beta$ -tubulin) was not only detecting the primary HA antibody used for western blotting but also the mouse HA antibody from the magnetic beads in the pull-down. Thus, in order to verify a pull-down an HA antibody raised in an animal besides mouse is required.



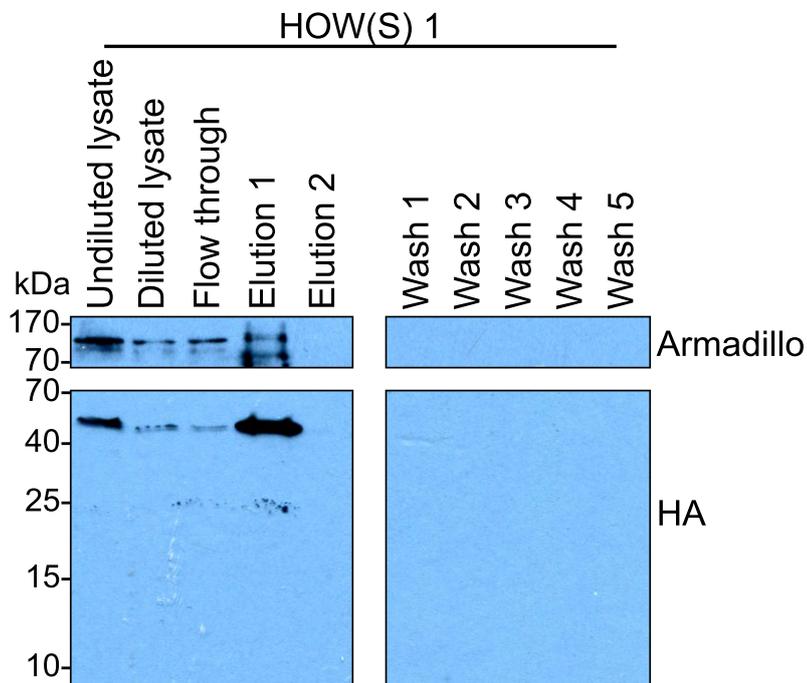
**Figure 3.10. Mouse anti-HA cannot verify the success of the HOW(S)-HA pull-down in western blots.** A) Western blot of a HOW(S)-HA pull experiment, which used 250 pairs of testes, with anti-PSI (1:5000) and anti-HA (1:1000). Elution 1 and 2 suggest that HOW(S)-HA and PSI are both pulled down. B) Western blot of the same samples from A but with anti-Armadillo (1:1000) and anti- $\beta$ -tubulin (1:5000). The anti- $\beta$ -tubulin for the elution 1 sample matches the same pattern as the banding for anti-HA in A. No Armadillo is present in the elution sample. All secondary antibody dilutions were 1:5000, see Table 2.9 for all antibody stock concentrations. Em dash (—) in B) indicates an empty lane.

While the success of the pull-down could not directly be confirmed with the mouse HA antibody, the other proteins probed does give some indication that the pull-down is selective. Most notably, Armadillo, a protein present in the somatic cells of the testis, is not present in the elution sample but is detected in the lysate and flow through samples,

### 3. Identifying the targets of HOW(S) in *Drosophila melanogaster* testis

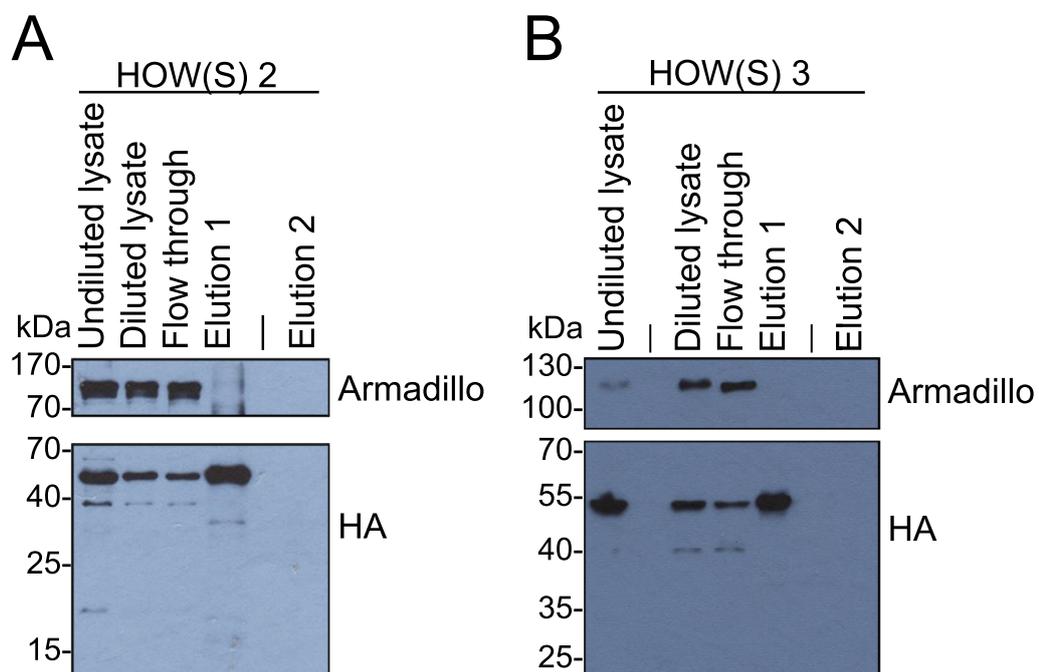
which suggests that there is isolation of specific proteins occurring in the pull-down (Fig 3.10B). PSI was co-precipitated in the pull-down elution (Fig 3.10B), given that this is also an RBP it seems plausible that it is being indirectly pulled-down via RNA that binds both HOW(S)-HA and PSI.

A second pull-down experiment was carried out to assess the ability to pull-down HOW(S)-HA, as before, but also to quantify the amount of RNA that can be extracted from a pull-down to determine how much material is required sequencing. Thus, a larger-scale experiment was carried out with 1000 pairs of dissected testes from *nanos*-GAL4>HOW(S)-HA flies. The cytoplasmic lysate from these testes was incubated with the anti-HA beads. A quarter of these beads were used for protein analysis using western blot, and three quarters were used for RNA extraction. Western blot analysis with a rabbit anti-HA antibody validated the pull-down of HOW(S)-HA protein in the ‘Elution 1’ sample (Fig 3.11), with no detection of the smaller molecular weight proteins seen with the mouse anti-HA (Fig 3.10A). This confirmed that HOW(S)-HA protein was not degrading during the pull-down protocol. The western blots also confirmed that no protein was being lost in the wash steps (Fig 3.11) and very little HOW(S)-HA left in the flow through. The corresponding RNA extractions from the lysate and elution samples yielded 7.3 µg and 190 ng of RNA, respectively. This was a sufficient quantity of RNA for next generation sequencing, and thus these samples formed the first replicate for the HOW(S)-HA RIP-seq.



**Figure 3.11. HOW(S)-HA pull-down confirmed with rabbit anti-HA.** Western blot of a HOW(S)-HA pull experiment from lysate, flow through and elution samples, as well as from the 5 wash steps carried out after the flow through is collected and prior to the protein elution. Samples were probed with mouse anti-Armadillo (1:1000) and rabbit anti-HA (1:5000). All secondary antibody dilutions were 1:1000, see Table 2.9 for all antibody stock concentrations. HOW(S)-HA is entirely eluted in the first round of elution (‘Elution 1’) and is not present in any of the wash steps.

Having confirmed the success of the pull-down of HOW(S)-HA via western blot, and that the quantity of starting material is appropriate for the downstream RNA sequencing, two further repeats of the HOW(S)-HA pull-down were performed with 1000 pairs of testes dissected for each replicate. Again, a quarter of the beads were used for protein elutions and three quarters for RNA extraction. Western blot analysis of these large-scale pull-downs validated the pull-down of HOW(S)-HA protein in the ‘Elution 1’ samples from both repeats (Fig 3.12). Additionally, no somatic Armadillo protein was present in the elution samples. The corresponding lysate RNA and elution RNA samples, henceforth referred to as ‘total RNA’ and ‘pull-down RNA’, formed the HOW(S) 2 and HOW(S) 3 replicates for the HOW(S)-HA RIP-seq.



**Figure 3.12. Successful enrichment of HOW(S)-HA in pull-down replicates 2 and 3.** Western blots of repeats 2 (A) and 3 (B) of HOW(S)-HA pull-downs. Each show an enrichment of HOW(S)-HA in the elution 1 lanes with no Armadillo present. Antibody concentrations used: 1:1000 for anti-Armadillo, 1:5000 for rabbit anti-HA, 1:1000 for secondary antibodies in A), 1:5000 for secondary antibodies in B). Em dashes (—) indicate empty lanes.

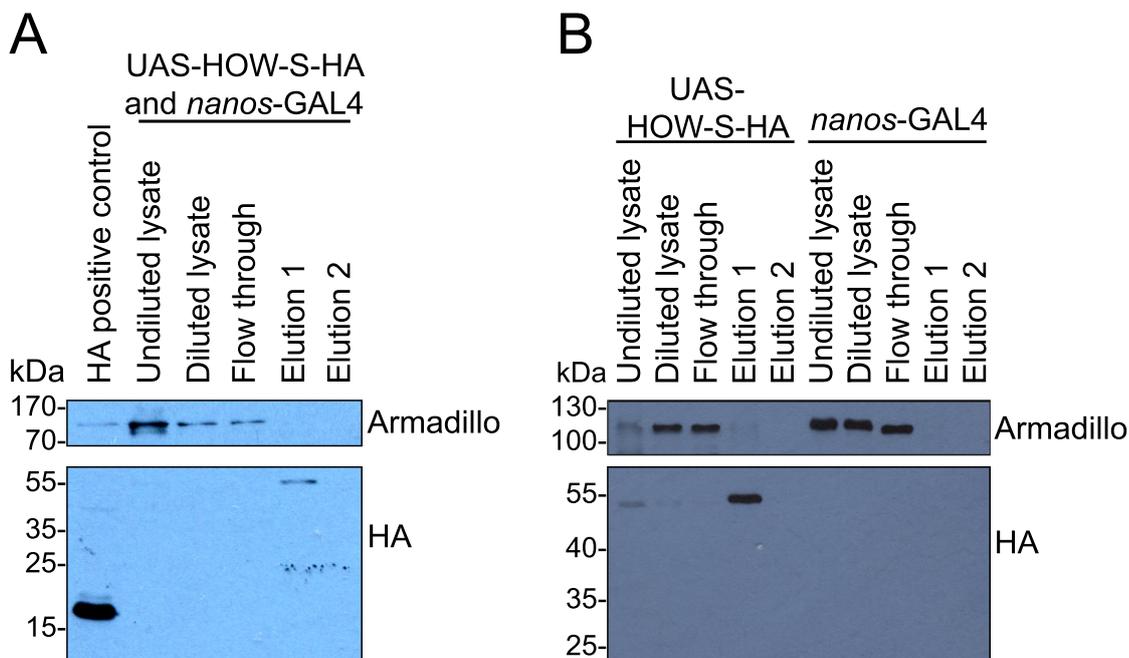
RIP-seq experiments can pick up non-specific RNA–protein interactions, such as RNAs that bind to the HA antibody. This makes developing an appropriate negative control crucial in improving the signal to noise ratio in the downstream computational analyses. To establish an appropriate negative control testes were dissected from the two parental lines, UAS-HOW-S-HA and *nanos*-GAL4. 125 pairs of testes from each of these fly lines were combined into one sample, to generate one background list from the negative control. A small-scale HA pull-down was performed with this combined parental sample, and protein was eluted from all the anti-HA beads, i.e. no RNA was extracted. The western blot analysis of the protein samples from the pull-down revealed that a faint band, the same

### 3. Identifying the targets of *HOW(S)* in *Drosophila melanogaster testis*

size as HOW(S)-HA, was present in elution 1 (Fig 3.13A). But signal was not detected in the lysate or flow through samples, suggesting that there might be low level expression of HOW(S)-HA in the UAS-HOW-S-HA line, or contamination had occurred between samples.

Separate large-scale pull-downs, with 1000 pairs of testes for the two parental lines confirmed that there was indeed leaky expression of HOW(S)-HA in the UAS-HOW-S-HA line but not in the *nanos*-GAL4 flies, which does not have a HOW(S)-HA CDS in its genome (Fig 3.13B). Thus, the *nanos*-GAL4 parental sample is more suitable to use as a negative control. The corresponding total RNA and pull-down RNA samples from the *nanos*-GAL4 only pull-down was used for subsequent sequencing and analysis.

RNA libraries were generated from the ‘total RNA’ and ‘pull-down RNA’ samples of the three HOW(S)-HA pull-down replicates and the *nanos*-GAL4 pull-down. 100 ng from each of these libraries were pooled and then sequenced using 75 bp single-end sequencing (see section 2.6.2. Sequencing for more details).



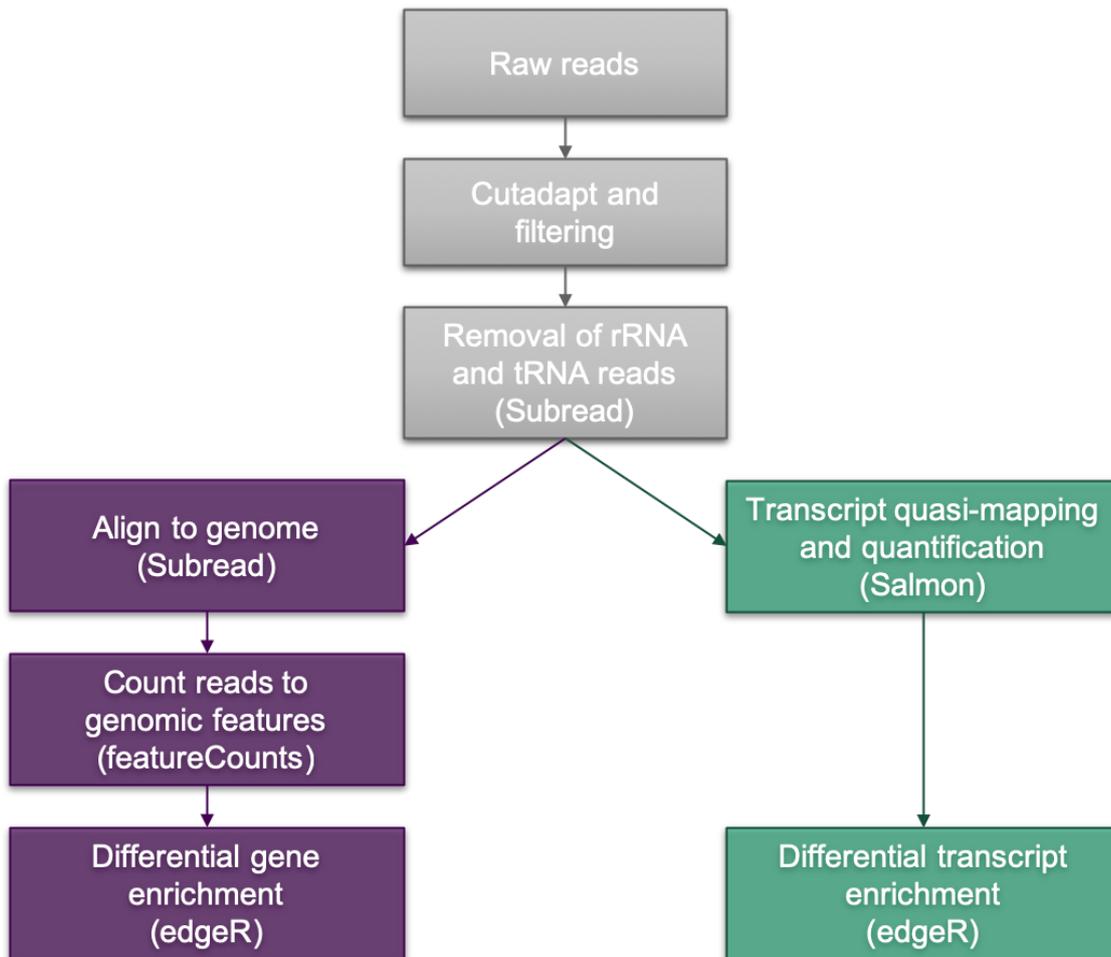
**Figure 3.13. UAS-HOW-S-HA line has leaky expression of HOW(S)-HA.** A) Western blot from an HA pull-down of the HOW(S)-HA parental lines shows that there is expression of HOW(S)-HA protein, as seen in the elution 1 sample. B) HOW(S)-HA is expressed at a low level in UAS-HOW-S-HA line but not in *nanos*-GAL4 line. Antibody concentrations used: 1:1000 for anti-Armadillo, 1:5000 for anti-HA, 1:5000 for secondary antibodies.

### 3.2.4. Genome aligning and transcriptome mapping of RIP-seq data

To identify RNAs that were pulled-down in the HOW(S)-HA RIP-seq, both gene-level and transcript-level differential enrichment analyses were implemented (Fig 3.14). This was done because while gene-level analysis is often more robust, it can hide transcript specific dynamics (Soneson et al., 2015). The adapter sequences were removed from the raw reads, then the reads were filtered based on their sequencing quality. Following this reads that map to tRNA and rRNA sequences were removed using Subread (Material and Methods 2.6.3.1). Here the analysis was split into two streams: gene-level and transcript-level analyses. In the gene-level analysis (purple outline in Fig 3.14) reads were aligned to the genome (*D. melanogaster* release 6.22) using Subread. featureCounts was used to count the reads to different genomic features. The counts for the feature ‘gene’ were used in the different gene enrichment analysis with edgeR (Material and Methods 2.6.3.2). Transcript quantification (green outline in Fig 3.14) was executed in Salmon, which quasi-mapped the reads to the transcriptome and produced count tables for each transcript (Material and Methods 2.6.3.3). These count tables were then used for differential transcript enrichment, also completed with edgeR (Material and Methods 2.6.3.4).

The starting number of raw reads ranged from 40–55 million reads across the 8 samples (Table 3.2). Similar percentages of low-quality reads (10–12%) were filtered out at the cutadapt and quality control filter stage (Table 3.2, ‘Poor quality’ column). All samples also had a very low amount of tRNA reads. The total RNA samples consistently had more rRNA reads than the corresponding pull-down sample, this is indicative that there is a level of specificity in the pull-down RNA compared to total RNA (Table 3.2). The total RNA for HOW(S) 1 had a particularly high amount of rRNA reads, and thus the smallest number of reads in the final alignments and mapping. All the other total RNA samples were within a million reads of each other in genome alignments and transcriptome mapping (see Table A.7 in Appendix I for an equivalent table of read numbers rather than percentage). The four different pull-down RNA samples all had a similar number of final mapped reads (20.4–24.1 million reads; Appendix I Table A.7).

3. Identifying the targets of HOW(S) in *Drosophila melanogaster* testis



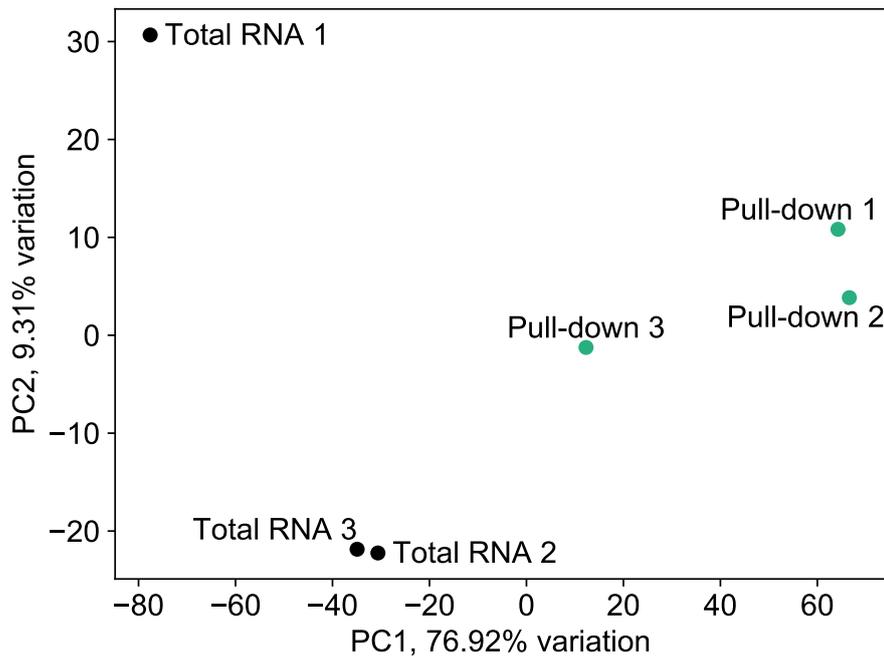
**Figure 3.14. Schematic of RIP-seq differential enrichment analysis pipeline.** Raw reads are processed to remove adapter sequences with cutadapt, and low quality reads are filtered out. rRNA and tRNA reads are removed using Subread. The analysis then splits into two based on gene- and transcript-level analysis. Genome alignment was carried out using Subread, followed by featureCounts for to generate count tables for the ‘gene’ feature. Transcript quasi-mapping and quantification was done using Salmon. Differential enrichment was carried out in edgeR using the count tables for genes and transcripts (see Materials and Methods 2.6.3 for details).

**Table 3.2. Breakdown of read counts through the pipeline by percentage.** Percentage of reads assigned in each step of the processing from the starting number of reads. In the genome columns ‘unaligned’ refers to reads that were not aligned by Subread, ‘aligned, unassigned’ refers to reads that were aligned to the genome but not assigned to the ‘gene’ feature by featureCounts, and ‘aligned, assigned’ refers to reads that were both aligned to the genome and assigned the ‘gene’ feature by featureCounts. The final reads used for different gene and transcriptome analysis are in bold.

Samples		Starting number of reads	Pre-processing			Genome alignment			Transcriptome quasi-mapping	
			Poor quality	rRNA	tRNA	Unaligned	Aligned, unassigned	Aligned, assigned	Unmapped	Mapped
<i>nanos</i> -GAL4	Total RNA	53743743	11.46	39.27	0.0003	11.11	3.57	<b>34.59</b>	12.17	<b>37.10</b>
	Pull-down	46955434	11.26	9.93	0.0018	25.17	6.82	<b>46.83</b>	29.43	<b>49.39</b>
HOW(S) 1	Total RNA	53660667	10.47	66.03	0.0003	6.19	1.59	<b>15.72</b>	6.48	<b>17.02</b>
	Pull-down	40300616	12.59	15.68	0.0003	13.02	8.08	<b>50.63</b>	20.89	<b>50.85</b>
HOW(S) 2	Total RNA	55018237	11.80	42.54	0.0003	6.62	3.65	<b>35.39</b>	7.90	<b>37.75</b>
	Pull-down	43895836	12.28	13.73	0.0018	12.42	8.39	<b>53.18</b>	19.27	<b>54.72</b>
HOW(S) 3	Total RNA	52141872	11.29	42.25	0.0003	6.73	3.67	<b>36.06</b>	8.10	<b>38.36</b>
	Pull-down	48005064	11.57	29.34	0.0014	10.75	5.76	<b>42.58</b>	13.96	<b>45.12</b>

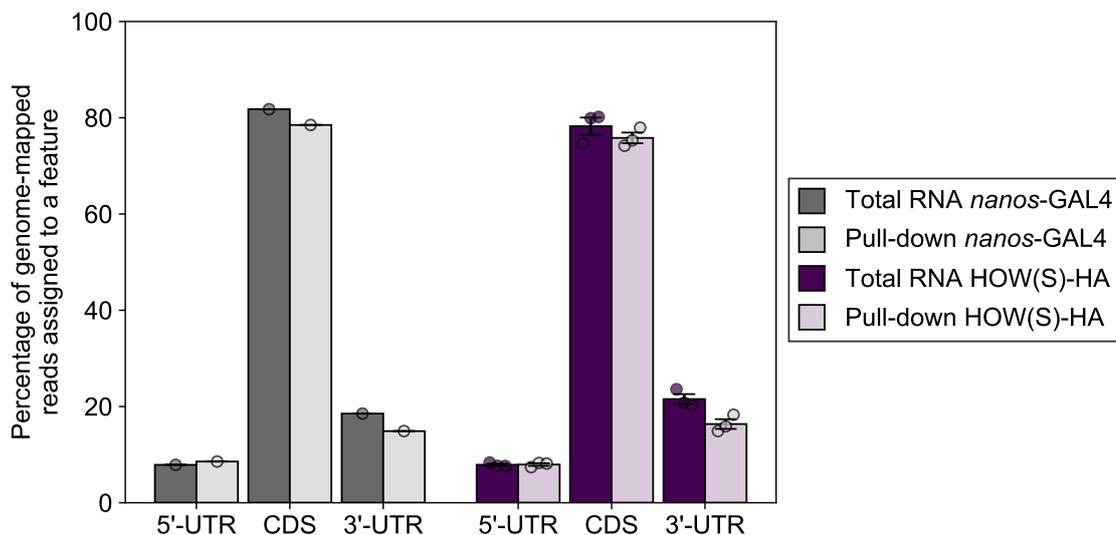
### 3. Identifying the targets of *HOW(S)* in *Drosophila melanogaster testis*

To evaluate the variation between the *HOW(S)*-HA replicates, a principal component analysis (PCA) was carried out using the transcript counts (Materials and Methods 2.6.3.5). The majority of the variation (76.92%) between the samples was accounted for by principal component 1 (PC1). The three *HOW(S)*-HA pull-down RNA samples clustered above 0 along the  $x$ -axis, while the three total RNA samples were below 0 (Fig 3.15). PC2 accounted for just 9.31% of the variation between samples. Again, the three pull-down samples were close together but the total RNA 1 sample was separated from total RNA replicates 2 and 3, along the  $y$ -axis (Fig 3.15).



**Figure 3.15. *HOW(S)*-HA pull-down samples are more similar to each other than the total RNA samples.** Biplot from the PCA using the  $\log_2(\text{Counts Per Million})$  of the transcriptome quasi-mapped reads. Principal component 1 (PC1) on the  $x$ -axis accounts for 76.92% of the variation between the six *HOW(S)*-HA samples, and PC2 on the  $y$ -axis accounts for 9.31% of the variation between samples. For PC1 the pull-down samples (green) all cluster above 0, and the total RNA samples (black) are all below 0. Total RNA 2 and 3 cluster very closely together with the total RNA 1 sample separated from them by PC2.

Unlike CLIP-seq, RIP-seq involves no digestion of the RNA prior to sequencing, so one would expect to pull down full-length RNA irrespective of the location of the RNA:protein interactions. To test if this was the case, featureCounts was used to assign reads to either the 5'-UTR, CDS or 3'-UTR from the genome-aligned reads. There was little difference in the distribution of the reads across these three features between the total RNA and the pull-down RNA samples in both the *nanos*-GAL4 and HOW(S)-HA samples (Fig 3.16). No statistical tests could be performed for the *nanos*-GAL4 samples (grey) as there are no replicates. For the HOW(S)-HA samples (Fig 3.16, purple), the percentage of reads assigned to each feature were compared between the total RNA and pull-down RNA samples. Wilcoxon signed-rank test or paired *t*-tests, depending on the normality of the data, were performed and none of the tests were statistically significant (see Table A.3 in Appendix I for test statistics). Suggesting that regardless of where HOW(S) is binding a transcript the full-length RNA is being pulled-down with little or no degradation of the RNA.



**Figure 3.16. No difference in the percentage of reads mapped to transcript features between total RNA and pull-down RNA samples.** Bar chart showing the percentages of genome-aligned reads assigned to a feature — 5'-UTR, CDS or 3'-UTR. Individual data points are shown with the standard error of the mean (SEM) as error bars. Statistics performed on the HOW(S)-HA data (purple) showed no difference between the total RNA and pull-down RNA samples (Table 3.16). No statistical tests were carried out on the *nanos*-GAL4 data (grey) as there are no replicates.

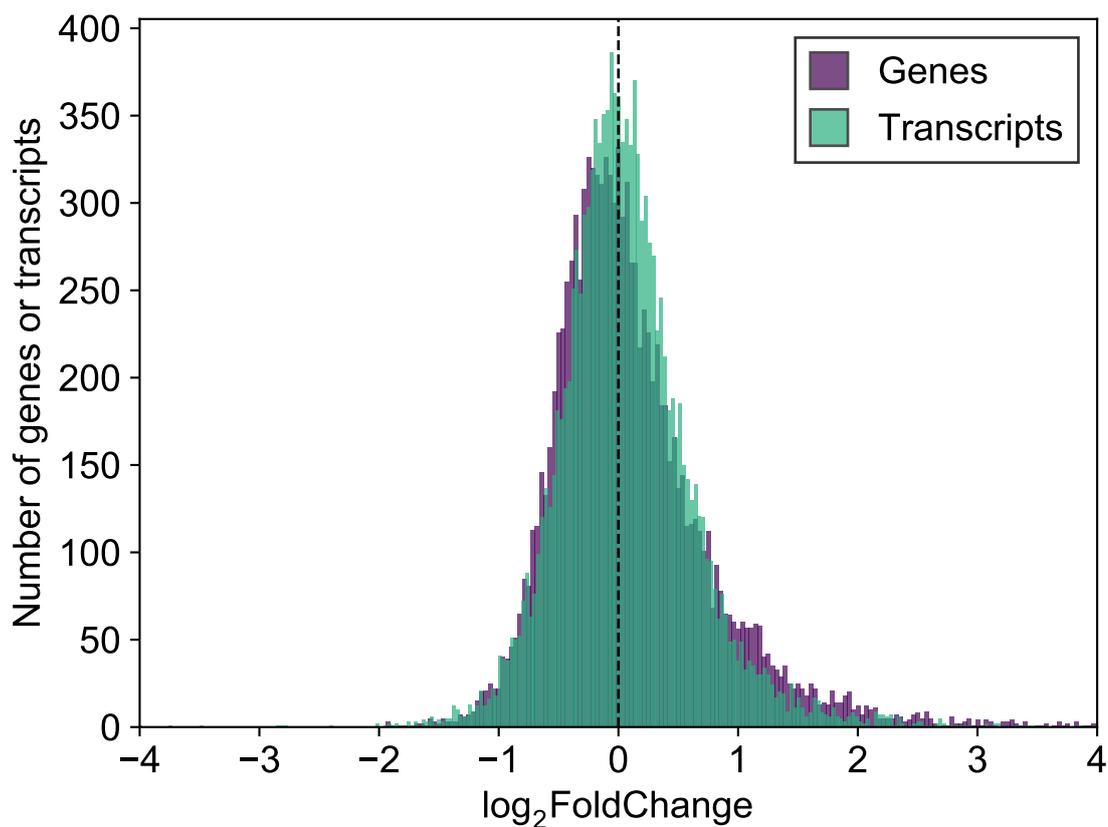
### 3. Identifying the targets of *HOW(S)* in *Drosophila melanogaster* testis

#### 3.2.5. Verifying the applicability of edgeR for RIP-seq differential enrichment analysis

To identify mRNAs that were specifically pulled down with HOW(S)-HA compared to background controls, differential gene and transcript enrichment was required. Several software packages are available to perform this analysis but it was necessary to evaluate suitability for this particular analysis strategy. edgeR was selected for the analysis as it is a leading differential enrichment analysis package that allows for samples with no replicates, which was the case for the *nanos*-GAL4 RIP-seq data (Robinson et al., 2010; McCarthy et al., 2012).

edgeR, as with many differential enrichment software, assumes that most genes or transcripts do not change between the two conditions one tests, which is often the case when comparing two total RNA samples. To determine if this was true when comparing the pull-down RNA to the total RNA, a histogram was plotted of the  $\log_2(\text{Fold Change})$  for the genes and transcripts from the replicated HOW(S)-HA data, where a value above 0 implies the gene or transcript is enriched in the pull-down and below 0 it is depleted (Fig 3.17). Most of the genes and transcripts are centred around 0; 64.7% of the genes and 70.1% of transcripts are within -0.5 and 0.5  $\log_2(\text{Fold Change})$ . The median  $\log_2(\text{Fold Change})$  of the genes is -0.006 and 0.019 for the transcripts. This means that the majority of genes and transcripts are not differentially enriched or depleted when comparing the HOW(S) total RNA and pull-down RNA. As expected for a pull-down, there is a larger number of genes/transcripts which are enriched than depleted, i.e. more genes and transcripts have a  $\log_2(\text{Fold Change})$  above 2 than below -2.

As acknowledged earlier, there were 3 replicates for the HOW(S)-HA pull-down but only one replicate for the *nanos*-GAL4 negative control pull-down, from which a background list of genes and transcripts needed to be generated. One of the statistical methods edgeR can implement to analyse unreplicated samples is the exact test. This exact test requires the biological coefficient of variation (BCV) to be set by the user, and the output of the test, the significance of differential enrichment values, is highly sensitive to the value of the user-defined BCV. The recommendation in the edgeR manual is for the BCV to be selected based on the BCV of similar data or at 0.1 for model organisms.

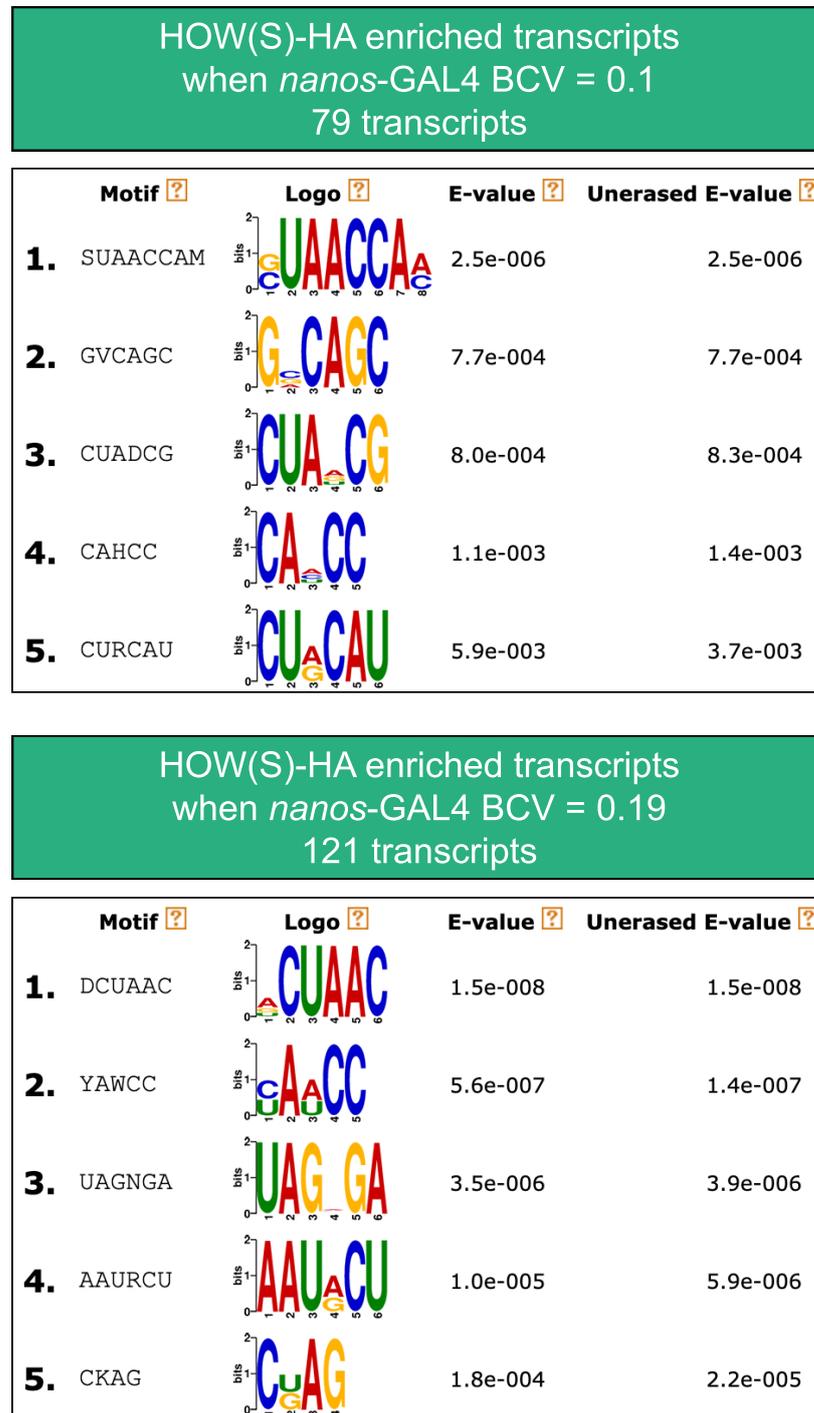


**Figure 3.17. Most genes and transcripts are neither enriched or depleted.** Histogram showing the  $\log_2(\text{Fold Change})$ , pull-down vs total RNA, distribution for genes (purple) and transcripts (green) from the HOW(S)-HA RIP-seq samples. Both samples are centered close to 0. 64.7% of genes and 70.1% of transcripts are within -0.5 and 0.5  $\log_2(\text{Fold Change})$ .

### 3. Identifying the targets of HOW(S) in *Drosophila melanogaster testis*

To establish an appropriate BCV value to be applied to the differential enrichment analysis of the *nanos*-GAL4 samples, two different BCVs were compared using the transcriptome-mapped data. The first was the BCV from the replicated HOW(S)-HA pull-down data, 0.1927707, which edgeR was able to calculate. The second BCV tested was 0.1, suggested by the edgeR manual. With these two different BCVs, two different background lists were generated of transcripts that were enriched in the *nanos*-GAL4 pull-down RNA compared to the *nanos*-GAL4 total RNA.

To compare these two background lists, and thus determine which BCV value is appropriate to use, the background lists were applied to the HOW(S)-HA differentially enriched transcript lists from the HOW(S)-HA data. Then, the downstream motif enrichment analysis, using DREME, was carried out on the 3'-UTRs of the two different HOW(S)-HA transcript lists and the top scoring motifs were compared. When the *nanos*-GAL4 BCV value is set to the same BCV as the HOW(S)-HA samples (0.1927707) the top scoring motif is much closer to the published motif (ACUAA) for HOW than the top scoring motif found when the *nanos*-GAL4 BCV is set to 0.1 (Fig 3.18). This suggested that using the significantly enriched transcripts from the *nanos*-GAL4 pull-down when the BCV is set to the same as the HOW(S)-HA samples is the most appropriate background list. This was then also applied for the gene-level differential enrichment analysis.



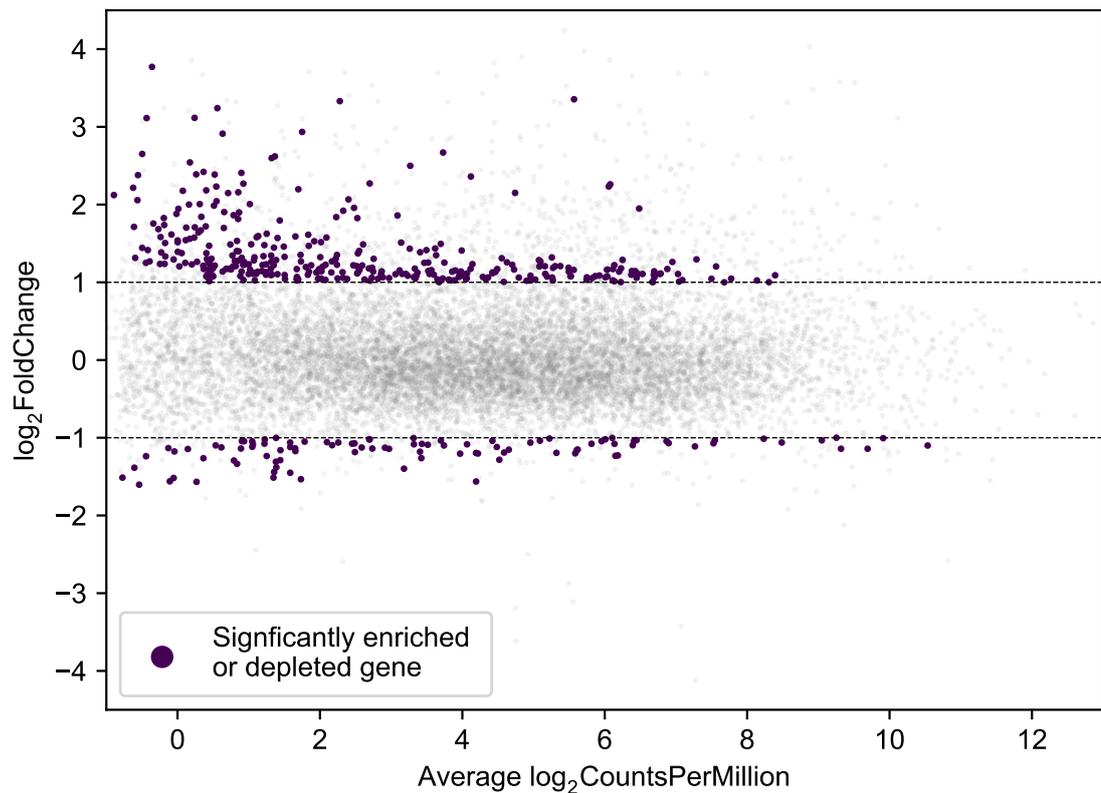
**Figure 3.18.** BCV for the *nanos*-GAL4 samples should be set to HOW(S)-HA samples' BCV. The top 5 enriched motifs, using DREME, from the 3'-UTRs of the HOW(S)-HA significantly differentially enriched transcripts when the negative control sample (*nanos*-GAL4) has a BCV of 0.1 (top) or 0.1927707 (bottom). The top hit for both of these are similar with UAAC present in both. The top hit when BCV = 0.019 is closer to the published HOW response element — ACUAA.

### 3. Identifying the targets of *HOW(S)* in *Drosophila melanogaster testis*

#### 3.2.6. Identifying the differentially enriched genes and transcripts from the *HOW(S)*-HA RIP-seq

Having established that edgeR is an appropriate package for RIP-seq data (Fig 3.17), and found the optimal BCV value to generate an appropriate background list from the *nanos*-GAL4 sample (Fig 3.18), the differential enrichment analysis was performed at both gene- and transcript-level. For the gene-level analysis the read counts assigned to the feature ‘gene’ using featureCounts were used as the input for the edgeR analysis (Materials and Methods 2.6.3.4).

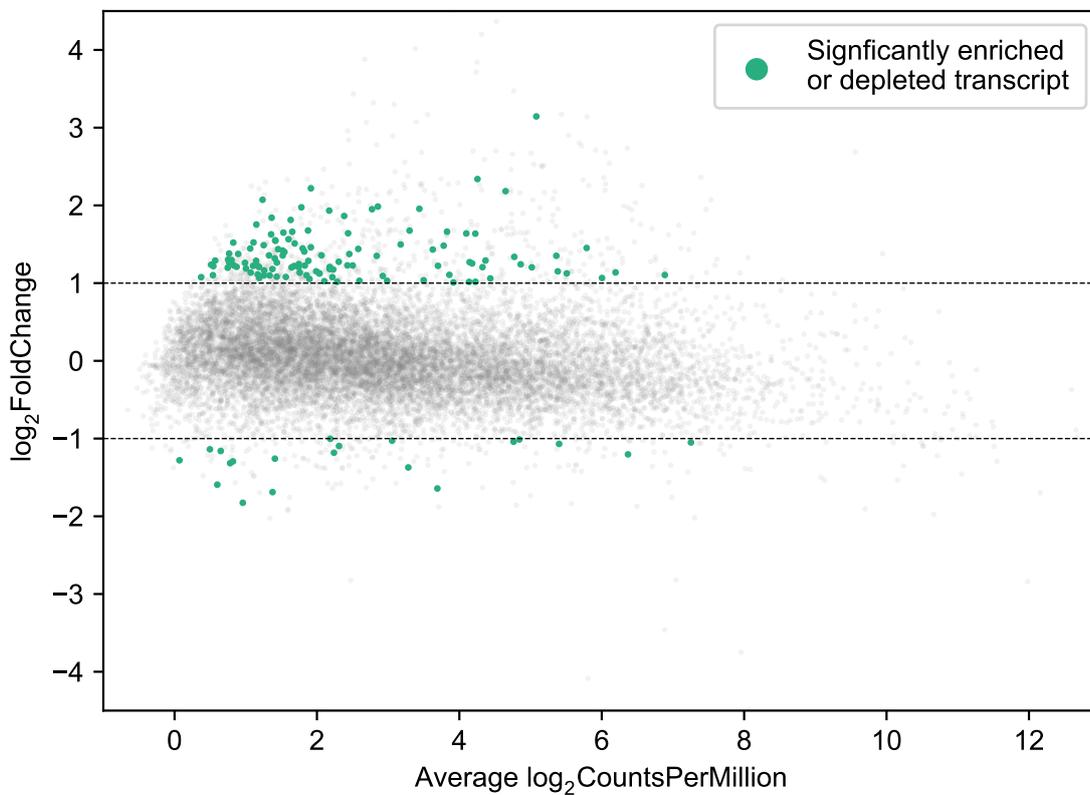
Enriched gene lists were generated for both the *HOW(S)*-HA pull-down and the *nanos*-GAL4 pull-down. A gene was determined to be significantly enriched or depleted if they met two criteria: 1) the gene had a  $\log_2$  fold change greater than 1 or below -1, and 2) it had an FDR corrected  $p$ -value below 0.05. Finally, any gene that was significantly enriched or depleted in both the *HOW(S)*-HA list and the *nanos*-GAL4 list was removed from the *HOW(S)*-HA pull-down list. This final list of genes from the *HOW(S)*-HA pull-down had 343 significantly enriched genes and 99 significantly depleted genes (Fig 3.19, full list of 343 genes in Appendix I Table A.8). From the scatter plot it is also evident that the data are noisy; many of the genes, over 1000 of them, meet the fold change threshold but were either not statistically significant or were also in the background list (Fig 3.19).



**Figure 3.19. 343 genes are enriched in HOW(S)-HA RIP-seq.** Scatter plot from the differential gene analysis of HOW(S)-HA pull-down RNA vs HOW(S)-HA total RNA. Average  $\log_2(\text{Counts Per Million})$  against  $\log_2(\text{Fold Change})$  for each gene is shown. Purple dots indicate genes that have: a  $p$ -value  $< 0.05$  following FDR correction; exceeded the 2-fold cut off; and did not meet these criteria in the negative control *nanos*-GAL4 RIP-seq. 343 genes were significantly enriched and 99 genes were significantly depleted.

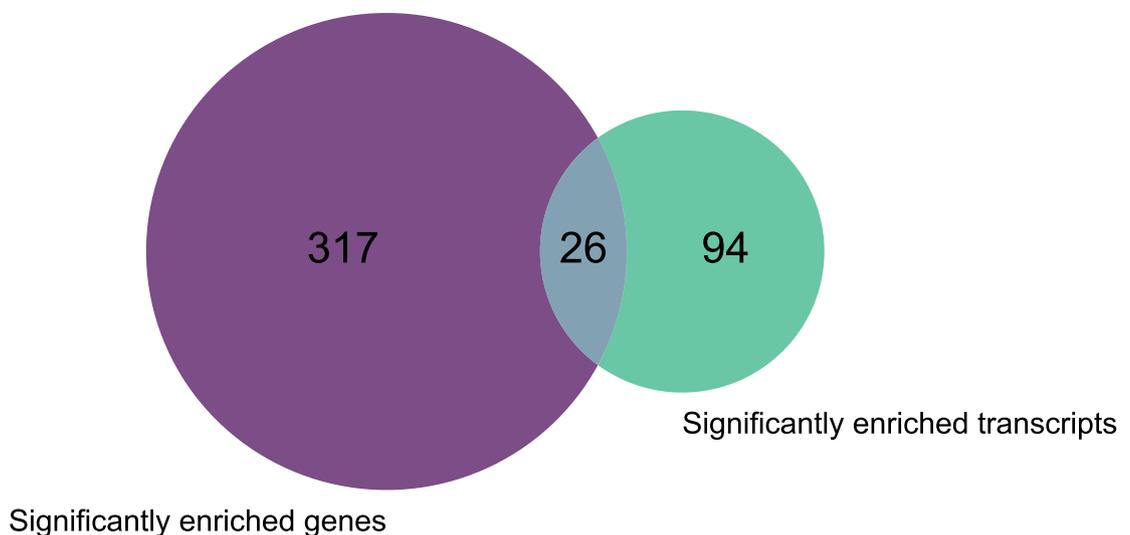
### 3. Identifying the targets of *HOW(S)* in *Drosophila melanogaster testis*

For the transcriptome-mapped reads, the Salmon count tables were imported into edgeR to generate the enriched list of transcripts for the *HOW(S)*-HA and *nanos*-GAL4 data (Materials and Methods 2.6.3.4). The same criteria used for the gene-level lists were used to identify the significantly enriched and depleted transcripts in the *HOW(S)*-HA pull-down. 121 transcripts were significantly enriched in *HOW(S)*-HA pull-down, compared to 20 transcripts that were significantly depleted (Fig 3.20, full list of 121 transcripts in Appendix I Table A.9). Similar to the gene-level analysis (Fig 3.19) the data are noisy; 731 transcripts of the transcripts have a  $\log_2(\text{Fold Change})$  above 1 or below -1 (Fig 3.20) but were either not statistically significant or they were also in the background list generated from the *nanos*-GAL4 differential transcript enrichment.



**Figure 3.20. 121 transcripts are enriched in *HOW(S)*-HA RIP-seq.** Scatter plot from the differential transcript analysis of *HOW(S)*-HA pull-down RNA vs *HOW(S)*-HA total RNA. Average  $\log_2(\text{Counts Per Million})$  against  $\log_2(\text{Fold Change})$  for each transcript is shown. Green dots indicate transcripts that have: a  $p$ -value  $< 0.05$  following FDR correction; exceeded the 2-fold cut off; and did not meet these criteria in the negative control *nanos*-GAL4 RIP-seq. 121 transcripts were significantly enriched and 20 transcripts were significantly depleted.

To compare the significantly enriched targets from the two parallel parts of the analysis (genes and transcripts), the transcript IDs were converted into gene IDs. Only 26 genes were in common between the gene and transcript lists (Fig 3.21), which is 7.58% of the gene list and 21.7% of the transcript list. To explore if this small overlap was due to the different mapping and aligning methods used, the differential gene enrichment analysis was carried out using Salmon to quasi-map reads to the genome rather than Subread. However, there were only 23 genes in common between the Salmon derived significantly enriched genes and transcripts lists. It is likely that the small overlap in the Subread derived significantly enriched genes and the Salmon derived significantly enriched transcripts is due to the noise of the RIP-seq experiment and/or from not having replicates for the *nanos*-GAL4 dataset, as the overlap between the lists increased when the background list was not applied.

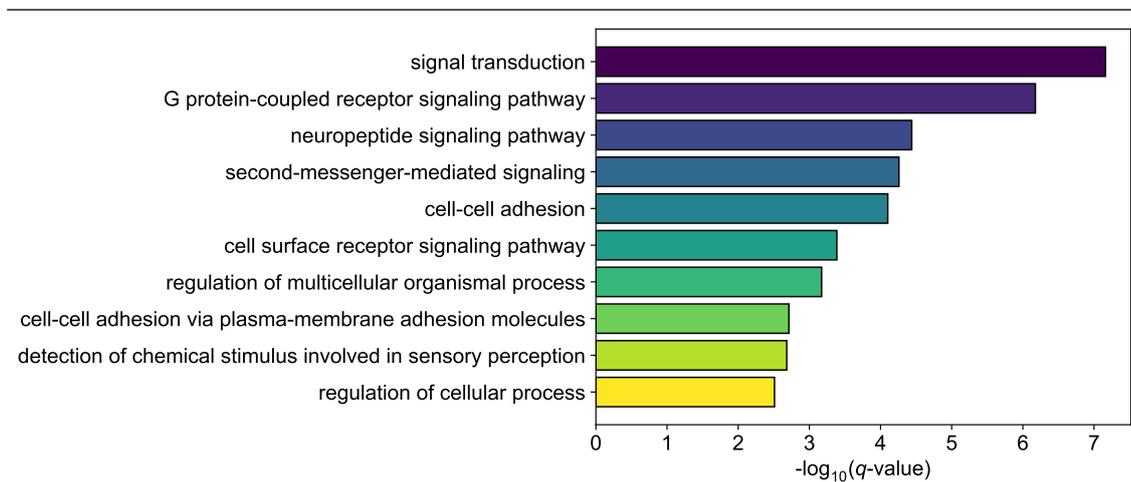


**Figure 3.21. Small overlap in significantly enriched genes and transcripts.** Venn diagram showing the overlap in significantly enriched genes and transcripts in the HOW(S)-HA RIP-seq. Transcripts were converted into genes for this analysis, and duplicate genes within the list were removed. Only 26 genes are in common between the two lists.

### 3. Identifying the targets of HOW(S) in *Drosophila melanogaster testis*

#### 3.2.7. ‘Signal transduction’ genes and transcripts are enriched in the HOW(S)-HA RIP-seq

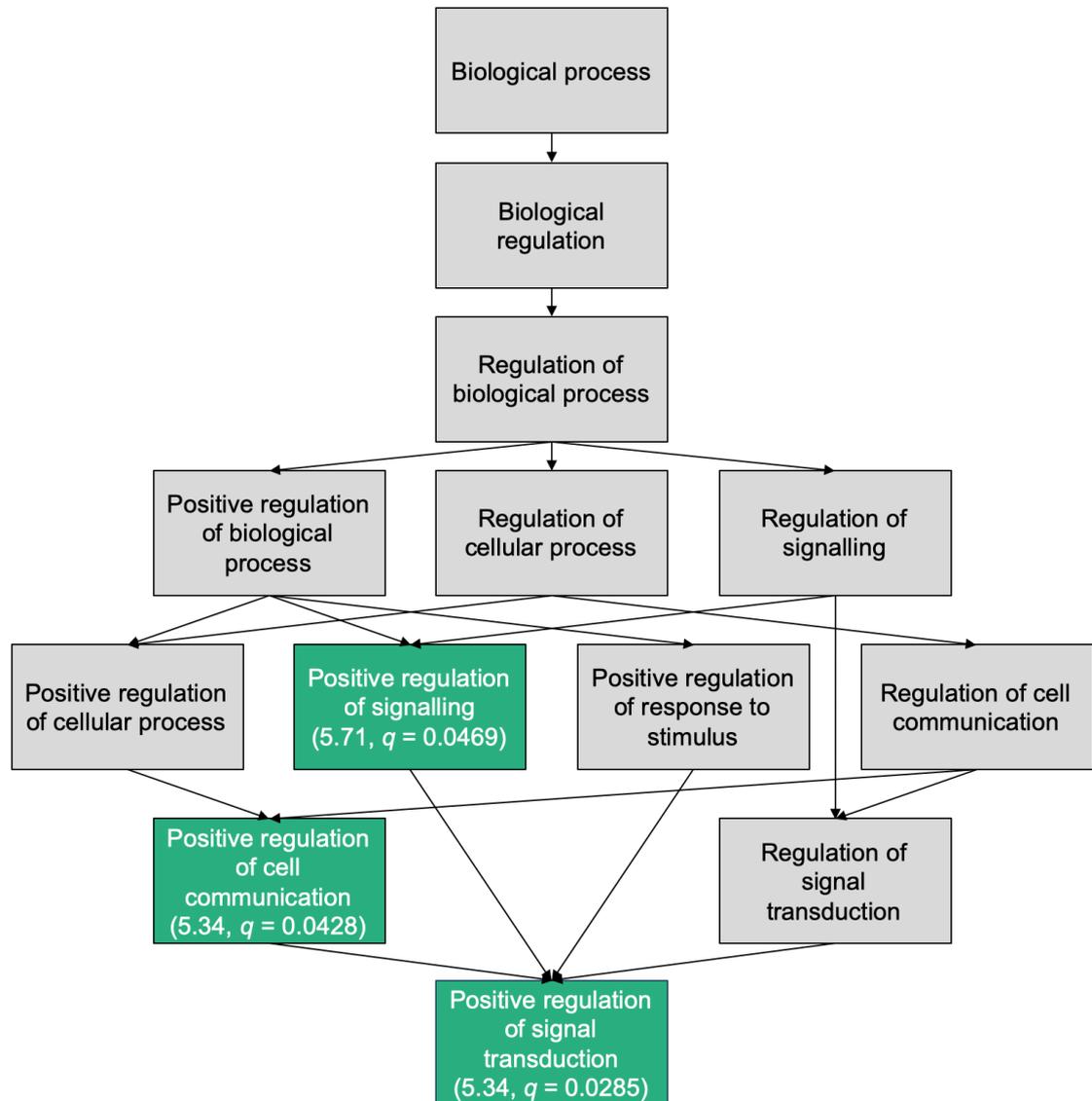
To understand the potential functional role HOW(S) plays in regulating mRNAs, GO term analysis was performed on the lists of significantly enriched genes and transcripts in the HOW(S)-HA pull-down using GOrilla (Methods and Materials 2.6.3.6). For the gene list, there were 53 significantly enriched GO terms under the ‘biological process’ domain (after FDR correction). The most significantly enriched of these was the ‘signal transduction’ GO term (GO:0007165; Fig 3.22). Several other signalling related terms were among the top ten enriched terms, as well as terms related to cell adhesion (Fig 3.22). For the transcript list, there were only three biological process GO terms that were significantly enriched: ‘positive regulation of signalling’, ‘positive regulation of cell communication’ and ‘positive regulation of signal transduction’ (Fig 3.23).



**Figure 3.22. Signal transduction is the most enriched GO term from the list of HOW(S)-HA pull-down genes.** The top ten significantly enriched GO terms from the 343 genes pulled-down in the HOW(S)-HA RIP-seq. GO terms from the ‘biological process’ domain were examined at levels 4–7 for this analysis. There were 27 significantly enriched terms in levels 4–7 (see Table A.10 in Appendix I). GO analysis was performed using GOrilla, and GO terms were filtered based on a  $q$ -value (FDR corrected  $p$ -value) cut-off of 0.05.

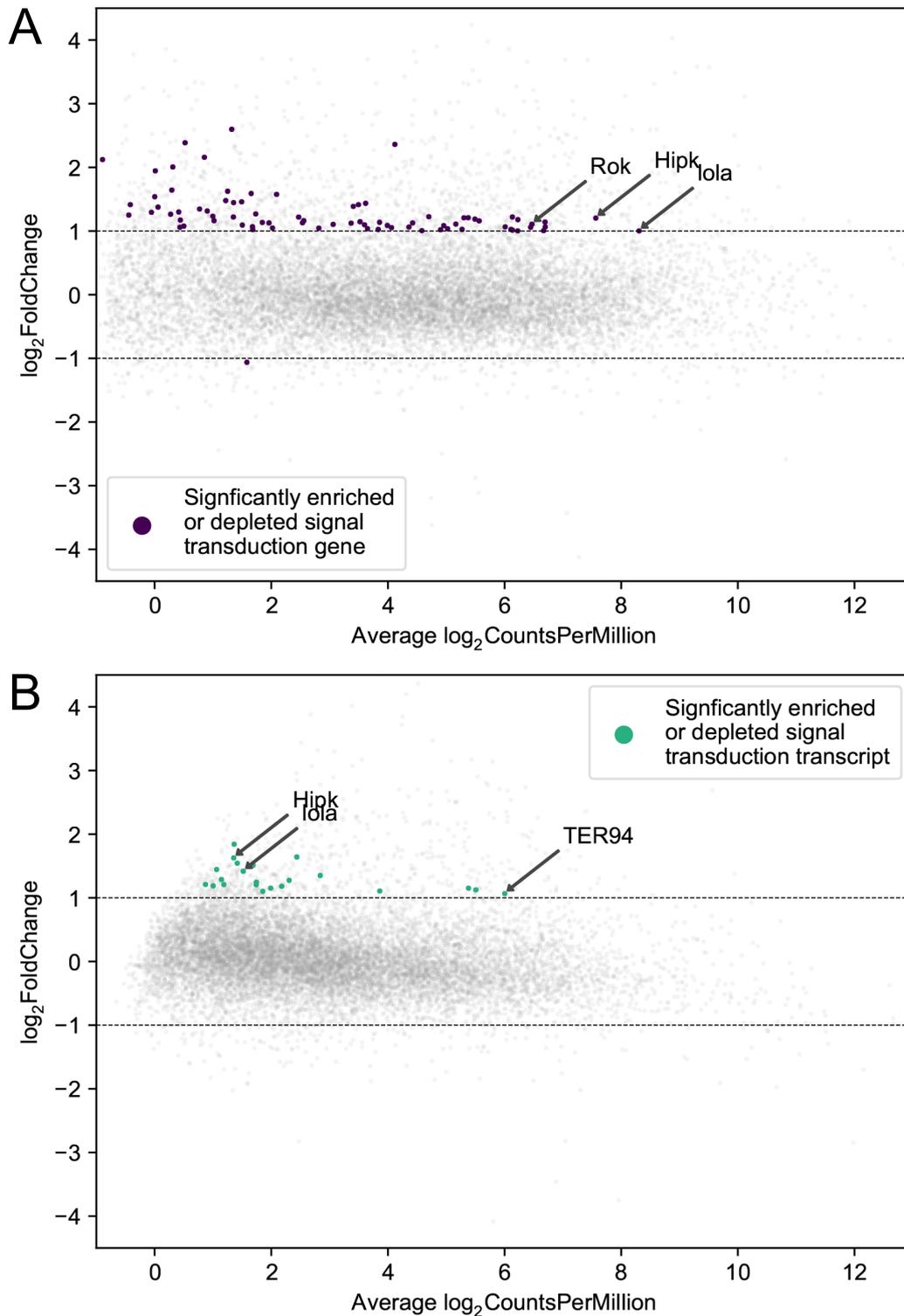
The enrichment of ‘signal transduction’ and closely related terms among the RNAs that were pulled-down with HOW(S) is highly consistent with HOW being a member of the signal transduction and activation of RNA (STAR) family of proteins. The fact this pattern is seen in both lists even though overlap between them is not that high indicates results are believable but noisy. 78 genes and 22 transcripts that were enriched in the HOW(S)-HA pull-down are connected to the signal transduction term (Fig 3.24). These include *homeodomain-interacting protein kinase (Hipk)*, *longitudinals lacking (lola)*, *Rho kinase (Rok)* and *transitional endoplasmic reticulum 94 (TER94)*. Of the 26 genes that are significantly enriched at both gene- and transcript-level (Fig 3.21), 5 of these are signal

transduction genes: *Hipk*, *Kul*, *lola*, *mbc* and *Rok*. Again, this is consistent with HOW being a member of the STAR family of proteins.



**Figure 3.23. GO terms related to signalling and communication are enriched among the HOW(S)-HA pull-down transcripts.** Flow chart of GO terms, boxes in green are terms significantly enriched in the list of 121 HOW(S)-HA pull-down enriched transcripts. In parentheses is the enrichment value followed by the FDR  $q$ -value. Terms with a  $q$ -value  $< 0.05$  were considered significantly enriched.

3. Identifying the targets of HOW(S) in *Drosophila melanogaster* testis



**Figure 3.24. 78 significant enriched genes and 22 transcripts are signal transduction related.** Scatter plots from the differential gene (A) and transcript (B) analysis of HOW(S)-HA pull-down RNA vs HOW(S)-HA total RNA. A) Purple dots indicate genes that were significantly enriched or depleted and come under the GO term ‘signal transduction’ (GO:0007165). 78 enriched genes and 1 depleted gene met this criteria, including the three labelled genes (*Rok*, *Hipk* and *lola*). B) Green dots indicate transcripts that were significantly enriched or depleted and come under the GO term ‘signal transduction’ (GO:0007165). 22 enriched transcripts and no depleted transcripts met this criteria, including the three labelled transcripts (*TER94*, *Hipk* and *lola*).

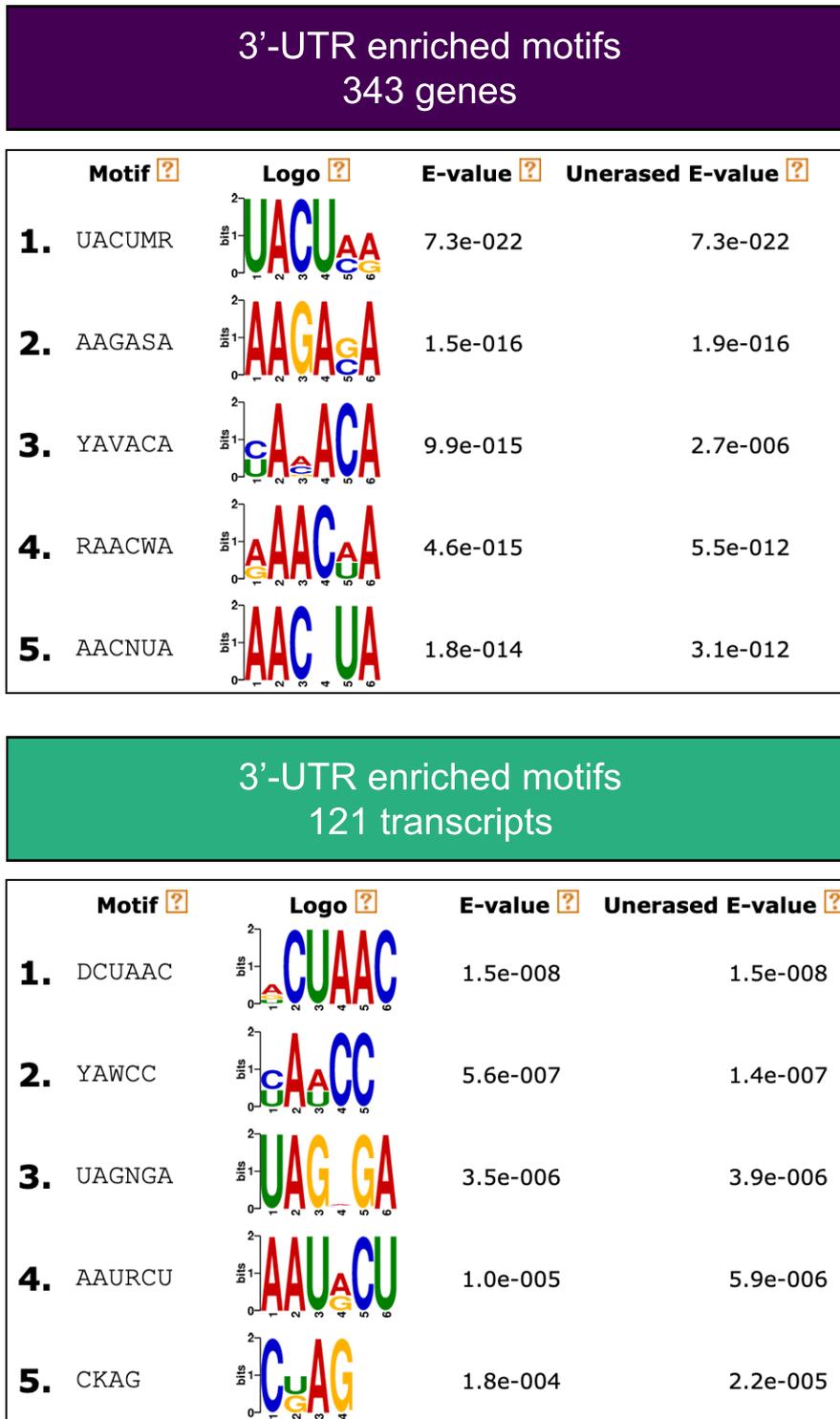
### 3.2.8. Finding binding motifs in the 3'-UTR of RNAs enriched in the HOW(S)-HA pull-down

The binding of Quaking subfamily proteins to 3'-UTRs has been well characterised (Jan et al., 1999; Nabel-Rosen et al., 1999; Ryder and Williamson, 2004), with a range of outputs on the RNAs they regulate from enhancing stability to translational repression (Jan et al., 1999; Saccomanno et al., 1999; Nabel-Rosen et al., 2002). 3'-UTRs generally possess higher AU content when compared to other parts of mRNAs, and in general the STAR proteins have binding motifs that exhibit high AU content. The HOW response element (HRE) is defined as ACUAA and was identified from HOW's interactions with the 3'-UTR of the *stripe* transcript (Israeli et al., 2007). However, no *in vivo* global assessment of HOW's RNA interactions and binding motifs has been undertaken until now. To identify potential HOW(S) binding sites within the significantly enriched transcripts and genes from the HOW(S)-HA RIP-seq, motif enrichment analysis was performed.

The initial search for a HOW(S) binding motif was carried out using the 3'-UTRs from the 121 HOW(S)-HA pull-down transcripts. For motif discovery with the 343 enriched genes the 3'-UTR from the most highly expressed transcript was used. DREME, from the MEME Suite, was used for the motif enrichment analysis, which allows users to input their own control sequences. Here, the control sequences used were the 3'-UTRs from all transcripts expressed in the testis based on the HOW(S)-HA RIP-seq data (Methods and Materials 2.6.3.7).

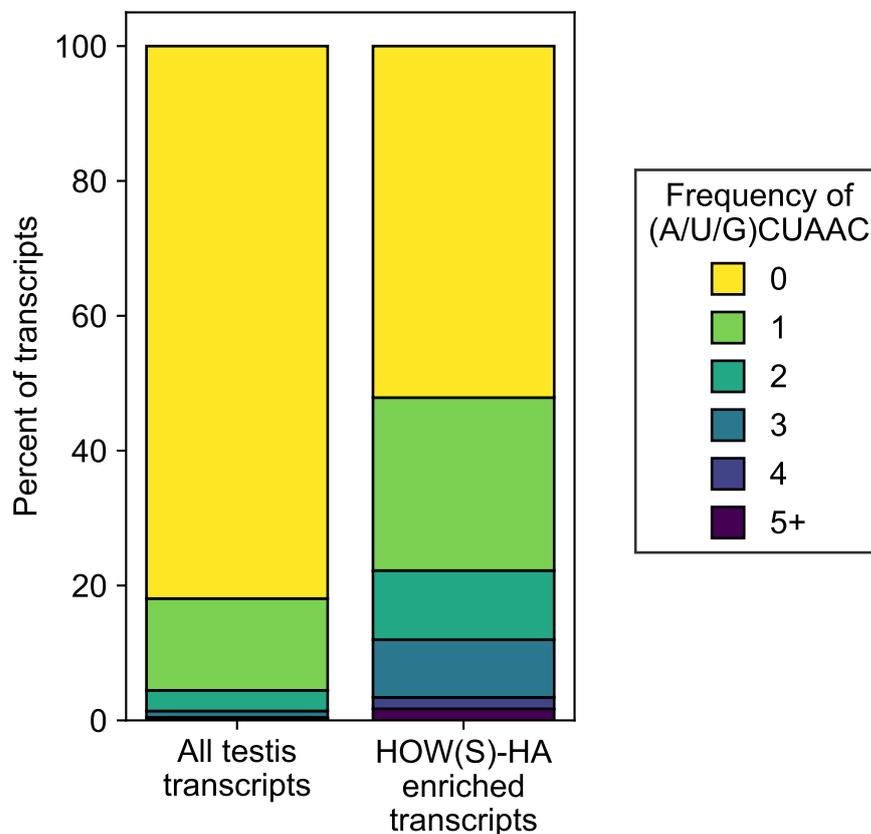
123 motifs with an *E*-value below 0.05 were detected by DREME from the 3'-UTRs of the 343 enriched genes, and 29 motifs from the 3'-UTRs of the 121 enriched transcripts. Despite the small amount of overlap between the gene and transcript enrichment lists (Fig 3.21), the motif hits with the highest *E*-value are remarkably similar: UACU(A/C)(A/G) for the enriched gene list, and (A/U/G)CUAAC for the enriched transcript list (Fig 3.25). These motifs, especially those found to be enriched at transcript-level, are also very similar to the previously identified HRE (Israeli et al., 2007).

3. Identifying the targets of HOW(S) in *Drosophila melanogaster testis*



**Figure 3.25. Motif enriched in the 3'-UTRs of HOW(S)-HA bound transcripts is very similar to the HRE defined within stripe mRNA.** The top 5 enriched motifs, using DREME, from the 3'-UTRs of differentially enriched genes (top) and transcripts (bottom). The 3'-UTRs for the genes were selected by choosing the highest expressing transcript for that gene. The top hit for both are similar, and with many of the top hits in each containing two adjacent adenosines. The transcript data is closer to the previously defined HOW binding site, found in *stripe* and *dpp* mRNA — ACUAA.

DREME found the top transcript motif in a greater percentage of HOW(S)-HA enriched 3'-UTRs compared to total testis 3'-UTRs. However, it does not describe how frequently the top motif appears within individual transcripts and if the frequency among the input sequences is different to the control sequences. To do this, the number of times a transcript's 3'-UTR contained the (A/U/G)CUAAC motif was compared between the HOW(S)-enriched list and the total testis RNA list. 2,254 out of the 12,473 total testis transcripts contain the motif at least once, and of these 2,254 just 24.7% have the motif more than once (Fig 3.26, left). Whereas of the 56 HOW(S)-HA enriched transcripts that contain this motif 46.4% of them contain the motif more than once (Fig 3.26, right). So, not only is this motif enriched among the HOW(S)-HA pull-down transcripts in their 3'-UTRs, it occurs more than once in those same 3'-UTRs at a higher frequency than the total testis 3'-UTR population.



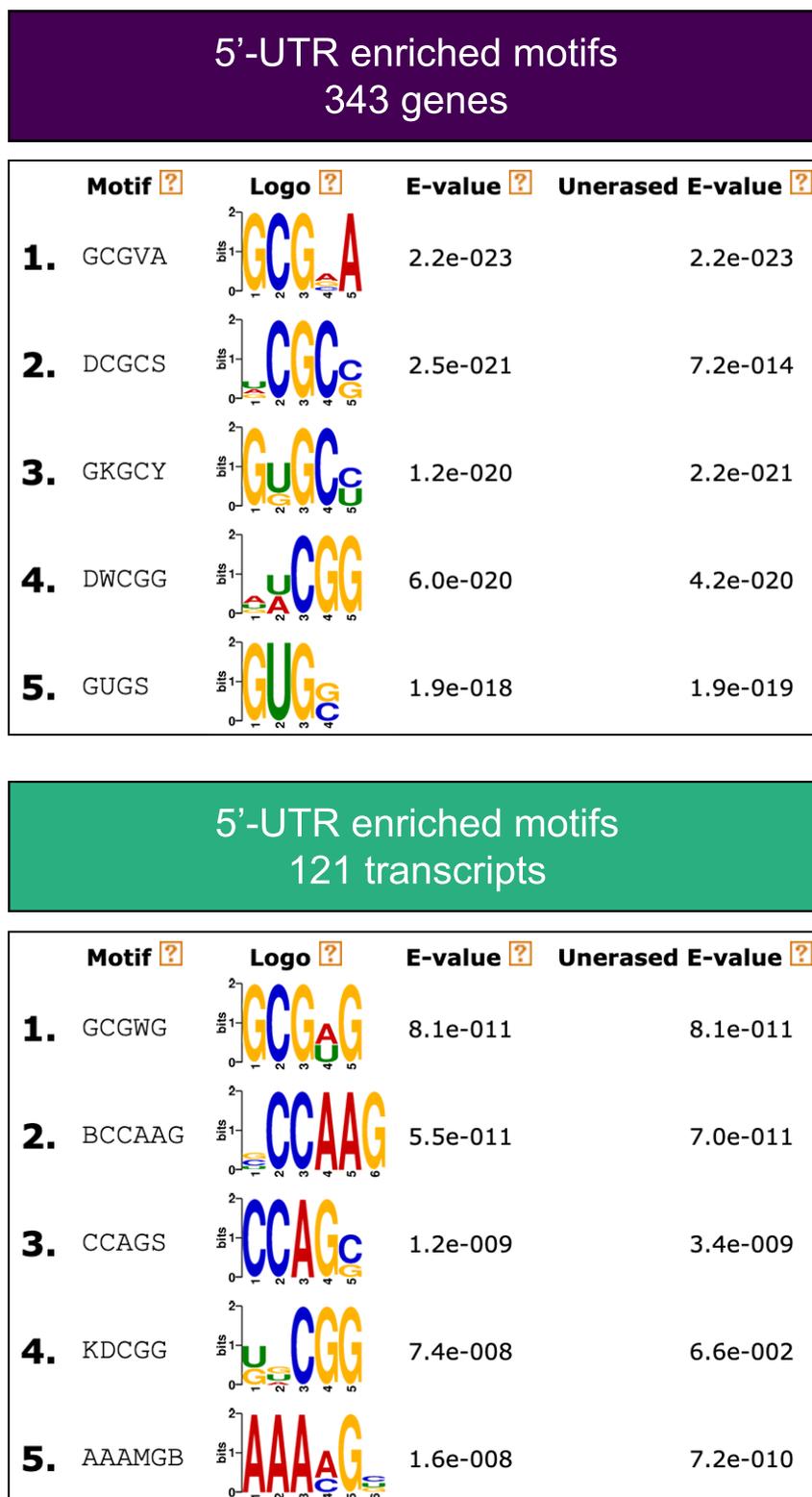
**Figure 3.26. HOW(S)-enriched transcripts are more likely to contain multiple (A/U/G)CUAAC motifs.** Among the 12,473 transcripts that are expressed in the testis, 18.1% have the (A/U/G)CUAAC motif in their 3'-UTR, and 4.5% have the motif more than once. 56 out of 117 (47.9%) HOW(S)-HA enriched transcripts have this motif in their 3'-UTR, and 22.2% have the motif more than once.

### 3. Identifying the targets of HOW(S) in *Drosophila melanogaster* testis

#### 3.2.9. Finding binding motifs in the 5'-UTR of RNAs enriched in the HOW(S)-HA pull-down

While the Quaking subfamily of STAR proteins are well established as 3'-UTR binding proteins, it has been shown that the *C. elegans* Quaking orthologue, GLD-1, can repress translation of a transcript when its binding site is in either UTR (Theil et al., 2018). Additionally, HOW(L) directly binds the 5'-UTR of *msl-2* with SXL to retain this transcript in the nucleus (Graindorge et al., 2013). Therefore, the 5'-UTRs of the HOW(S)-HA enriched transcripts and genes were searched to determine if there were enriched motifs in this region too.

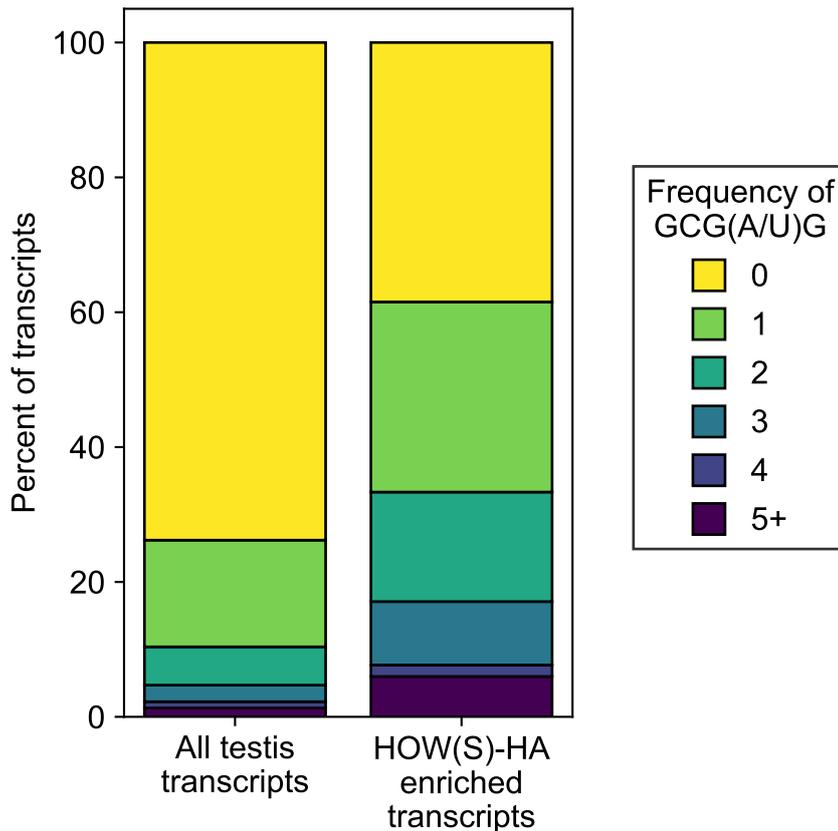
To identify sequences enriched in the 5'-UTR of HOW(S) targets, DREME was employed with the 5'-UTR sequences from the same gene and transcript lists used for the 3'-UTR analysis. 70 motifs with an *E*-value below 0.05 were discovered from the 343 enriched genes, and 44 motifs were discovered from the 121 enriched transcripts. Once again, the top hit from the gene and transcript lists are similar to one another: GCG(A/G/C)A for the genes and GCG(A/U)G for the transcripts (Fig 3.27). These 5'-UTR motifs show little similarity to the 3'-UTR top hits while having very similar *E*-values. This could suggest that HOW(S) has a much broader RNA-binding capacity than previously indicated from gene by gene studies. Alternatively, this 5'-UTR motif could present a binding motif of another RBP that interacts with HOW(S) in the context of 5'-UTRs. HOW has been shown to interact with other RBPs previously, for example, SXL and Hrp48 (Graindorge et al., 2013; Szostak et al., 2018).



**Figure 3.27.** Motif enriched in the 5'-UTR of HOW(S)-HA RIP-seq transcripts is very different to the 3'-UTR motifs. The top 5 enriched motifs, using DREME, from the 5'-UTRs of differentially enriched genes (top) and transcripts (bottom). The 5'-UTRs for the genes were selected by choosing the highest expressing transcript for that gene. The top hit for both are very similar.

### 3. Identifying the targets of *HOW(S)* in *Drosophila melanogaster* testis

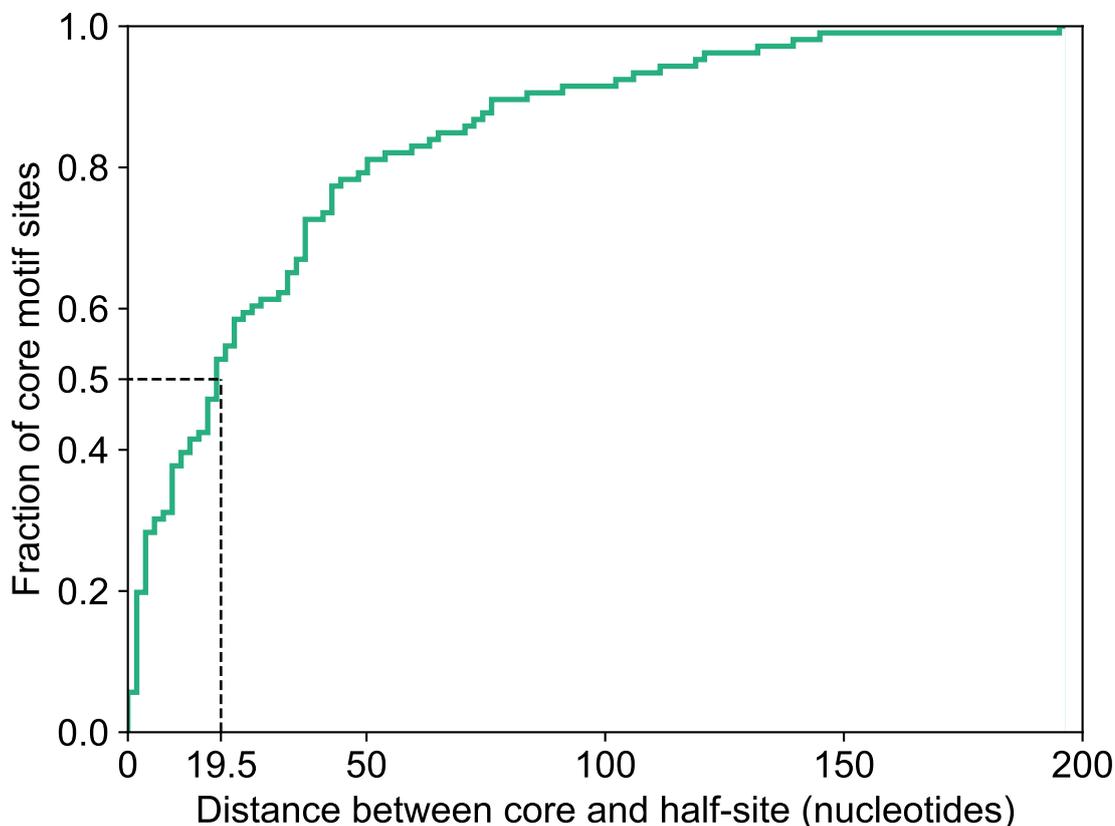
As with the 3'-UTR top transcript motif, the frequency of the top 5'-UTR transcript motif — GCG(A/U)G — within individual 5'-UTRs was quantified in the *HOW(S)*-HA pulled-down transcripts and the total testis transcript population. In the total testis, out of the 3,240 5'-UTRs which contain the motif, 39.6% have this motif more than once (Fig 3.28, left). Whereas of the 72 *HOW(S)*-HA enriched transcripts that contain the motif in their 5'-UTRs, 54.2% of them have the motif more than once (Fig 3.28, right). This increase in frequency is not as pronounced as the 3'-UTRs; a  $\sim 1.4$  times increase compared to  $\sim 1.9$  times.



**Figure 3.28. *HOW(S)*-enriched transcripts are more likely to contain multiple GCG(A/U)G motifs.** Among the transcripts that are expressed in the testis 26.2% have the GCG(A/U)G motif in their 5'-UTR, and 10.4% contain the motif in their 5'-UTR more than once. 72 out of 117 (61.5%) *HOW(S)*-HA enriched transcripts have this motif in their 5'-UTR, and 33.3% have the motif more than once.

### 3.2.10. Finding the nearest half-sites to the 3'-UTR core motif

STAR proteins have been shown to bind asymmetric bipartite motifs, i.e. a core motif with a neighbouring half-site (Galarneau and Richard, 2009). However, this has not been explored for any of the HOW protein isoforms. To determine if the core motifs in the HOW(S) pull-down transcripts have a half-site nearby the 3'-UTRs were searched for the nearest half-site. Here, the 3'-UTR top hit motif — (A/U/G)CUAAC — was used as the core motif, and the half-site was defined as UAAY (where Y is either of the pyrimidines), which was generated from Quaking and GLD-1 binding studies (Galarneau and Richard, 2005; Galarneau and Richard, 2009). Of the 106 core motifs found across the 3'-UTRs of 56 HOW(S) enriched transcripts all of them had a half-site within the 3'-UTR, and more than half (58/106) were 20 nt or closer to the core motif (Fig 3.29).

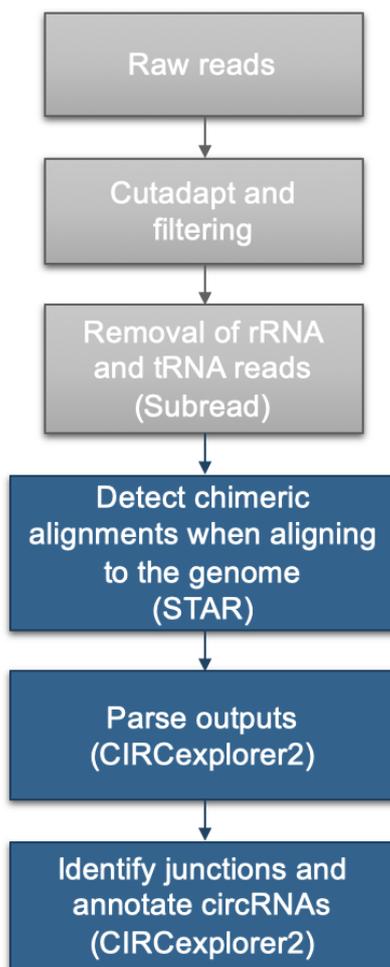


**Figure 3.29.** The majority of 3'-UTR core motif sites have a UAAY half-site less than 20 nucleotides away. Cumulative frequency of the distances between the 106 (A/U/G)CUAAC core motifs and their nearest UAAY half-site. The median distance, as marked on the plot, is 19.5 nucleotides. No maximum distance threshold was used and all core motifs had a neighbouring half-site within the same 3'-UTR.

### 3. Identifying the targets of HOW(S) in *Drosophila melanogaster* testis

#### 3.2.11. Identifying circRNAs pulled-down by HOW(S)-HA

Finally, Quaking, the human orthologue of HOW, has been found to have a role in circRNA biogenesis during embryonic development (Conn et al., 2015). Therefore, I sought to determine whether any circular RNAs (circRNAs) had been pulled down in the HOW(S)-HA RIP-seq. In order to perform such analysis, reads were re-mapped with the STAR aligner and CIRCexplorer2 was used to align and annotate circRNAs from the RIP-seq data (Materials and Methods 2.6.3.8; Fig 3.30). Though the RIP-seq experimental design is not optimised for circRNA detection, 6 circRNAs were identified that had more than 10 reads aligned to them (Table 3.3). 10 reads was used as a cut-off as this was the lowest average read count for a transcript that was kept using the edgeR `filterByExpr` function in the differential enrichment analysis. 3 of these circRNAs (highlighted in blue) contain the 3'-UTR (A/U/G)CUAAC motif. The circRNA from *muscleblind* is a well-characterised circRNA, 'dme\_circ\_0001328' in circBase (Ashwal-Fluss et al., 2014; Glažar et al., 2014). This circRNA, along with the circRNA from *CG30395*, have the greatest potential as HOW(S) targets due to the read counts being higher in the 'pull-down' samples than the 'total RNA' for two of the HOW(S) replicates but not the *nanos*-GAL4 samples, unlike the other four identified (Table 3.3).



**Figure 3.30. Schematic of circRNA analysis pipeline.** The pre-processing stages (grey) are the same as with the main RIP-seq analysis (Fig 3.14). Raw reads are processed to remove adapter sequences with cutadapt, and low quality reads are filtered out. Chimeric reads were detected using STAR and these were then used by CIRCexplorer2 to identify and annotate circRNAs.

**Table 3.3. circRNAs from HOW(S)-HA RIP-seq with 10 reads or more mapped.** 6 circRNAs have 10 or more reads aligned from at least one of the 8 samples. 3 of these, highlighted in blue, have the (A/U/G)CUAAC motif within them.

Location	Transcript ID	Name	<i>nanos</i> -GAL4		HOW(S)-HA 1		HOW(S)-HA 2		HOW(S)-HA 3	
			Total	Pull-down	Total	Pull-down	Total	Pull-down	Total	Pull-down
X:12411364-12411589	FBtr0347270	CR32652	63	129	74	82	102	69	89	31
2R:17275409-17276063	FBtr0310346	muscleblind	57	45	26	53	30	45	19	9
2R:21769766-21769904	FBtr0300226	CG30395	65	44	24	59	33	12	8	73
2R:24771724-24772795	FBtr0072433	Eps-15	3	9	12	4	7	1	6	2
4:1022241-1023680	FBtr0309865	Plexin A	3	9	3	5	9	10	12	2
2L:20648188-20648248	FBtr0300249	Uhg3	1	5	4	5	9	1	8	7

### 3.3. Discussion

#### 3.3.1. Overexpression of HOW(S)-HA in GSCs and early spermatogonia does not alter testis morphology or male fertility

The work presented here relied upon establishing the expression of HOW(S)-HA in GSCs and early spermatogonia for the *in vivo* RIP-seq. To understand the consequences of overexpressing HOW(S) in this manner the effects on testis morphology and male fertility were studied. Immunofluorescence experiments showed no differences in morphology between testes that expressed HOW(S)-HA and those that did not (Fig 3.4), in line with previous observations (Monk et al., 2010). Confocal microscopy also showed that HOW(S)-HA was predominantly localised in the cytoplasm, which is consistent with previous observations in male germ cells and neurons (Monk et al., 2010; Rodrigues et al., 2012). This is also in line with HOW(S) lacking the nuclear localisation signal that HOW(L) contains in its isoform specific C-terminal end (Nabel-Rosen et al., 1999). In tendons and glial cells HOW(S) is more evenly distributed between the nucleus and cytoplasm (Nabel-Rosen et al., 1999; Edenfeld et al., 2006; Rodrigues et al., 2012). In glial cells it was shown that HOW(S)'s nuclear localisation is dependent on its interaction with Crooked neck protein (Edenfeld et al., 2006). In the testis, it appears that there is no promotion of HOW(S) localisation to the nucleus, via a protein binding partner or otherwise.

Overexpressing HOW(S)-HA had little effect on male fertility, however there were multiple confounding factors in drawing conclusions from the sperm competition assays. Firstly, the data measuring the sperm 'offence' ability were inconclusive because the two parental lines, which were intended to be the controls to compare to the overexpressing HOW(S)-HA males, were different from one another making comparisons challenging (Fig 3.8). An additional complication was that UAS-HOW-S-HA parental flies exhibited leaky expression of HOW(S)-HA in the testes, as revealed by western blots of HOW(S)-HA pull-downs from these flies (Fig 3.13). Finally, the *nanos*-GAL4 males and the *nanos*-GAL4>HOW(S)-HA males have bar-shaped eyes, which could affect male mating behaviour and copulation success, similar to that seen with white-eyed males (Xiao et al., 2017). This could have contributed to the low remating rates of these flies (Table 3.1), which led to a small *n*-number for the *nanos*-GAL4 males, which reduced the statistical power of the experiment. Altogether, our results did not indicate that males overexpressing HOW(S)-HA had any fertility defects, but further experiments are required to provide a more robust assessment.

#### 3.3.2. Identification of RNAs that HOW(S) binds in GSCs and spermatogonia

RIP-seq was employed to determine which RNAs HOW(S) binds *in vivo*, with the aim of improving the understanding of the role of HOW(S) in spermatogenesis. While this

### 3. Identifying the targets of HOW(S) in *Drosophila melanogaster* testis

method does not identify specific binding sites, unlike more recently developed methods such as CLIP-seq and TRIBE (Licatalosi et al., 2008; McMahon et al., 2016), 121 transcripts and 343 genes were identified as enriched in the HOW(S)-HA pull-down and the 3'-UTRs of these transcripts and genes were enriched in motifs similar to the previously established HOW response element (HRE; Israeli et al., 2007). This indicates that the results from the RIP-seq analysis are robust enough to counter the limitations of the method.

The lists of transcripts and genes bound to HOW(S) did not include *bam*, an RNA that HOW(L) binds in early spermatogonia (Monk et al., 2010). While it has been shown that HOW(S) and HOW(L) are able to bind the same RNAs (Nabel-Rosen et al., 1999), HOW(L) is localised to the nucleus and HOW(S) to the cytoplasm. Thus, the lack of enriched *bam* is an indication that this HOW(S)-HA RIP-seq captured true cytoplasmic RNA targets of HOW(S) and that there is minimal post-lysis re-assortments of the *in vivo* RNA-protein interactions, which can occur during immunoprecipitation based experiments (Mili and Steitz, 2004). The enriched genes and transcripts lists did not include the three RNAs — *stripe*, *nrx-IV*, and *dgrasp* — that have previously been shown to bind HOW(S) in other tissues (Nabel-Rosen et al., 2002; Edenfeld et al., 2006; Giuliani et al., 2014). This suggests that these RNAs are either not expressed in the GSCs and early spermatogonia, or are not bound by HOW(S) in these cells. Investigating the RNAs that are bound by HOW(S) in the GSCs and spermatogonia is the focus of Chapter 4.

#### 3.3.3. Sequence drivers of HOW(S)-RNA binding

Despite the limited overlap between the enriched genes and transcripts pulled-down by HOW(S) (Fig 3.21), the downstream motif analysis identified similar motifs from these two different lists. It is likely that the motif from the differentially enriched transcript list reflects a more accurate HOW(S) binding sequence because the 3'-UTRs used for the gene-level analysis were generated from the highest expressed transcript for each of the enriched genes. Therefore, the gene-level analysis might not reflect the transcripts actually bound by HOW(S) or the transcript driving the noted enrichment of that gene. This might explain why the transcript list motif — (A/U/G)CUAAC — is more similar to the published HRE (ACUAA), than the gene-level motif — UACU(A/C)(A/G) (Fig 3.25).

Surprisingly, the 5'-UTR enriched motifs were very different to the 3'-UTR motifs, and had a comparable *E*-value (Fig 3.27). Again, the gene and transcript lists generated very similar top hit motifs (GCG(A/G/C)A and GCG(A/U)G, respectively), suggesting that these motifs are not random or from the noise of the RIP-seq experiment. Given the conserved nature of the consensus binding sequences of STAR proteins (Table 1.1), it is likely that 3'-UTR motif identified here is close to the optimum binding sequence for HOW. However, these 5'-UTR motifs could be lower affinity binding sites, or they could indicate that HOW(S) interacts with another RBP that binds to these 5'-UTR motifs.

The sequences that drive direct HOW(S)-RNA interactions will be explored further in Chapter 5.

RBPs can have a greater impact, or are more likely to impact, the RNAs they bind that contain multiple binding sites than RNAs with just one binding site (Mukherjee et al., 2011; Teplova et al., 2013; Sharma et al., 2021). For example, RNAs with multiple intronic HuR binding sites exhibit greater destabilisation upon HuR knockdown in a combinatorial manner (Mukherjee et al., 2011). So, in addition to identifying motifs enriched in the transcripts that HOW(S) binds, the number of motifs within each enriched transcript was compared to all testis transcripts. Both the (A/U/G)CUAAC 3'-UTR motif and the GCG(A/U)G 5'-UTR motif appeared at a greater frequency in the HOW(S)-HA transcript list than in the all testis transcript list (Figs 3.26 and 3.28). Giving further support that these transcripts identified are bound by HOW(S) *in vivo*.

### 3.3.4. Conclusions

In summary, this chapter reports the identification of RNAs bound to HOW(S) in GSCs and early spermatogonia in *D. melanogaster* testis. 121 transcripts were identified in the transcript-level analysis (Fig 3.20), 56 of which contained the top hit motif, (A/U/G)CUAAC, in their 3'-UTRs (Figs 3.25 and 3.26). This motif strongly resembles the previously identified HRE, ACUAA (Israeli et al., 2007), as well as the consensus binding sequences of other STAR family proteins (Table 1.1). 343 genes were identified from the gene-level analysis. These genes were strongly enriched for the 'signal transduction' GO term and the sequences enriched in their 3'-UTRs also resembled the HRE.



## 4. Understanding the role of HOW(S) regulated events in germ cells

### 4.1. Introduction

At the apical tip of the *D. melanogaster* testis, where HOW is expressed, there are two key regulatory processes occurring — GSC homeostasis and the control of the mitotic divisions prior to meiosis (Davies and Fuller, 2008). As is typical for adult stem cells, the former relies on the stem cell niche, the microenvironment in which the stem cell is situated. In the testis this niche comes in the form of a cluster of somatic hub cells that reside at the apical tip of the testis, and the GSCs form contacts with the hub cells (Fig 1.2; de Cuevas and Matunis, 2011). Additionally, each GSC is surrounded by two somatic cyst stem cells. When the GSC divides one cell retains the GSC identity and the other forms a new gonialblast. The two cyst stem cells also divide, and the two new cyst cells encapsulate the gonialblast and will continue to do so throughout the differentiation process (Davies and Fuller, 2008; de Cuevas and Matunis, 2011).

The extrinsic signalling between the somatic cells and germ cells are essential for both maintaining the GSC population and the proliferation of the germ cells. Doubtless, intrinsic cellular signals are also important for proper spermatogenesis. As the spermatogonia within a cyst divide the cells from the original gonialblast do not complete cytokinesis and remain linked via intercellular bridges, also referred to as ring canals, so that they remain synchronised in the timing of their mitotic and meiotic divisions (Greenbaum et al., 2011). The number of mitotic divisions is tightly regulated with 99% of the 112 cysts counted undergoing exactly four rounds of mitosis (Insko et al., 2009). One of the intrinsic signals important for the timing of divisions is Bam protein, which must accumulate to a critical level to trigger the switch from mitosis to meiosis (Insko et al., 2009).

STAR proteins provide a direct link between signalling and RNA regulation, as the STAR proteins themselves are regulated by post-translational modifications. For example, phosphorylation of Sam68 by ERK enhances its splicing activity to include a particular exon of CD44 (Matter et al., 2002). Similarly, phosphorylation of HOW by ERK2 enhances HOW dimerisation and binding, thus linking MAPK/ERK signalling to RNA regulation (Nir et al., 2012).

In *D. melanogaster* testis, HOW loss of function mutants exhibit a loss of GSCs. Both HOW(L) and HOW(S) can partially rescue this phenotype when they're expressed in the germ cells of these HOW mutant flies (Monk et al., 2010). Additionally, HOW(L) was shown to directly bind *bam* RNA. Given that overexpression of HOW(L) results in an extra mitotic division, it was proposed that by suppressing *bam* HOW(L) contributes to the regulation of the mitosis to meiosis switch (Monk et al., 2010). This example also

#### 4. Understanding the role of *HOW(S)* regulated events in germ cells

demonstrates how understanding the RNAs that RBPs bind to improves the understanding of the function of RNA binding protein.

In the previous chapter, the RNAs that *HOW(S)* binds in GSCs and early spermatogonia were identified via RIP-seq. To better understand why *HOW* is essential for spermatogenesis, this chapter aims to uncover which of the RNAs that *HOW(S)* binds are critical for proper spermatogenesis by examining the impact on testis morphology and male fertility when knocking down these genes.

## 4.2. Results

### 4.2.1. Conditional knockdown of *HOW* in the testis results in large morphological changes

Having identified 121 transcripts and 343 genes bound by *HOW(S)* via RIP-seq, the next step was to understand how these transcripts and genes relate to the function of *HOW* in the testes. It has been previously observed that when *HOW* is knocked down in *D. melanogaster* testes, using a *nanos*-GAL4 driver, there is a loss of germ cells and an expansion of the somatic cells (Monk et al., 2010). However, the published microscopy images presented with these observations were only of the apical tip of the testis (Monk et al., 2010). To understand the effects of the knockdown of *HOW* on the whole testis, a whole testis phenotype assessment was performed. Tiled confocal microscopy images were collected for WT and *HOW* knockdown testes.

*HOW* was knocked down in the testes using an RNAi line that targets all six annotated *HOW* RNA isoforms (BDSC #55665) and crossed with the same *nanos*-GAL4 driver (#64227) used for overexpressing *HOW(S)* in the RIP-seq experiment (Chapter 3). Both the female and male progeny from this cross were placed in vials with a WT partner but produced no progeny, suggesting these *HOW* knockdown flies were sterile. Additionally, the testes of the *HOW* knockdown males were noticeably smaller than WT when observed using a light microscope.

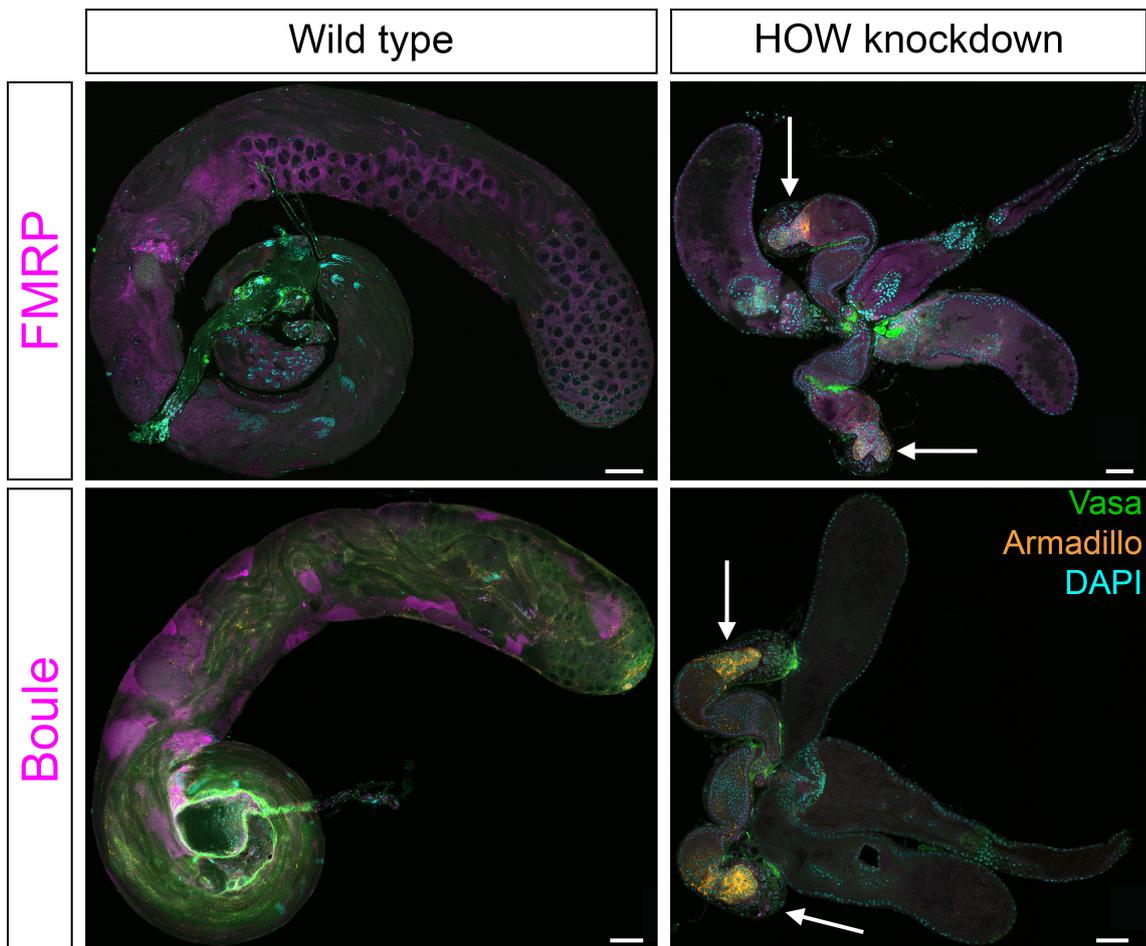
The *HOW* knockdown males were dissected for confocal microscopy. Due to their small size, the testes were not separated from the accessory glands or ejaculatory duct to minimise damage to the tissue when dissecting. WT testes were dissected for comparison and these were separated from the accessory glands and ejaculatory duct. Several antibodies were used, across both WT and *HOW* knockdown samples, to examine the cell populations in the tissues across the three phases of spermatogenesis. Three germ cell markers were used: 1) Vasa — abundant in spermatogonia (mitosis phase) and also present throughout spermatogenesis, 2) FMRP (also known as *Fmr1*) — abundant in spermatocytes (meiosis phase) and also present in early spermatids (spermiogenesis phase), 3) Boule — abundant in round and elongating spermatids and also present in late spermatocytes. Armadillo

was used as a somatic cell marker. All samples were stained with anti-Armadillo, anti-Vasa and counter-stained with DAPI. Tissues were then additionally stained with either anti-FMRP (Figs 4.1 and 4.2) or with anti-Boule (Figs 4.1 and 4.3).

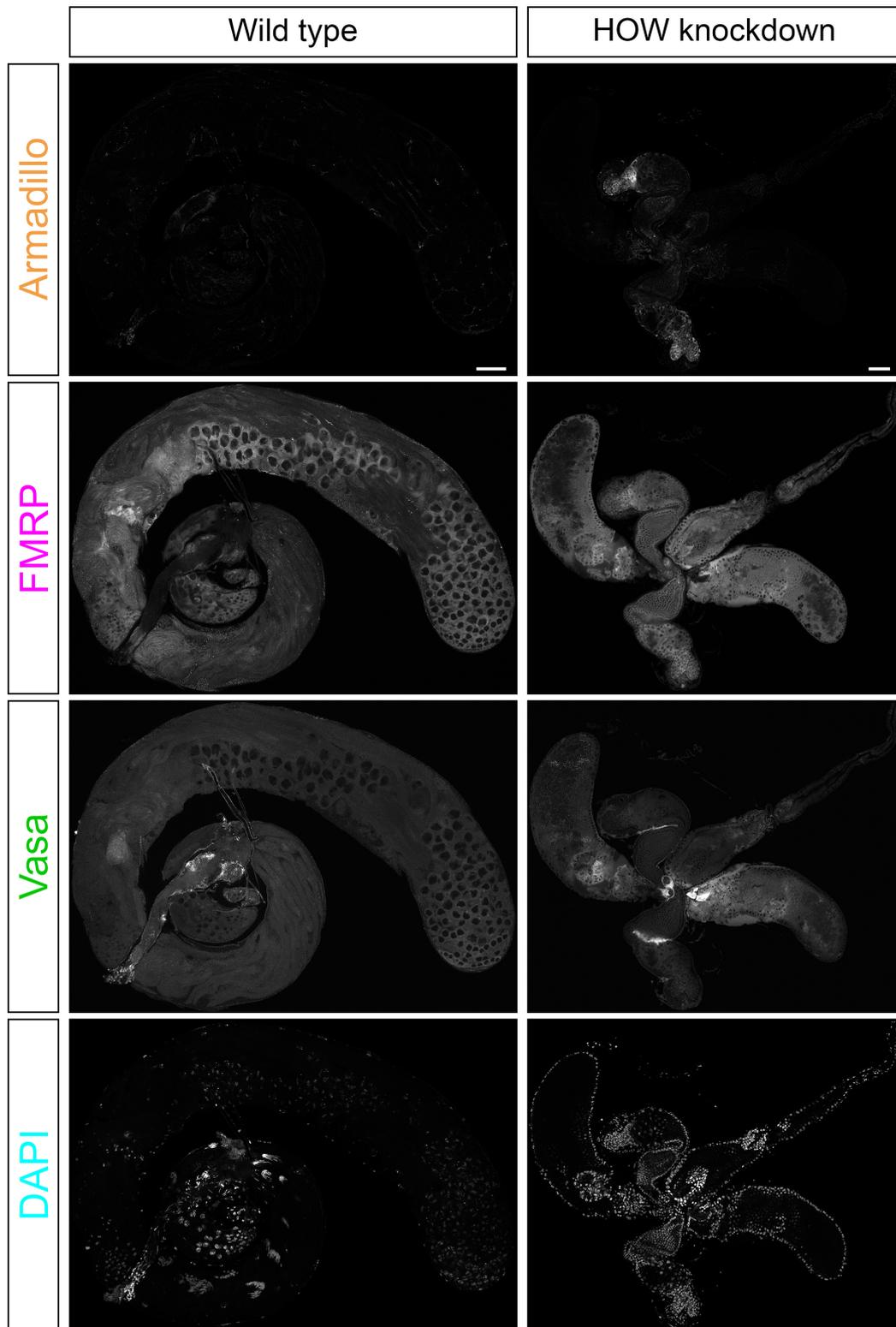
The most striking difference between the HOW knockdown and WT testes, which can be seen in both sets of staining, is the difference in size; the HOW knockdown testes are substantially smaller (Fig 4.1). Secondly, the aforementioned overproliferation of the somatic cells is also very clear. The small cluster of Armadillo positive cells in the apical tip of the Boule-stained WT testes are the somatic hub cells (Figs 4.1 and 4.3). By contrast, in the HOW knockdown testes the bulk of the cells at the apical tip are Armadillo positive, and in the FMRP-stained testes there is a fork-like structure at the apical tip of the bottom testis (Figs 4.1 and 4.2).

There is no specific staining for any of the germ cell markers in the HOW knockdown testes. No spermatocytes with their characteristic large nuclei are visible with the FMRP staining (Figs 4.1 and 4.2). And no elongating spermatids are highlighted by anti-Boule (Figs 4.1 and 4.3). During the normal progression of spermatogenesis nuclei elongate in the spermatids into distinctive structures (Fig 1.3), which here can be seen in the basal end of the WT testes by the DAPI staining, however, all the nuclei are round in shape in the HOW knockdown testes (Figs 4.1–4.3). Altogether, these images have confirmed that upon knockdown of HOW in the GSCs and early spermatogonia there is an overproliferation of the somatic cells and a significant loss of germ cells. Additionally, there are no indications of any germ cells in the meiosis or spermiogenesis phases of spermatogenesis.

4. Understanding the role of *HOW(S)* regulated events in germ cells

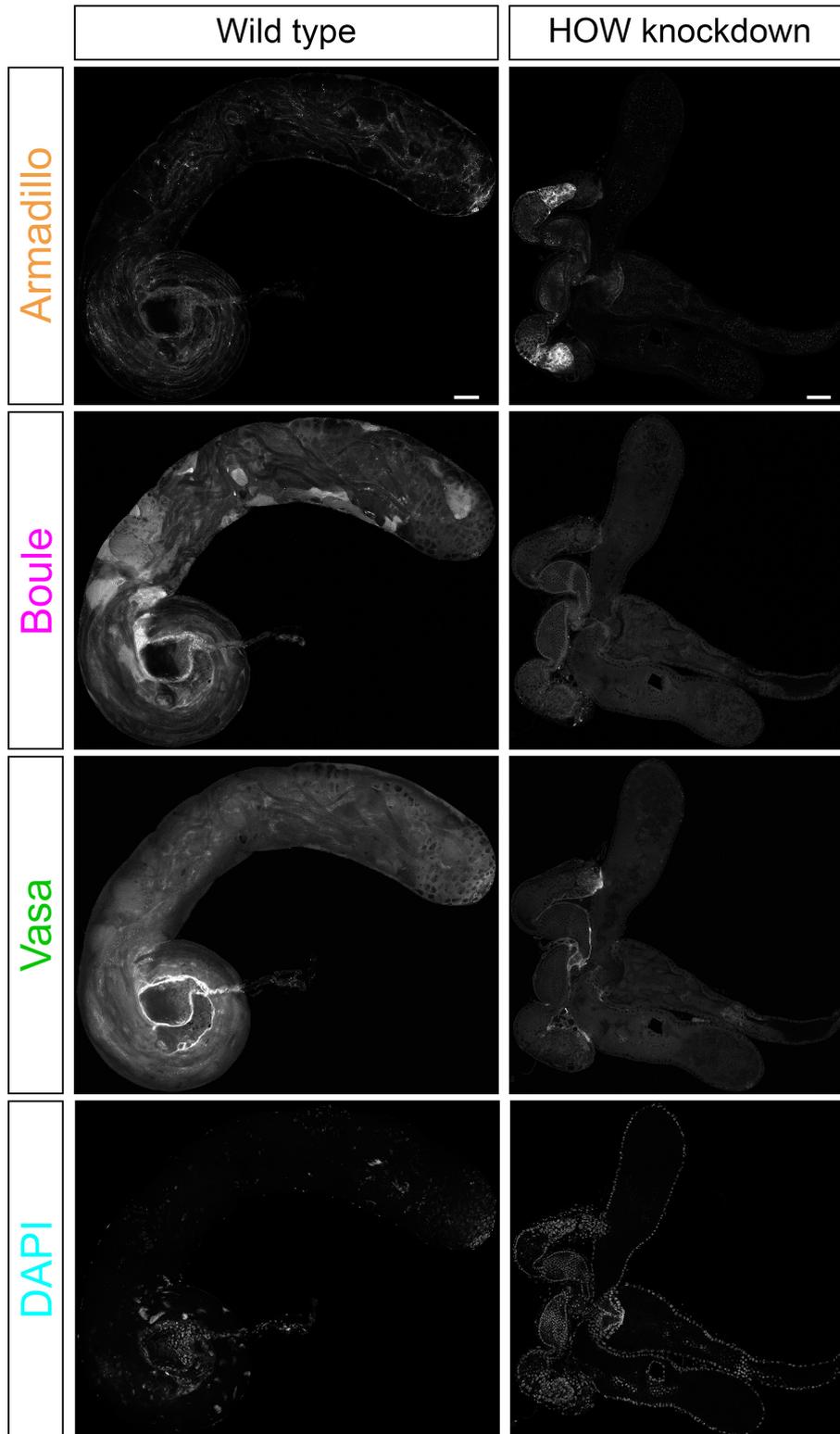


**Figure 4.1.** *nanos-GAL4>HOW-RNAi* testes are small and have few germ cells. Confocal microscopy images of testes from WT (left-hand panels) and male reproductive organs from *nanos-GAL4>HOW-RNAi* flies (right-hand panels). Tissues in all samples were stained with anti-Armadillo (orange), anti-Vasa (green), and counterstained with DAPI (cyan). Top panels were stained with anti-FMRP (magenta), bottom panels with anti-Boule (magenta). *HOW* knockdown testes show an overproliferation of the Armadillo positive somatic cells, and no specific staining for the germ cell markers (Vasa, FMRP and Boule). Arrows indicate testes in right-hand panels. Bar = 50  $\mu$ m.



**Figure 4.2. No meiotic cells in *nanos-GAL4*>*HOW*-RNAi testes.** Confocal microscopy images of testes from WT (left-hand panels) and male reproductive organs from *nanos-GAL4*>*HOW*-RNAi flies (right-hand panels). Tissues were stained with anti-Armadillo, anti-FMRP, anti-Vasa, and counter-stained with DAPI. *HOW* knockdown testes show an overproliferation of the Armadillo positive somatic cells, and no specific staining for the germ cell markers Vasa or FMRP, which in WT testes are particularly abundant in spermatogonia and spermatocytes cells, respectively. Bar = 50  $\mu$ m.

4. Understanding the role of *HOW(S)* regulated events in germ cells



**Figure 4.3. Loss of elongating spermatids in *nanos-GAL4*>*HOW*-RNAi testes.** Confocal microscopy images of testes from WT (left-hand panels) and male reproductive organs from *nanos-GAL4*>*HOW*-RNAi flies (right-hand panels). Tissues were stained with anti-Armadillo, anti-Boule, anti-Vasa, and counter-stained with DAPI. *HOW* knockdown testes show an overproliferation of the Armadillo positive somatic cells, and no specific staining for the germ cell markers Vasa or Boule, which in WT testes are particularly abundant in spermatogonia and elongating spermatids cells, respectively. Bar = 50  $\mu$ m.

#### 4.2.2. Selecting transcripts enriched in the HOW(S)-HA RIP-seq for phenotypic analysis

To relate the impact of HOW knockdown in the testes with the 121 transcripts enriched in the HOW(S)-HA pull-down, a selection of these transcripts were investigated to find whether they, like HOW, were essential or important for spermatogenesis.

A maximum of 12 transcripts could be studied within the time constraints. To select these 12 transcripts for phenotypic experiments, a range of features from the earlier RIP-seq analysis and the literature were considered. Two of these features formed specific criteria that the transcript of interest had to meet. The first was whether the top 3'-UTR motif identified using DREME, (A/U/G)CUAAC (Fig 3.25), was present in either of the UTRs. 70 of the 121 enriched transcripts met this first criterion. The second excluded 22 transcripts (from the remaining 70) that had previously been screened and found to not have an impact on GSC maintenance or differentiation in *D. melanogaster* spermatogenesis (Yu et al., 2016). Other features from the RIP-seq analysis that went into consideration but did not form strict cut-offs were: if the transcript was from a gene that was also differentially enriched, the distance between 3'-UTR core motif and its nearest half-site (UAAY), whether the transcript had the top 5'-UTR motif (GCG(A/U)G; Table 4.1).

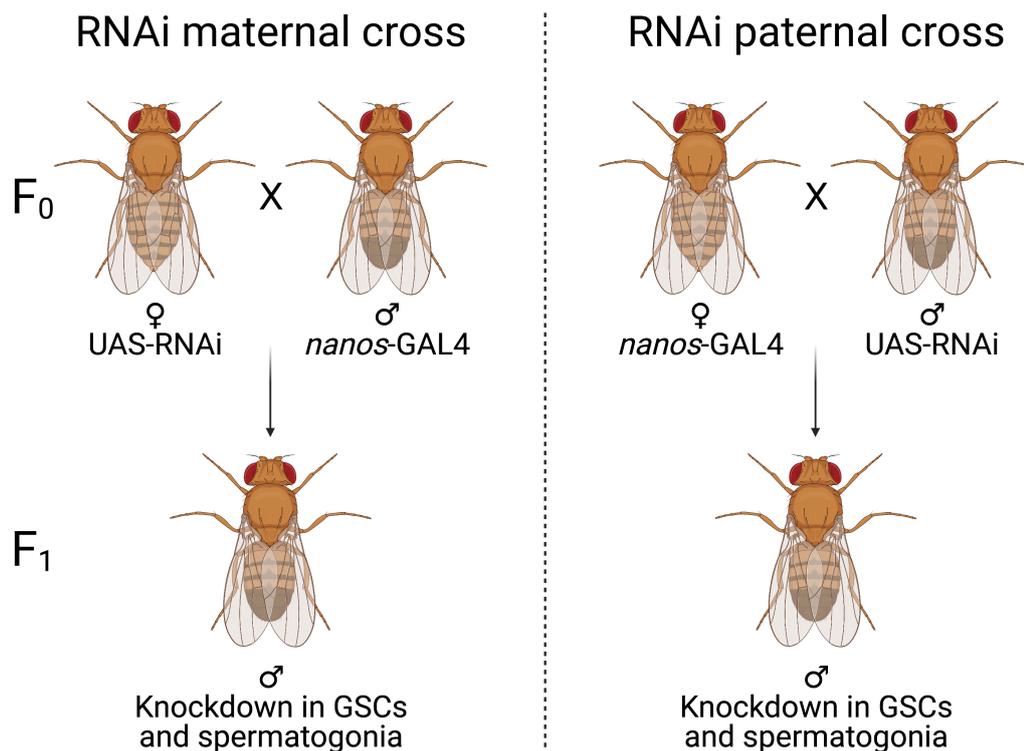
Data from the literature was also taken into account when selecting the 12 transcripts. For example, a limitation of immunoprecipitation based methods is that RNAs and proteins can reassociate post-lysis (Mili and Steitz, 2004). Therefore, some of the RNAs identified as enriched in the pull-down might not be expressed in the same cells as HOW(S). To remedy this, Dr Dapeng Wang (LeedsOmic) analysed published single-cell RNA-seq data from *D. melanogaster* testis (Witt et al., 2019), and the gene expression level from the 'GSCs and early spermatogonia' category were taken into account in the selection process (Table 4.1). Additionally, gene-specific literature searches were carried out for any previous reports related to germ cell development in either male or female flies. For example, the enriched *lola* transcript (*lola*-RJ, FBtr0089360) corresponds to the *lola*-O protein isoform (Goeke et al., 2003), which when specifically knocked out via CRISPR resulted in semi-sterile females (Dinges et al., 2017). No phenotype was reported for knockout *lola*-O males, however another study has implicated *lola* in the control of transit-amplifying mitotic divisions in the male germ line — notably in an opposite manner to HOW(L) (Monk et al., 2010; Davies et al., 2013). From this combination of computational analysis and literature searches 12 transcripts that were enriched in the HOW(S)-HA RIP-seq were selected for investigation into their roles in spermatogenesis (Table 4.1).

**Table 4.1. Expression and sequence features of 12 transcripts enriched in the HOW(S)-HA RIP-seq selected for phenotypic experiments.** The 12 selected transcripts are listed with their FlyBase gene symbol and transcript ID. The second column contains the  $\log_2$ (Fold Change) value from the HOW(S) RIP-seq, i.e. the enrichment level for each transcript. For  $\log_2$ (CPM) and  $p$ -value see Table A.9 in Appendix I. Columns 3–6 are shaded in green if the transcript has the feature described by the column header. Column 7 contains the percentile of gene expression from the ‘GSCs and early spermatogonia’ cell type extracted from single-cell testis data (analysed by Dr Dapeng Wang, using data from Witt et al., 2019).

Gene symbol (transcript ID)	Transcript $\log_2$ (Fold Change) from RIP-seq	Differentially enriched gene	(A/U/G)CUAAC motif in 3'-UTR (number of motifs)	(A/U/G)CUAAC motif in 5'-UTR (number of motifs)	GCG(A/U)G motif in 5'-UTR (number of motifs)	Percentile expression in GSCs and early spermatogonia
Df31 (FBtr0085919)	2.339	Y	Y (2)	Y (1)	Y (1)	66
smash (FBtr0308850)	1.264	Y	Y (1)	Y (1)	Y (3)	51
jvl (FBtr0305694)	1.099	Y	Y (2)	Y (1)	Y (1)	72
Hipk (FBtr0072552)	1.628	Y	Y (4)	N	Y (1)	57
lola (FBtr0089360)	1.418	Y	Y (1)	N	N	79
CycG (FBtr0085803)	1.107	N	Y (1)	Y (1)	Y (2)	87
bif (FBtr0073526)	1.181	N	Y (1)	N	Y (1)	43
CLIP-190 (FBtr0332452)	1.951	N	Y (4)	N	Y (1)	78
Syx1A (FBtr0392904)	1.564	N	Y (2)	N	Y (6)	78
toc (FBtr0077662)	1.754	N	Y (3)	N	Y (1)	74
Syx6 (FBtr0088102)	1.261	N	Y (2)	N	N	89
TER94 (FBtr0343852)	1.066	N	N	Y (1)	N	89

To determine whether any of these 12 genes are required for proper spermatogenesis, each gene was knocked down in germ cells and were screened for major defects in either testes morphology or fertility. RNAi stocks were acquired for each of the genes to perform knockdown experiments in the germ cells. Two RNAi stocks, which target different regions of the genes, were acquired per gene of interest. The exception to this was for *toucan* (*toc*) for which there was only one RNAi stock available (see Table 2.1 in Materials and Methods for all stock details).

For knockdown, these 23 RNAi lines, were each crossed with the *nanos*-GAL4 driver to knockdown the expression of the gene targets in the GSCs and early spermatogonia. These crosses were carried out in both directions, i.e. a cross where the female was the RNAi line and the male was the GAL4 driver (referred to as the maternal cross) and vice versa for the opposite cross direction (the paternal cross; Fig 4.4). The testes from F<sub>1</sub> males were dissected and viewed under a microscope to look for any morphological defects. The F<sub>1</sub> males and F<sub>1</sub> females were placed in vials for several days to see if they produced any F<sub>2</sub> progeny.



**Figure 4.4. Schematic of the RNAi crosses.** The RNAi fly lines, targeting one of the twelve genes of interest, contain a UAS followed by the RNAi sequence. The F<sub>0</sub> crosses were between the RNAi lines and the *nanos*-GAL4 driver in both directions, when the RNAi line is the female in the cross (maternal cross, left) and when the RNAi line is the male in the cross (paternal cross, right). The F<sub>1</sub> male progeny should have a knockdown in the expression of the gene of interest in the GSCs and spermatogonia.

#### 4. Understanding the role of HOW(S) regulated events in germ cells

The Hipk and lola knockdowns showed morphology and/or fertility defects (Table 4.2). The Hipk RNAi 1 line targets three of the four Hipk RNA isoforms, including Hipk-RA (FBtr0072552) the specific transcript that was enriched in the HOW(S) RIP-seq, and Hipk RNAi 2 targets all four Hipk RNA isoforms. The lola RNAi 1 and 2 lines both target all 25 lola RNA isoforms. The Updated Targets of RNAi Reagents (UP-TORR) tool predicts that none of the Hipk and lola RNAi lines used should have any off target effects (Hu et al., 2013).

Morphology defects were observed in progeny from the maternal crosses for Hipk RNAi 1 and Hipk RNAi 2 knockdowns (Table 4.2). Testes from both of these crosses displayed pinched apical tips. The testes from the Hipk RNAi 2 maternal knockdown cross were much smaller than WT testes, which was not the case for the F<sub>1</sub> males from the Hipk RNAi 1 maternal cross. The F<sub>1</sub> males from both the maternal and paternal lola RNAi 1 crosses appeared to have wider apical tips (Table 4.2), however, it should be noted that these light microscopy observations were carried out with the knowledge of the genotype of the flies and that this is a published phenotype for *lola* mutants (Davies et al., 2013). No other testis morphology defects were observed when knocking down genes with the other 20 RNAi lines.

**Table 4.2. Morphology and fertility defects observed from Hipk and lola conditional knockdowns.**

Knockdown cross	Morphological phenotypes of F <sub>1</sub> testes from light microscopy	Do F <sub>1</sub> flies produce F <sub>2</sub> progeny?
♀ Hipk RNAi 1 ♂ <i>nanos</i> -GAL4	Yes — F <sub>1</sub> testes from the maternal RNAi cross have pinched apical tip	Yes
♀ <i>nanos</i> -GAL4 ♂ Hipk RNAi 1	No	Yes
♀ Hipk RNAi 2 ♂ <i>nanos</i> -GAL4	Yes — F <sub>1</sub> testes are very small and have pinched apical tip	Few — lots of eggs produced, very few adults
♀ <i>nanos</i> -GAL4 ♂ Hipk RNAi 2	No	Yes
♀ lola RNAi 1 ♂ <i>nanos</i> -GAL4	Slightly wider apical tip	Yes
♀ <i>nanos</i> -GAL4 ♂ lola RNAi 1	Slightly wider apical tip	Yes

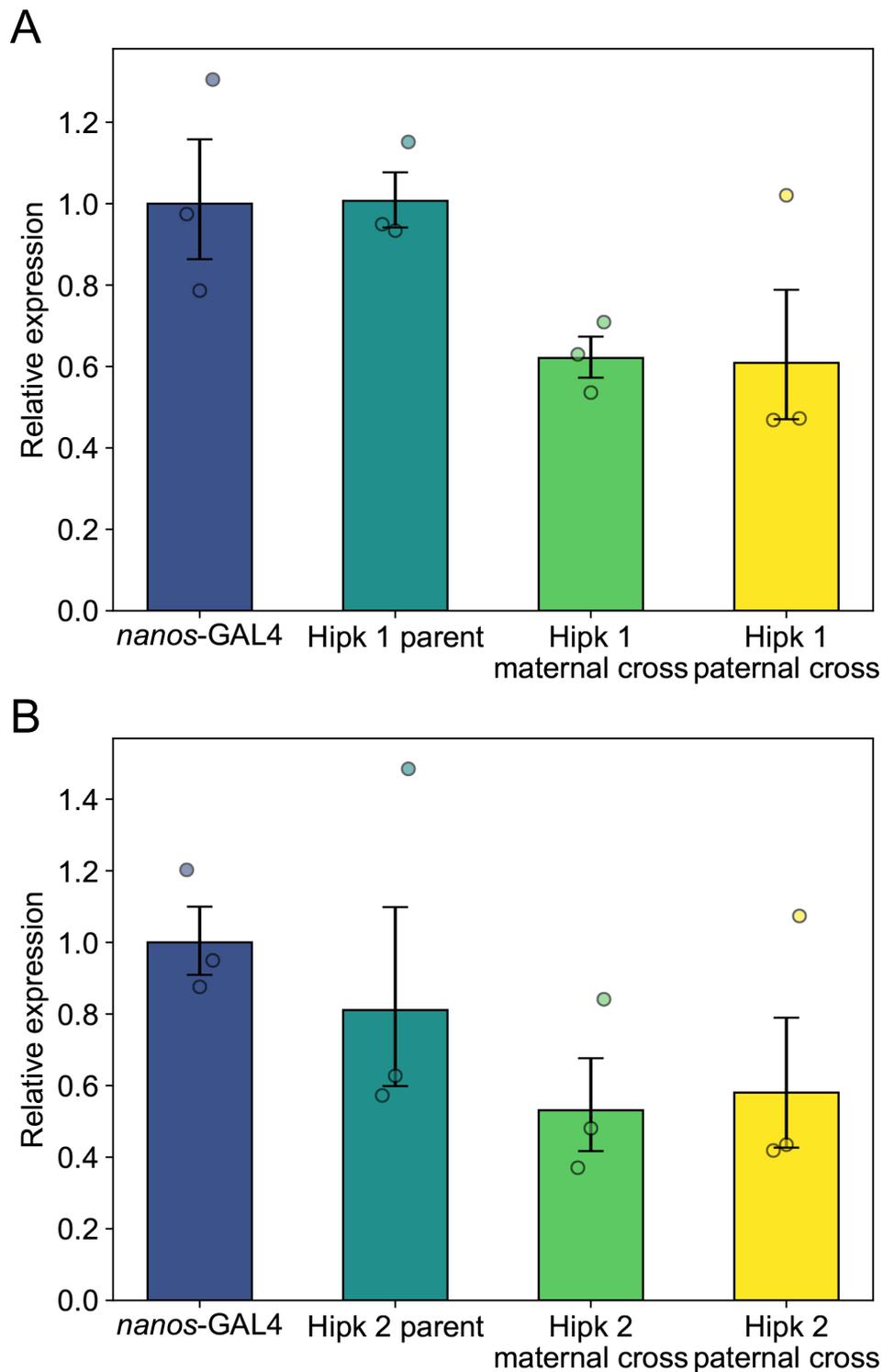
The male F<sub>1</sub> progeny, with the relevant gene of interest knocked down, from the crosses of all 23 RNAi lines with *nanos*-GAL4 were placed in vials with the female F<sub>1</sub> flies from the same cross for several days. Each vial was then examined to see if adult F<sub>2</sub> flies were produced. The only flies that displayed an obvious fertility defect were the F<sub>1</sub> progeny from the Hipk RNAi 2 maternal cross (Table 4.2). At this stage of the screening no formal fertility assays were set up or quantified but the F<sub>1</sub> flies produced noticeably few F<sub>2</sub> progeny. Many eggs (or embryos) were laid by the F<sub>1</sub> females but few of these hatched into larvae (Table 4.2).

### 4.2.3. Hipk transcript expression in the testes of Hipk RNAi knockdowns and parental stocks

Following these screens for major morphology and fertility defects, homeodomain interacting protein kinase (Hipk) was the strongest candidate as a gene crucial for spermatogenesis. Hipk is involved in many signalling pathways, including the Notch and JAK/STAT pathways (Lee et al., 2009; Tettweiler et al., 2019), and it is important for the development of eyes and wings, among other tissues (Lee et al., 2009; Huang et al., 2011). In *C. elegans*, the Hipk orthologue Hpk-1 is required for germline development (Berber et al., 2013), and HIPK4 in mice is required for spermiogenesis (Crapster et al., 2020). Studying the impact of Hipk knockdown in the testis further could shed light onto the role of HOW(S) and why HOW is vital for proper sperm development.

Prior to carrying out further studies into the impact of knocking down Hipk, qRT-PCR was conducted to quantify the expression of *hipk* mRNA in the testes of the parental fly lines and the knockdown progeny from the maternal and paternal RNAi crosses. The qRT-PCR results, which are relative to the *nanos*-GAL4 samples, indicate that average *hipk* transcript levels in the F<sub>1</sub> knockdown testes are decreased compared to the parental lines (Fig 4.5). In the fertility observations, the Hipk 2 maternal cross had a strong negative impact, the qRT-PCR results show that the testes from the Hipk 2 maternal cross had the lowest *hipk* expression (mean of 53% relative to *nanos*-GAL4). However, when using the Kruskal–Wallis *H* test to compare differences in *hipk* expression, neither sets of samples from the Hipk 1 or Hipk 2 knockdowns were significantly different (Fig 4.5; see Table A.4 in Appendix I for full test statistics). This is not surprising as the qRT-PCR results show large variation between the biological triplicates in the Hipk RNAi 2 samples, which could reflect normal biological variation. Overall, we were satisfied that on average *hipk* RNA abundance was lower in the testes of knockdown samples than any of the parental lines for both Hipk RNAi 1 and Hipk RNAi 2 fly lines.

4. Understanding the role of HOW(S) regulated events in germ cells



**Figure 4.5. Quantification of Hipk knockdown in the testes via qRT-PCR.** Bar graphs of the mean relative expression of *hipk* in testes for the parental lines and progeny of the two different RNAi lines crossed with the *nanos-GAL4* driver in both directions. The two RNAi lines are A) Hipk RNAi 1 and B) Hipk RNAi 2. Each set of data were normalised to the *nanos-GAL4 hipk* expression, and the data points of the three biological replicates for each sample are displayed. Error bars are SEM. None of the samples were significantly different from one another (see Table A.4).

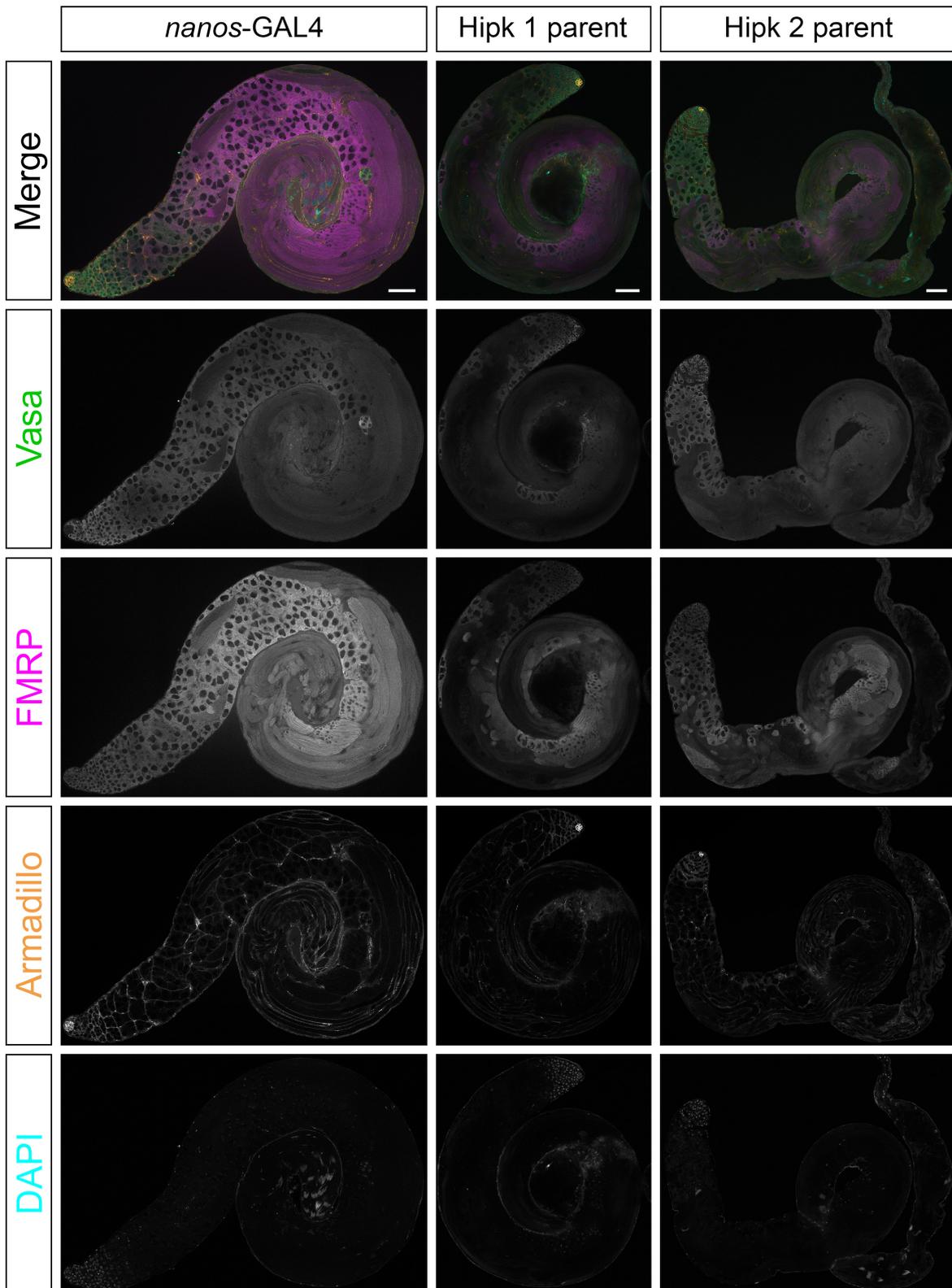
#### 4.2.4. Assessing testis morphology phenotypes of the conditional Hipk knockdowns

In the initial screening for major morphology defects light microscopy had been used, and testes of the male progeny from the Hipk 2 maternal cross were noticeably smaller (Table 4.2). To further probe this phenotype, more testes were dissected from all the Hipk knockdown crosses, and the parental stocks, to study the impact of Hipk knockdown in more detail with immunofluorescence confocal microscopy. As well as observing morphological changes, the presence of different germ cell populations were examined with the following antibodies: anti-Vasa (for spermatogonia), anti-FMRP (for spermatocytes and early spermatids), anti-Armadillo (for somatic cells). Additionally, nuclei were stained with DAPI. Tiled images, with the hub cells in plane, were taken of whole testes and the genotypes of the testes on each slide were not known at the time of imaging. Images were taken of 5–10 individual testes per sample, and 1 representative image is shown for each sample (Figs 4.6, 4.7, and 4.8).

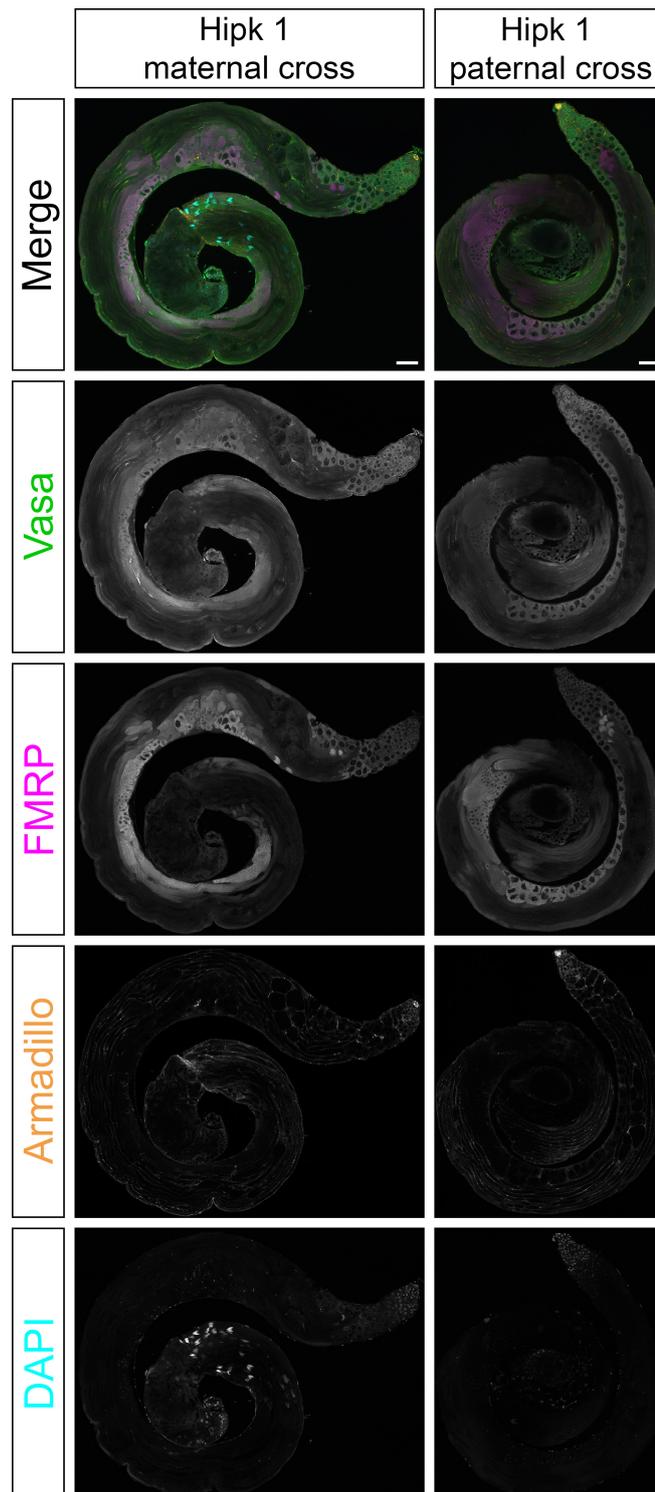
Testes from the three parental lines (*nanos*-GAL4, Hipk RNAi 1 and Hipk RNAi 2) were imaged alongside the testes from the Hipk knockdown flies for comparison between the morphology and cell populations present. Both Hipk RNAi parental lines appear similar to WT testes (Fig 4.2) with round apical tips and the presence of spermatogonia, spermatocytes and spermatids were observed (Fig 4.6, middle and right-hand panels). The testes from *nanos*-GAL4 flies also showed these three stages of differentiating germ cells, however, some of these also had apical tips that were narrow with a large number of hub cells (identified as the cluster of Armadillo positive cells at the apical tip; Fig 4.6, left-hand panels). Some of the Hipk RNAi parental testes also had hubs that seemed larger than WT but without pinching of the apical tip.

Several of the knockdown testes from Hipk RNAi 1 had pinched apical tips, in males from both the maternal and paternal crosses (Fig 4.7), although this appeared more consistently in the progeny from the Hipk 1 maternal crosses. In terms of cell types observed, germ cells from all three spermatogenesis phases were present and the hub appeared larger than WT but no more so than the hubs in the *nanos*-GAL4 parental testes (Figs 4.6 and 4.8).

4. Understanding the role of *HOW(S)* regulated events in germ cells



**Figure 4.6. Parental lines of the Hipk conditional knockdowns show all three phases of spermatogenesis.** Confocal microscopy images of testes stained with anti-FMRP (magenta), anti-Vasa (green), anti-Armadillo and counter-stained with DAPI (cyan). The testes were dissected from the parental lines used for conditional knock-down of Hipk: *nanos-GAL4* (left-hand panels), Hipk RNAi 1 (middle panels) and Hipk RNAi 2 (right-hand panels). Bar = 50  $\mu$ m.



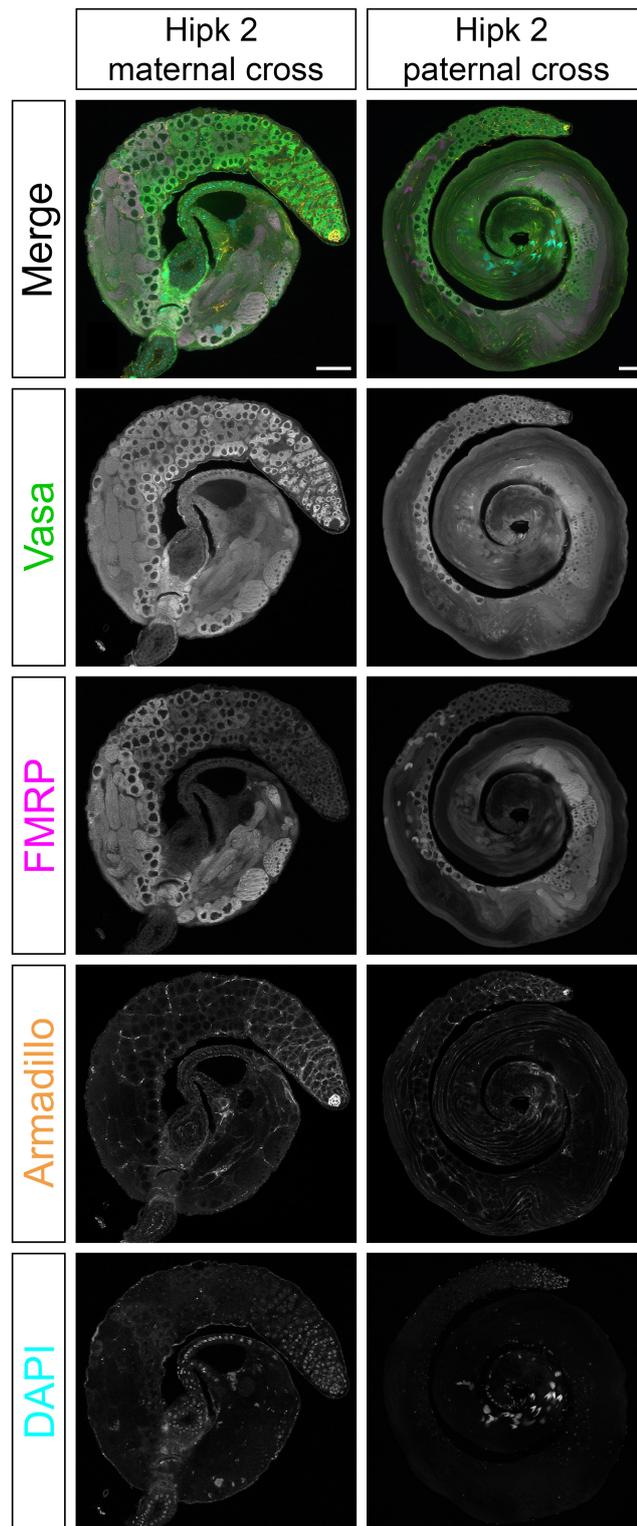
**Figure 4.7. Testes from conditional Hipk RNAi 1 knockdowns have populations of germ cells from all three phases of spermatogenesis.** Confocal microscopy images of testes stained with anti-FMRP (magenta), anti-Vasa (green), anti-Armadillo and counter-stained with DAPI (cyan). The testes were dissected from progeny of crosses between Hipk RNAi 1 flies and the *nanos*-GAL4 driver. The male progeny from the cross where the RNAi line was female in the cross is referred to as the ‘Hipk 1 maternal cross’ (left-hand panels) and when the RNAi line was male in the cross the resulting progeny sample is referred to as ‘Hipk 1 paternal cross’ (right-hand panels). Bar = 50  $\mu$ m.

#### 4. Understanding the role of *HOW(S)* regulated events in germ cells

The aforementioned small testes of the males from the *Hipk* RNAi 2 maternal cross was evident in the confocal images of these testes (Fig 4.8, left-hand panels). Typically, when a testis has remained in its characteristic spiral shape it is long enough to coil round twice (see Fig 4.2 for a WT example). In the *Hipk* 2 maternal cross testes were often only coiled around once, even when very tightly coiled (Fig 4.8). The apical tips were also pinched in shape and the hub appeared large. Spermatocytes and elongating spermatids with short tails were present, made visible by the *Vasa* and *FMRP* staining. However, either no or very few late spermatids, which have longer tails, were observed for most of these testes. Lastly, few images had nuclei that had developed into the elongated shape at the basal end of the testis. Overall, the testes from the *Hipk* 2 maternal cross produce no, or few, mature sperm cells.

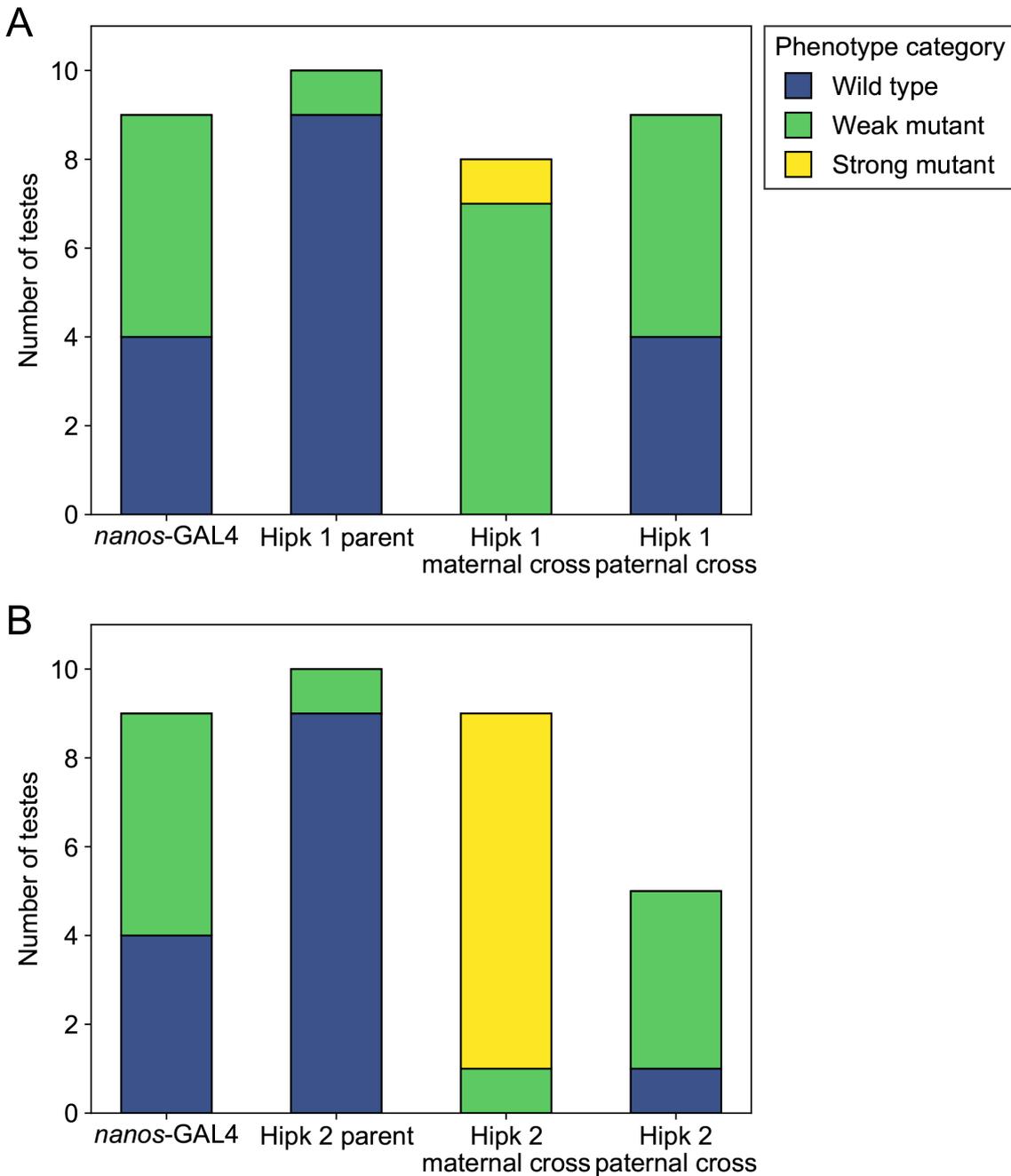
The testes from the *Hipk* 2 paternal cross did not show severe defects in morphology (Fig 4.8, right-hand panels). However, as well as a pinched apical tip, they did, in general, appear much thinner for a greater portion of the testis than either of its parental lines (Fig 4.6). Unlike the *Hipk* 2 maternal cross testes, the paternal ones had late spermatids with long tails and elongated nuclei, and when coiled tightly were long enough for two complete spirals (Fig 4.8).

Across these microscopy images there were a combination of morphological and cellular differences. To compare the samples beyond the descriptions above, a categorising approach was taken to examine both morphological and cellular differences simultaneously. Five volunteers independently classified each image, without knowing the genotypes, into one of the three phenotype categories. The following features were highlighted to define each category: 1) ‘wild type’ — testes with rounded apical tips and late spermatids, 2) ‘weak mutant’ — testes with pinched apical tips and late spermatids, 3) ‘strong mutant’ — small testes with pinched apical tips and no late spermatids. The mode assignment from the five volunteers was taken as the consensus and final categorisation of each image (see Materials and Methods 2.5.3.2).



**Figure 4.8. Testes from conditional Hipk RNAi 2 maternal knockdowns are small and do not have late spermatids.** Confocal microscopy images of testes stained with anti-FMRP (magenta), anti-Vasa (green), anti-Armadillo and counter-stained with DAPI (cyan). The testes were dissected from progeny of crosses between Hipk RNAi 2 flies and the *nanos*-GAL4 driver. The male progeny from the cross where the RNAi line was female in the cross is referred to as the ‘Hipk 2 maternal cross’ (left-hand panels) and when the RNAi line was male in the cross the resulting progeny sample is referred to as ‘Hipk 2 paternal cross’ (right-hand panels). Bar = 50  $\mu$ m.

4. Understanding the role of HOW(S) regulated events in germ cells



**Figure 4.9. Maternal Hipk RNAi 2 knockdown produce strong mutant phenotypes in the testes of their progeny.** Bar charts showing the number of confocal images for each sample classified into one of three categories: wild type (blue), weak mutant (green) and strong mutant (yellow). The confocal images of whole testes were from 7 samples: the *nanos*-GAL4 parental line (represented in A and B), the Hipk RNAi 1 parental line and the two Hipk 1 knockdowns (A), the Hipk RNAi 2 parental line and the two Hipk 2 knockdowns (B). Images were taken with the hub cells in plane, and were categorised by 5 individuals, the consensus for each of the images gave the final category allocation. The majority of the Hipk 2 maternal images were classed as strong phenotype (B), and the Hipk 1 maternal sample was the only other one to not have any images classed as ‘wild type’ (A).

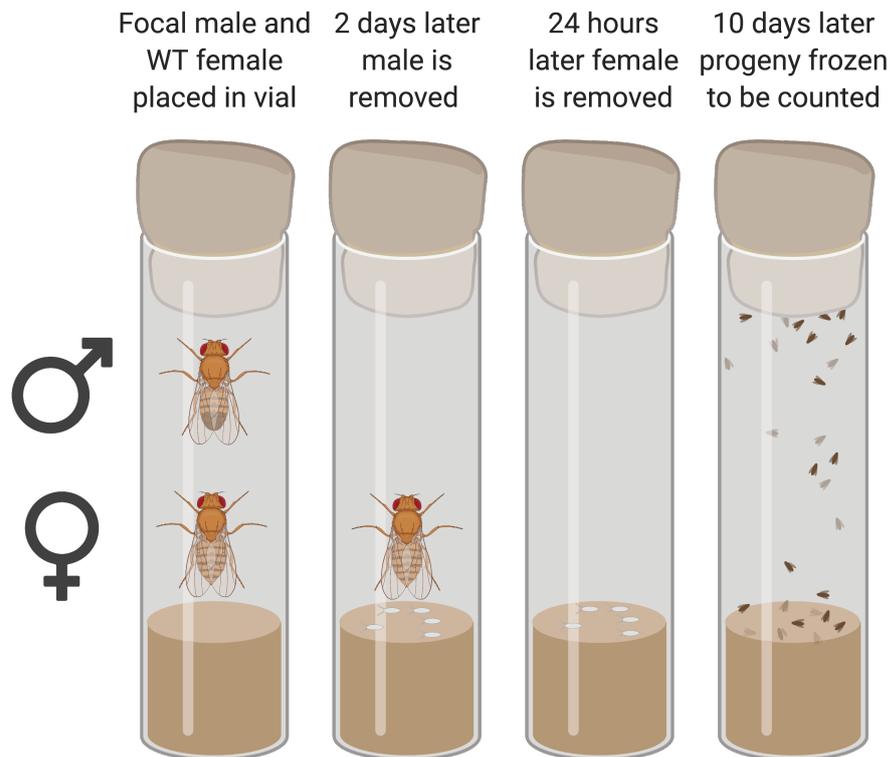
The majority of testis images from both of the Hipk RNAi parental lines were categorised as ‘wild type’ (Fig 4.9). While 5 out of 9 of the *nanos*-GAL4 testes were classed as ‘weak mutant’ and the rest as ‘wild type’, reflecting the earlier observations of the pinched apical tip. The Hipk 1 paternal cross had the same number of images classified as ‘wild type’ and ‘weak mutant’ as the *nanos*-GAL4 testes. However, the Hipk 1 maternal cross images were consistently classified as ‘weak mutant’ and one testis was in the ‘strong mutant’ category (Fig 4.9A). Similarly, the Hipk 2 maternal cross was more different from the parental lines than Hipk 2 paternal cross (Fig 4.9B). Suggesting that the direction of the cross is having an impact on how severely the knockdown affects the male progeny. The majority (8 out of 9) of the Hipk 2 maternal crosses images were classed as ‘strong mutant’. Only one other image (from the Hipk 1 maternal cross) was placed into this category, reiterating earlier observations that the testes from the progeny of the Hipk 2 maternal cross are morphologically different from the other samples.

After observations of large hubs in some of the samples, stacked images that captured the full depth of the hub were taken at the apical tips. Stacked images were taken of 5–8 testes per sample, and the number of hub cells were counted. All the transgenic fly lines had more hub cells than WT testes (Fig 4.10). However, neither of the Hipk 1 and 2 knockdown samples were significantly different to their parental lines (Fig 4.10; Table A.5 in Appendix I).



#### 4.2.5. The impact of conditional knockdown of Hipk on male fertility

Knocking down Hipk in the testes caused morphological changes, particularly in the males from the Hipk 2 maternal cross. To study if these differences had an impact on the fertility of these males 3-day fertility assays were carried out. In this experiment a male fly with a genotype of interest (a focal male) was placed in a vial with a WT female fly. Unlike the sperm competition assay carried out in Chapter 3 (section 3.2.2), the flies were not watched to confirm matings, instead they pairings were placed in the vials for 2 days. 2 days was deemed sufficient time for mating to occur, as 91% of WT males mated within 4 hours in the sperm competition assay (Table 3.1). After 2 days, the focal males were removed and the female was left to lay embryos for one more day. After the female was removed, the vials remained in the 25 °C room for 10 days before being frozen and the adult progeny counted (Fig 4.11).

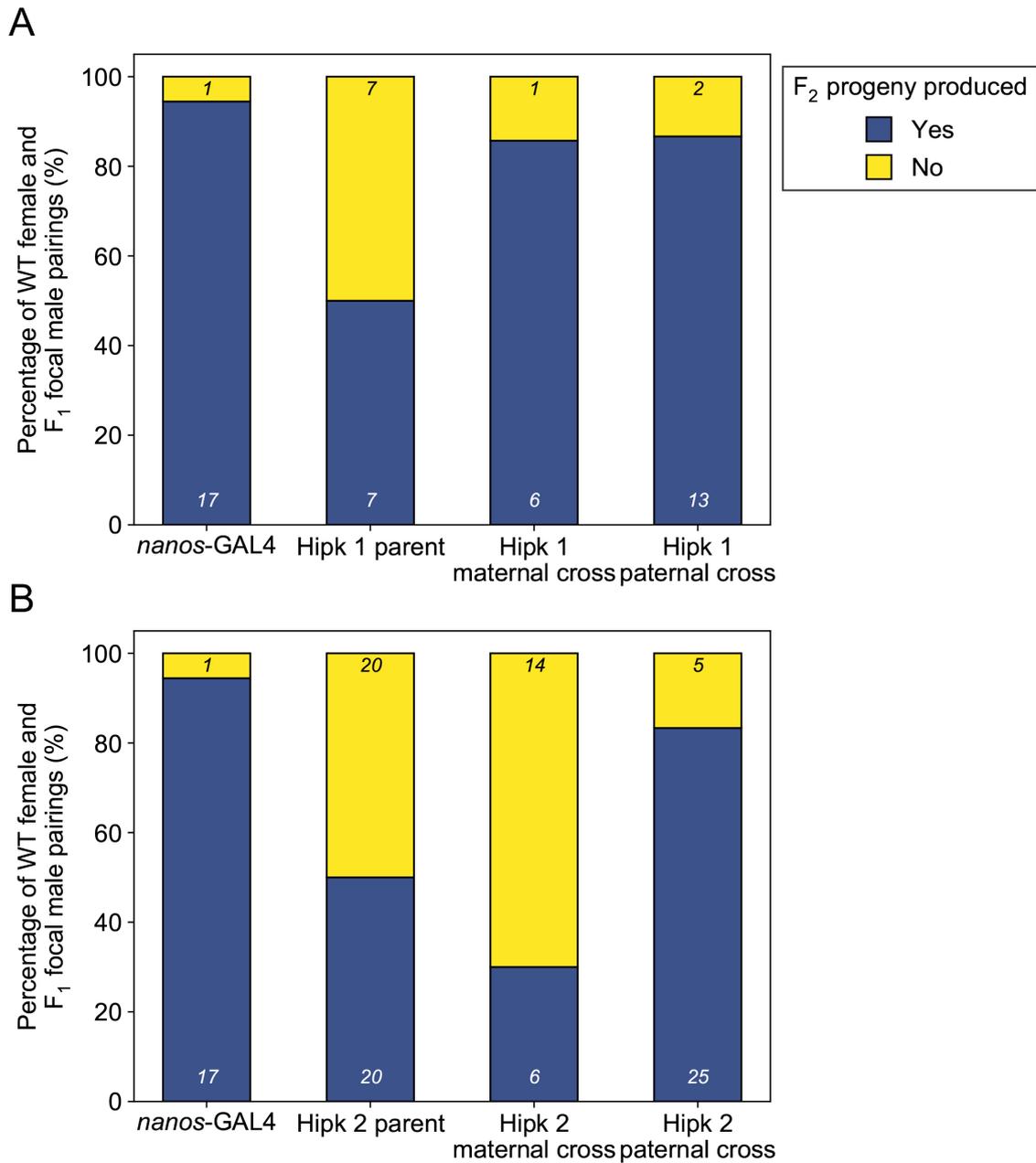


**Figure 4.11. Schematic of 3-day fertility assay.** A focal male is placed in a vial with a WT female for 2 days before being removed. The female is given a further 24 hours to lay embryos before also being removed from the vial. 10 days later the vials are frozen and the number of adult progeny are counted.

#### 4. Understanding the role of *HOW(S)* regulated events in germ cells

The focal males used in the 3-day fertility assay were  $F_1$  males from 7 different  $F_0$  crosses. There were three parental line crosses: *nanos*-GAL4, Hipk RNAi 1 and Hipk RNAi 2, where the  $F_1$  males have the same genotype as their  $F_0$  parents. To achieve knockdown, the two Hipk RNAi lines were crossed with the *nanos*-GAL4 driver in both directions. The first observation made was the percentage of the WT females which produced  $F_2$  progeny with each focal male type. As the vials were not watched to confirm matings, if no  $F_2$  were produced from a pairing this could be because either the  $F_1$  focal male is sterile or because no mating occurred.

Fertility was assessed both as whether or not a pair produced offspring and if they did, the number of offspring produced. Differences in proportion of pairs producing offspring were analysed with chi-squared tests. For the Hipk 1 set of crosses, the proportion of pairs not producing offspring differed across genotypes ( $\chi^2 = 10.063$ ,  $df = 3$ ,  $p = 0.014$ ; Fig 4.12A). The *nanos*-GAL4 parental line had the highest percentage of vials that produced  $F_2$  progeny. With the other two parental lines, Hipk 1 parent and Hipk 2 parent, only 50% of the vials produced any  $F_2$  progeny (Fig 4.12). The Hipk 1 knockdown males from both the maternal and paternal crosses had a similar progeny producing rate as the *nanos*-GAL4 (Fig 4.12A). Thus, the difference between the Hipk 1 set of crosses is driven by the Hipk 1 parent results (Fig 4.12A; see Table A.6 for pairwise test statistics). For the Hipk 2 knockdowns, the Hipk 2 maternal cross resulted in the smallest percentage of vials (30%) with  $F_2$  progeny out of all 7 focal male types (Fig 4.12B). While the Hipk 2 paternal cross results were not dissimilar to the *nanos*-GAL4 parental line (Fig 4.12B). Again, chi-squared tests were carried out and the observed frequency was significantly different from the expected values ( $\chi^2 = 25.189$ ,  $df = 3$ ,  $p < 0.001$ ). While pairwise chi-squared tests show contributions to this difference from both the Hipk 2 parent and the Hipk 2 maternal cross, these two results are not significantly different to one another ( $\chi^2 = 1.434$ ,  $df = 1$ ,  $p = 0.231$ ; see Table A.6 for pairwise test statistics).



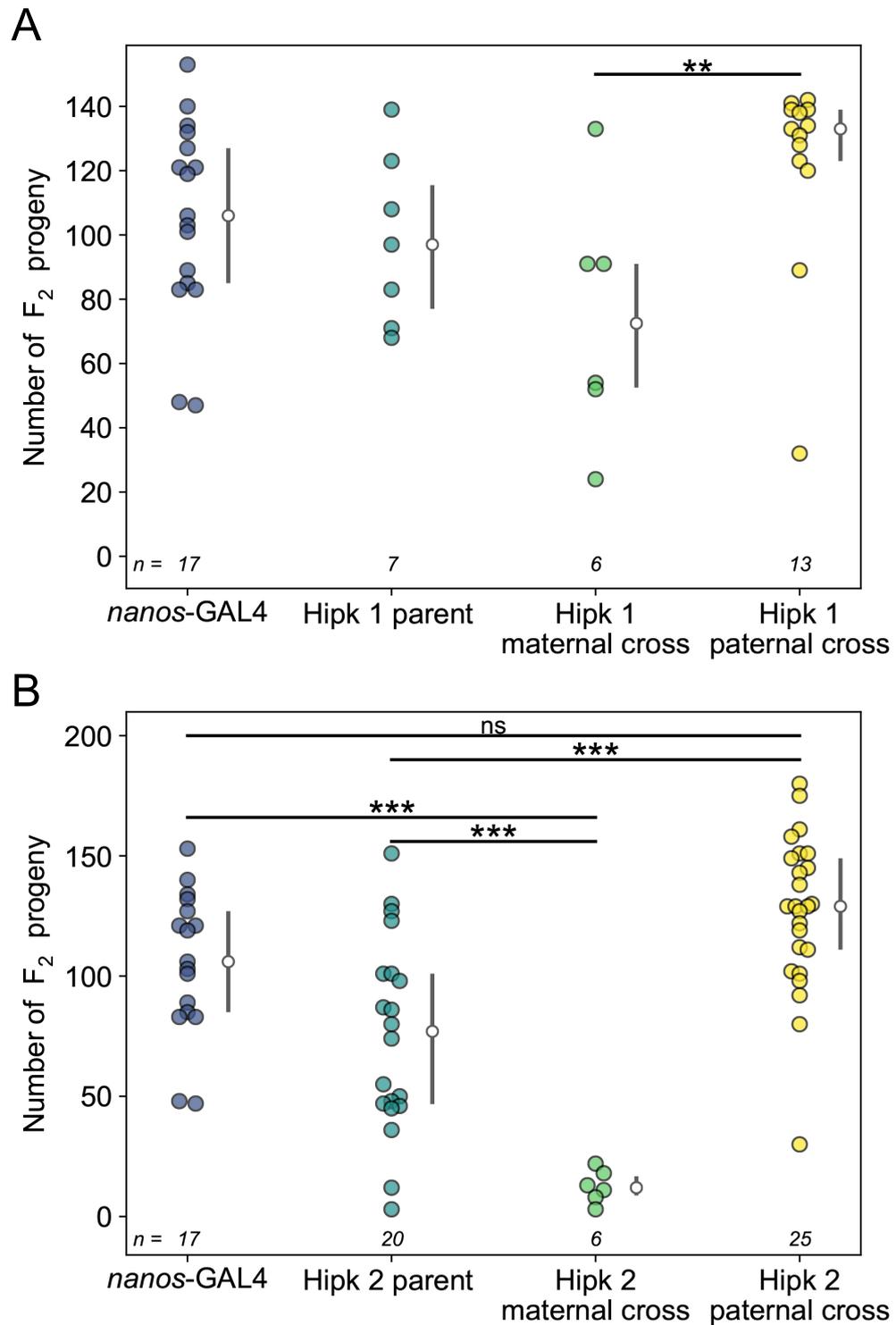
**Figure 4.12.** The percentage of pairings that produced  $F_2$  progeny was lowest between WT females and the  $F_1$  males from the Hipk 2 maternal cross. Stacked bar chart showing the percentage of pairings between WT females and  $F_1$  focal males that did (blue,  $n$  number in white text) or did not (yellow,  $n$  number in black text) produce  $F_2$  progeny. The seven types of focal males were: A) *nanos*-GAL4 parental line, Hipk RNAi 1 parental line, and the  $F_1$  Hipk 1 knockdown males from two crosses (maternal and paternal) between *nanos*-GAL4 and Hipk RNAi 1 parental lines. B) *nanos*-GAL4 parental line, Hipk RNAi 2 parental line, and the  $F_1$  Hipk 2 knockdown males from two crosses (maternal and paternal) between *nanos*-GAL4 and Hipk RNAi 2 parental lines.

#### 4. Understanding the role of *HOW(S)* regulated events in germ cells

The adult F<sub>2</sub> progeny from each mating pair in the 3-day fertility assay were counted. The males from the Hipk 1 maternal cross had the lowest median number of progeny compared to its parental lines and the Hipk 1 paternal cross. However, it was only significantly different to the Hipk 1 paternal cross sample and not to either of the parental lines. Similarly, the Hipk 1 paternal cross was not significantly different to either of its parental lines (Table 4.3; Fig 4.13A). This suggests that the Hipk RNAi 1 knockdown did not have an effect on male fertility.

For the Hipk RNAi 2 knockdowns, the F<sub>1</sub> males from the Hipk 2 maternal cross produced significantly fewer progeny than either of the parental genotypes (Table 4.3; Fig 4.13B). Not only was the median very low but the maximum number of progeny from a single vial was 22 flies. And similar to the Hipk 1 paternal cross, the F<sub>1</sub> males from the Hipk 2 paternal cross had the highest median number of F<sub>2</sub> progeny, which was significantly more than the Hipk 2 parent data but not the *nanos*-GAL4 parental line (Table 4.3; Fig 4.13B).

Altogether, the data from the 3-day fertility assay suggests that the ‘weak mutant’ phenotype observed in the morphology analysis (Fig 4.9) does not correlate to any fertility defects in the 3-day mating assay. The ‘strong mutant’ phenotype, with a lack or depletion of late spermatids, was assigned to most of the testes images from the Hipk 2 maternal cross knockdown. Unsurprisingly, this coincides with impaired fertility of these male flies. Many of these flies produced no progeny in the 3-day assay (Fig 4.12B), and of the flies that did produce progeny the median number sired was very low (Fig 4.13B).



**Figure 4.13. Fertile Hipk 2 maternal cross males sire very few progeny.** Scatter bar graphs of the number of progeny from matings between WT female flies and different focal males: *nanos*-GAL4 (A and B), Hipk 1 parental line and the Hipk 1 knockdowns (A), Hipk 2 parental line and the Hipk 2 knockdowns (B). White circles represent the median, and the grey lines represent the interquartile ranges. See Table 4.3 for details of the statistical tests.  $p < 0.05$  (\*).  $p < 0.05$  (\*),  $p < 0.01$  (\*\*),  $p < 0.001$  (\*\*\*)

4. Understanding the role of HOW(S) regulated events in germ cells

**Table 4.3. Statistics for the number of F<sub>2</sub> progeny produced.** The data for the Hipk 1 sample set (the two parental lines and two knockdown crosses) were all normal except for the Hipk 1 paternal cross (tested via the Shapiro–Wilk test). Thus, the non-parametric Kruskal–Wallis  $H$  test was used to test for differences between the knockdown samples for the Hipk RNAi 1 sample set. The test statistic ( $H$ ) with the degrees of freedom in subscript are reported with the  $p$ -value. Post-hoc pairwise comparisons were carried out using Dunn’s test, the standardised test statistic ( $Z$ ) with the degrees of freedom in subscript are reported with the Bonferroni corrected  $p$ -value. The data for the Hipk 2 sample set were all normal, so a one-way ANOVA was performed. The test statistic ( $F$ ) with the degrees of freedom for the between groups and within groups in subscript are reported with the  $p$ -value. Post-hoc pairwise comparisons were carried out using Tukey’s honestly significant difference test, the mean difference ( $M$ ) is reported with the adjusted  $p$ -value.

Sample comparison	Test statistic	$p$ -value
Hipk 1 sample set	$H_3 = 9.007$	0.029
Hipk 1 maternal cross and <i>nos</i> -GAL4	$Z_1 = 8.417$	0.158
Hipk 1 maternal cross and Hipk 1 parent	$Z_1 = 5.631$	0.420
Hipk 1 maternal cross and Hipk 1 paternal cross	$Z_1 = 17.109$	0.006
Hipk 1 paternal cross and <i>nos</i> -GAL4	$Z_1 = 8.692$	0.060
Hipk 1 paternal cross and Hipk 1 parent	$Z_1 = -11.478$	0.051
<i>nos</i> -GAL4 and Hipk 1 parent	$Z_1 = -2.786$	0.621
Hipk 2 sample set	$F_{3,64} = 22.699$	$3.994 \times 10^{-10}$
Hipk 2 maternal cross and <i>nos</i> -GAL4	$M = -92.912$	$9.597 \times 10^{-7}$
Hipk 2 maternal cross and Hipk 2 parent	$M = -62.500$	$8.343 \times 10^{-4}$
Hipk 2 maternal cross and Hipk 2 paternal cross	$M = -113.980$	$1.271 \times 10^{-9}$
Hipk 2 paternal cross and <i>nos</i> -GAL4	$M = 21.068$	0.194
Hipk 2 paternal cross and Hipk 2 parent	$M = 51.480$	$1.537 \times 10^{-5}$
<i>nos</i> -GAL4 and Hipk 2 parent	$M = 30.412$	0.036

### 4.3. Discussion

#### 4.3.1. Screening RNAs bound by HOW(S) for function in spermatogenesis

12 RNAs that are potentially bound by HOW(S) in the *D. melanogaster* testis were selected from the 121 transcripts and 343 genes that were identified in the HOW(S) RIP-seq (Table 4.1). The expression of these 12 genes were knocked down via RNAi in GSCs and early spermatogonia, when HOW(S) is also expressed, to investigate whether they are essential for normal spermatogenesis. The initial screen indicated that knockdown of one of these genes, *hipk*, caused morphological and fertility defects (Table 4.2). However, this is not to say that the regulation of other RNAs identified in the RIP-seq experiment do not have roles in spermatogenesis, as this initial screening only detected large or obvious defects. Further, the RIP-seq results indicate that HOW(S) may bind to a large number of RNAs and so the synergistic effect of this regulation would not be picked up on in these single gene knockdown experiments.

From this initial screening using light microscopy to examine morphology and mating between the F<sub>1</sub> knockdown progeny, *hipk* stood out as a gene important for spermatogenesis. The testis from the Hipk RNAi 2 knockdown appeared to be much smaller than WT testis and the F<sub>1</sub> female flies that were left to mate with the F<sub>1</sub> males from the Hipk RNAi 2 maternal cross laid many eggs but few hatched (Table 4.2), this suggested that Hipk knockdown might specifically be affecting male gametogenesis. *Hipk* was both an enriched gene and had an enriched transcript (Hipk-RA). All of the *hipk* 3'-UTRs have two of the RIP-seq identified 3'-UTR motif — (A/U/G)CUAAC — but Hipk-RA has the most with four sites. Only 3 other transcripts from the HOW(S) RIP-seq had 4 or more of these sites (Fig 3.26). The specific binding of HOW(S) to the *hipk* transcripts, and others, is explored further in Chapter 5.

#### 4.3.2. Hipk function in spermatogenesis

Hipk is a signal transduction kinase important in the development of several organs, including eye (Lee et al., 2009) and wing development (Huang et al., 2011). It is involved in many signalling pathways including: Notch (Lee et al., 2009), Hippo (Chen and Verheyen, 2012; Poon et al., 2012), and the JAK-STAT pathway (Tettweiler et al., 2019). The latter pathway is crucial to GSC homeostasis, as the secretion of Unpaired from the hub cells activates the JAK-STAT pathway in the adjacent GSCs which is required for their self-renewal (Kiger et al., 2001; Tulina and Matunis, 2001). Additionally, it has recently been shown in mice that HIPK4, a mammalian specific Hipk family member, is required for spermiogenesis (Crapster et al., 2020). Thus, it is unsurprising that knockdown of Hipk in the GSCs and early spermatogonia would disrupt normal sperm development in *D. melanogaster*.

#### 4. Understanding the role of HOW(S) regulated events in germ cells

To knockdown *Hipk* expression in the testis two different *Hipk* RNAi lines were used that had differing impacts on testis morphology and fertility. *Hipk* RNAi line 1 targets three of the four *hipk* transcripts and was generated using the pVALIUM22 vector, which, at the time of use, had been understood to work well in the germline, however, it is actually best suited for the female germline only (see Methods 2.1.2). While *Hipk* RNAi 2 targets all four transcripts, and was generated with the pVALIUM20 vector which is suited for the soma, male and female germlines. The maternal knockdown cross using *Hipk* RNAi 2 resulted in a strong fertility defect (Fig 4.13), and the *Hipk* 1 maternal cross appeared to have a larger effect on testis morphology than the *Hipk* 1 paternal cross (Fig 4.7A) even though both samples had the same mean *hipk* mRNA expression (Fig 4.5A), suggesting that the directionality of the RNAi crosses is important. Of all the *Hipk* knockdown samples, the *Hipk* RNAi 2 maternal cross also had the lowest *hipk* expression in the testis, as measured by qRT-PCR (Fig 4.5). However, the *hipk* mRNA expression in the knockdown samples was not significantly lower than any of the parental lines, probably due to the large degree of biological variation, especially with the *Hipk* 2 samples. For the RNAi parental lines, the variation could come from the leaky expression of the UAS-RNAi hairpin. While the RNAi hairpin targeting *hipk* should not be expressed without the GAL4 transcription factor present to bind to the UAS, it was observed that the testes from UAS-HOW-S-HA flies were expressing low levels of HOW(S)-HA protein without being crossed to a GAL4 driver (Fig 3.13B). Variation between triplicates could also come from sensitivity of qRT-PCR, genomic DNA contamination was seen in some of the *Hipk* 2 parental and knockdown samples (but not in any of the *Hipk* 1 samples), this was accounted for in the qRT-PCR experimental design by using exon junction spanning primers for *hipk* and the two reference genes (Table 2.5), but could still have interfered with the quantification. Consequently, it is challenging to correlate the strength of the knockdown with the strength of the phenotypic defects observed, and more consistent knockdown or knockout of *Hipk* in the testes could help to clarify its role in spermatogenesis.

The strong morphological phenotype, characterised by a thin apical tip and few or no late spermatids, coincided with a dramatic decrease in male fertility in the *Hipk* 2 maternal knockdown flies (Figs 4.8, 4.12 and 4.13B). The decrease in *hipk* expression, assuming that there were no off-target effects of the RNAi, in these flies did not always result in a total block of spermatogenesis. While the majority of these flies produced no progeny with WT female flies (Fig 4.12B), there were some that were fertile, though these produced very few progeny (Fig 4.13B). From the confocal microscopy images, spermatogonia, spermatocytes and early spermatids were clearly present in the *Hipk* 2 maternal cross testes but there were few full-length late spermatids (Fig 4.8), which probably contributed to the small size of these testes. This suggests that the decrease in *Hipk* could be important for the later stages of spermiogenesis, similar to HIPK4 function in mice (Crapster et al., 2020). However, these testes also were thin or ‘pinched’ at the apical tip (Fig 4.8). It is unclear from the images available if this is caused by a low number of GSCs and early spermatogonia not filling up the normal amount of space, or if the opposite is occurring where there are more small spermatogonial cysts (2- or 4-cell cysts) than normal pushing the larger ones (8-

and 16-cell cysts) further away from the apical tip. If these morphological changes are from a loss of Hipk, more investigations into the cell types and quantities of these cell types present in these testes would improve our understanding of the function of Hipk in spermatogenesis.

The weak morphological phenotype, characterised by a thin apical tip and the presence of late spermatids, did not appear to impact fertility (Figs 4.7 and 4.13). However, it is worth noting that the 3-day mating assay was carried out using young, unmated males and that WT aged, mated males have a thin testis phenotype (Boyle et al., 2007; Chang et al., 2019). This thin phenotype in aged flies comes with a decrease in the number of germ cells in the testis, as there is a decrease in Unpaired signal from the hub cells to the germ cells with age and this is correlated with a decrease in the number of GSCs in the testis and a slower rate of GSC division (Boyle et al., 2007). This presumably contributes to the decrease in fertility seen in older male flies (Snook and Promislow, 2003), and this depletion in mature sperm with age is also exacerbated by mating (Chang et al., 2019; Prowse and Partridge, 1997). Thus, further studies examining whether the weak morphological phenotype is caused by a smaller number of GSCs and/or a slower rate of cell division is warranted. Additionally, studies into other aspects of fertility that incorporate age and/or multiple matings could reveal fertility defects in the Hipk knockdown flies that were not observed here.

### 4.3.3. Conclusions

In summary, this chapter reports on the knockdown of *hipk*, an RNA which was identified as bound by HOW(S) in Chapter 3, in *D. melanogaster* testis using two independent RNAi lines. The maternal cross from the Hipk RNAi line 2 resulted in male flies with small testes that were either sterile or produced very few progeny when mated with female WT flies. Due to the differences between the Hipk knockdown flies, additional work is needed to verify this effect is from Hipk knockdown alone. If Hipk is necessary for normal spermatogenesis, then this could provide part of the explanation as to how HOW is essential for spermatogenesis. Establishing a direct link between HOW(S) binding to *hipk* and the effect on *hipk* mRNA expression will be required in further work.



# 5. Characterising the molecular mechanism of RNA binding by HOW

## 5.1. Introduction

RNA binding proteins are diverse in their structures, functions and modes of binding (Hentze et al., 2018; Corley et al., 2020). Proteomics and transcriptomics based approaches have hugely advanced and expanded our understanding of RBPs in the last decade (Hentze et al., 2018; Van Nostrand et al., 2020). However, omics based methods can have low signal to noise ratios and most cannot measure binding constants, unlike lower throughput methods. Experiments that measure direct binding, such as electrophoretic mobility shift assays (EMSA) and fluorescence anisotropy (FA), are well-established methods to calculate binding constants and allow one to probe the kinetics of RBP–RNA interactions. Additionally, solving the structures of RBPs, especially when bound to RNA, has improved our knowledge of mechanisms and sequence specificities of RNA binding (Corley et al., 2020). For example, there is now at least one high resolution protein structure for each STAR subfamily, which has improved our understanding of how the different regions of the STAR domain impact the sequence specificity of the different, but related, proteins (Liu et al., 2001; Teplova et al., 2013; Feracci et al., 2016).

A range of binding experiments, including EMSAs, FA and CLIP-seq, have been carried out on the STAR proteins to examine how these proteins specifically bind RNA (Table 1.1). These have found that the proteins in the SF1 and Quaking subfamilies have very similar consensus binding sequences, and most of the proteins bind these sequences in the low nanomolar range (Garrey et al., 2006; Carmel et al., 2010). *S. cerevisiae* BBP, orthologue of human SF1, has an optimal heptamer sequence, UACUAAC, which it binds in the low nanomolar affinity range (Garrey et al., 2006). While SF1’s optimal binding site is a very similar, but less constrained, six nucleotide sequence: ACUNAC (Corioni et al., 2011). Similar to SF1, the optimal binding sequences for the Quaking subfamily proteins are typically hexamer sequences, NA(C/A)U(A/C)A, where the preference for the first position and the preferences between the adenosines and cytidines in the third and fifth positions vary between the different members of the Quaking subfamily (Table 1.1; Israeli et al., 2007; Carmel et al., 2010). While in the Sam68 subfamily, binding and structural experiments carried out on Sam68 and T-STAR revealed that the QUA2 region, which the other subfamilies use to extend their RNA binding surface, do not interact with RNA (Feracci et al., 2016). Thus, the optimal binding sequences for these Sam68 subfamily proteins is the much shorter trinucleotide sequence (A/U)AA (Feracci et al., 2016).

Dimerisation is another characteristic of STAR proteins that is important for their RNA binding. The Quaking and Sam68 subfamilies dimerise via the QUA1 regions of their

## 5. Characterising the molecular mechanism of RNA binding by HOW

STAR domains (Beuck et al., 2010; Meyer et al., 2010), which the SF1 proteins do not have. This dimerisation has been shown to enhance RNA binding of Quaking (Beuck et al., 2012), HOW (Nir et al., 2012), Sam68 and T-STAR (Feracci et al., 2016). However, the details of how STAR proteins bind bipartite binding sites are disputed. SELEX experiments indicated that several of the STAR proteins bind their full consensus sequences with a nearby half-site 2–20 nucleotides away, and that mutations to either of these sites were detrimental to binding (Galarneau and Richard, 2005; Galarneau and Richard, 2009). While FA and EMSAs have shown that STAR proteins can bind a single full consensus site with nanomolar affinity, the affinity increases with a second full site 10 or more nucleotides away, depending on the protein (Carmel et al., 2010; Feracci et al., 2016).

Previous experiments investigating HOW’s binding to the 3’-UTRs of *stripe* and *dpp* transcripts, found that HOW can specifically bind the pentamer ACUAA (Israeli et al., 2007). From the experiments with these two RNAs, this pentamer was termed the HOW response element (HRE). These experiments were done in a ‘dipping’ style experiment, where biotinylated RNA was coupled to streptavidin beads and then different concentrations of HOW(L) (200, 20 or 2 nM) was incubated with the beads, binding was then verified by western blotting. It was found that, like GLD-1, HOW prefers cytidine rather than an adenosine in the second position of the HRE (Israeli et al., 2007). Additionally, the only time HOW was detected at the 2 nM concentration was when HOW was incubated with the ACUAA pentamer was in the loop, of at least 12 nucleotides, of a stem-loop structure (Israeli et al., 2007). However, these observations of HOW’s affinity for RNA were not quantified, as the equilibrium dissociation constant ( $K_D$ ) could not be calculated from this experimental design.

HOW binding was also explored in a high throughput *in vitro* RNAcompete experiment, which looked at heptamers, and defined the HOW consensus heptamer as NCUAACN, with a slight preference for first nucleotide to be an adenosine (Ray et al., 2013). This difference in the first nucleotide position and the strong preference for a second cytidine in the sixth position (Ray et al., 2013), highlights that the HRE might not be the definitive optimal binding sequence for HOW. Though similar to the dipping experiments, RNAcompete does not calculate the affinity of RBPs for different RNA sequences.

In Chapter 3 two different motifs were enriched in the 5’- and 3’-UTRs of the 121 transcripts bound by HOW(S) in *D. melanogaster* male germ cells. The motif found in the 3’-UTRs, (A/U/G)CUAAC, is very similar to the HRE pentamer and the heptamer defined by RNAcompete. The 5’-UTR motif, GCG(A/U)G, is very different to the HRE and to the other STAR protein binding sequences. To understand how these sequences drive HOW(S)–RNA interactions, this chapter aims to establish HOW’s affinity for different RNAs based on the motifs and transcripts identified from the HOW(S) RIP-seq.

## 5.2. Results

### 5.2.1. Optimisation of expression and purification of HOW's STAR domain

#### 5.2.1.1. Defining the STAR domain for protein purification

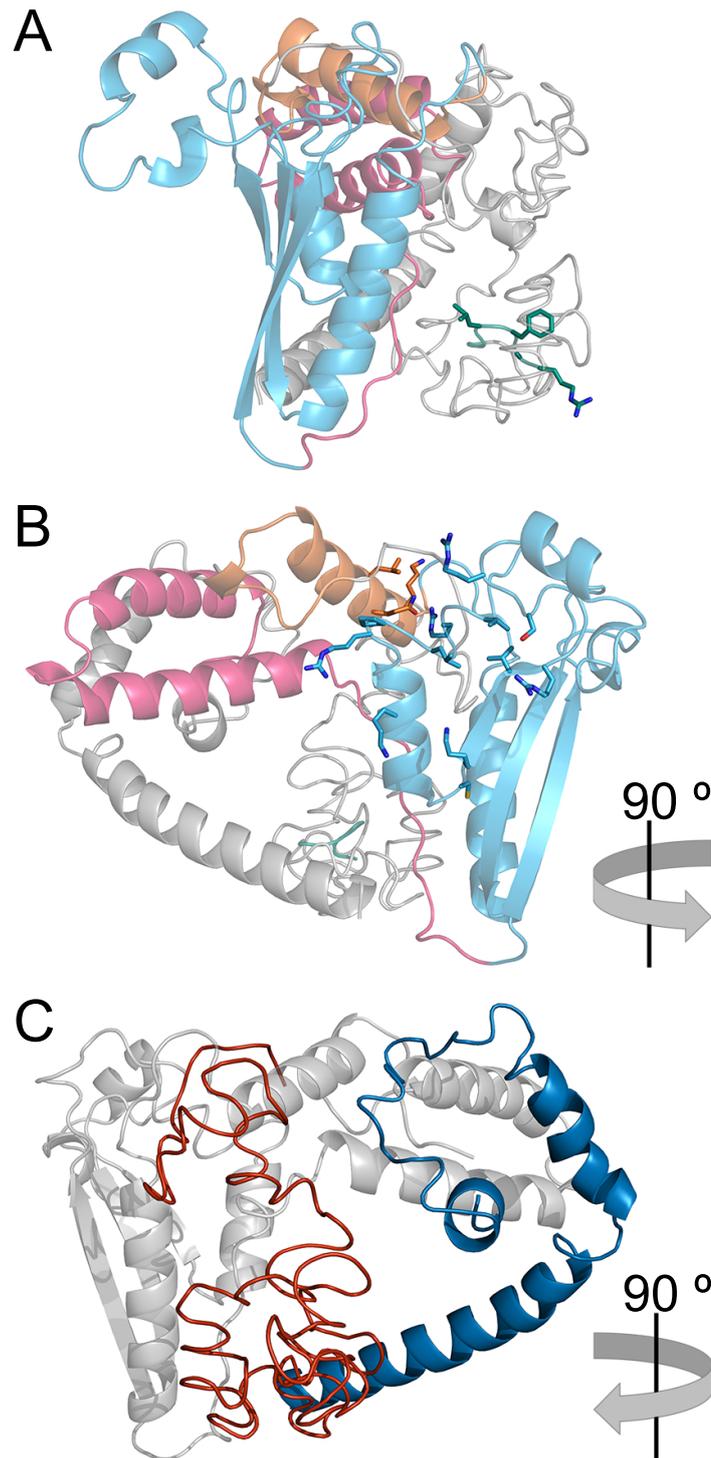
To investigate the RNA-binding properties of HOW(S) via fluorescence anisotropy, pure soluble protein was required. To increase the likelihood of achieving this several factors were considered in the design of the recombinant protein: information from other STAR protein structures, the HOW(S) amino acid sequence and HOW(S)'s predicted structure.

Only endogenous, full-length HOW protein has previously been used to examine its RNA-binding activities (Israeli et al., 2007; Giuliani et al., 2014). However, several studies have isolated the STAR domains of GLD-1, Sam68, and other STAR proteins to show that this domain confers the RNA-binding ability onto the proteins to which they belong (Vernet and Artzt, 1997; Ryder and Williamson, 2004; Galarneau and Richard, 2009). HOW's STAR domain is common to all of the annotated protein isoforms of HOW, including the two most-studied isoforms: HOW(L) and HOW(S) (Fig 5.1A).

To inspect whether it is likely that HOW's STAR domain is responsible for its RNA-binding activities, it was compared to GLD-1, the *C. elegans* orthologue of HOW. GLD-1 was selected because its STAR domain, bound to RNA, has the highest resolution crystal structure in the Quaking subfamily (Teplova et al., 2013). The amino acid sequences of GLD-1, HOW(S) and HOW(L) were aligned with Clustal Omega (Fig 5.1B). The STAR domain shows high conservation, and importantly all the GLD-1 RNA-binding residues identified from the crystal structure are conserved in HOW's STAR domain (Fig 5.1B). Notably, none of the RNA-interacting residues are close in the primary sequence to the C-terminal end where the HOW(L) and HOW(S) isoforms differ from one another.

To see if HOW(S)'s 6 unique C-terminal amino acids might interact with RNA in 3D space, a predicted 3D structure of HOW(S) was generated with I-TASSER (Fig 5.2; Roy et al., 2010). In this model the last 6 amino acids at the C-terminal end (Fig 5.2A, teal) are not close to the conserved RNA-binding residues (Fig 5.2B, blue and orange). An additional observation from the model is that there are no secondary structure features predicted after the STAR domain (Fig 5.2C, red), which could prove challenging to purify in a bacterial system, and based on other STAR proteins it is unlikely to contribute to RNA-binding. From the insights gained from the 3D model and combined with knowledge that HOW(L) and HOW(S) can bind the same RNA (Nabel-Rosen et al., 2002), it was decided to recombinantly express and purify just the STAR domain of HOW (residues 72–266) for the RNA-binding experiments.





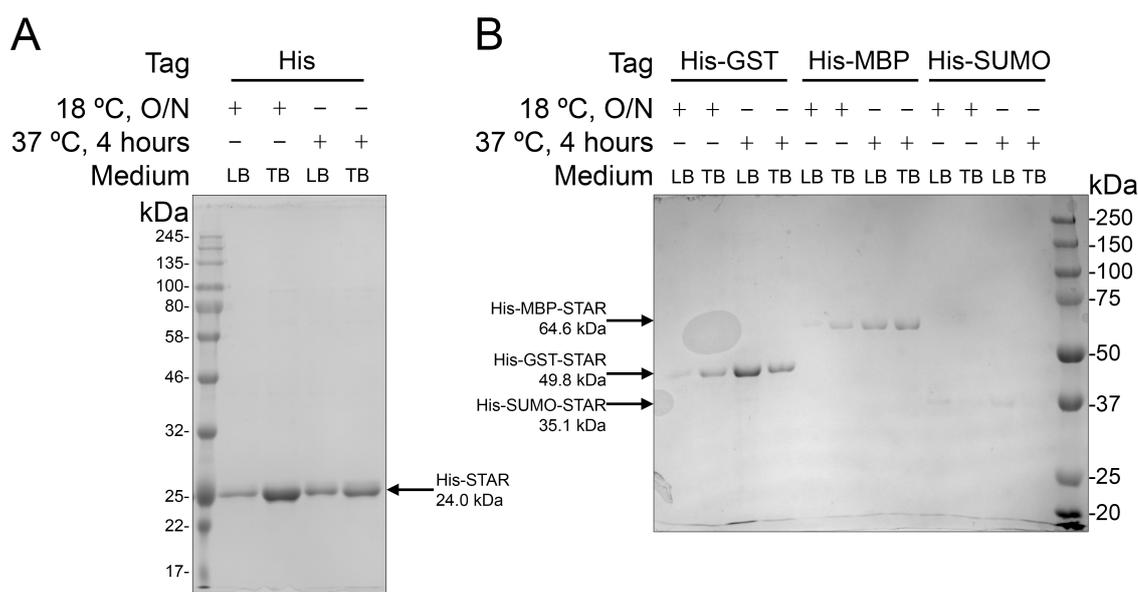
**Figure 5.2. HOW(S)'s unique C-terminal amino acids are not close to the RNA-binding region.** The I-TASSER predicted structure of HOW(S) in three different views. In A) and B) the QUA1 region is highlighted in pink, maxi-KH domain is blue, QUA2 region is orange, and the six isoform specific residues are in teal. A) The side chains of the C-terminal HOW(S) specific residues (**GGLFAR**) are visible as 'sticks' in the bottom right of the model. B) The side chains of the 17 conserved RNA-interacting residues from GLD-1 are shown as 'sticks'. The model has been rotated 90° to the right from the view in A. C) The STAR domain is in grey, the QA-rich region (blue) forms three  $\alpha$  helices, and the C-terminal region after the STAR domain (red) has no predicted secondary structures. The model has been rotated 90° to the left from the view in A.

## 5. *Characterising the molecular mechanism of RNA binding by HOW*

### 5.2.1.2. **Small-scale screening of STAR domain constructs and growth conditions**

To find a construct for the purification of HOW's STAR domain that would give us a reasonable yield and high purity, small-scale screens were carried out with different affinity tag constructs and growth conditions. A codon-optimised gene of the STAR domain was synthesised (Genewiz; Sequence A.3, Appendix I) and cloned into four vectors from the pOPIN Vector Suite to generate four different N-terminal tags: 1) a 6 histidine tag (His-tag), 2) a 6 histidine and glutathione-S-transferase tag (His-GST-tag), 3) a 6 histidine and maltose binding protein tag (His-MBP-tag), and 4) a 6 histidine and small ubiquitin-like modifier tag (His-SUMO-tag). For amino acid sequences of the 4 constructs see Sequences A.4–A.7 (Appendix 1).

All 4 plasmids were transformed into BL21(DE3) cells for small-scale screens (Material and Methods 2.9). Two induction conditions were tested: either an overnight induction at 18 °C or for 4 hours at 37 °C. Both of these conditions were carried out in two different media: either LB or TB. The latter is richer with a higher concentration of yeast and tryptone than LB and also contains glycerol (Table 2.13). The resulting protein samples from the automated MagneHis purification were separated by SDS-PAGE, which indicated that the His-STAR and His-GST-STAR constructs both produced a high amount of protein (Fig 5.3). However, the best growth conditions for these constructs were different; the most His-STAR protein was produced from an overnight induction while growing in TB (Fig 5.3A), whereas the most His-GST-STAR was produced from a 4 hour induction in LB (Fig 5.3B).

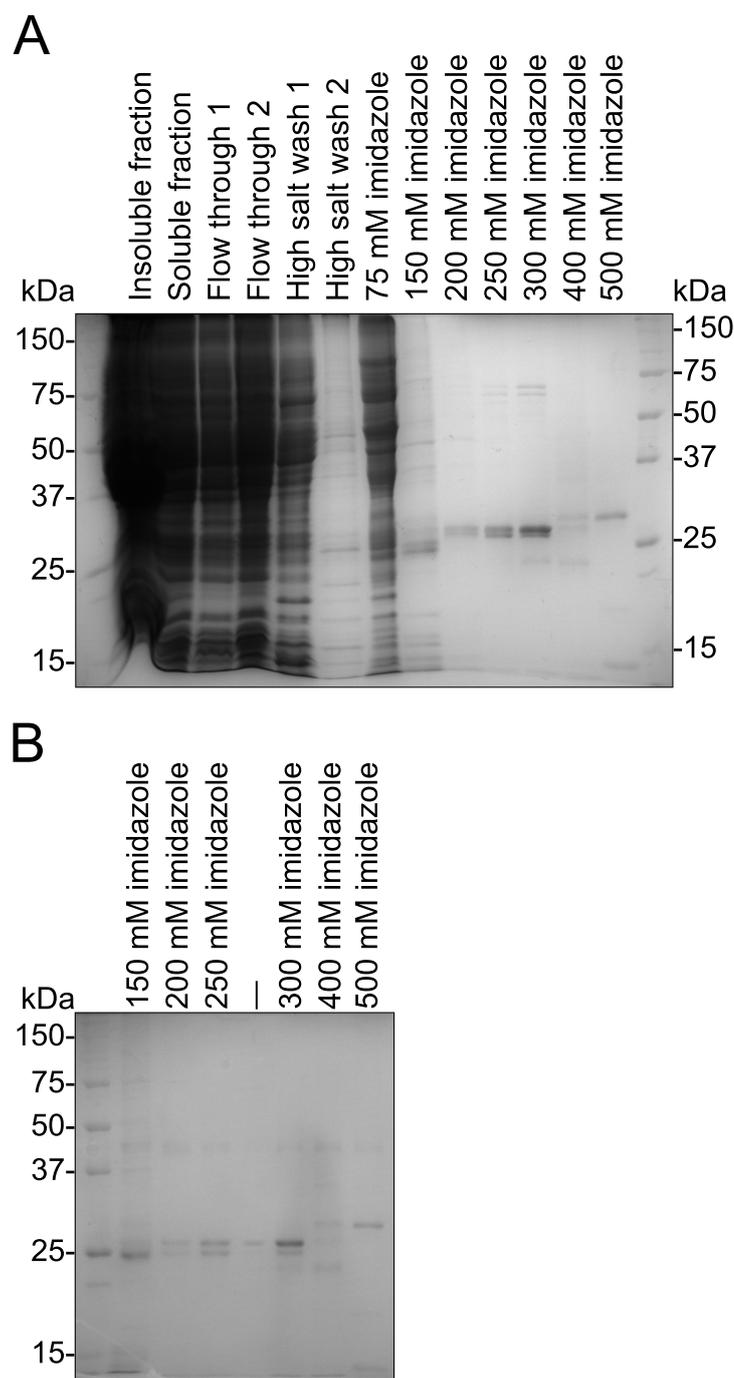


**Figure 5.3. His- and His-GST-tags express STAR domain best in small scale expression screens.** Coomassie stained gels of small-scale protein expression screens of the STAR domain. 4 different conditions were tested for each of the 4 different tags: Overnight (O/N) induction at 18 °C, 4 hour induction at 37 °C, and growing in either LB or TB media. Construct names and their theoretical mass (kDa) are indicated with arrows. A) Purified protein samples from the His-STAR construct screening. 18 °C induction in TB media gave the highest yield for this construct. B) Purified protein samples from the small-scale screening for the His-GST-STAR, His-MBP-STAR and His-SUMO-STAR constructs. His-GST-STAR construct produced more protein than the other two constructs, and the best condition for the His-GST-STAR construct was a 4 hour induction at 37 °C in LB media.

**5.2.1.3. Large-scale expression and purification of the STAR domain.**

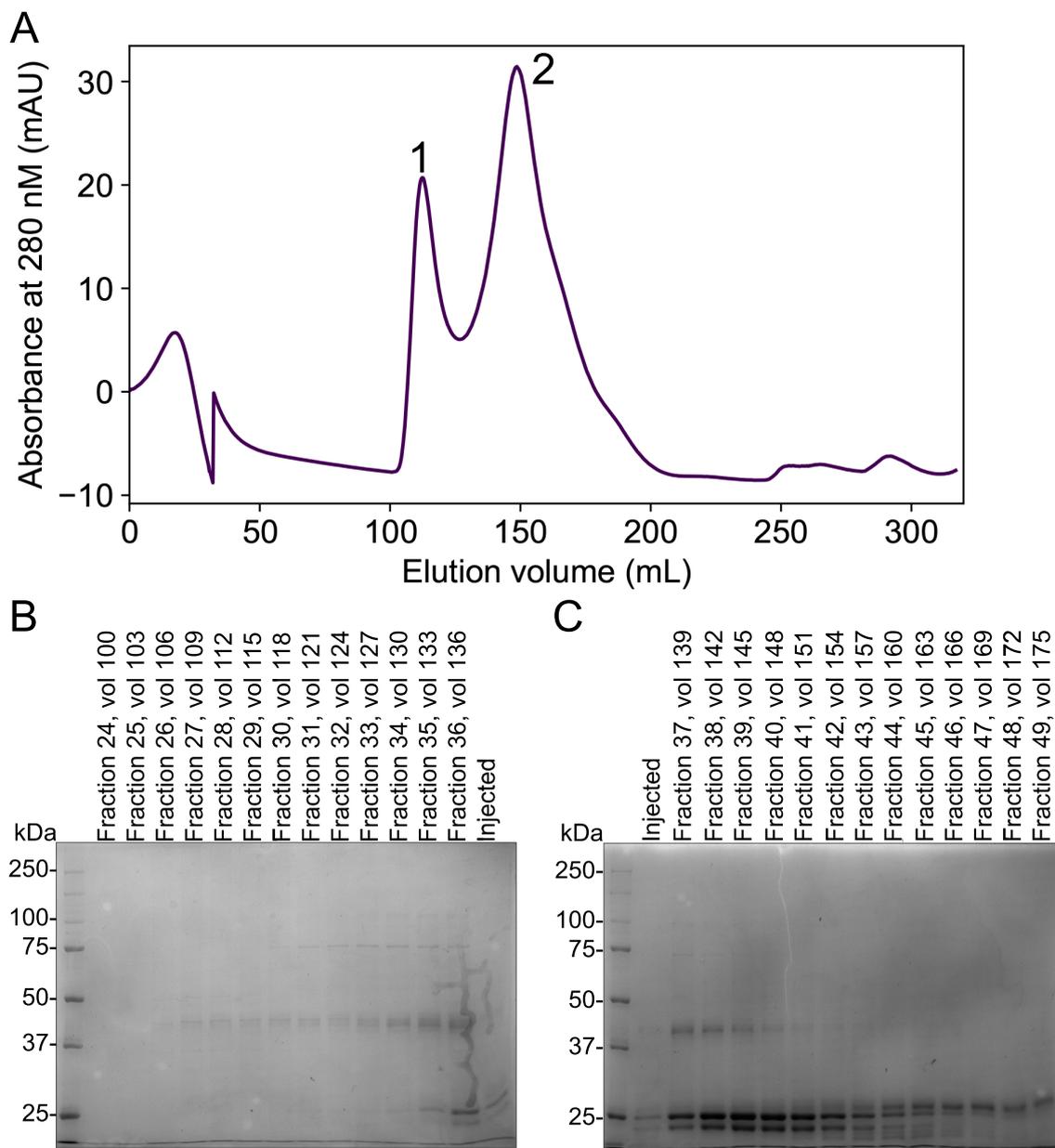
The His-STAR construct gave the highest yield in the small-scale screens, so this was used for the first large-scale purification (Materials and Methods 2.9.4). His-STAR was expressed in BL21(DE3) cells and after cell lysis and centrifugation, the supernatant (also referred to as the ‘soluble fraction’) was passed through a HisTrap affinity chromatography column. This column was washed with a high salt buffer (1M NaCl) to remove any bacterial RNA potentially bound to the protein, followed by increasing concentrations of imidazole to elute the His-STAR off the column. As expected from the small-scale screen, protein around the 25 kDa marker was clearly present in the 150–300 mM elutions (Fig 5.4A). All of the imidazole elutions, except the 75 mM elution because of the many contaminants in that sample, were then cleaved overnight with PreScission protease (Fig 5.4B).

Following protease cleavage, the 150–300 mM fractions were combined for size exclusion chromatography (SEC), though with closer inspection after the SEC had been performed, the 150 mM fraction should not have been included as it had several contaminants (Fig 5.4B). From the chromatogram, two large peaks were observed (Fig 5.5A). Samples collected from fractions across these peaks were run on SDS-PAGE gels. Most of the samples from peak 1 (fractions 24–33) contained a band between the 37 and 50 kDa markers (Fig 5.5B), most likely the PreScission protease that is around 46 kDa. Fractions 37–42 contained two bands around the 25 kDa marker. This is where we would expect the purified STAR protein (~22 kDa), suggesting that some truncation or degradation of the protein had occurred.



**Figure 5.4. HisTrap purification and PreScission cleavage of His-STAR.** Coomassie stained gels from the first two stages of large-scale His-STAR protein purification. A) Protein samples from HisTrap purification of His-STAR. Insoluble and soluble fractions post-bacterial cell lysis and centrifugation. The soluble fraction was passed through the HisTrap column and protein samples were collected from the flow through, high salt washes and increasing concentrations of imidazole elutions. His-STAR (runs at ~25 kDa) present in the 150–300 mM imidazole elutions. B) Protein samples after PreScission protease cleavage from the 150–500 mM imidazole. Protein bands around the 25 kDa marker, presumed to be His-STAR, in the 150–300 mM imidazole elutions. Dash (—) in B) indicates an empty lane with overflow from adjacent lanes.

5. Characterising the molecular mechanism of RNA binding by HOW

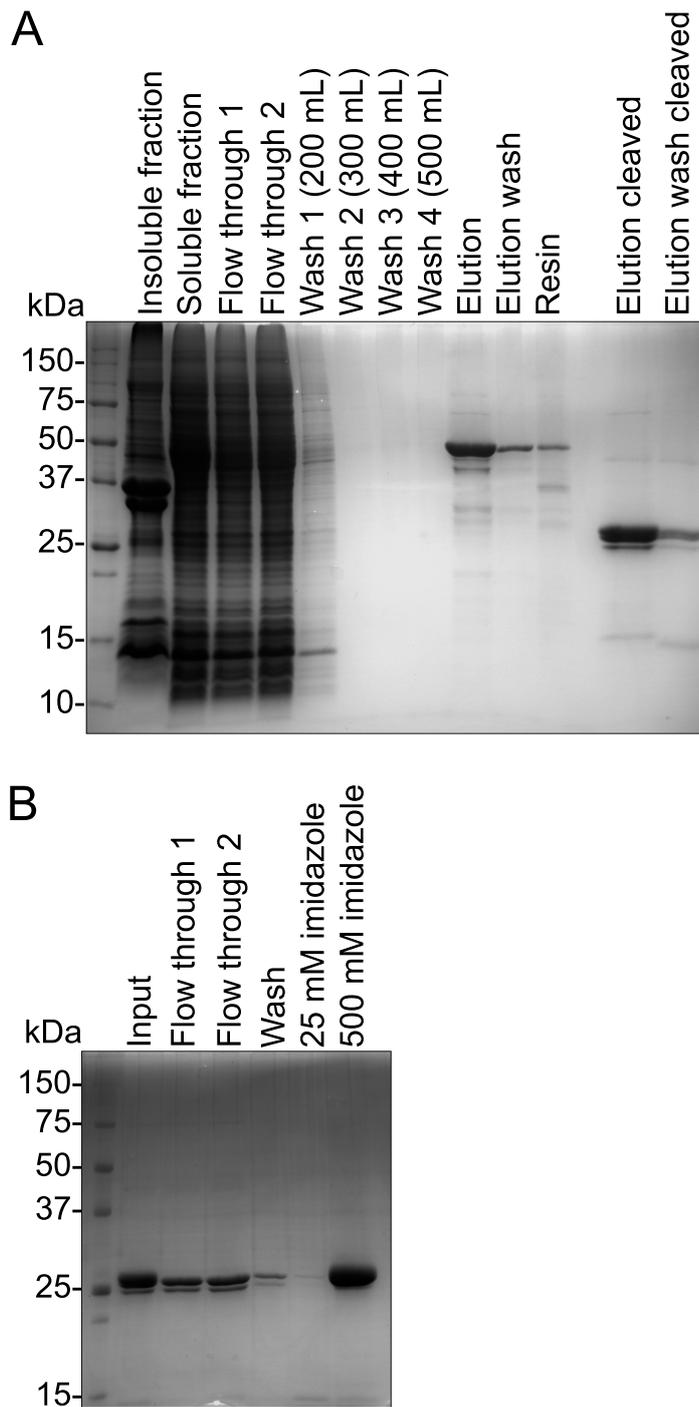


**Figure 5.5. Gel filtration of His-STAR suggests protein degradation.** A) Chromatogram from SEC of the combined 150–300 mM imidazole HisTrap elutions. With two absorbance peaks between 100 and 200 mL. B–C) Coomassie stained SDS-PAGE gels of fractions covering the 100–175 mL elution volume. B) Samples from peak 1 and partially into peak 2 contain a protein between 37–50 kDa. C) Samples from peak 2 (fractions 37–42) reveal two products around the 25 kDa marker, the size we expect for His-STAR.

It is highly likely that HOW forms a dimer via its QUA1 region, in the same fashion as the other STAR family members that have a QUA1 region. GST is known to dimerise and when used as an affinity tag it can help stabilise proteins that also dimerise. The His-GST-STAR construct also gave the second highest yield after the His-STAR construct in the small-scale screens (Fig 5.3). So, the His-GST-STAR construct was used to attempt to purify HOW's STAR domain without the degradation seen with the His-STAR construct (Materials and Methods 2.9.5).

The His-GST-STAR plasmid was transformed into BL21(DE3) cells and protein expression induced. Protein samples were collected through the purification process and separated by SDS-PAGE. First, the bacterial cells were lysed and centrifuged. The supernatant (Fig 5.6A, soluble fraction) was incubated with Glutathione Sepharose 4B resin, before being loaded onto a chromatography column. The flow through and wash steps removed most of the contaminating proteins (Fig 5.6A). After 4 washes with buffer with normal NaCl concentration (150 mM), the column was eluted with glutathione and washed once more. The elution and post-elution wash samples were cleaved with PreScission protease overnight. The theoretical molecular weight of His-GST-STAR is 49.8 kDa and the prominent band in the elution and elution wash samples around the 50 kDa marker are most likely our protein of interest (Fig 5.6A). The resin sample shows that some small amount of protein remained bound to the column after elution (Fig 5.6A). When cleaved the His-GST-tag and STAR protein are both similar sizes at ~25 kDa (Fig 5.6A, 'elution cleaved' and 'elution wash cleaved').

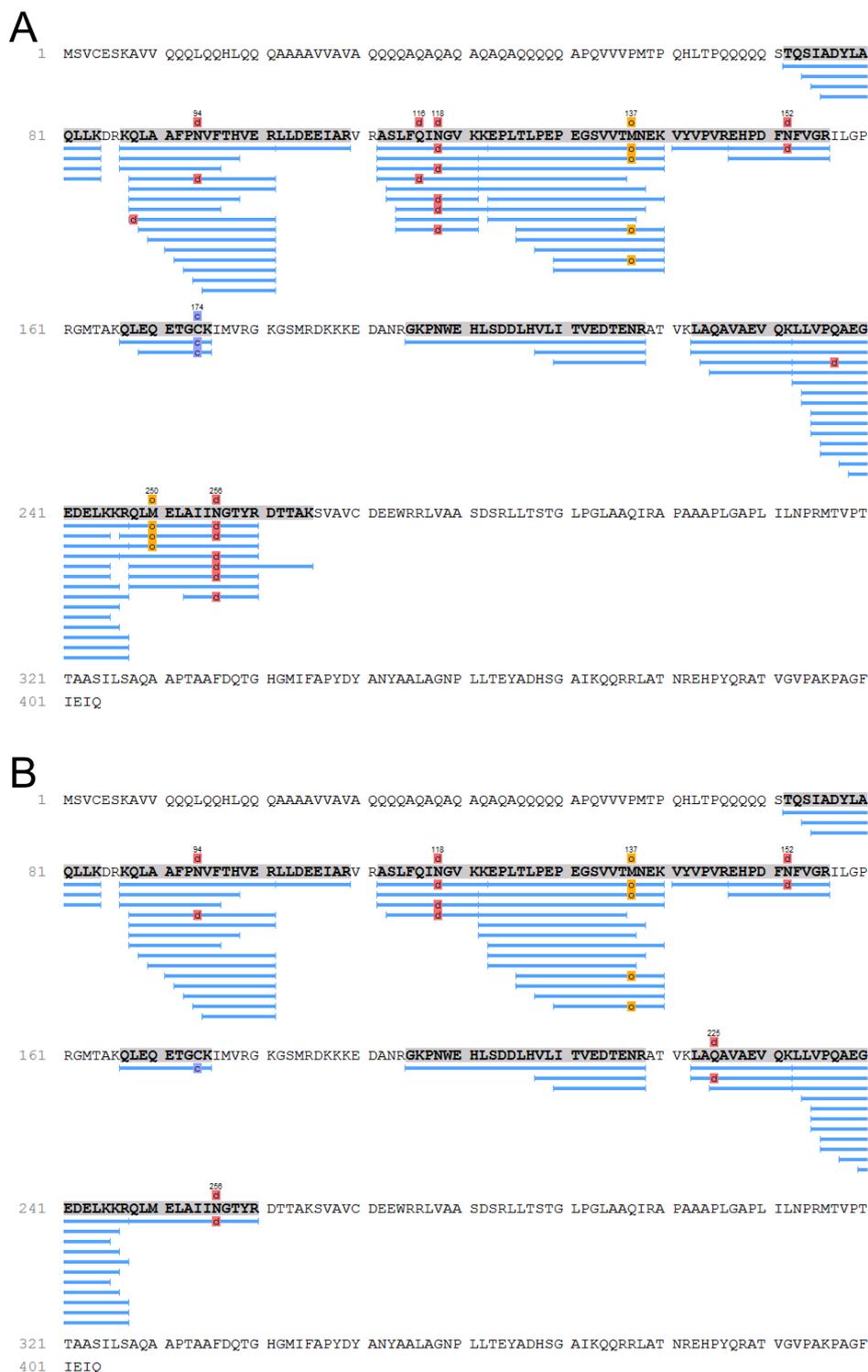
To separate the His-GST-tag from the STAR domain the two cleaved samples were combined and passed through a HisTrap column. The STAR domain flowed through the column and the His-GST-tag was eluted with 500 mM imidazole solution (Fig 5.6B). However, the expression and purification of the STAR domain again appeared to have produced two products around 25 kDa.



**Figure 5.6. Purification of His-GST-STAR also suggests degradation.** Coomassie stained SDS-PAGE gels from large-scale His-GST-STAR expression and purification. A) Protein samples from glutathione affinity chromatography and PreScission cleavage. The insoluble and soluble fractions from after bacterial cell lysis and centrifugation. Flow through and wash samples from chromatography steps. Elution and elution wash show a large amount of protein around the 50 kDa marker, the size we expect for His-GST-STAR. Cleaved samples protein samples are from after PreScission protease cleavage, both His-GST and STAR domain are around 25 kDa. B) Protein samples after HisTrap affinity chromatography to separate the cleaved His-GST tag (500 mM imidazole sample) from the STAR protein (flow through 1 and 2).

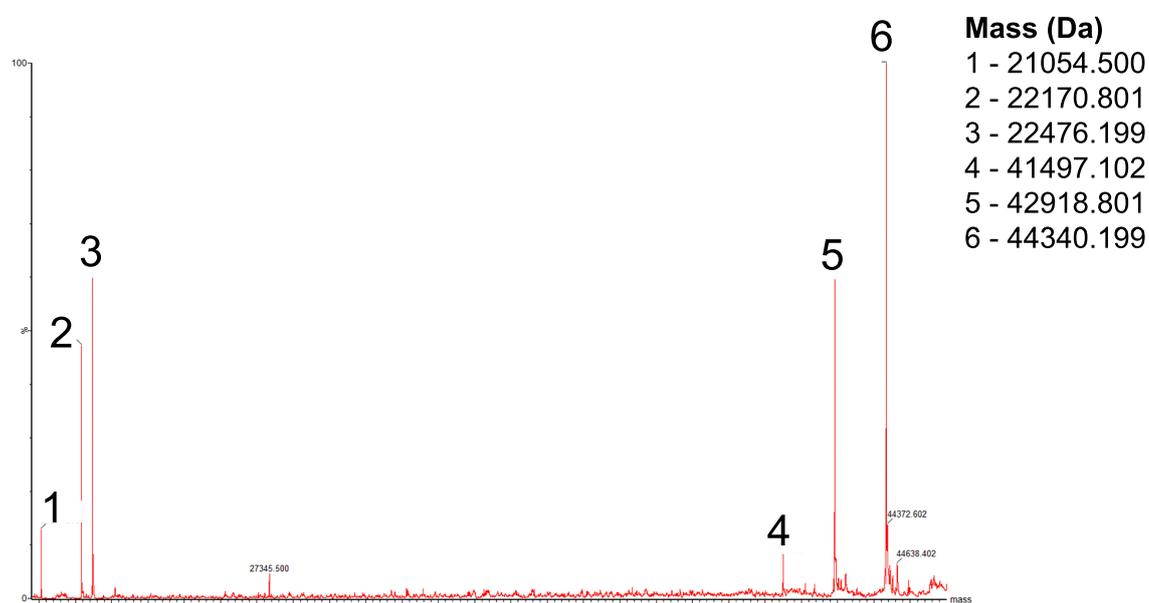
To identify the two protein products of  $\sim 25$  kDa from the His-GST-STAR expression and purification, samples were sent for mass spectrometry (MS). Peptide identification (via MS/MS) was carried out by the Biomolecular Mass Spectrometry Facility (University of Leeds) with protein bands excised from the gel after HisTrap purification (Fig 5.6B). The MS/MS results for the top band had 83% coverage for the STAR domain (Fig 5.7A). The MS/MS results for the bottom band had 82% coverage (Fig 5.7B). This confirmed that the bottom band was not a contaminant. The His-GST-tag is an N-terminal tag and this smaller protein was eluted in the glutathione affinity chromatography process means that it was likely that a C-terminal truncation was occurring. Peptides in the MS/MS data cover both the N- and C-terminus of the expressed construct in the top band of the gel, confirming that this is the full length protein (Fig 5.7A). MS/MS peptides from the bottom band cover the N-terminus but only extend to residue 261 (Fig 5.7B). The truncated protein could be residues 72–261, though this would likely run closer to the full length STAR domain (72–266) on the SDS-PAGE than the observed band (Fig 5.6B). Another alternative is that the unique peptide extending to residue 261 (Fig 5.7B) is from full length protein in the higher band cross contaminating this band. Multiple peptides extend to residue 248, thus the C-terminally truncated protein could be residues 72 to  $\sim 248$ . This hypothesis correlates better with the sizes of the bands on the gel.

## 5. Characterising the molecular mechanism of RNA binding by HOW



**Figure 5.7. MS/MS analysis suggests the presence of full length STAR domain and a C-terminal degradation.** MS/MS analysis of two gel-excised protein bands from the His-GST-STAR purification. The HOW sequence is shown with peptides detected in blue bars below. Small boxes represent different modifications to the amino acid residues: red is deamidation, orange is oxidation, purple is carbamidomethylation. A) The peptides identified from the higher molecular weight band cover almost the entire STAR domain (72–266) including the C-terminus of this construct. B) The peptides identified from the lower molecular weight band results show coverage for most of the STAR domain but only extends to residue 261, confirming a C-terminal truncation.

To isolate where the truncation is in the C-terminal end, protein sample from the HisTrap flow through 1 and 2 (Fig 5.6B) were combined and submitted for mass determination (via liquid chromatography MS). Six masses were identified; peak 2 at 22,170 Da closely matches the theoretical molecular weight of the cleaved STAR protein (22,171 Da; Fig 5.8). Peak 6, the highest in abundance, is double this mass and is probably the dimerised form of the full STAR domain. Peaks 4 and 5 could be dimers formed from two truncated STAR protomers or a combination of full STAR and truncated STAR. Peak 1 is 1,116 Da smaller than peak 2, likely the C-terminal truncated version of the STAR domain. The last 11 amino acids of the STAR domain are: NGTYRDTTAKS (Sequence A.8, Appendix I), which has a molecular weight of 1213 Da. The molecular weight of the final 10 amino acids is 1099 Da. This narrowed down the location of the truncation to after either the last isoleucine or asparagine in the STAR domain.



**Figure 5.8. Truncated STAR is 10 or 11 amino acids shorter than the full-length STAR protein.** Mass spectrum, with  $m/z$  on the  $x$ -axis and relative abundance on the  $y$ -axis, for mass determination of the products from the His-GST-STAR purification. Peak 2 is 22,170 Da and cleaved STAR's theoretical molecular weight is 22,171 Da. STAR protein is expected to dimerise and peak 6 is double the mass of peak 2. Peak 1 is 1,116 Da smaller than peak 2, it is likely the C-terminal truncated version of STAR, short 10 or 11 amino acids. Provided by the Biomolecular Mass Spectrometry Facility (University of Leeds).

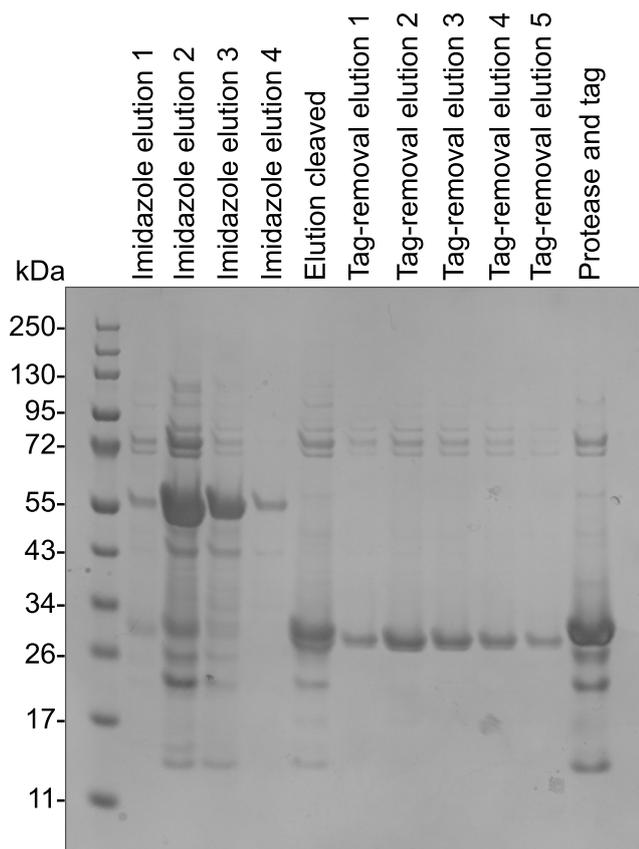
## 5. *Characterising the molecular mechanism of RNA binding by HOW*

To investigate whether there was a known protease that cleaves amino acids in the region of this truncation, which could then be considered in the purification process, the suspected region of truncation with five amino acids flanking on either side (MELAI**I**NGTYRD) were submitted to PeptideCutter (Gasteiger et al., 2005). However, the only enzyme that cleaved in this region of interest was proteinase K, which is an improbable candidate as the whole protein would have been degraded if this was present. Thus, there was no indication that altering the expression and/or purification protocols would help to purify only the full-length STAR domain.

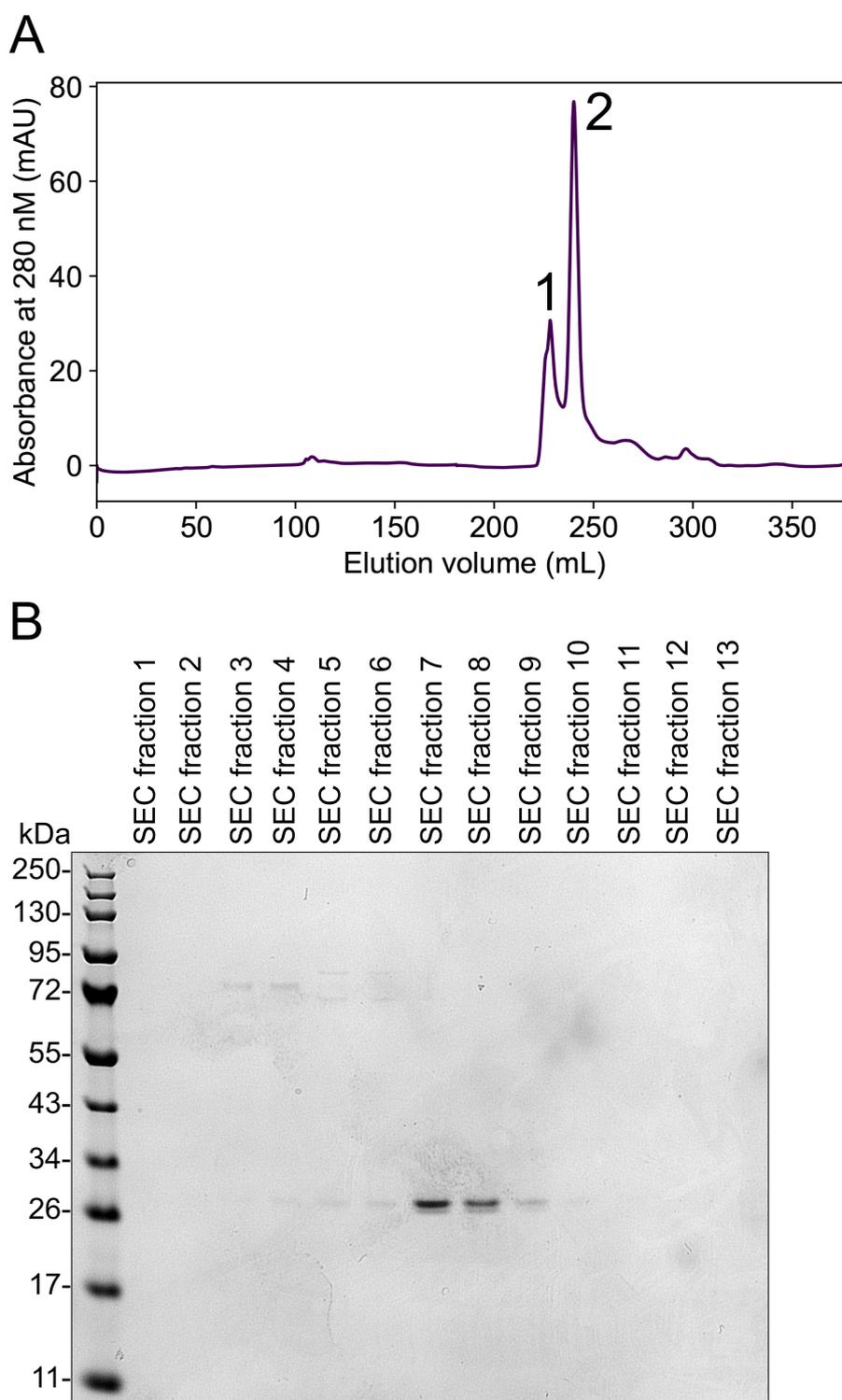
Several attempts were made to purify a truncated version of the STAR domain (residues 72–257) in order to purify non-degraded protein. However, these were unsuccessful due to contamination issues, and because of time constraints it was decided to move forward with the His-GST-STAR construct for producing protein for the RNA-binding experiments. GST-tagged protein was chosen because the His-GST-STAR construct appeared to produce a higher ratio of full-length STAR domain to truncated STAR domain than the His-STAR construct (Figs 5.6B and 5.5C).

The STAR domain was purified by the Protein Production Facility (University of Leeds) from bacterial pellets from the same growth and expression batch as the previous His-GST-STAR purification (Fig 5.6; Materials and Methods 2.9.6). SDS-PAGE was used to evaluate samples throughout the purification. A large quantity of protein was detected around the 55 kDa marker (His-GST-STAR is ~50 kDa) in the imidazole elutions 2 and 3 (Fig 5.9). The His-GST-tag was cleaved from the STAR protein (Fig 5.9, ‘Elution cleaved’). The tag and protease were separated from the STAR domain with another round of Ni<sup>2+</sup> affinity chromatography (Fig 5.9, ‘Protease and tag’). The STAR protein was present in the ‘Tag-removal elution 1–5’ samples (Fig 5.9).

Finally, to purify the STAR domain from the higher molecular weight contaminants, the ‘Tag-removal elution’ samples were combined for SEC. SEC fractions 7 and 8 from the SDS-PAGE correspond to peak 2 (Fig 5.10B). Again, two protein products have been isolated around the ~26 kDa marker, however, the smaller molecular weight band appears to be much less abundant than the higher band (Fig 5.10B). SEC fractions 7 and 8 were combined and sent for peptide identification (via MS/MS). The presence of the full-length STAR domain was confirmed with 94% coverage, including coverage of both N- and C-terminal regions. This purified protein sample was used for the subsequent RNA-binding experiments.



**Figure 5.9. Ni<sup>2+</sup> affinity chromatography of His-GST-STAR.** Coomassie stained SDS-PAGE gel from the first stages of the purification of the STAR domain using the His-GST-STAR construct. The imidazole elutions are from the first round of Ni<sup>2+</sup> affinity chromatography. These samples were combined and protease cleaved to remove the His-GST-tag ('Elution cleaved'). Another round of Ni<sup>2+</sup> affinity chromatography was carried out to separate the STAR protein (in the 'Tag-removal elution' samples) from the His-GST-tag and protease ('Protease and tag'). Provided by the Protein Production Facility (University of Leeds).



**Figure 5.10. Purification of His-GST-STAR produces predominantly full-length STAR domain.** A) Chromatogram from SEC of the combined tag-removal elutions (Fig 5.9), showing to absorbance peaks between 200–250 mL. B) Coomassie stained SDS-PAGE gel from several SEC fractions. SEC fractions 7–8 have the strongest STAR protein signal at the 26 kDa marker, with no higher molecular weight contaminants seen in fractions 3–6. A faint band, under the predominate one, can be seen in SEC fractions 7–8. SEC data and gel image provided by the Protein Production Facility (University of Leeds).

## 5.2.2. Assessing the RNA-binding capacity of the HOW STAR domain

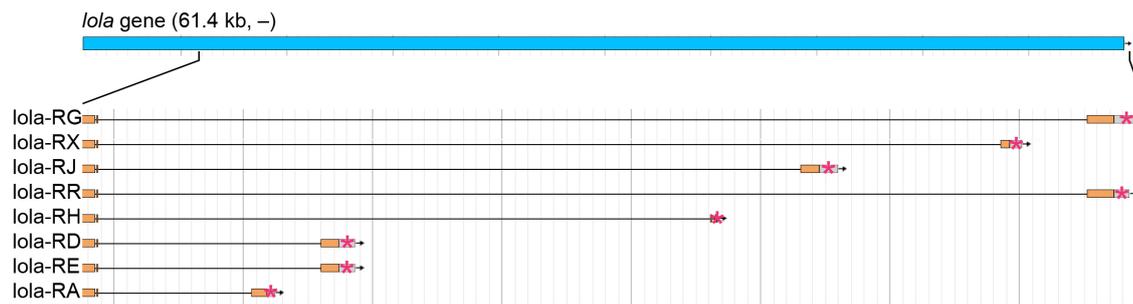
### 5.2.2.1. Designing RNA oligos for RNA-binding experiments

To assess the RNA-binding capacity of the HOW STAR domain, nine 3'-fluorescein labelled RNA oligos were designed for fluorescence anisotropy (FA) experiments (Materials and Methods 2.11). The oligos were designed to examine the STAR domain's interactions with generic sequences as well as sequences identified from the HOW(S)-HA RIP-seq experiments in Chapter 3.

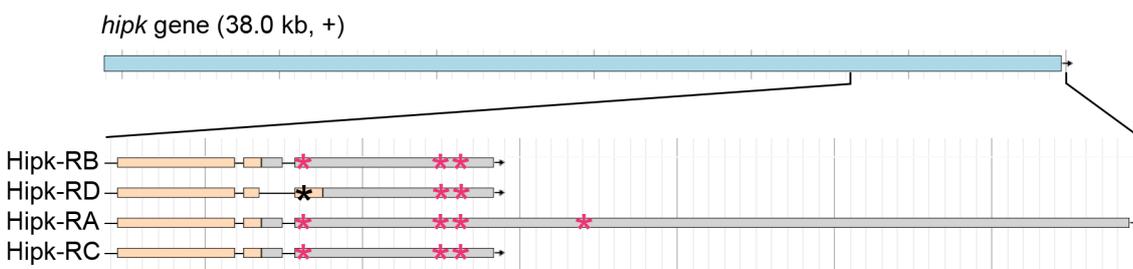
The two generic oligos used in the FA experiments were: 10 adenosine nucleotides (poly(A)<sub>10</sub>) and 10 uridine nucleotides (poly(U)<sub>10</sub>; Table 5.1). The other seven RNA oligos each had the top 5'- or 3'-UTR motif (Figs 3.25 and 3.27) from the DREME motif enrichment analysis — GCG(A/U)G and (A/U/G)CUAAC — incorporated into their sequence. Each of these oligos also included 4 nucleotides flanking the sequences of interest. Based on a study that STAR proteins bind asymmetric bipartite motifs (Galarneau and Richard, 2009), the RNA sequences that featured the 3'-UTR core motif were each paired with an oligo that also featured its closest half-site.

Two oligos were sequences from *lola*-RJ's 3'-UTR, which was the differentially enriched transcript from the HOW(S)-HA RIP-seq. The first oligo had the only ACUAAC motif found in the *lola*-RJ 3'-UTR and flanked by four nucleotides either side (Fig 5.11, Table 5.1). The second had this ACUAAC core motif and the nearest half-site (UAAC), which is three nucleotides away (Table 5.1). Four oligos were designed around the 3'-UTRs of *hipk*'s transcripts. The differentially enriched *hipk* transcript from the HOW(S)-HA RIP-seq was *Hipk*-RA (FBtr0072552), which has the longest 3'-UTR of the four *hipk* transcripts (Fig 5.12). *Hipk*-RA has four (A/U/G)CUAAC motif sites in its 3'-UTR. One site is common to *Hipk*-RA, -RB and -RC, two of the sites are separated by just one nucleotide and common to all four transcripts, and the fourth motif site is unique to *Hipk*-RA. So, one oligo was designed around the two ACUAAC motifs common to all *hipk* transcripts, referred to as the 'Hipk double core', and a second was designed that included this double motif and the nearest half-site (UAAC), which was three nucleotides downstream of the double core (Table 5.1). An oligo for the motif specific to *Hipk*-RA (GCUAAC) was designed, alongside an oligo with the half-site (UAAU) four nucleotides downstream (Table 5.1). The ninth RNA oligo was a sequence from the 5'-UTR of *jvl*-RF (FBtr0305694), which is common to three of the seven *jvl* transcripts, and included the GCGUG 5'-UTR motif (Fig 5.13, Table 5.1).

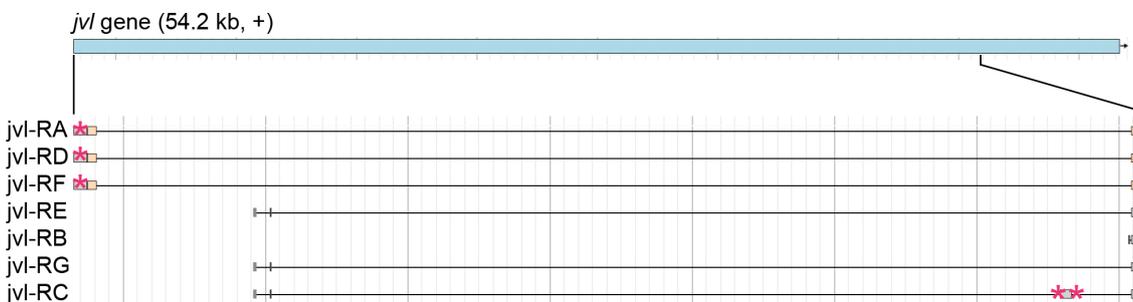
5. Characterising the molecular mechanism of RNA binding by HOW



**Figure 5.11. Schematic of the 3'-UTRs of eight *lola* transcripts and the location of (A/U/G)CUAAC motifs.** Top: the gene span of *lola*, which is 61.4 kb and on the antisense strand. Bottom: Zoom in of the 3'-UTRs of eight *lola* transcripts, the coding regions are in beige and the 3'-UTRs are in grey. The location of (A/U/G)CUAAC motifs in the 3'-UTRs of the eight transcripts are indicated with pink asterisks. There are 17 other *lola* transcripts that are not shown and do not have an (A/U/G)CUAAC motif. Bold vertical lines are every 6,250 nt. Adapted from JBrowse (Buels et al., 2016).



**Figure 5.12. Schematic of the 3'-UTRs of the four *hipk* transcripts and the location of (A/U/G)CUAAC motifs.** Top: the gene span of *hipk*, which is 38.0 kb and on the sense strand. Bottom: Zoom in of the 3'-UTRs of all four *hipk* transcripts, the coding regions are in beige and the 3'-UTRs are in grey. The location of the (A/U/G)CUAAC motifs in the 3'-UTRs of the four transcripts are indicated with pink asterisks. The black asterisk on Hipk-RD indicates a motif in the CDS rather than 3'-UTR. Bold vertical lines are every 1,250 nt. Adapted from JBrowse (Buels et al., 2016).



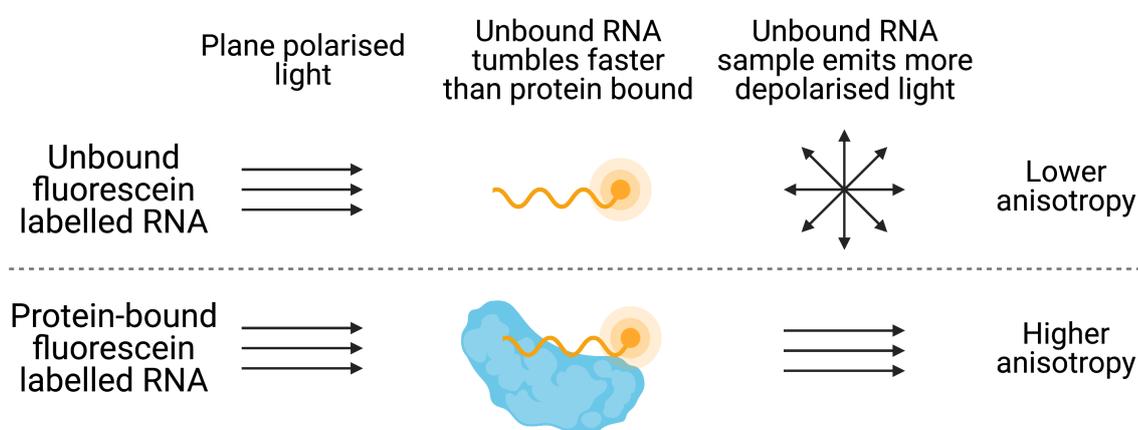
**Figure 5.13. Schematic of the 5'-UTRs of the seven *jvl* transcripts and the location of GCG(A/U)G motifs** Top: the gene span of *jvl*, which is 54.2 kb and on the sense strand. Bottom: Zoom in of the 5'-UTRs of all seven *jvl* transcripts, the coding regions are in beige and the 5'-UTRs are in grey. The location of the GCG(A/U)G motifs in the 5'-UTRs of the seven transcripts are indicated with pink asterisks. Bold vertical lines are every 6,250 nt. Adapted from JBrowse (Buels et al., 2016).

**Table 5.1. RNA oligos designed for RNA-binding experiments with HOW’s STAR domain.** Name and sequence of RNA oligos with the transcript IDs and transcript nucleotide positions that the sequences were designed from. All oligos were ordered with 3’ fluorescein labels. The 3’-UTR motifs are in pink, half-sites in orange, and the 5’-UTR motif is in blue.

Oligo name	Transcript ID	Oligo sequence	Length (nt)	Transcript region
lola core	FBtr0089360	ACACACUAAACUCGU	14	3121–3134
lola core + half-site	FBtr0089360	ACACACUAAACUCGUAAACUAUG	21	3121–3141
Hipk double core	FBtr0072552	UACAACUAACAACUAACAGAU	21	5731–5751
Hipk double core + half-site	FBtr0072552	UACAACUAACAACUAACAGAUAAACAAUU	28	5731–5758
Hipk-RA core	FBtr0072552	AGCAGCUAACAAUU	14	6955–6968
Hipk-RA core + half-site	FBtr0072552	AGCAGCUAACAAUUGUAAUUGUA	23	6955–6977
jv1 5’-UTR core	FBtr0305694	GUUGGCGUGUUUU	13	164–176
Poly(A) <sub>10</sub>	—	AAAAAAAAAAAA	10	—
Poly(U) <sub>10</sub>	—	UUUUUUUUUU	10	—

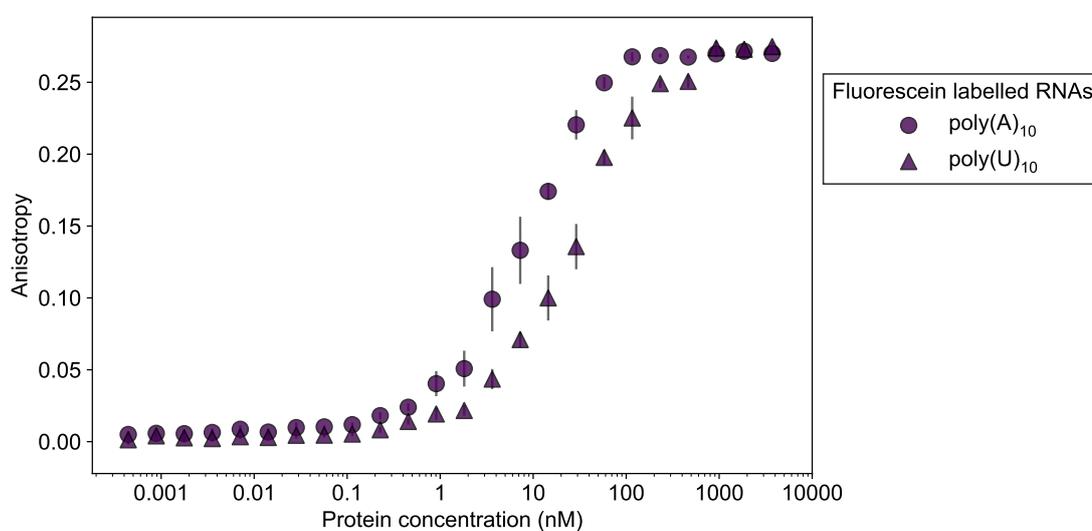
### 5.2.2.2. Measuring the RNA-binding affinity of the HOW STAR domain

To measure the affinity between the HOW STAR domain and the 3'-fluorescein labelled RNA oligos FA was used. In these FA experiments protein was two-fold serially diluted in a 384-well plate. 5 nM of the relevant oligo was added to the appropriate wells and the plate was left to equilibrate for at least 40 minutes before being read by the plate reader (see Material and Methods 2.11.2 for more details). In these direct binding assays, the higher anisotropy measurements reflect an increased amount of RNA bound to protein than lower anisotropy results (Fig 5.14). However, different protein–RNA complexes will have different maximal anisotropy values.



**Figure 5.14. Schematic of unbound vs protein-bound RNA in fluorescence anisotropy assays.** Plane polarised light is used to excite a 3' fluorescein labelled RNA. RNA that is unbound (top row) tumbles faster in solution than protein-bound RNA (bottom row). The faster tumbling samples emit more depolarised light than the slower samples. Thus, unbound RNA will have lower anisotropy values than bound RNA.

Simbu orthobunyavirus nucleoprotein (SIMV NP) was used as a positive control in the FA experiments. SIMV NP has been shown to bind poly(A), poly(U) and poly(C), with a slight preference for poly(A) (Pangratiou, 2020). Thus, it was likely to bind all of the oligos designed to examine the STAR domain's binding affinity (Table 5.1). To confirm the experimental setup was working correctly, the first oligos tested against SIMV NP were the generic poly(A)<sub>10</sub> and poly(U)<sub>10</sub> oligos, both of which showed nanomolar affinity and the expected slight preference for poly(A)<sub>10</sub> (Fig 5.15). The other seven oligos used also bound to SIMV NP (Appendix I, Fig A.3).

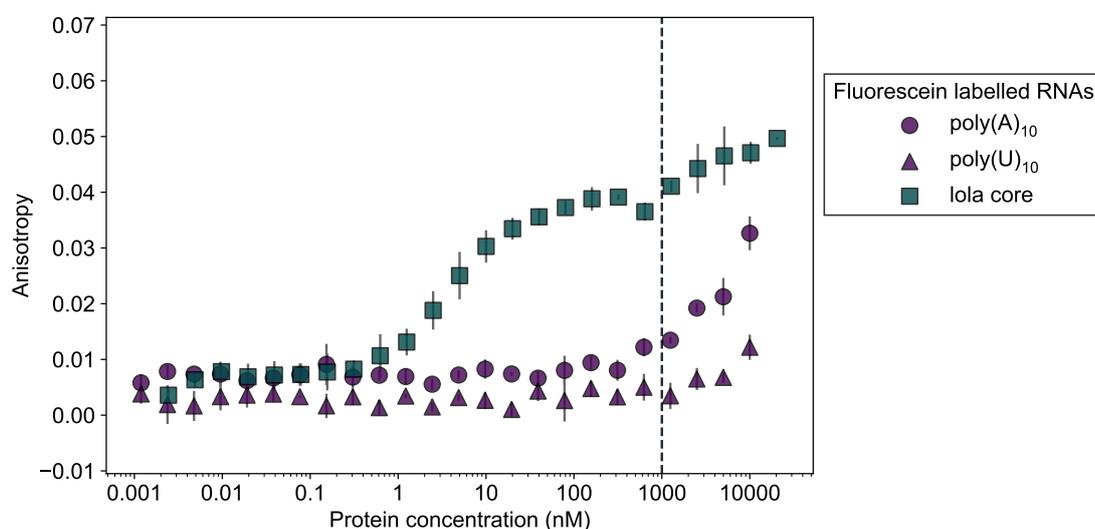


**Figure 5.15. The positive protein control, SIMV NP, binds poly(A)<sub>10</sub> and poly(U)<sub>10</sub>.** Scatter plot of SIMV NP protein concentration against mean anisotropy for two oligos: poly(A)<sub>10</sub> and poly(U)<sub>10</sub>. Error bars reflect the standard deviation of the three technical replicates.

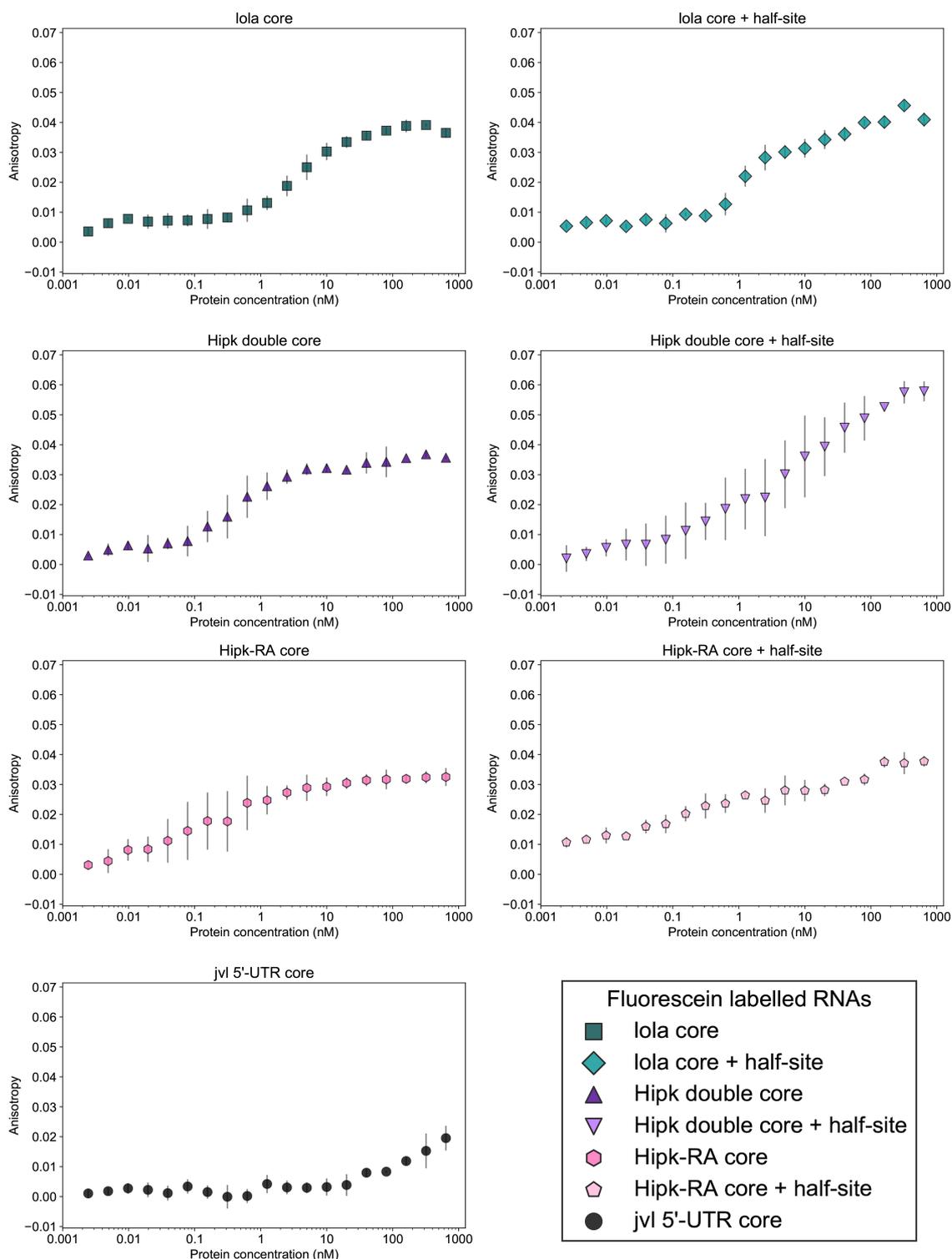
## 5. Characterising the molecular mechanism of RNA binding by HOW

Having confirmed the FA experimental set-up was functioning as expected, the purified HOW STAR domain was plated with a starting concentration of either 20.5 or 10  $\mu\text{M}$  and two-fold serially diluted to 250 or 120 pM, respectively. The first oligos compared were the generic poly(A)<sub>10</sub> and poly(U)<sub>10</sub> oligos to the ‘lola core’ RNA, which has a single ACU AAC core motif. There was no indication of poly(A)<sub>10</sub> or poly(U)<sub>10</sub> binding to the STAR domain until the protein concentration was above 1  $\mu\text{M}$  (Fig 5.16, purple). The ‘lola core’ binding curve showed a more typical sigmoid-shaped binding curve, with RNA-binding occurring in the nanomolar range (Fig 5.16, green). However, after an initial plateau around 100–1000 nM the anisotropy increased again. Given that no binding between the STAR domain and the generic oligos was observed above 1  $\mu\text{M}$ , this implies that non-specific RNA-binding is occurring above this protein concentration (Moerke, 2009). For the subsequent results, the graphs displayed and curve fitting calculations made were carried out on data with a 1  $\mu\text{M}$  protein concentration cut-off.

The FA results from all six of the RNA oligos that featured a 3'-UTR core motif indicated that they bound to the STAR domain in the nanomolar range (Fig 5.17). The ‘jvl 5'-UTR core’ oligo only showed binding at much higher protein concentrations than the 3'-UTR motif oligos, and did not form a sigmoid curve as no plateau was reached (Fig 5.17, bottom left).



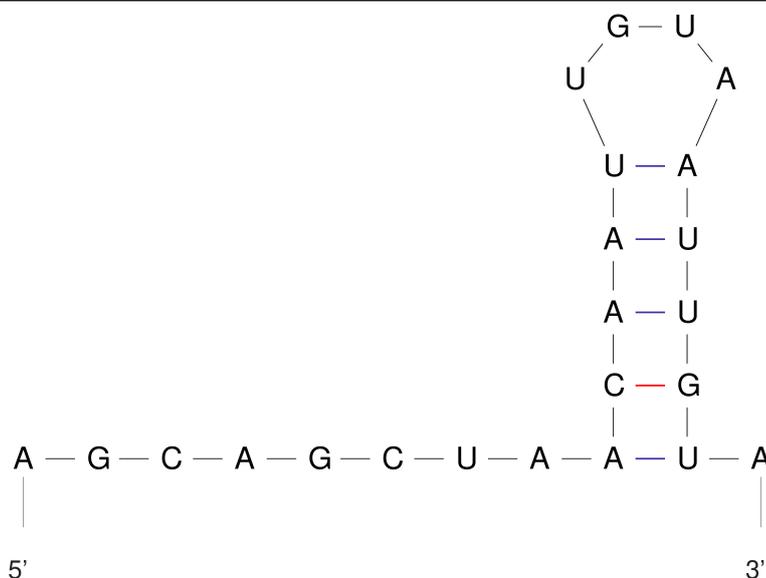
**Figure 5.16. The STAR domain binds the lola core oligo selectively.** Scatter plot of STAR protein concentration against mean anisotropy for three oligos: poly(A)<sub>10</sub>, poly(U)<sub>10</sub> and lola core. Error bars reflect the standard deviation of the three technical replicates.



**Figure 5.17. STAR domain binds the 3'-UTR motif derived oligos.** Scatter plots of STAR protein concentration against mean anisotropy for seven oligos (listed in key). Error bars reflect the standard deviation of the three technical replicates. The six 3'-UTR core motif based oligos all bind STAR in the nanomolar range. The 'jvl 5'-UTR core' oligo only bound at much higher concentrations than the other six.

## 5. Characterising the molecular mechanism of RNA binding by HOW

To calculate the apparent  $K_D$  values and compare the STAR domain's affinity for these different oligos, the anisotropy values were converted into fraction bound (see Materials and Methods 2.11.3 for details). The  $K_D$  values were taken from the logistic equation used to fit the binding curves (Table 5.2). While the data from all the RNAs were fitted successfully with the logistic equation, the 'jvl 5'-UTR core' and 'Hipk-RA + half-site' fraction bound data did not result in typical S-shaped binding curves (Fig 5.19A–B). This was not surprising for the 'jvl 5'-UTR core', which in the anisotropy data did not show a plateau (Fig 5.17). The 'Hipk-RA core + half-site' logistic fit resulted in the curve appearing almost linear (Fig 5.19B). This could be because it might form a secondary structure; it was the only oligo to have a predicted fold with a negative  $\Delta G$  value (-1.50) when each oligo was passed through Mfold (Zuker, 2003). The structure features one stem-loop, some of which incorporates the GCUAAC motif into the stem (Fig 5.18), thus the ability for this oligo to form a structured RNA could be interfering with the binding to the STAR domain.



**Figure 5.18.** The predicted fold of the 'Hipk-RA core + half-site' oligo has a stem-loop. The predicted fold of the 'Hipk-RA core + half-site' oligo from Mfold (Zuker, 2003) with a  $\Delta G$  value of -1.50. The fold has a stem-loop structure which contains the UAAU half-site and part of the GCUAAC core motif.

The fraction bound data from the other five 3'-UTR motif oligos produced more typical S-shaped curves with the logistic equation (Fig 5.19A). The lowest apparent  $K_D$  value (highest affinity) from these oligos was the 'Hipk-RA core' at 0.10 nM (Table 5.2). The 'Hipk-RA core' oligo is most similar in sequence to 'lola core', they are both 14 nucleotides long and feature the 3'-UTR core motif once with no half-site. They differ by one nucleotide in the core motif sequence; the 'Hipk-RA core' motif starts with a guanosine, while the 'lola core' motif begins with an adenosine (GCUAAC vs ACUAAC). If this motif is the sequence that determines the STAR domain's affinity to the RNA, then it is striking that the 'lola core' oligo has an apparent  $K_D$  nearly 50 times higher, at 4.80 nM, than the 'Hipk-

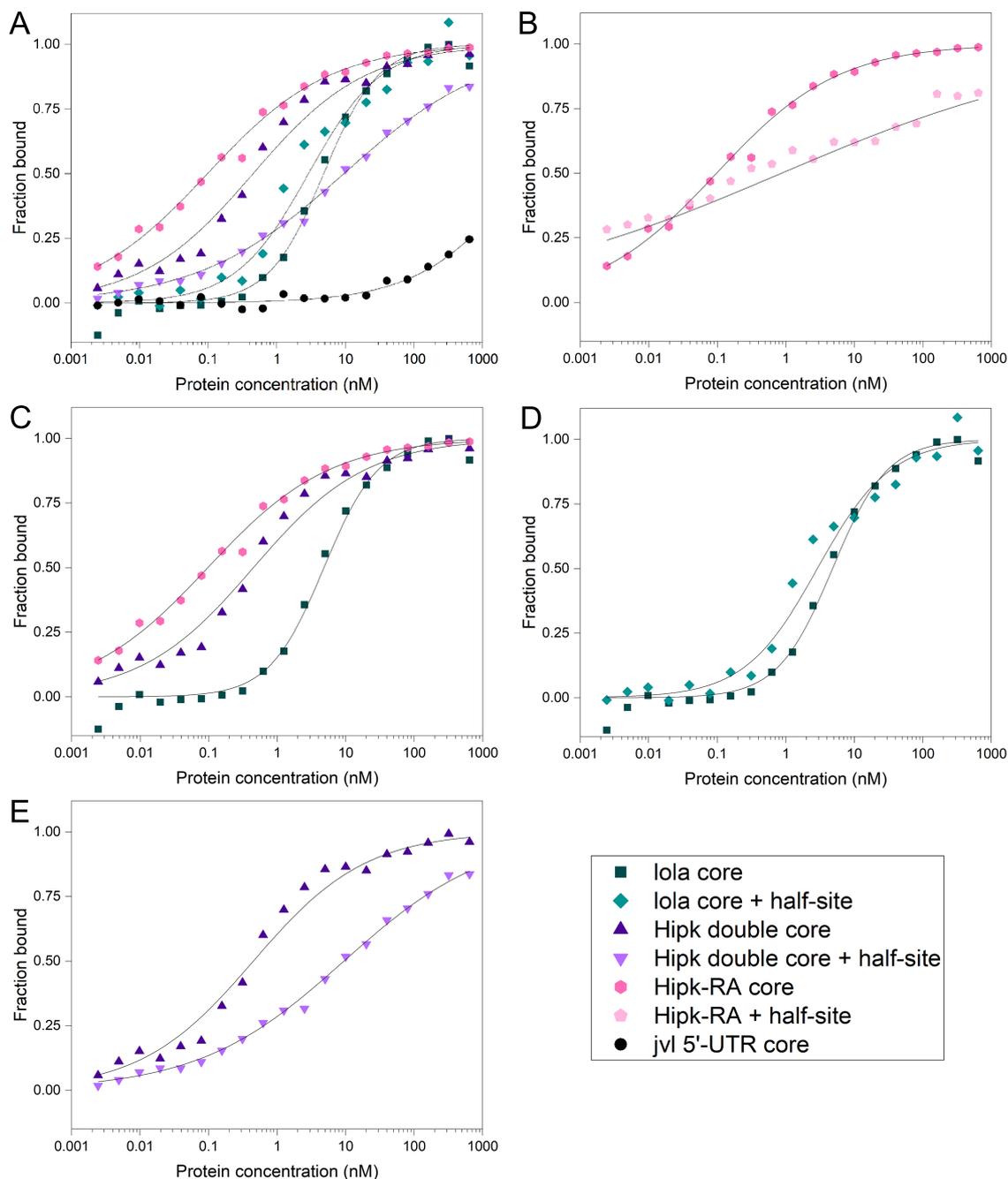
RA core' oligo (Table 5.2 and Fig 5.19C). However, some of this difference in affinity could also be influenced by the differing flanking sequences between these two oligos. The third oligo with the 3'-UTR core motif was the 'Hipk double core' which featured two ACUAAC motifs. The STAR domain had an apparent  $K_D$  of 0.45 nM with this oligo, closer to its affinity for the 'Hipk-RA core' than the 'lola core' (Table 5.2 and Fig 5.19C). Without further experiments to investigate whether it is the additional length of the sequence or the extra core motif that is having an impact, it is hard to conclude what is driving the increased affinity from the 'lola core' to the 'Hipk double core', while the affinity for the latter is still lower than the 'Hipk-RA core'.

As described earlier, these three RNA oligos that featured a 3'-UTR core motif, (A/U/G)CUAAC, were each compared to an RNA oligo that contained the core motif as well as its closest half-site (UAAY). These were designed to measure whether these half-sites increased HOW's STAR domain affinity for the RNAs. Given the aforementioned poor fit of the 'Hipk-RA core + half-site' data, potentially because of secondary structure formation, it cannot be used to assess whether half-sites impact the STAR domain's affinity for an RNA (Fig 5.19B). The 'lola core + half-site' RNA had an apparent  $K_D$  of 2.89 nM, more than 1.5 times lower than the 'lola core' alone (Table 5.2 and Fig 5.19D). In contrast to this, the 'Hipk double core + half-site' oligo had an apparent  $K_D$  of 9.42 nM, more than 20 times greater than the 'Hipk double core' oligo (Table 5.2 and Fig 5.19E), and in general the STAR domain had the lowest affinity for this oligo out of all the 3'-UTR based RNA oligos. However, the data from the 'Hipk double core + half-site' has a large amount of standard error. It also plateaued at a higher anisotropy value, around 0.06, than the other 3'-UTR based oligos (around 0.03–0.04; Fig 5.17), which could indicate multiple binding sites, aggregation or other confounding factors. Overall, unfortunately, these data do not enable us to conclude whether a nearby half-site improves or hampers the STAR domain's affinity for an RNA.

**Table 5.2. Apparent  $K_D$  of the STAR domain for RNAs containing the 5'- and 3'-UTR binding motifs.** The apparent  $K_D$  and related standard error for nine oligos. The  $K_D$  was determined from the logistic equation used to fit the fraction bound data from fluorescence anisotropy assays. The 3'-UTR motifs are in pink, half-sites in orange, and the 5'-UTR motif is in blue.

Oligo name	Apparent $K_D$ (nM)	Standard error ( $\pm$ )	Oligo sequence
lola core	4.80	0.93	ACACACUAACUCGU
lola core + half-site	2.89	0.74	ACACACUAACUCGUAAUCUAUG
Hipk double core	0.45	0.16	UACAACUAACAACUAACAGAU
Hipk double core + half-site	9.42	2.13	UACAACUAACAACUAACAGAUAAACAAUU
Hipk-RA core	0.10	0.02	AGCAGCUAACAAUU
Hipk-RA core + half-site	0.86	1.92	AGCAGCUAACAAUUGUAAUUGUA
jv1 5'-UTR core	4352	23109	GUUGGCGUGUUUU
Poly(A) <sub>10</sub>	DNB	—	AAAAAAAAAA
Poly(U) <sub>10</sub>	DNB	—	UUUUUUUUUU

## 5. Characterising the molecular mechanism of RNA binding by HOW



**Figure 5.19. STAR domain binds the ‘Hipk-RA core’ oligo with highest affinity.**

Scatter plots of STAR protein concentration against the mean fraction of RNA bound for seven oligos (listed in key) with the logistic curves shown. A) Fraction bound data of six of the oligos with their logistic curve fits. B) Comparison of the fraction bound data from the Hipk-RA core motif with and without a half-site. C) Comparison of the fraction bound data from the three oligos designed around the (A/U/G)CUAAC 3'-UTR core motif without a half-site. D) Comparison of the fraction bound data from the lola core oligo with and without a half-site. E) Comparison of the fraction bound data from the Hipk double core oligo with and without a half-site.

## 5.3. Discussion

### 5.3.1. HOW's STAR domain binds the 3'-UTR motif, but not the 5'-UTR motif, with high affinity

From the FA experiments presented in this chapter, we have shown that HOW's STAR domain binds to specific RNAs with high affinity. High affinity binding was demonstrated with all RNAs containing (A/G)CUAAC, the motif enriched in the 3'-UTRs of HOW(S) bound RNAs (Fig 5.19 and Table 5.2). But the STAR domain did not bind to the generic poly(A) and poly(U) RNAs (Fig 5.16) and it had a very low affinity for the oligo derived from *juv*'s 5'-UTR, which contained a GCGUG motif (Table 5.2). This motif was enriched in the 5'-UTRs of RNAs that were bound by HOW(S) (Fig 3.27). It has been shown that the strength of *in vitro* binding does correlate with the strength of binding *in vivo* (Taliaferro et al., 2016), making it unlikely that the GCGUG sequence is a binding site for HOW(S).

Given that the two different UTR motifs had similar *E*-values from the DREME analysis that detected their enrichment (Figs 3.25 and 3.27), the 5'-UTR motif could present a binding motif of another RNA-binding protein that interacts with HOW(S). In the *in vitro* RNAcompete experiments that defined heptamer consensus binding sequences for *D. melanogaster* RBPs, SRA stem-loop interacting RNA binding protein 1 (SLIRP1) heptamer was defined as NNGCG(U/C)(G>A/C/U) (Ray et al., 2013), which was the closest match to the 5'-UTR motif enriched in HOW(S) bound RNAs — GCG(A/U)G. While the cellular localisation of SLIRP1 has not been reported in the literature, modENCODE tissue data shows that the highest expression of SLIRP1 is in the testis (Brown et al., 2014). Hence, it is possible that SLIRP1 and HOW are co-expressed in germ cells.

### 5.3.2. Guanosine in the first position of the HOW binding motif may be preferred over adenosine

The oligos that incorporated the motif enriched in the 3'-UTR of HOW(S) bound transcripts — (A/U/G)CUAAC — were designed from the 3'-UTRs of *lola* and *hipk*. The oligos that contained a single motif from the *lola*-RJ and *Hipk*-RA transcripts were each 14 nucleotides long and were bound to the STAR domain with high affinity,  $4.8 \pm 0.93$  nM and  $0.10 \pm 0.02$  nM, respectively (Fig 5.19 and Table 5.2). The apparent  $K_D$  for the 'Hipk-RA core' oligo was the lowest for all nine RNAs tested. While previous papers on HOW-RNA binding did not establish  $K_D$  values, the dipping experiments with 200 nM, 20 nM and 2 nM of HOW protein found that HOW bound to RNA at 2 nM when the ACUAA binding sequence was embedded within a loop structure (Nir et al., 2012), suggesting the strength of affinity reported here is comparable. While it has been reported that STAR

## 5. Characterising the molecular mechanism of RNA binding by HOW

proteins require a half-site for sufficient binding (Galarneau and Richard, 2009), the results here support previous findings that multiple STAR proteins can bind a single site alone in the low nanomolar range (Garrey et al., 2006; Carmel et al., 2010).

In comparing the ‘lola core’ and ‘Hipk-RA core’ it was surprising that the HOW STAR domain had an affinity for the Hipk-RA based oligo that was one order of magnitude greater than the lola-RJ based oligo (Table 5.2). If the binding specificity is entirely driven by the hexamer motif then the only difference between these two oligos is that the lola-RJ motif begins with an adenosine and the Hipk-RA motif start with a guanosine. If a guanosine is equivalent, or even preferred, this would be a unique among the STAR protein consensus binding sequences, which have consistently demonstrated preference for an adenosine in this position (Table 1.1; Garrey et al., 2006; Carmel et al., 2010; Corioni et al., 2011). However, from the experiments carried out here we cannot rule out the influence from the flanking regions on HOW–RNA binding when comparing these oligos. Thus, further experiments to test the other three nucleotides (i.e. adenosine, cytidine and uradine) in the first position of the motif in the background of the Hipk-RA core oligo would help define how the different nucleotides drive HOW binding specificity in this first position.

Both of the core motifs in the ‘Hipk double core’ RNA, found in the 3’-UTR of all *hipk* transcripts (Fig 5.12), began with an adenosine and were separated by just one nucleotide. HOW’s STAR domain had a greater affinity for this RNA than the ‘lola core’ (Table 5.2). Given how close these the two motifs are in the double core oligo it seems unlikely that this improved affinity would be due to HOW functioning as a dimer, as the crystal structure of GLD-1’s STAR domain suggests that the two sites need to be at least 10 nucleotides apart (Teplova et al., 2013). It is also difficult to make a direct comparison between the single core oligos and the double core oligo due to the difference in their lengths — 14 and 21 nucleotides, respectively.

The enriched motif, and in the oligos tested here, contain a cytidine in the sixth position, which is not included in the consensus binding motifs of the other proteins in the Quaking subfamilies (Carmel et al., 2010). However, it is consistent with the optimal heptamer sequence found in the *in vitro* RNAcompete experiment: (A>U/G/C)CUAACN. Further, the optimal binding sequence for SF1 and BBP, both in the SF1 subfamily, have a cytidine in the equivalent position (Table 1.1; Garrey et al., 2006; Corioni et al., 2011). Additionally, the crystal structures of Quaking and GLD-1 bound to RNAs showed amino acid residues as interacting with cytidines in the equivalent position (Teplova et al., 2013). Collectively with our RIP-seq identified motif, it is likely that residues from the HOW(S) STAR domain directly interacts with at least 6 nucleotides and that the HRE, previously defined as ACUAA, could be extended, though further experimentation is required.

### 5.3.3. The effect of half-sites and RNA structure on HOW–RNA interactions

As mentioned in the introduction to this chapter, the details of how STAR proteins bind bipartite motifs is somewhat unclear. SELEX experiments indicated that STAR proteins can bind full sites with a half-site 2–20 nucleotides away (Galarneau and Richard, 2009), while other studies have indicated that STAR proteins' affinity for RNA increases when the two full sites are 10 or more nucleotides away, depending on the specific protein (Carmel et al., 2010; Feracci et al., 2016). Here, the three core oligos were extended to include the nearest UAAY half-sites, which were three nucleotides downstream the for the lola core and Hipk double core oligos and five nucleotides for the Hipk-RA core (Table 5.1). Unfortunately, our results were inconclusive as to whether the presence UAAY half-site increases HOW's affinity for an RNA because the addition of the half-sites had differing effects on the STAR domain's apparent  $K_D$  in the three different oligos tested (Fig 5.19 and Table 5.2).

The effect of HOW–RNA binding from the addition of a half-site to the Hipk-RA core could not be assessed because the 'Hipk-RA core + half-site' oligo was predicted to have a secondary structure where the core motif was in a stem structure, rendering it partially inaccessible (Fig 5.18). The binding curve did not adhere to a normal sigmoid shape, and while a logistic equation was able to fit a curve to this data the calculated  $K_D$  is not reliable. However, this atypical binding curve does suggest that binding is affected by the accessibility of the core motif, if the secondary structure prediction is accurate, in line with previous experiments that found that HOW could not bind the ACUAA sequence when embedded in a stem structure (Israeli et al., 2007).

### 5.3.4. Conclusions

In summary, HOW's STAR domain can specifically, and with high affinity, bind RNAs with an (A/G)CUAAC motif but does not bind the GCGUG motif. The lowest apparent  $K_D$  was with the oligo derived from the Hipk-RA sequence, providing further evidence that *hipk* mRNA may be bound, and regulated, by HOW(S) in *D. melanogaster* testis.



## 6. Discussion

Spermatogenesis is a complex differentiation process and the regulation of RNA metabolism by RNA-binding proteins is needed throughout the process for the proper production of mature sperm cells. Held out wings is one of the RBPs that is essential for spermatogenesis in *D. melanogaster*. When *how* is knocked down in the testis there is a dramatic loss of germ cells (Fig 4.1; Monk et al., 2010). HOW is important for both the survival of GSCs in the testis and for regulating the mitotic divisions in the first phase of spermatogenesis (Monk et al., 2010). For the latter function, HOW's longer and nuclear isoform, HOW(L), is important, as overexpression of HOW(L) in GSCs and early spermatogonia result in extra rounds of mitotic divisions (Monk et al., 2010). This same phenotype is seen in testes lacking in *bam*, a key regulator of mitotic divisions and the entry into meiosis (Insko et al., 2009). HOW has been shown to bind *bam* RNA, and given HOW(L)'s previously established roles in mRNA repression and nuclear retention, it is likely that this HOW isoform suppresses the expression of *bam* in the early spermatogonia to help maintain a balance between proliferation and differentiation of the spermatogonia (Monk et al., 2010).

Although HOW is also expressed in the GSCs, its role in GSC maintenance in the testis is not well understood. In the *how<sup>r17/r4</sup>* mutant flies, which combines a total loss of function *how* allele and a partial loss of function *how* allele, there is a decrease in the number of GSCs, and sometimes a total loss, in the testes (Monk et al., 2010). Overexpression of either HOW(L) or HOW(S) can partially rescue this loss, suggesting that both isoforms have functions in GSC maintenance (Monk et al., 2010). It is unlikely that they are working redundantly as HOW(L) is localised to the nucleus and HOW(S) to the cytoplasm (Fig 3.5; Monk et al., 2010).

In many cell types where HOW is expressed, including germ cells, HOW(S)'s cytoplasmic function is poorly understood compared to HOW(L)'s in the nucleus. For example, there are ten mRNAs that have previously been shown to specifically bind HOW(L), while there are only three for HOW(S) (Table 1.2). Moreover, one of the three HOW(S) targets, *nrx-IV*, is related to HOW(S)'s role in alternative splicing in the nucleus of glial cells, rather than a cytoplasmic function (Edenfeld et al., 2006). Given the incomplete understanding of both HOW's role in spermatogenesis and of HOW(S)'s cytoplasmic function, this thesis focused on determining HOW(S)'s RNA-binding activity in spermatogenesis.

## 6.1. HOW(S) binds RNAs associated with signal transduction functions

The first aim of this thesis was to identify the targets of HOW(S) in germ cells, which was achieved by using RIP-seq. 343 genes and 121 transcripts were identified using parallel genome and transcriptome based analyses (Figs 3.19 and 3.20). This is the first time an omics approach has been used to identify HOW-bound RNAs in any tissue, and, given the large number that have been identified here, this dataset could be useful for studying the role of HOW in contexts outside of germ cells. As RIP-seq does not identify RNAs via direct binding sites, some of these targets may not be directly bound by HOW(S) but instead might exist in RNP complexes with HOW(S). However, many of the RNAs identified contained motifs in their UTRs that are very similar to other STAR protein consensus sequences, making them strong candidates as RNAs directly bound by HOW(S) (Fig 3.25 and Table 1.1).

The majority of the work in this thesis focused on the transcript-level data as this allowed for specific 5'- and 3'-UTRs to be studied in motif enrichment analysis and the direct binding assays. While motif enrichment was carried out on the gene-level data, these results were probably less accurate as the 5'- and 3'-UTRs selected for the analysis were from the highest expressed transcript for each gene. However, for quantifying differential enrichment gene-level analysis can be more robust than transcript-level as it is less granular (Love et al., 2014). Therefore, future work exploring the potential HOW(S) RNA targets from the gene-level list could be insightful. Insulin receptor (InR), which was a differentially enriched gene but none of its transcripts were significantly enriched (Tables A.8 and A.9), stands out as a gene worth further investigation as a HOW(S) target because loss of functional InR can result in a loss of GSCs, and constitutively active InR can prevent the typical loss of GSCs observed in starved male flies (Ueishi et al., 2009; McLeod et al., 2010). If HOW(S) regulates the expression of InR in GSCs, this could contribute to why HOW is important for the maintenance of GSCs.

The GO terms enriched in the lists of HOW(S) bound genes and transcripts were related to cell signalling and signal transduction (Figs 3.22 and 3.23). As a STAR protein, HOW has been shown to be responsive to cellular signalling, for example its phosphorylation by the MAPK/ERK can enhance its ability to bind RNA (Nir et al., 2012). From this GO term analysis, it also appears that HOW(S) might itself regulate signalling pathways via its RNA-binding activity. One of the most promising candidates as an RNA that may require HOW(S) regulation for proper spermatogenesis is Hipk, which is annotated with GO terms corresponding to 'signal transduction' (Fig 3.24).

## 6.2. *Hipk* mRNA could be an important target of HOW(S) for spermatogenesis

*Hipk* is a kinase involved in several signalling pathways (Lee et al., 2009; Tettweiler et al., 2019), and was one of the RNAs identified that was significantly enriched in both the transcript-level and gene-level analyses (Table 4.1). It was one of the 12 transcripts that was selected for further investigation into its role in spermatogenesis. The flies from the maternal *Hipk* RNAi 2 cross, which knocked down expression of *Hipk* in the GSCs and spermatogonia, had small testes with abnormal morphology and were either sterile or had significantly impaired fertility. However, as discussed in Chapter 4, more work is needed to verify the effect of loss of *Hipk* on germ cells, as the results were not consistent between the paternal and maternal crosses or the two different RNAi lines used. If *Hipk* is required for spermatogenesis, these effects on morphology and fertility need to be characterised further to understand *Hipk*'s function in the testis. For example, from the immunofluorescence experiments we can see early elongating spermatids are present in all of the testes categorised as having strong morphological defects, so is the main reason for infertility due to issues with spermiogenesis? Additionally, the apical tip is narrow in the testes in the strong and weak morphological phenotype categories; is this because of a loss of GSCs or an overproliferation of early spermatogonia or another reason?

In the direct binding assays, HOW's STAR domain had the highest affinity for the 14 nucleotide sequence from *Hipk*-RA's 3'-UTR out of all the oligos containing an (A/U/G)C UAAC motif (Table 5.2). This further supports the RIP-seq data to indicate that *Hipk* mRNA is a specific target of HOW *in vivo*. However, further work is needed to determine the effect of HOW(S) binding to *hipk* transcripts, and whether it is required for *hipk* mRNA regulation during spermatogenesis. Previously, HOW(S) has been shown to stabilise its RNA targets (Nabel-Rosen et al., 2002; Giuliani et al., 2014), so reporter assays that examined whether HOW(S) has a stabilising effect on *hipk* mRNA, and whether this was dependent on the (A/U/G)CUAAC sites in *hipk*'s 3'-UTRs (Fig 5.12), will be required to dissect the mechanistic relationship between them. If HOW(S) is regulating *Hipk* expression, via increasing the *hipk* transcript stability or by other means, then additional experiments to explore a causal link between HOW(S)'s RNA-binding activity and *Hipk*'s role in spermatogenesis would be necessary. Such experiments could include mutating the (A/U/G)CUAAC sites in *hipk* transcripts and examining whether this changes *Hipk* protein expression and/or recapitulates any of the testis morphology and male fertility phenotypes seen with the *Hipk* RNAi experiments reported here.

### 6.3. Expanding the consensus sequence that drives HOW(S)–RNA interactions

STAR proteins in the Quaking and SF1 subfamilies have similar consensus sequences that are six or seven nucleotides long (Table 1.1). Experiments that have defined the optimal binding sequences for several of these proteins have revealed subtle differences between them, for example, *S. cerevisiae* BBP is less flexible than its human orthologue, SF1, and also binds to RNA with a higher affinity (Garrey et al., 2006; Corioni et al., 2011). HOW’s previously reported binding sequence is A(C>A)UAA (Israeli et al., 2007), however, this was originally identified from a single mRNA target of HOW (*stripe*’s 3’-UTR). Though this pentamer has been found in other HOW mRNA targets, other experiments have indicated that HOW has a broader binding capability (e.g. *msl-2*; Graindorge et al., 2013). Additionally, the affinity HOW has for RNAs could not be quantified from the type of binding experiments that have been carried out prior to this study.

The transcriptome-wide approach to identify HOW(S) targets in Chapter 3 provided a much bigger pool with which to examine HOW(S)’s consensus sequence than any previous experiments. However, one of the limitations of RIP-seq is that it does not identify direct RNA-binding sites, which experiments such as CLIP-seq are able to. RIP-seq also makes identifying a consensus sequence more challenging as indirect RNA targets can also be pulled-down if the protein of interest is part of larger RNP complexes. Despite these limitations, RIP-seq was employed because it does not have UV cross-linking or RNA digestion steps and so requires less starting material than CLIP-seq. Thus RIP-seq provided a balance between the time it takes to acquire enough tissue for an experiment while still being able to explore HOW(S)–RNA interactions in a transcriptome-wide manner. To identify potential binding sites in the RNAs bound by HOW(S), DREME was used for motif enrichment analyses. This revealed two very different motifs from the 5’- and 3’-UTRs of the RNAs bound by HOW(S) — GCG(A/U)G and (A/U/G)CUAAC, respectively (Figs 3.25B and 3.27B). While the motifs from the enriched gene list are probably less accurate than the ones found using the transcript list (Figs 3.25A and 3.27A), they were remarkably similar, increasing our confidence in the motifs discovered from the list of transcripts enriched in the HOW(S)-HA RIP-seq.

From the FA assays carried out, I found that HOW’s STAR domain has strong nanomolar affinity for RNA oligos containing (A/G)CUAAC motifs but weak affinity for the ‘jvl 5’-UTR core’ oligo with a GCGUG motif (Table 5.2). While not definitive, our initial results indicate that a guanosine in the first position of the motif might be preferred over an adenosine (Fig 5.19C), this would be a unique preference in this nucleotide position for STAR proteins (Table 1.1). Additional questions remain over what constitutes HOW’s optimal binding sequence, such as, does a sixth nucleotide, and specifically a cytidine, increase HOW’s affinity for an RNA? A sixth nucleotide was not previously reported as part of the HOW binding sequence (Israeli et al., 2007), however, the STAR domains from

other Quaking proteins have hexamers as their optimal sequences so it could certainly be possible that its important for HOW binding too (Table 1.1).

## 6.4. Additional features may contribute to HOW(S)–RNA interactions

In combination with HOW’s consensus sequence other features, such as RNA secondary structure and interactions with other RBPs, could affect HOW(S)–RNA interactions. Previous dipping experiments showed that HOW has a higher affinity for binding sites in a hairpin loop rather than unstructured RNAs (Fig 1.10; Israeli et al., 2007). Finding the secondary structure of the 3’-UTRs could help to explain why some RNAs are bound by HOW and not others. Unfortunately, there are currently no *D. melanogaster* datasets in the RNA Atlas of Structure Probing database, which collates data from experiments such as SHAPE-seq and DMS-seq that can determine the secondary structure of RNAs in a transcriptome-wide manner (Li et al., 2021). However, if this data became available in the future, it could help to indicate which of the RNAs are more likely to be directly bound by HOW(S) from the hundreds of potential transcripts and genes identified here.

HOW has been shown to directly interact with other proteins which affect its RNA-binding activity. HOW(S) is bound by Crooked neck in glial cells where the complex then translocates to the nucleus, there HOW(S) is able to affect the alternative splicing of *nrx-IV* (Edenfeld et al., 2006; Rodrigues et al., 2012). Further, HOW can interact with other RBPs, such as SXL and Hrp48 (Graindorge et al., 2013; Szostak et al., 2018), and here we found that PSI co-immunoprecipitated with HOW(S)-HA in the pull-downs performed for RIP-seq (Fig 3.10). Interactions with another RBP could be responsible for the strong enrichment of the GCG(A/U)G motif seen in the 5’-UTRs of the transcripts bound by HOW(S) (Fig 3.27), despite the fact that HOW’s STAR domain only has weak affinity to this motif (Table 5.2). A possible candidate is SLIRP1, which is has a similar binding motif to the 5’-UTR enriched motif discovered here (Ray et al., 2013). Investigating whether SLIRP1, or other RBPs, interact with HOW(S) in germ cells via pull-downs followed by mass spectrometry or western blots, could also contribute to understanding why certain RNAs are bound by HOW(S) and not others. Solving the structures of such complexes might reveal how these interactions occur and how they may affect RNA-binding activity because currently there are no structures of any of the Quaking proteins with their protein binding partners.

## 6.5. HOW(S) function in the testis

In germ cells HOW is required for stem cell maintenance and for regulating the transit-amplifying mitotic divisions. Overexpression of HOW(S) does not interfere with the mitotic divisions, unlike overexpression of HOW(L), but overexpression of HOW(S) in mu-

## 6. Discussion

tant *how* flies can partially rescue the loss of GSC phenotype (Monk et al., 2010). From the results of this thesis it has not been possible to understand the mechanism by which HOW(S) contributes to maintaining the GSC population, but many RNAs that HOW(S) binds have been identified here, which will hopefully help future studies answer this question. For example, *hipk* and *InR* show potential as targets of HOW(S) that are important for spermatogenesis.

As for the way in which HOW(S) regulates RNA metabolism, the only previously demonstrated function for HOW(S) in the cytoplasm is in increasing the stability of the RNAs it binds. This has been shown with HOW(S)'s *stripe* and *dgrasp* mRNA targets (Nabel-Rosen et al., 2002; Giuliani et al., 2014). A role in stabilising mRNAs suggests that HOW(S) is a positive regulator of protein expression. Consequently, one would expect a loss of HOW(S) to result in a decrease in the protein expression of its RNA targets, which is what occurs with the *Hipk* RNAi and *InR* mutants which have defects in spermatogenesis. To understand the function of HOW(S) in the testis, future work will need to determine whether it does confer a stabilising effect on *hipk* and the other RNAs identified here. However, alongside mRNA stability several of the cytoplasmic STAR proteins also regulate mRNA translation. For example, GLD-1 is a translational repressor in gametogenesis (Marin and Evans, 2003; Scheckel et al., 2012), in mouse spermatogenesis Sam68 a translational activator (Paronetto et al., 2006; Paronetto et al., 2009), and the short cytoplasmic isoform of Quaking, Quaking-6, can autoregulate the translation of the *quaking* mRNA isoforms and translationally repress other RNAs (Saccomanno et al., 1999; Zhao et al., 2010b; Fagg et al., 2017). HOW(S) may also autoregulate the *how* mRNAs, as HOW-RC, which translates to HOW(M), is one of the 121 differentially enriched transcripts (Table A.9). Thus, future experiments that explored whether HOW(S) regulates the translation of the RNAs it binds, as well as its own transcripts, is merited and would help to better describe its role in spermatogenesis, and potentially other processes too.

## 6.6. Conclusions

HOW is an essential RNA binding protein in *D. melanogaster* spermatogenesis and this thesis presents hundreds of potential RNAs bound by HOW(S) in the GSCs and early spermatogonia. This includes 121 transcripts and 343 genes from the parallel transcript- and gene-level analyses. This is the first time a transcriptome-wide approach has been taken to identify the RNAs bound by any isoform of HOW and this large dataset enabled the identification of two motifs enriched in the transcripts bound by HOW(S) — (A/U/G)CUAAC and GCG(A/U)G. FA results showed that HOW's RNA binding domain, the STAR domain, can directly bind the (A/U/G)CUAAC motif with high affinity (a  $K_D$  of 0.10 nM), the first time HOW's RNA-binding activity has been quantified. The weak binding to the GCG(A/U)G motif suggests HOW(S) interacts with other proteins in the testis to increase its specificity for its RNA targets, however, more work is needed to confirm this. The STAR domain binds with particularly high affinity to an oligo with

a GCUAAC motif from a *hipk* transcript. Finally, initial experiments investigating the testis morphology and fertility of *Hipk* knockdown flies indicate that *hipk* mRNA could be a target of HOW(S) that is important for proper spermatogenesis.



# A. Appendix I

## A.1. DNA and amino acid sequences

‘UAS fwd + HOW rev’, ‘HOW fwd + HA rev’ and ‘HA fwd + SV40 rev’ primer combinations were used to sequence the full HOW(S)-HA coding sequence from UAS-HOW-S-HA flies. The genomic sequence showed a few minor difference to the sequence on FlyBase (Sequence A.1). The corresponding amino acid sequence (Sequence A.2) indicates that there are two differences (noted in pink) between the amino acid in the HOW(S)-HA transgene and the annotated HOW(S) coding sequence (FBpp0083576) are all outside of the STAR domain. The first is an extra glutamine (Q) and alanine (A) in the QA-rich region. The second difference is an additional 14 amino acids that in FlyBase is part of the HOW-PF isoform (FBpp0307182; often referred to as HOW(M)). Importantly, this HOW(S)-HA transgene does not contain the nuclear localisation signal found in the HOW(L) isoform but does have the final 6 amino acids (GGLFAR), which are unique to the HOW-PB isoform, which corresponds to HOW(S), are present here.

```
ATGAGTGTCTGTGAGAGCAAAGCCGTTGTGCAACAGCAACTGCAGCAGCACTTGCAGCAACAGGCAGCCGCAGC
AGTTGTTGCGGTTCGCGCAACAGCAGCAGGCTCAAGCCCAAGCTCAAGCCAGGCTCAGGCAGGCACAGCAGC
AGCAACAGGCGCCGAGGTGGTGGTCCCCATGACCCCGCAGCACTTGACCCACAGCAGCAGCAGCAGAGCACA
CAGAGCATCGCCGACTATCTGGCCAGTTGCTCAAGGACCGCAAGCAAGCTGGCCGCTTCCCCAACGTCTTCAC
CCACGTGCAACGCCTGCTGGACGAAGAAATTGCACGCGTGCAGCCCTCACTGTTCCAGATCAATGGGGTCAAGA
AGGAGCCGCTCACTCTGCCCAGCCGAGGGCTCTGTGGTGACGATGAACGAGAAGGTTTATGTGCCAGTCCGC
GAGCATCCAGATTTCAACTTTGTGCGTGCATTTTGGGACCCCGTGGCATGACCGCCAAGCAATTGGAACAGGA
GACCGGCTGCAAGATTATGGTCCGAGGCAAGGGTTCCATGCGCGACAAGAAGAAGGAGGACGCCAACCCTGGCA
AGCCAACTGGGAGCATCTGTCCGATGACCTGCATGTCTGATAACCGTCGAGGACACCGAGAACCCTGCCACA
GTGAAGTTGGCCCAGGCCGTCGCCGAAGTACAGAAGTTGCTCGTGCCGCAAGCCGAAGGCGAAGATGAGCTAAA
GAAACGTCAACTCATGGAATTGGCGATTATTAATGGCACTTATAGGGACACAACAGCGAAATCTGTGCGAGCTT
TCTCATGCGTTGGCTCTGCTTCTTATCTGTATCCCGCAGTGTGCGATGAGGAGTGGCGCCGCTGTTGCCGCC
TCTGATAGCCGCTGCTGACATCCACCGCCTGCCCGCCCTTGCCGCCAGATCCGTGCACCCGCCGCCGCC
GCTTGGCGCCCCATTGATCCTGAATCCCCGATGACCGTCCCCACAACGGCGGCCAGCATATTGTCCGCCAGG
CCGCTCCGACAGCCGCTTCGACCAGACCGCCATGGAATGATCTTCGCACCGTACGATTATGCGAACTACGCC
GCCCTAGCCGCAATCCTCTGCTCACGGAATATGCTGATCATAGCGGTGGGTGTTTTGCCAGACACCCAGCTTT
CTTGTAACAAGTGGTGAGCTCCGCCACCATGGATCTCCACCGGTTGGAGGCCGCATCTTTTACCCATACGATG
TTCCTGACTATGCGGGCTATCCCTATGACGTCCCGACTATGCAGGATCCTATCCATATGACGTTCCAGATTAC
GCTGCTCATGGCGGAC
```

**Sequence A.1. Genomic sequence of the HOW(S)-HA coding region in the UAS-HOW-S-HA fly.** HA-tag is coloured in blue. Differences between this sequence and the HOW(S) FlyBase annotation (HOW-RB; FBtr0084178) that change the protein sequence (HOW-PB; FBpp0083576) are coloured in pink, synonymous substitutions are coloured in orange.



MAHHHHHSSGMKIEEGKLVWINGDKGYNGLAEVGKKFEKDTGIKVTVEHPDKLEEKFPQVAATGDGPDIIFW  
AHRDFGGYAQSGLLAEITPKAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSIIYNKDLLPNPPKTWEEIPALD  
KELKAKGKSALMFNLQEPYFTWPLIAADGGYAFKYENGYDIKDVGVNAGAKAGLTFLVDLIKHKHMNADTDY  
SIAEAAFNKGETAMTINGPWAWSNIDTSKVNYGVTVLPTFKGQPSKPFVGVLSAGINAASPNKELAKEFLENYL  
LTDEGLEAVNKDKPLGAVALKSYYEELAKDPRIAATMENAQKGEIMPNIQMSAFWYAVRTAVINAASGRQTV  
EALKDAQTSSGLEVLFGGPTQSIADYLAQLLKDRKQLAAFPVNFTHVERLLDEEIARVRASLFQINGVKKEPLT  
LPEPEGSVVTMNEKVYVPVREHPDFNFVGRILGPRGMTAKQLEQETGCKIMVRGKGSMDKKKEDANRGKPNWE  
HLSDDLHVLITVEDTENRATVKLAQAVAQVQKLLVPAEGEDELKKRQLMELAIINGTYRDTTAKS

**Sequence A.6. Amino acid sequence of the His-MBP-STAR construct.** 584 amino acids, theoretical molecular weight of 64,711 Da. His-tag coloured in blue, MBP in purple, 3C cleavage sequence in green, HOW STAR domain in black.

MAHHHHHSGSDSEVNQEAKEVKEVKEPETHINLKVSDGSSEIFFKIKKTTPLRRLMEAFKRQKEMDSLRF  
YDGIRIQADQTPEDLDMEDNDIIEAHREQISSGLEVLFGGPTQSIADYLAQLLKDRKQLAAFPVNFTHVERLLD  
EEIARVRASLFQINGVKKEPLTLPEPEGSVVTMNEKVYVPVREHPDFNFVGRILGPRGMTAKQLEQETGCKIMV  
RGKGSMDKKKEDANRGKPNWEHLSDDLHVLITVEDTENRATVKLAQAVAQVQKLLVPAEGEDELKKRQLMEL  
AIINGTYRDTTAKS

**Sequence A.7. Amino acid sequence of the His-SUMO-STAR construct.** 310 amino acids, theoretical molecular weight of 35,213 Da. His-tag coloured in blue, SUMO in purple, 3C cleavage sequence in green, HOW STAR domain in black.

GPTQSIADYLAQLLKDRKQLAAFPVNFTHVERLLDEEIARVRASLFQINGVKKEPLTLPEPEGSVVTMNEKVYV  
PVREHPDFNFVGRILGPRGMTAKQLEQETGCKIMVRGKGSMDKKKEDANRGKPNWEHLSDDLHVLITVEDTEN  
RATVKLAQAVAQVQKLLVPAEGEDELKKRQLMELAIINGTYRDTTAKS

**Sequence A.8. STAR domain after cleavage by PreScission protease.** 197 amino acids, theoretical molecular weight of 22,171 Da. Remaining two amino acids from the 3C cleavage site is in green, HOW STAR domain in black.

## A.2. Test statistic tables

**Table A.1. Statistics for the number of progeny produced after first mating in sperm competition assay.** Only the *nanos*-GAL4 data were normal (tested via the Shapiro–Wilk test), so the non-parametric Kruskal–Wallis  $H$  test was used to compare the number of progeny produced after the first mating in the sperm competition assay between male genotypes. The test statistic ( $H$ ) with the degrees of freedom in subscript are reported with the  $p$ -value. The null hypothesis was rejected (i.e.  $p < 0.05$ ), so post-hoc pairwise comparisons were carried out using Dunn’s test, the standardised test statistic ( $Z$ ) with the degrees of freedom in subscript are reported with the Bonferroni corrected  $p$ -value. Graphical representation of data in Fig 3.7.

Sample comparison	Test statistic	$p$ -value
All male genotypes	$H_4 = 15.631$	0.004
Scarlet and WT	$Z_1 = -3.436$	0.006
Scarlet and UAS-HOW-S-HA	$Z_1 = -1.284$	1.000
Scarlet and <i>nanos</i> -GAL4	$Z_1 = 0.585$	1.000
Scarlet and <i>nanos</i> -GAL4>HOW(S)-HA	$Z_1 = 0.987$	1.000
WT and UAS-HOW-S-HA	$Z_1 = 1.429$	1.000
WT and <i>nanos</i> -GAL4	$Z_1 = 2.035$	0.419
WT and <i>nanos</i> -GAL4>HOW(S)-HA	$Z_1 = -3.125$	0.018
UAS-HOW-S-HA and <i>nanos</i> -GAL4	$Z_1 = 1.150$	1.000
UAS-HOW-S-HA and <i>nanos</i> -GAL4>HOW(S)-HA	$Z_1 = 1.705$	0.882
<i>nanos</i> -GAL4 and <i>nanos</i> -GAL4>HOW(S)-HA	$Z_1 = -0.003$	1.000

**Table A.2. Statistics for the sperm competition assay.** None of the data were normal (tested via the Shapiro–Wilk test), so the non-parametric Kruskal–Wallis  $H$  test was used to compare the results within the P1 and P2 categories. The test statistic ( $H$ ) with the degrees of freedom in subscript are reported with the  $p$ -value. If the null hypothesis was rejected (i.e.  $p < 0.05$ ) post-hoc pairwise comparisons were carried out using Dunn’s test, the standardised test statistic ( $Z$ ) with the degrees of freedom in subscript are reported with the Bonferroni corrected  $p$ -value. Graphical representation of data in Fig 3.8.

Sample comparison	Test statistic	$p$ -value
Focal males in P1	$H_3 = 31.575$	$6.433 \times 10^{-7}$
WT and UAS-HOW-S-HA	$Z_1 = 4.960$	$4.222 \times 10^{-6}$
WT and <i>nanos</i> -GAL4	$Z_1 = 3.450$	0.003
WT and <i>nanos</i> -GAL4>HOW(S)-HA	$Z_1 = 4.211$	$1.523 \times 10^{-4}$
UAS-HOW-S-HA and <i>nanos</i> -GAL4	$Z_1 = 0.240$	1.000
UAS-HOW-S-HA and <i>nanos</i> -GAL4>HOW(S)-HA	$Z_1 = 1.407$	0.956
<i>nanos</i> -GAL4 and <i>nanos</i> -GAL4>HOW(S)-HA	$Z_1 = -0.774$	1.000
Focal males in P2	$H_3 = 4.686$	0.196

**Table A.3. Statistics for transcript feature assignments.** Genome-aligned reads were assigned to different features using featureCounts. Paired comparisons were made between the three replicates of the total RNA and pull-down RNA from the HOW(S)-HA RIP-seq. The percentage of total RNA reads assigned to 5’-UTR were not normal (tested via Shapiro–Wilk test), so a Wilcoxon signed-rank test was carried out, the test statistic ( $W$ ) with the number of paired samples ( $N$ ) are reported with the  $p$ -value. All other samples were normal, so paired  $t$ -test was carried out and the test statistic ( $t$ ) with the degrees of freedom in subscript are reported with the  $p$ -value. Graphical representation of data in Fig 3.16.

Transcript feature	Test statistic	$p$ -value
5’-UTR	$W = 3.0, N = 3$	1.000
CDS	$t_2 = 2.0360$	0.179
3’-UTR	$t_2 = 2.656$	0.117

**Table A.4. Statistics for Hipk expression determined via qRT-PCR.** The data for all samples in both sets of knockdowns (Hipk RNAi 1 and Hipk RNAi 2) were normal (tested via the Shapiro–Wilk test), except for the maternal knockdown crosses. Thus, the non-parametric Kruskal–Wallis  $H$  test was used to test for differences between the sets of knockdown samples for Hipk RNAi 1 and Hipk RNAi 2. Each set includes four samples: the two parental lines and the two knockdown crosses. The test statistic ( $H$ ) with the degrees of freedom in subscript are reported with the  $p$ -value. Graphical representation of data in Fig 4.5.

Sample comparison	Test statistic	$p$ -value
Hipk 1 sample set	$H_3 = 5.051$	0.168
Hipk 2 sample set	$H_3 = 4.128$	0.248

**Table A.5. Statistics for the number of hub cells.** The data for all samples in both sets of knockdowns (Hipk RNAi 1 and Hipk RNAi 2) were normal (tested via the Shapiro–Wilk test). Thus, a one-way ANOVA was performed. The test statistic ( $F$ ) with the degrees of freedom for the between groups and within groups in subscript are reported with the  $p$ -value. Post-hoc pairwise comparisons were carried out using Tukey’s honestly significant difference test, the mean difference ( $M$ ) are reported with the adjusted  $p$ -value. Graphical representation of data in Fig 4.10

Sample comparison	Test statistic	$p$ -value
Hipk 1 sample set	$F_{4,29} = 6.513$	0.002
WT and <i>nos</i> -GAL4	$M = -5.250$	0.001
WT and Hipk 1 parent	$M = -5.800$	0.001
WT and Hipk 1 maternal cross	$M = -3.571$	0.046
WT and Hipk 1 paternal cross	$M = -3.750$	0.026
<i>nos</i> -GAL4 and Hipk 1 parent	$M = -0.550$	0.992
<i>nos</i> -GAL4 and Hipk 1 maternal cross	$M = 1.679$	0.577
<i>nos</i> -GAL4 and Hipk 1 paternal cross	$M = 1.500$	0.646
Hipk 1 parent and Hipk 1 maternal cross	$M = 2.229$	0.422
Hipk 1 parent and Hipk 1 paternal cross	$M = 2.050$	0.478
Hipk 1 maternal cross and Hipk 1 paternal cross	$M = -0.179$	1.000
Hipk 2 sample set	$F_{4,29} = 5.334$	0.002
WT and <i>nos</i> -GAL4	$M = -5.250$	0.003
WT and Hipk 2 parent	$M = -5.143$	0.005
WT and Hipk 2 maternal cross	$M = -4.250$	0.022
WT and Hipk 2 paternal cross	$M = -2.600$	0.405
<i>nos</i> -GAL4 and Hipk 2 parent	$M = 0.107$	1.000
<i>nos</i> -GAL4 and Hipk 2 maternal cross	$M = 1.000$	0.920
<i>nos</i> -GAL4 and Hipk 2 paternal cross	$M = 2.650$	0.327
Hipk 2 parent and Hipk 2 maternal cross	$M = 0.893$	0.951
Hipk 2 parent and Hipk 2 paternal cross	$M = 2.543$	0.394
Hipk 2 maternal cross and Hipk 2 paternal cross	$M = 1.650$	0.752

**Table A.6. Statistics for the number of vials with  $F_2$  progeny in the 3-day mating assay.** Pearson's chi-squared test was carried out on the data, first comparing all four focal  $F_1$  males within each of the Hipk knockdown sets. Then pairwise chi-squared tests were carried out. The test statistic ( $\chi^2$ ) with the degrees of freedom in subscript are reported with the  $p$ -value. If there was only 1 degree of freedom (i.e. in the pairwise analysis) then Yates's correction for continuity was applied. Graphical representation of data in Fig 4.12.

Sample comparison	Test statistic	$p$ -value
All $F_1$ males related to the Hipk 1 knockdowns	$\chi_3^2 = 10.063$	0.014
<i>nanos</i> -GAL4 and Hipk 1 parent	$\chi_1^2 = 6.095$	0.014
<i>nanos</i> -GAL4 and Hipk 1 maternal cross	$\chi_1^2 = 0.000$	1.000
<i>nanos</i> -GAL4 and Hipk 1 paternal cross	$\chi_1^2 = 0.028$	0.868
Hipk 1 parent and Hipk 1 maternal cross	$\chi_1^2 = 1.237$	0.266
Hipk 1 parent and Hipk 1 paternal cross	$\chi_1^2 = 2.997$	0.083
Hipk 1 maternal cross and Hipk 1 paternal cross	$\chi_1^2 = 0.000$	1.000
All $F_1$ males related to the Hipk 2 knockdowns	$\chi_3^2 = 25.189$	$1.410 \times 10^{-5}$
<i>nanos</i> -GAL4 and Hipk 2 parent	$\chi_1^2 = 8.779$	0.003
<i>nanos</i> -GAL4 and Hipk 2 maternal cross	$\chi_1^2 = 13.881$	$1.948 \times 10^{-4}$
<i>nanos</i> -GAL4 and Hipk 2 paternal cross	$\chi_1^2 = 0.457$	0.499
Hipk 2 parent and Hipk 2 maternal cross	$\chi_1^2 = 1.434$	0.231
Hipk 2 parent and Hipk 2 paternal cross	$\chi_1^2 = 6.908$	0.009
Hipk 2 maternal cross and Hipk 2 paternal cross	$\chi_1^2 = 12.313$	$4.499 \times 10^{-4}$

### A.3. RIP-seq data tables

Table A.7 is the number of reads at each of the main processing stages, in the same format as Table 3.2. The sample with the lowest number of final reads is HOW(S) 1's total RNA (8.4 million for genome, 9.1 for transcriptome). The sample with the highest number of final reads is HOW(S) 2's pull-down (23.3 million genome aligned reads, 24.0 transcriptome mapped reads).

Table A.8 contains all 343 genes that were significantly differentially enriched in the HOW(S)-HA RIP-seq. Each gene is listed with the  $\log_2(\text{Fold Change})$ , i.e. the level of enrichment, the  $\log_2(\text{Counts per Million})$  and the FDR corrected  $p$ -value.

Table A.9 contains all 121 transcripts that were significantly differentially enriched in the HOW(S)-HA RIP-seq. Each transcript is listed with the  $\log_2(\text{Fold Change})$ , i.e. the level of enrichment, the  $\log_2(\text{Counts per Million})$  and the FDR corrected  $p$ -value.

Table A.10 contains the 27 GO terms from levels 4–7 that were significantly enriched when comparing the significantly enriched genes from the HOW(S)-HA RIP-seq to all genes expressed in the testis. The GO term is listed with the  $q$ -value (the FDR corrected  $p$ -value), the amount of enrichment, and the level of the GO term. GO term analysis was performed with GOrilla, as described in Material and Methods 2.6.3.6.

**Table A.7. Breakdown of read counts through the pipeline.** Number of reads assigned in each step of the processing from the starting number of reads. In genome columns the alignment refers to the genome alignment carried out by Subread and the assignment refers to reads assigned to features by featureCounts. The final reads used for different gene and transcriptome analysis are in bold.

Samples		Starting number of reads	Pre-processing			Genome alignment			Transcriptome quasi-mapping	
			Poor quality	rRNA	tRNA	Unaligned	Aligned, unassigned	Aligned, assigned	Unmapped	Mapped
<i>nanos-GAL4</i>	Total RNA	53743743	6159559	21104610	156	5971135	1920673	<b>18587610</b>	6538563	<b>19940855</b>
	Pull-down	46955434	5286714	4660684	834	11817993	3201061	<b>21988148</b>	13817860	<b>23189342</b>
HOW(S) 1	Total RNA	53660667	5619168	35430765	148	3323901	853295	<b>8433390</b>	3479500	<b>9131086</b>
	Pull-down	40300616	5073550	6318625	137	5245985	3257661	<b>20404658</b>	8417260	<b>20491044</b>
HOW(S) 2	Total RNA	55018237	6494077	23405220	153	3641607	2006176	<b>19471004</b>	4347585	<b>20771202</b>
	Pull-down	43895836	5390068	6027698	802	5452372	3682211	<b>23342685</b>	8458802	<b>24018466</b>
HOW(S) 3	Total RNA	52141872	5887989	22028079	133	3506584	1915307	<b>18803780</b>	4225684	<b>19999987</b>
	Pull-down	48005064	5556163	14084693	677	5159028	2762913	<b>20441590</b>	6701253	<b>21662278</b>

**Table A.8. 343 significantly enriched genes from the HOW(S)-HA RIP-seq.** Each gene is listed with the  $\log_2$ (Fold Change), i.e. the level of enrichment, the  $\log_2$ (Counts per Million) and the FDR corrected  $p$ -value.

Gene ID	Gene Symbol	$\log_2$ FC	$\log_2$ CPM	$p$ -value
FBgn0266095	lncRNA:CR44832	7.026426683	-0.746560917	0.01490027
FBgn0030829	CG12998	3.771104694	-0.358146878	0.007262219
FBgn0022987	qkr54B	3.354918598	5.568768862	0.001032975
FBgn0031306	CG4577	3.33099249	2.279271682	0.000946507
FBgn0039239	CG13641	3.24024439	0.55925858	0.007210046
FBgn0259861	Su(Ste):CR42430	3.115702214	0.23891225	0.013569186
FBgn0266633	asRNA:CR45140	3.11280643	-0.433845973	0.021934725
FBgn0027934	$\alpha$ -Est4aPsi	2.934818091	1.74980478	0.001439173
FBgn0265649	lncRNA:CR44456	2.912556513	0.632703639	0.003786349
FBgn0266005	lncRNA:CR44779	2.669814287	3.729451703	0.00930888
FBgn0265058	asRNA:CR44169	2.651562681	-0.49651603	0.02193523
FBgn0035495	CG14989	2.620357536	1.366684545	0.002664506
FBgn0032336	AstC	2.597835207	1.321473642	0.006747634
FBgn0032178	Spn31A	2.542522201	0.174033748	0.019864279
FBgn0031945	CG7191	2.498895703	3.267937491	0.001272149
FBgn0028863	stol	2.420628385	0.365134236	0.008985711
FBgn0036287	CG10663	2.407915795	0.896409776	0.004154954
FBgn0039593	Sid	2.388998345	0.262375005	0.015305025
FBgn0029834	CG5937	2.386477412	0.521528309	0.005596723
FBgn0086693	iav	2.379049164	-0.555345924	0.041651337
FBgn0026439	Eaat1	2.360723465	4.119573526	0.004404925
FBgn0031473	CG3104	2.272089714	2.69850065	0.005880365
FBgn0267242	asRNA:CR45682	2.269313249	0.924460631	0.011900061
FBgn0022893	Df31	2.260629342	6.076657494	0.00129774
FBgn0010222	Nmdmc	2.232937151	6.05687234	0.000949987
FBgn0036044	Zasp67	2.231213388	0.544410147	0.0153082
FBgn0051427	CR31427	2.216185084	-0.623057831	0.039633318
FBgn0038380	CG14877	2.196966551	1.696454085	0.016102278
FBgn0039064	CG4467	2.183079584	0.402539728	0.017220735
FBgn0033133	Tsp42Ek	2.177079018	0.074380754	0.024448732
FBgn0029723	Proc-R	2.157950797	0.852640893	0.00633326
FBgn0085427	CG34398	2.150058075	4.741017021	0.001834543
FBgn0085431	Sol1	2.148567208	0.710137428	0.016522579
FBgn0035385	FMRFaR	2.122213723	-0.893420666	0.041560579
FBgn0266323	lncRNA:CR44988	2.066045927	2.400542632	0.019976646
FBgn0086675	fne	2.056542303	-0.562919142	0.019939067
FBgn0267810	asRNA:CR46135	2.042846799	0.547404565	0.022401346
FBgn0036544	sff	2.006085845	0.311662418	0.013460022

Gene ID	Gene Symbol	log <sub>2</sub> FC	log <sub>2</sub> CPM	p-value
FBgn0036319	Ent3	2.005927322	1.015333151	0.02677852
FBgn0034276	Sardh	2.00003844	0.161045602	0.011204723
FBgn0038126	CG8483	1.986086651	0.511470343	0.020891038
FBgn0038479	CG17477	1.957846607	2.479114981	0.003775636
FBgn0037796	CG12814	1.948900511	6.483141576	0.000949987
FBgn0011829	Ret	1.945424911	0.011653692	0.025286213
FBgn0250908	beat-VII	1.920709238	2.327392189	0.019281917
FBgn0033668	exp	1.904065683	0.655229695	0.00885767
FBgn0036282	Smyd4-2	1.8988549	0.864066988	0.013683195
FBgn0085483	CG34454	1.880132617	-0.014963164	0.0173184
FBgn0033708	CG8850	1.863665256	0.787086803	0.025745864
FBgn0085382	CG34353	1.860299938	3.08917215	0.028102816
FBgn0037396	CG11459	1.842493577	0.333765573	0.036493294
FBgn0261611	CG42700	1.838623259	2.228870746	0.00263381
FBgn0262599	SmydA-3	1.826428866	-0.192348226	0.019408679
FBgn0032021	CG7781	1.825654841	2.525840188	0.010587891
FBgn0034182	SmydA-7	1.814504412	0.850526271	0.008192811
FBgn0053294	CR33294	1.795435319	1.435557562	0.022880333
FBgn0051386	lncRNA:CR31386	1.755973737	-0.339095086	0.029974609
FBgn0035694	CG13299	1.736622806	-0.186780483	0.041298056
FBgn0029523	CR18275	1.719795699	0.395884748	0.01205057
FBgn0039736	CG7912	1.712969135	-0.611629152	0.041898991
FBgn0265081	asRNA:CR44192	1.707935495	0.242784487	0.039633318
FBgn0266401	asRNA:CR45041	1.703656658	0.121722253	0.016493327
FBgn0033696	Cyp6g2	1.698837187	0.649786566	0.048269626
FBgn0267243	asRNA:CR45683	1.683650078	-0.269109912	0.04173501
FBgn0028879	CG15270	1.668206063	0.325465087	0.047851979
FBgn0028422	GluRIID	1.643046245	0.293807665	0.037448728
FBgn0038199	CCHa1	1.624151855	1.251784375	0.015844693
FBgn0038828	CG17270	1.61487792	-0.085318588	0.041413585
FBgn0031116	CG1695	1.61196936	1.836128096	0.005927682
FBgn0266222	lncRNA:CR44917	1.594824977	1.133240466	0.009938472
FBgn0037408	NPFR	1.58960845	1.655517757	0.008590193
FBgn0034715	Oatp58Db	1.582359571	-0.237983507	0.030580678
FBgn0050489	Cyp12d1-p	1.579914537	1.110692325	0.016943458
FBgn0039031	Gbp3	1.575602176	2.091135802	0.013452015
FBgn0000451	ect	1.575469929	0.234450401	0.029671819
FBgn0037730	CG9444	1.570117268	1.404278584	0.011687383
FBgn0034473	Or56a	1.559821968	0.17187616	0.02340874
FBgn0039101	CG16710	1.541662308	0.113981105	0.035215961
FBgn0034568	CG3216	1.538869956	0.001688705	0.046614011

A. Appendix I

Gene ID	Gene Symbol	log <sub>2</sub> FC	log <sub>2</sub> CPM	p-value
FBgn0263446	asRNA:CR43469	1.532131548	0.8347119	0.019939067
FBgn0267457	asRNA:CR45807	1.528286575	1.920070014	0.010939185
FBgn0267160	asRNA:CR45600	1.515793003	2.015226846	0.027486759
FBgn0034906	CG13561	1.515492204	0.00198354	0.02653417
FBgn0051216	Naam	1.510592168	3.140322016	0.016646998
FBgn0266359	asRNA:CR45007	1.505234781	-0.174699608	0.039965956
FBgn0040832	CG8012	1.493976394	3.696669945	0.016140836
FBgn0267322	lncRNA:CR45758	1.488770232	1.033689295	0.011186193
FBgn0266029	lncRNA:CR44794	1.487335521	2.606261651	0.003770353
FBgn0039033	Or94a	1.486575673	0.741253553	0.019302626
FBgn0004514	Oct-TyrR	1.478086034	1.219024619	0.037553895
FBgn0265540	asRNA:CR44390	1.470003999	0.443468737	0.032279314
FBgn0263442	asRNA:CR43465	1.468402978	1.843137191	0.011322219
FBgn0003733	tor	1.458896397	1.491763621	0.022265758
FBgn0033744	Dh44-R2	1.445724191	1.354085981	0.016161813
FBgn0039453	CG6403	1.443159353	-0.496055885	0.041761931
FBgn0038147	CCHa2	1.435371236	3.618201609	0.010744324
FBgn0265169	asRNA:CR44238	1.432483262	3.261084858	0.003065789
FBgn0053513	Nmdar2	1.414249928	-0.417451622	0.041561689
FBgn0265002	CG44153	1.413942208	3.494904652	0.010746502
FBgn0051460	CG31460	1.408781974	3.990605742	0.016943458
FBgn0031939	CG13796	1.403779255	2.551402214	0.008814583
FBgn0034151	CG15617	1.401253031	0.830319738	0.015836431
FBgn0053203	CG33203	1.398824383	0.857060691	0.01765128
FBgn0263233	robls54B	1.392235278	0.003233982	0.031284127
FBgn0028956	mthl3	1.386751489	3.406526841	0.013683195
FBgn0036259	CG9760	1.385047378	-0.228635091	0.032507555
FBgn0051660	smog	1.375446509	0.058615632	0.035215961
FBgn0267996	asRNA:CR46263	1.374339227	0.404462074	0.020971721
FBgn0024189	sns	1.371595599	2.533466553	0.004649833
FBgn0264606	Fife	1.352336509	1.727741213	0.012308592
FBgn0039944	CG17162	1.352317333	2.23936157	0.01729574
FBgn0050263	stum	1.348308586	3.631585794	0.040770268
FBgn0032048	Dh31	1.345218655	0.770681459	0.021342442
FBgn0260964	Vmat	1.343872842	1.315044417	0.025362497
FBgn0267228	lncRNA:CR45668	1.336604448	1.205988032	0.04756522
FBgn0031302	CG14340	1.332814515	-0.170796028	0.045592322
FBgn0032879	CarT	1.327783201	1.502746795	0.016839117
FBgn0004513	Mdr65	1.323474957	2.005022841	0.036493294
FBgn0263093	CR43361	1.320794827	1.080043428	0.038401301
FBgn0032773	fon	1.319518655	5.257412658	0.01424633

Gene ID	Gene Symbol	log <sub>2</sub> FC	log <sub>2</sub> CPM	p-value
FBgn0051720	mthl15	1.314851417	0.902030617	0.049981068
FBgn0266624	asRNA:CR45131	1.314154899	-0.594283957	0.04857724
FBgn0000542	ec	1.309846903	4.553650542	0.022086894
FBgn0032470	Ttc30	1.303551377	0.476342546	0.035661884
FBgn0050147	Hil	1.30238029	2.754927497	0.018932368
FBgn0039844	CG1607	1.301612336	2.221625028	0.008636088
FBgn0263257	Cngl	1.299316248	0.41267423	0.02746087
FBgn0263346	smash	1.294428745	7.288581086	0.015794242
FBgn0037754	CG8500	1.29370392	-0.054828671	0.039869655
FBgn0041789	Pax	1.287821493	6.249538297	0.019477187
FBgn0020248	stet	1.286400457	1.384259833	0.033206213
FBgn0037726	Dhc1	1.284842293	5.075679271	0.023730679
FBgn0037807	CG6293	1.282097476	1.071760273	0.016845148
FBgn0038833	CG15696	1.279615945	0.858050462	0.027906091
FBgn0063449	Uhg2	1.275820838	0.804580304	0.03926312
FBgn0265807	lncRNA:CR44596	1.269096323	-0.391503847	0.043462502
FBgn0267614	asRNA:CR45952	1.268781109	2.168251674	0.01424633
FBgn0250871	pot	1.268777633	5.096813366	0.004992365
FBgn0032151	nAChR $\alpha$ 6	1.267776405	1.733848809	0.026649733
FBgn0037130	Syn1	1.263991674	1.31595756	0.014929438
FBgn0045443	mthl11	1.263556493	0.273856108	0.033277247
FBgn0000535	eag	1.262410655	0.43079285	0.027766286
FBgn0035612	frm	1.261650133	0.935225866	0.021567243
FBgn0266756	btsz	1.261514428	6.949802905	0.017270782
FBgn0030052	CG12065	1.257351278	4.626365985	0.039869708
FBgn0034002	CG8079	1.253706264	3.519491309	0.043698253
FBgn0023531	CG32809	1.252888056	3.747138505	0.013460022
FBgn0051324	CG31324	1.250595582	3.465204723	0.012122426
FBgn0030358	CG10362	1.249965598	-0.443432579	0.045768824
FBgn0264862	asRNA:CR44053	1.249600203	0.108579222	0.031061904
FBgn0266635	asRNA:CR45142	1.23912092	0.451455148	0.042662457
FBgn0037697	GstZ2	1.237277778	-0.180993496	0.039633614
FBgn0283651	CG46301	1.236088696	-0.245954267	0.042270729
FBgn0053556	form3	1.234201292	4.141647126	0.015844693
FBgn0036368	CG10738	1.231597581	0.998376857	0.018245199
FBgn0034530	Rcd6	1.231156651	1.243843013	0.021194625
FBgn0036143	CG14142	1.229600671	0.200596558	0.034920651
FBgn0050046	CG30046	1.227863497	0.941205842	0.018839369
FBgn0015572	$\alpha$ -Est4	1.225539716	2.078566208	0.011695334
FBgn0051388	CG31388	1.225062913	0.922667902	0.029283884
FBgn0038886	Ugt49B2	1.224902788	2.239204791	0.034824626

A. Appendix I

Gene ID	Gene Symbol	log <sub>2</sub> FC	log <sub>2</sub> CPM	p-value
FBgn0003091	Pkc53E	1.224799456	4.701027545	0.011204022
FBgn0037601	Cyp313b1	1.22255294	1.193553962	0.014147607
FBgn0260486	Ziz	1.220400881	6.134232643	0.008590193
FBgn0285896	btl	1.219940122	1.350426526	0.031861022
FBgn0052645	CG32645	1.218868068	1.876800307	0.010744324
FBgn0022800	Cad96Ca	1.2184791	2.469999761	0.04026197
FBgn0083975	Nlg4	1.216855961	2.593919406	0.013391051
FBgn0039656	CG11951	1.215772435	0.522854154	0.04659018
FBgn0032082	CG18088	1.212585799	0.361015138	0.048054655
FBgn0034990	CG11406	1.209193288	0.818075311	0.041326517
FBgn0039486	CAH9	1.20849588	2.515380505	0.007448296
FBgn0264324	spg	1.208290168	5.376712172	0.017967955
FBgn0039688	Kul	1.207809109	5.303587731	0.003352859
FBgn0032731	Swip-1	1.207002249	5.159959592	0.037290185
FBgn0037292	plh	1.20445311	0.025493903	0.035163644
FBgn0035142	Hipk	1.203472288	7.564400342	0.004058875
FBgn0039620	wat	1.203385464	5.083243244	0.019302626
FBgn0035578	CG13707	1.203016874	0.776057162	0.023811144
FBgn0263079	CG43338	1.200405234	2.037337559	0.02074903
FBgn0038917	CG6678	1.196668179	1.802351501	0.017301669
FBgn0040491	Buffy	1.186991794	5.491731464	0.015715006
FBgn0259242	CG42340	1.186882144	2.732433915	0.030928223
FBgn0264894	CG44085	1.185972149	5.274845454	0.004024331
FBgn0000635	Fas2	1.178883427	6.228612325	0.02832244
FBgn0262613	CG43139	1.178316135	2.687700238	0.041430019
FBgn0027929	NimB1	1.178107893	0.49613538	0.047430357
FBgn0036646	CR18217	1.175375999	3.332871254	0.005803952
FBgn0038076	Cyp313a4	1.173671415	1.310548459	0.021079162
FBgn0034364	CG5493	1.173174001	0.682049839	0.023407959
FBgn0000153	b	1.173072886	0.714544486	0.029675169
FBgn0037519	CG3014	1.172836316	4.11483823	0.004696366
FBgn0053542	upd3	1.171504332	0.441359531	0.048522971
FBgn0026438	Eaat2	1.171144239	0.367672049	0.034757933
FBgn0265988	mv	1.171083904	6.878112628	0.020161383
FBgn0014396	tim	1.170363288	6.114577949	0.020242175
FBgn0038819	Cpr92F	1.170296919	1.674124596	0.040068551
FBgn0053111	CG33111	1.170205611	4.490316943	0.008227088
FBgn0032800	CG10137	1.169206716	1.987906138	0.019975601
FBgn0033904	CG18327	1.165664227	0.751247961	0.030244158
FBgn0033076	CG15233	1.164348522	4.726404102	0.013318885
FBgn0044028	Notum	1.163194824	1.018642475	0.025436205

Gene ID	Gene Symbol	log <sub>2</sub> FC	log <sub>2</sub> CPM	p-value
FBgn0032120	CG33298	1.162589758	6.038016086	0.004992365
FBgn0026319	Traf4	1.161841317	2.557021221	0.023782102
FBgn0000163	baz	1.160898665	6.031750398	0.016112915
FBgn0260660	Mp	1.159232281	5.564227078	0.035734881
FBgn0032536	Ance-3	1.156581461	3.564728322	0.036159519
FBgn0036821	CG3961	1.15304454	3.118402552	0.016102278
FBgn0261269	conv	1.152410774	3.901391976	0.01424633
FBgn0036732	Oatp74D	1.145901158	3.521390524	0.011964163
FBgn0028371	jbug	1.145115723	6.380293919	0.010744324
FBgn0031034	CG14205	1.144724125	1.184054412	0.034548876
FBgn0261804	CG42750	1.143286195	2.022168826	0.021726227
FBgn0013718	nuf	1.143002926	6.428673839	0.009301989
FBgn0264959	Src42A	1.141249908	6.693999602	0.026295725
FBgn0028622	qsm	1.139966156	2.86744825	0.012697318
FBgn0000594	Est-P	1.139750331	0.899025377	0.039001704
FBgn0039925	Kif3C	1.139745753	3.980883209	0.020694543
FBgn0051148	Gba1a	1.139066067	2.747494688	0.034038276
FBgn0259241	CG42339	1.138744388	0.633121423	0.031016107
FBgn0033504	CAP	1.136807609	6.20157568	0.00764811
FBgn0051183	CG31183	1.136652861	3.856829561	0.020661156
FBgn0024321	NK7.1	1.136118153	5.495782759	0.040622528
FBgn0085422	CG34393	1.134955336	1.846755992	0.022676742
FBgn0030077	CG15365	1.134255034	1.124838495	0.022534398
FBgn0036144	GlcAT-P	1.132719808	4.463058276	0.01729574
FBgn0267733	lncRNA:CR46064	1.131777794	1.239076103	0.035588257
FBgn0283531	Duox	1.129718878	2.534718916	0.012308592
FBgn0033395	Cyp4p2	1.129313169	2.317653274	0.040068551
FBgn0028658	Adat1	1.127657773	1.950653408	0.041898991
FBgn0260004	Snmp1	1.127192997	1.960506204	0.017683396
FBgn0030808	RhoGAP15B	1.127156605	4.423462571	0.039533893
FBgn0054056	CG34056	1.12630099	2.160323413	0.018921391
FBgn0028644	beat-Ic	1.12532518	3.958986056	0.009536379
FBgn0004910	Eip63F-1	1.123641209	3.368722164	0.025953991
FBgn0034137	CG4945	1.121092355	1.068904126	0.032507555
FBgn0039667	FipoQ	1.120413888	1.236229808	0.022582288
FBgn0035755	CG14830	1.117838568	4.424901709	0.034600455
FBgn0005660	Ets21C	1.115714658	3.99267711	0.012901773
FBgn0286198	LKRSDH	1.114078178	4.064793944	0.02949781
FBgn0010473	tut1	1.113532741	2.52370998	0.034915527
FBgn0023095	caps	1.112086214	6.785070359	0.035449429
FBgn0260945	Atg1	1.109055446	5.163973809	0.016085847

A. Appendix I

Gene ID	Gene Symbol	log <sub>2</sub> FC	log <sub>2</sub> CPM	p-value
FBgn0264792	lncRNA:CR44022	1.108772394	1.060549907	0.048554634
FBgn0261999	Ca-Ma2d	1.108663559	2.857619347	0.010744324
FBgn0011674	insc	1.10801323	3.20658081	0.015715006
FBgn0003068	per	1.107716388	3.153642336	0.031861022
FBgn0035772	Sh3beta	1.107417484	6.768215103	0.010012953
FBgn0085644	CR41423	1.107273125	1.102541645	0.038352804
FBgn0036030	Prps	1.107145893	6.530817915	0.013547365
FBgn0037448	CG15186	1.107080967	3.53869765	0.008192811
FBgn0026181	Rok	1.107071502	6.468560478	0.003949982
FBgn0031981	Megf8	1.106412484	7.029235798	0.005545917
FBgn0032895	twit	1.106172497	0.998814904	0.032162776
FBgn0052677	X11Lbeta	1.104623956	1.403018633	0.02923331
FBgn0039932	fuss	1.104223024	3.060943129	0.010478876
FBgn0051514	asRNA:CR31514	1.101237567	0.477823213	0.046341223
FBgn0034095	CG15701	1.10060635	0.420406941	0.035755515
FBgn0039073	CG4408	1.098938329	1.326680576	0.024933355
FBgn0264000	GluRIB	1.098043036	3.599056815	0.018181674
FBgn0042185	MCU	1.097399896	6.372969241	0.012308592
FBgn0285917	sbb	1.097292286	6.688397871	0.02331868
FBgn0037238	CG1090	1.097132727	1.24094371	0.035474245
FBgn0035158	CG13895	1.096249826	1.830655931	0.022610749
FBgn0259683	Ir40a	1.094000295	1.503987147	0.037115146
FBgn0037713	CG16790	1.092943923	1.311034604	0.022401346
FBgn0265487	mbl	1.090556893	8.393145453	0.028102816
FBgn0010300	brat	1.087897919	5.924437806	0.00739829
FBgn0015399	kek1	1.086653295	3.991490994	0.005449628
FBgn0265804	lncRNA:CR44593	1.084883189	3.354482118	0.024410892
FBgn0029939	CG9650	1.084576628	3.233318631	0.022731268
FBgn0266418	wake	1.084437568	4.957696115	0.014280408
FBgn0052255	Gr64f	1.081612242	0.881133059	0.04215822
FBgn0265140	Meltrin	1.079650693	5.865704701	0.019549871
FBgn0031571	bark	1.078662035	5.517727545	0.022676742
FBgn0264953	Piezo	1.077367444	5.524226517	0.016701565
FBgn0040571	CG17193	1.077313121	0.403408725	0.04221514
FBgn0030613	Rab3-GEF	1.077155717	0.500754141	0.035026344
FBgn0033636	tou	1.075012586	6.035753583	0.022066431
FBgn0033936	Achl	1.073557103	0.485149262	0.032319461
FBgn0263258	chas	1.073525918	5.189442579	0.015555402
FBgn0032147	IP3K1	1.069017308	4.287776265	0.015823069
FBgn0062978	CG31808	1.068971788	3.949833003	0.012033237
FBgn0039747	AdoR	1.068054981	1.677350016	0.020188113

Gene ID	Gene Symbol	log <sub>2</sub> FC	log <sub>2</sub> CPM	p-value
FBgn0259140	CG42255	1.067084362	3.319892577	0.011204022
FBgn0085421	Epac	1.064988608	6.010458882	0.030597635
FBgn0035880	Culd	1.064901015	2.882354551	0.037074836
FBgn0045852	ham	1.06414734	5.748705213	0.024298326
FBgn0033313	Cirl	1.064075919	6.696819133	0.035762373
FBgn0033652	ths	1.062070765	4.360303079	0.020029232
FBgn0039431	plum	1.061131708	5.26295409	0.038364832
FBgn0033919	CG8547	1.060233584	4.793237427	0.00522354
FBgn0005558	ey	1.058207917	0.433980442	0.036453287
FBgn0263353	CG11000	1.057957145	2.728311616	0.015349187
FBgn0259212	cno	1.057589981	6.446636437	0.04215822
FBgn0261446	CG13377	1.057400557	2.325062317	0.010744324
FBgn0030090	fend	1.053999265	2.970959208	0.011186193
FBgn0052062	Rbfox1	1.052736445	5.821179542	0.012823089
FBgn0087012	5-HT2A	1.052687297	4.061654739	0.010477857
FBgn0030017	CG2278	1.049810295	2.248474241	0.032987251
FBgn0259173	corn	1.049427952	3.724378409	0.011560038
FBgn0033659	Damm	1.047711798	2.005141012	0.029416292
FBgn0085386	CG34357	1.047586316	2.021076435	0.031437838
FBgn0263929	jvl	1.046820388	7.776005892	0.027526149
FBgn0052850	Rnf11	1.046607025	3.339007341	0.022759556
FBgn0050377	CG30377	1.045323119	1.974063429	0.015038595
FBgn0259178	5PtaseI	1.045082794	2.815436863	0.016701565
FBgn0020304	drongo	1.044772399	7.499067783	0.004980764
FBgn0033657	Sln	1.04294329	1.68976121	0.036764812
FBgn0000179	bi	1.04289313	2.381795884	0.012938388
FBgn0264822	asRNA:CR44030	1.042583403	1.509354548	0.044319776
FBgn0000422	Ddc	1.040197696	4.329678219	0.005545917
FBgn0038897	CG5849	1.040015109	1.672855961	0.029739297
FBgn0037956	CG6959	1.039503759	4.051276332	0.032085691
FBgn0003997	hid	1.037233929	3.649761176	0.019844418
FBgn0035699	CG13300	1.036594142	2.625895732	0.035215961
FBgn0003353	sei	1.035251027	3.807043409	0.011750191
FBgn0262562	CG43102	1.035222525	5.024075187	0.016112915
FBgn0266330	asRNA:CR44995	1.034743253	1.693500766	0.019635708
FBgn0039590	CG10011	1.033244374	5.744724623	0.039633614
FBgn0267815	asRNA:CR46140	1.030940493	2.078469826	0.047458109
FBgn0267766	asRNA:CR46097	1.029898389	0.925389165	0.034980545
FBgn0000721	for	1.029184084	7.089571472	0.031152582
FBgn0050345	CG30345	1.026378862	3.889306434	0.01647033
FBgn0264089	sli	1.025760892	5.27099832	0.020029232

A. Appendix I

Gene ID	Gene Symbol	log <sub>2</sub> FC	log <sub>2</sub> CPM	p-value
FBgn0085443	spri	1.025500376	6.111219496	0.031437838
FBgn0036896	wnd	1.02503664	3.833987318	0.021816475
FBgn0260442	rhea	1.024072675	8.137942614	0.005770742
FBgn0262139	trh	1.023854431	1.084521231	0.036918734
FBgn0031055	et	1.023326069	1.681709561	0.030808937
FBgn0004369	Ptp99A	1.022008259	4.952554281	0.014599682
FBgn0264087	Slob	1.021180189	2.703499109	0.038220489
FBgn0004583	ex	1.02000359	4.902001451	0.011186193
FBgn0020378	Sp1	1.0142015	0.445596783	0.044051518
FBgn0283499	InR	1.013742029	6.135840217	0.021320427
FBgn0046704	Liprin- $\alpha$	1.008641209	7.043173259	0.003817175
FBgn0031299	CG4629	1.005514478	4.582112207	0.0326285
FBgn0035802	Pura	1.005512645	6.668744749	0.026640691
FBgn0004656	fs(1)h	1.003007274	6.674917374	0.006199799
FBgn0015513	mbc	1.002595937	6.225806529	0.022245401
FBgn0283521	lola	1.001398478	8.305846529	0.01647033
FBgn0034396	CG15097	1.000994068	3.673245133	0.011252172
FBgn0266801	CG45263	1.000089632	7.678778083	0.035013315

**Table A.9. 121 significantly enriched transcripts from the HOW(S)-HA RIP-seq.** Each transcript is listed with the  $\log_2$ (Fold Change), i.e. the level of enrichment, the  $\log_2$ (Counts per Million) and the FDR corrected  $p$ -value.

Transcript ID	Gene Symbol	$\log_2$ FC	$\log_2$ CPM	$p$ -value
FBtr0339087	CG44243-RB	3.144750005	5.080781043	0.020612844
FBtr0085919	Df31-RA	2.338730103	4.254508992	0.037737095
FBtr0083418	CG17477-RA	2.219222929	1.915350068	0.014835422
FBtr0091836	His2A:CG33832-RA	2.183347179	4.649188891	0.023910157
FBtr0329941	Liprin- $\alpha$ -RC	2.072332077	1.236436578	0.034842641
FBtr0086574	CG15109-RC	1.985098661	2.855782944	0.04598503
FBtr0113261	CG7907-RB	1.975252532	1.781614049	0.048826119
FBtr0110786	CG12814-RB	1.956117203	3.438244077	0.028746433
FBtr0332452	CLIP-190-RQ	1.950878134	2.773730506	0.017958811
FBtr0073403	CG1582-RA	1.932556658	2.171597542	0.040945786
FBtr0072668	Ctr9-RB	1.863575025	2.382491371	0.022840971
FBtr0304729	mbc-RB	1.841544115	1.364080863	0.020612844
FBtr0343581	l(1)G0196-RN	1.812808475	1.631339311	0.023873452
FBtr0077662	toc-RA	1.754127551	1.149569625	0.040945786
FBtr0072044	CG30412-RA	1.676796909	1.87470829	0.020427227
FBtr0332928	bbc-RE	1.675491283	3.300347261	0.014835422
FBtr0081997	Nmdmc-RA	1.66173103	3.826997681	0.012863403
FBtr0346773	PRY-RA	1.660239372	1.653531456	0.031187771
FBtr0076912	BBS1-RA	1.649699021	1.529396044	0.039023254
FBtr0081513	crc-RB	1.641167187	2.437619551	0.018142462
FBtr0300773	CG1317-RC	1.638566982	4.097405899	0.029202843
FBtr0081873	CG31460-RA	1.636423291	4.224406267	0.045359515
FBtr0072552	Hipk-RA	1.627561315	1.356540214	0.04233416
FBtr0392904	Syx1A-RC	1.564347157	1.600698869	0.046721525
FBtr0307021	CG43347-RB	1.55135003	1.414233695	0.04233416
FBtr0304935	Pitslre-RE	1.544292933	1.416870493	0.035007555
FBtr0084173	ND-42-RB	1.523327922	1.109103686	0.042444228
FBtr0088780	sand-RA	1.521392941	0.825194679	0.036077268
FBtr0306565	mthl10-RD	1.510124552	1.687899636	0.028178935
FBtr0310273	Gpo3-RE	1.497113976	3.175843646	0.037209138
FBtr0344624	CG3014-RC	1.487542395	1.251945445	0.039573716
FBtr0076699	CG8012-RA	1.48125455	3.779630856	0.039282885
FBtr0081166	Ddc-RB	1.460045923	1.914408785	0.039282885
FBtr0070639	cib-RA	1.452153755	5.787467532	0.028634969
FBtr0084214	PyK-RA	1.446233552	1.799505971	0.043071528
FBtr0112890	fz2-RC	1.444724922	1.062936015	0.04598503
FBtr0112860	Nc73EF-RI	1.439978953	2.580104858	0.021896875
FBtr0078509	Ten-m-RB	1.437085466	1.454666617	0.034034636

A. Appendix I

Transcript ID	Gene Symbol	log <sub>2</sub> FC	log <sub>2</sub> CPM	p-value
FBtr0082833	Dip-B-RB	1.431984373	3.626241256	0.022659527
FBtr0089360	lola-RJ	1.418010938	1.519864158	0.034034636
FBtr0073519	rho-4-RA	1.40591421	1.822270682	0.039282885
FBtr0074932	CG32212-RA	1.400184653	1.543651601	0.035278255
FBtr0302710	CG42673-RC	1.380394246	0.76669312	0.045664061
FBtr0339797	Lrch-RC	1.375505089	0.89377151	0.041557263
FBtr0301287	fon-RC	1.374884427	2.454322253	0.038113429
FBtr0076972	MCU-RB	1.357194942	2.063442515	0.028178935
FBtr0077657	FASN2-RA	1.35626016	1.326532792	0.037277868
FBtr0100514	how-RC	1.35547278	1.519302856	0.032486682
FBtr0303860	CG42837-RA	1.351343424	5.361933117	0.028746433
FBtr0071785	Swim-RA	1.350190405	2.839518482	0.04598503
FBtr0087640	CG30485-RA	1.337524595	4.771958385	0.026387587
FBtr0085133	CAH9-RA	1.318378501	1.407567029	0.034034636
FBtr0302164	galectin-RE	1.302941473	0.752164708	0.037277868
FBtr0345429	Ets21C-RD	1.297038261	0.801400038	0.037526008
FBtr0088550	ced-6-RA	1.290125419	0.57093248	0.039351246
FBtr0333395	Psa-RI	1.290051363	4.367534978	0.039573716
FBtr0085460	Kul-RA	1.287880231	1.145099292	0.036379293
FBtr0330611	navy-RC	1.286529312	1.879322677	0.031303839
FBtr0333959	RhoGEF3-RJ	1.27430352	2.306235365	0.046225922
FBtr0332864	CG13700-RC	1.269577141	4.14409576	0.049441736
FBtr0081069	CG10211-RA	1.266957718	0.766934252	0.043376773
FBtr0308850	smash-RF	1.264146467	1.437316876	0.039917202
FBtr0088102	Syx6-RE	1.261374105	0.989613678	0.045824312
FBtr0305958	CG15233-RB	1.252907641	4.179815588	0.044079784
FBtr0075568	RhoGAP71E-RA	1.245971116	1.74483044	0.040692767
FBtr0100647	CG9133-RD	1.241742141	4.861441778	0.047539941
FBtr0305181	tyf-RI	1.23636042	0.514218694	0.04740428
FBtr0347066	CG13917-RD	1.229468871	1.154352204	0.039282885
FBtr0301925	CR18217-RB	1.228383992	2.428009399	0.043356258
FBtr0305303	pyd-RL	1.228013538	0.830512783	0.04740428
FBtr0344788	Megf8-RB	1.226684976	2.502486277	0.04134474
FBtr0334128	CG44001-RE	1.226045986	1.829898328	0.036014282
FBtr0302276	CG42635-RA	1.222507697	3.700208867	0.022112751
FBtr0080946	Trpgamma-RB	1.222028939	1.102273344	0.047697463
FBtr0343137	CG44774-RC	1.220042193	1.679774979	0.040634959
FBtr0075203	Eip74EF-RC	1.218164438	0.544879623	0.041453576
FBtr0331353	Mekk1-RD	1.209276453	0.872037897	0.044939203
FBtr0308589	Dgk-RH	1.207704701	1.187595149	0.047539941
FBtr0301203	CG8547-RC	1.207335824	2.179168219	0.035007555

Transcript ID	Gene Symbol	log <sub>2</sub> FC	log <sub>2</sub> CPM	p-value
FBtr0307300	Tsp26A-RC	1.205723065	1.742793373	0.040707416
FBtr0076872	Sh3beta-RA	1.205413817	4.325924395	0.021544633
FBtr0080735	CG4161-RA	1.203393985	5.01739879	0.04740428
FBtr0306679	Abca3-RB	1.201759027	1.644603093	0.039282885
FBtr0333051	Hsc70Cb-RI	1.197866254	0.746815669	0.04740428
FBtr0071610	cv-2-RA	1.18795942	1.002158851	0.04233416
FBtr0073526	bif-RB	1.181448214	2.17800399	0.04598503
FBtr0334404	beta-Spec-RB	1.180641639	1.374141897	0.049618347
FBtr0080198	Gr32a-RA	1.174579785	2.22803061	0.04556176
FBtr0082208	Mical-RE	1.164470311	1.259492265	0.04740428
FBtr0340151	Rok-RB	1.151639792	1.989592507	0.040634959
FBtr0077096	Usp47-RB	1.151010529	5.382181573	0.022840971
FBtr0087152	CG6262-RA	1.137229116	6.19359555	0.03592593
FBtr0304992	CG32369-RC	1.135318138	1.069762625	0.04556176
FBtr0071691	CG10433-RB	1.135167198	1.758316571	0.043665411
FBtr0300128	Cyp6a9-RB	1.126872212	2.031472116	0.047305486
FBtr0083999	Calx-RB	1.124980186	5.507936266	0.026327155
FBtr0088970	so-RA	1.123881082	2.035704036	0.038996724
FBtr0300248	CG42458-RA	1.111900062	1.168476294	0.044961743
FBtr0085803	CycG-RC	1.106708373	3.860050088	0.031187771
FBtr0076027	yps-RA	1.104364473	6.885681164	0.026326741
FBtr0433535	AOX3-RB	1.102174355	0.539457486	0.04926829
FBtr0333284	skd-RF	1.101592328	1.85393949	0.041591242
FBtr0305694	jvl-RF	1.099280051	1.255637267	0.049618347
FBtr0302670	CG42668-RG	1.098827531	1.336369795	0.049618347
FBtr0305210	rhea-RB	1.092040061	2.924096338	0.038192285
FBtr0071065	Atg5-RA	1.082804498	1.434939184	0.049246198
FBtr0452101	lilli-RI	1.07908856	1.56360983	0.04422332
FBtr0077502	TTL4B-RA	1.076467855	2.217362475	0.042779514
FBtr0303874	Apoltp-RC	1.076075365	0.373295299	0.049618347
FBtr0080514	B4-RB	1.066582261	1.188352284	0.049324136
FBtr0343852	TER94-RE	1.066014673	6.003393255	0.022964103
FBtr0332865	CG13700-RD	1.060241211	4.434930396	0.020612844
FBtr0088872	ACC-RB	1.053090351	1.895588582	0.042415754
FBtr0347413	CCY-RB	1.03630472	3.498048095	0.034806273
FBtr0082387	CG5270-RB	1.030029811	2.595996947	0.047423233
FBtr0310322	CG42531-RB	1.029336358	2.985954157	0.039282885
FBtr0084715	CG11168-RA	1.024805179	2.10359656	0.045540605
FBtr0091837	His3:CG33833-RA	1.017161844	4.225642875	0.035373266
FBtr0085189	pins-RA	1.016138499	2.283922016	0.041686429
FBtr0079046	CG8892-RB	1.015849154	4.137025023	0.04233416

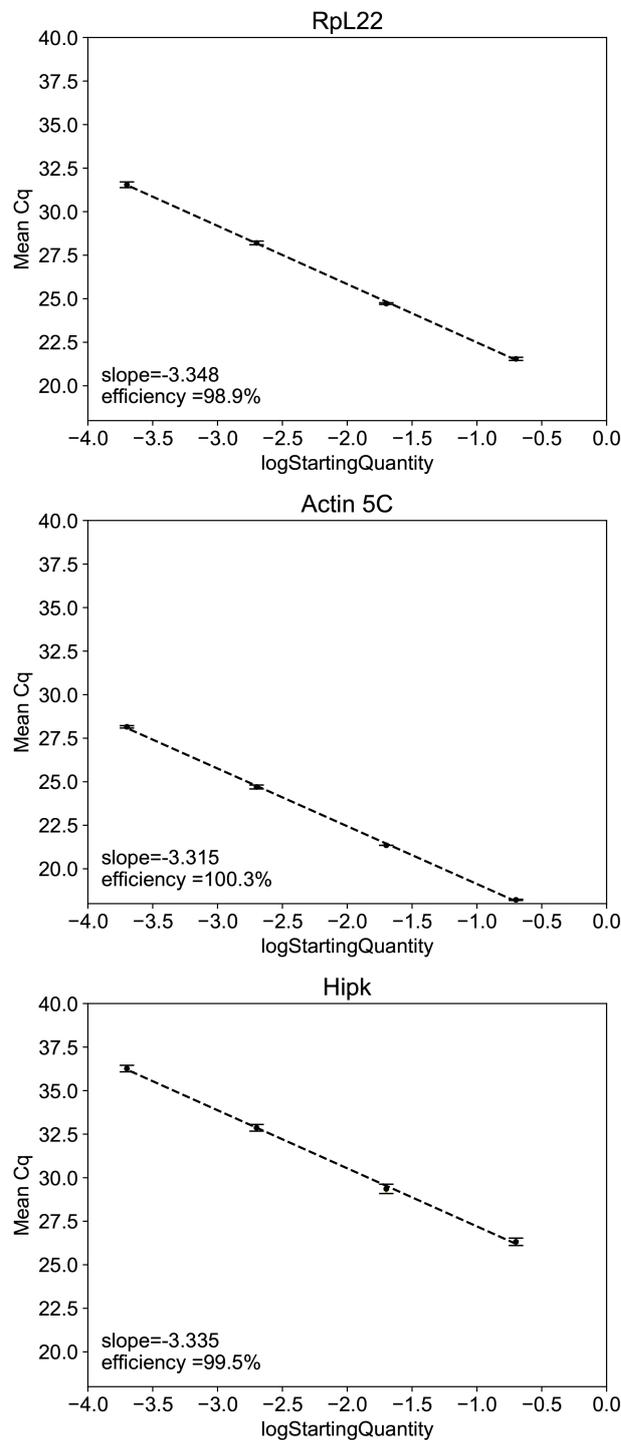
A. Appendix I

<b>Transcript ID</b>	<b>Gene Symbol</b>	<b>log<sub>2</sub>FC</b>	<b>log<sub>2</sub>CPM</b>	<b>p-value</b>
FBtr0088254	CG33144-RA	1.007765664	3.916204141	0.027760065

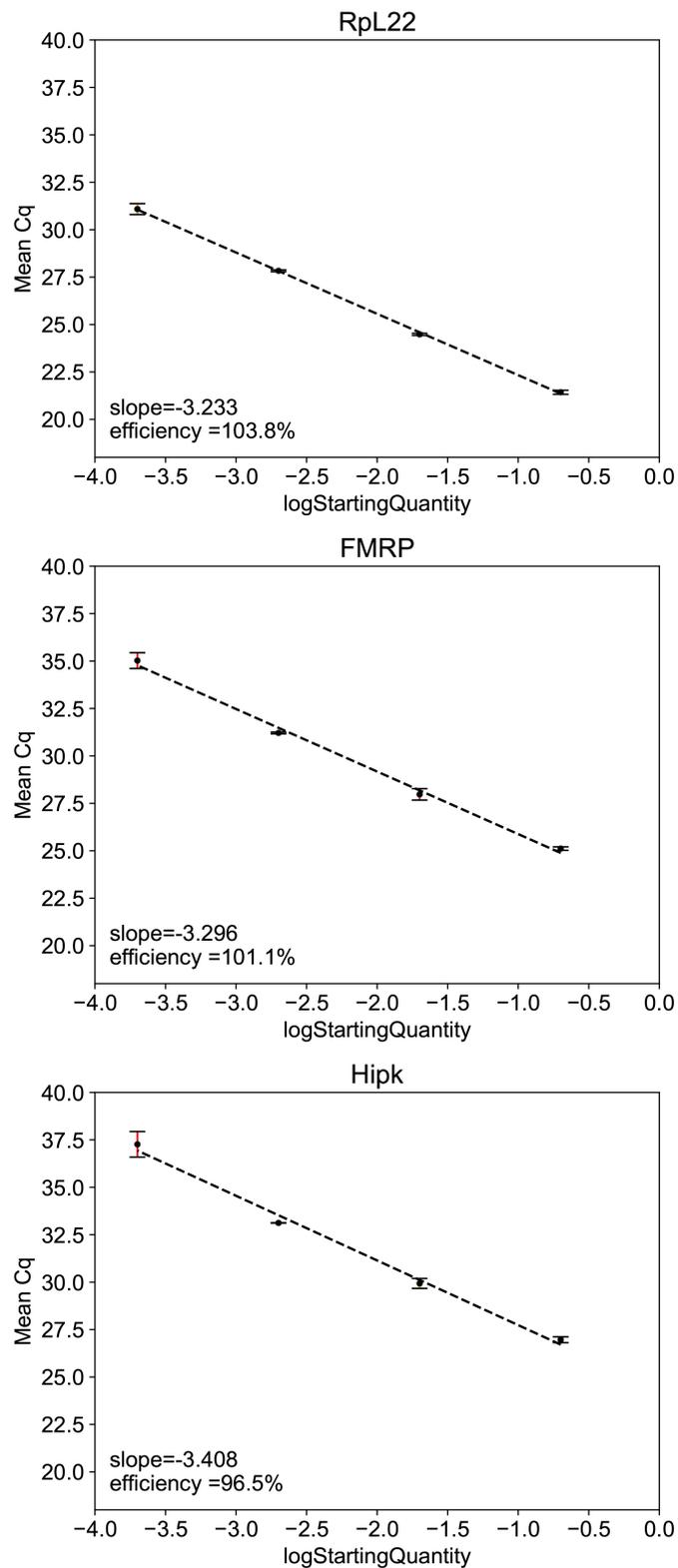
**Table A.10. GO terms from the enriched genes identified in the HOW(S)-HA RIP-seq.** The 27 GO terms from levels 4–7 that were significantly enriched when comparing the significantly enriched genes from the HOW(S)-HA RIP-seq to all genes expressed in the testis. The GO term is listed with the  $q$ -value (the FDR corrected  $p$ -value), the amount of enrichment, and the level of the GO term.

GO term	Description	$q$ -value	Enrichment	Level
GO:0007165	signal transduction	6.90E-08	2.9	5
GO:0007186	G protein-coupled receptor signaling pathway	6.66E-07	7.52	6
GO:0007218	neuropeptide signaling pathway	3.64E-05	19.2	7
GO:0019932	second-messenger-mediated signaling	5.51E-05	9.68	7
GO:0098609	cell-cell adhesion	7.91E-05	8.05	4
GO:0007166	cell surface receptor signaling pathway	4.11E-04	3.42	6
GO:0051239	regulation of multicellular organismal process	6.73E-04	2.32	4
GO:0098742	cell-cell adhesion via plasma-membrane adhesion molecules	1.94E-03	9.6	5
GO:0050907	detection of chemical stimulus involved in sensory perception	2.08E-03	16.46	5
GO:0050794	regulation of cellular process	3.09E-03	1.42	4
GO:0008045	motor neuron axon guidance	3.97E-03	7.12	6
GO:0050877	nervous system process	4.01E-03	3.07	4
GO:0017085	response to insecticide	4.92E-03	13.71	5
GO:0035556	intracellular signal transduction	6.83E-03	2.84	6
GO:0050906	detection of stimulus involved in sensory perception	7.24E-03	9.4	4
GO:0009593	detection of chemical stimulus	9.17E-03	11.76	4
GO:0007187	G protein-coupled receptor signaling pathway, coupled to cyclic nucleotide second messenger	1.02E-02	8.59	7
GO:0042391	regulation of membrane potential	1.03E-02	6.98	4
GO:0023051	regulation of signaling	1.21E-02	1.95	5
GO:0010646	regulation of cell communication	1.25E-02	1.95	5
GO:0007156	homophilic cell adhesion via plasma membrane adhesion molecules	1.99E-02	9.68	6
GO:0006811	ion transport	2.14E-02	2.34	5
GO:0050793	regulation of developmental process	3.30E-02	1.97	4
GO:0050804	modulation of chemical synaptic transmission	3.32E-02	3.92	7
GO:0099177	regulation of trans-synaptic signaling	3.39E-02	3.92	6
GO:0007611	learning or memory	4.48E-02	3.45	6
GO:0050890	cognition	4.57E-02	3.45	5

### A.4. qRT-PCR primer efficiencies

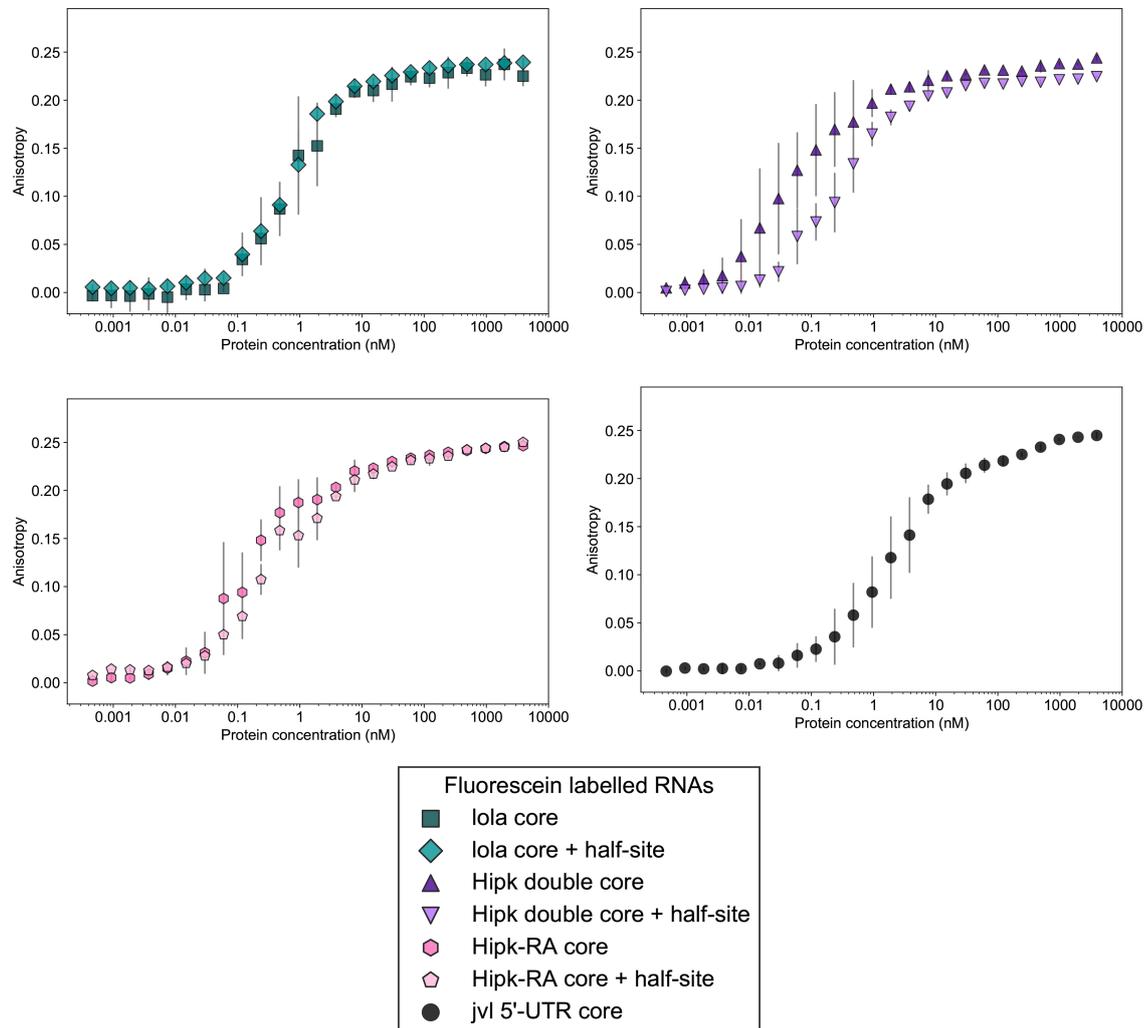


**Figure A.1. Primer efficiencies for primers used to evaluate Hipk RNAi 1 knockdown.** Mean Cq values for the standard curves from three primer pairs used to evaluate Hipk RNAi 1 knockdown. cDNA was pooled from three biological replicates of *nanos*-GAL4, Hipk 1 parental, Hipk 1 maternal cross and Hipk 1 paternal cross for the Rpl22 and actin 5C primers. The maternal and paternal cross samples were not used in the cDNA pool for the Hipk primer standard curve. Error bars are SEM.



**Figure A.2. Primer efficiencies for primers used to evaluate Hipk RNAi 2 knockdown.** Mean Cq values for the standard curves from three primer pairs used to evaluate Hipk RNAi 2 knockdown. cDNA was pooled from three biological replicates of *nanos*-GAL4, Hipk 2 parental, Hipk 2 maternal cross and Hipk 2 paternal cross for the Rpl22 and FMRP primers. The maternal and paternal cross samples were not used in the cDNA pool for the Hipk primer standard curve. Error bars are SEM.

### A.5. SIMV NP fluorescence anisotropy



**Figure A.3. SIMV NP binds all seven oligos derived from lola, Hipk and jvl transcripts.** Scatter plot of SIMV NP protein concentration against mean anisotropy for seven oligos listed in legend. Three technical replicates were carried out except for the ‘Hipk double core’ and ‘Hipk-RA core + half’ where only two replicates were done. Error bars reflect the standard deviation of the replicates.

## References

- Alphey, L., Jimenez, J., White-Cooper, H., Dawson, I., Nurse, P., and Glover, D. M. (1992). “twine, a *cdc25* homolog that functions in the male and female germline of *Drosophila*”. *Cell* 69(6), pp. 977–988.
- Ashwal-Fluss, R., Meyer, M., Pamudurti, N. R., Ivanov, A., Bartok, O., Hanan, M., Evtal, N., Memczak, S., Rajewsky, N., and Kadener, S. (2014). “CircRNA Biogenesis competes with Pre-mRNA splicing”. *Molecular Cell* 56(1), pp. 55–66.
- Babic, I., Jakymiw, A., and Fujita, D. J. (2004). “The RNA binding protein Sam68 is acetylated in tumor cell lines, and its acetylation correlates with enhanced RNA binding activity”. *Oncogene* 23(21), pp. 3781–3789.
- Backe, P. H., Messias, A. C., Ravelli, R. B., Sattler, M., and Cusack, S. (2005). “X-Ray Crystallographic and NMR Studies of the Third KH Domain of hnRNP K in Complex with Single-Stranded Nucleic Acids”. *Structure* 13(7), pp. 1055–1067.
- Baehrecke, E. H. (1997). “who encodes a KH RNA binding protein that functions in muscle development”. *Development* 124(7), pp. 1323–32.
- Bailey, T. L. (2011). “DREME: motif discovery in transcription factor ChIP-seq data”. *Bioinformatics* 27(12), pp. 1653–1659.
- Baker, C. C. and Fuller, M. T. (2007). “Translational control of meiotic cell cycle progression and spermatid differentiation in male germ cells by a novel eIF4G homolog”. *Development* 134(15), pp. 2863–2869.
- Barreau, C., Benson, E., Gudmannsdottir, E., Newton, F., and White-Cooper, H. (2008). “Post-meiotic transcription in *Drosophila* testes”. *Development* 135(11), pp. 1897–1902.
- Bass, T. M., Grandison, R. C., Wong, R., Martinez, P., Partridge, L., and Piper, M. D. (2007). “Optimization of dietary restriction protocols in *Drosophila*”. *Journals of Gerontology - Series A Biological Sciences and Medical Sciences* 62(10), pp. 1071–1081.
- Beadell, A. V. and Haag, E. S. (2014). “Evolutionary dynamics of GLD-1-mRNA complexes in *Caenorhabditis* nematodes”. *Genome Biology and Evolution* 9(7), pp. 314–35.
- Beck, A. R., Miller, I. J., Anderson, P., and Streuli, M. (1998). “RNA-binding protein TIAR is essential for primordial germ cell development.” *Proceedings of the National Academy of Sciences of the United States of America* 95(5), pp. 2331–6.
- Berber, S., Llamas, E., Thaivalappil, P., Boag, P. R., Crossley, M., and Nicholas, H. R. (2013). “Homeodomain interacting protein kinase (HPK-1) is required in the soma for robust germline proliferation in *C. elegans*”. *Developmental Dynamics* 242(11), pp. 1250–1261.
- Berglund, J. A., Chua, K., Abovich, N., Reed, R., and Rosbash, M. (1997). “The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC”. *Cell* 89(5), pp. 781–787.

## References

- Beuck, C., Qu, S., Fagg, W. S., Ares Jr., M., and Williamson, J. R. (2012). “Structural analysis of the quaking homodimerization interface”. *Journal of Molecular Biology* 423(5), pp. 766–81.
- Beuck, C., Szymczyna, B. R., Kerkow, D. E., Carmel, A. B., Columbus, L., Stanfield, R. L., and Williamson, J. R. (2010). “Structure of the GLD-1 homodimerization domain: insights into STAR protein-mediated translational regulation.” *Structure* 18(3), pp. 377–89.
- Blighe, K. (2019). *PCAtools: PCAtools: Everything Principal Components Analysis*.
- Boyle, M., Wong, C., Rocha, M., and Jones, D. L. (2007). “Decline in Self-Renewal Factors Contributes to Aging of the Stem Cell Niche in the *Drosophila* Testis”. *Cell Stem Cell* 1(4), pp. 470–478.
- Brand, A. and Perrimon, N. (1993). “Targeted gene expression as a means of altering cell fates and generating dominant phenotypes”. *Development* 118(2), pp. 401–415.
- Brawley, C. and Matunis, E. (2004). “Regeneration of Male Germline Stem Cells by Spermatogonial Dedifferentiation in Vivo”. *Science* 304(5675), pp. 1331–34.
- Brown, J. B., Boley, N., Eisman, R., May, G. E., Stoiber, M. H., Duff, M. O., Booth, B. W., Wen, J., Park, S., Suzuki, A. M., et al. (2014). “Diversity and dynamics of the *Drosophila* transcriptome”. *Nature* 512(7515), pp. 393–399.
- Buels, R., Yao, E., Diesh, C. M., Hayes, R. D., Munoz-Torres, M., Helt, G., Goodstein, D. M., Elsik, C. G., Lewis, S. E., Stein, L., et al. (2016). “JBrowse: A dynamic web platform for genome visualization and analysis”. *Genome Biology* 17(1), p. 66.
- Cagney, G., Park, S., Chung, C., Tong, B., O’Dushlaine, C., Shields, D. C., and Emili, A. (2005). “Human tissue profiling with multidimensional protein identification technology”. *Journal of Proteome Research* 4(5), pp. 1757–67.
- Carmel, A. B., Wu, J., Lehmann-Blount, K. A., and Williamson, J. R. (2010). “High-affinity consensus binding of target RNAs by the STAR/GSG proteins GLD-1, STAR-2 and Quaking”. *BMC Molecular Biology* 11(1), p. 48.
- Chang, Y. C., Tu, H., Chen, J. Y., Chang, C. C., Yang, S. Y., and Pi, H. (2019). “Reproduction disrupts stem cell homeostasis in testes of aged male *Drosophila* via an induced microenvironment”. *PLoS Genetics* 15(7), e1008062.
- Chen, J. and Verheyen, E. M. (2012). “Homeodomain-interacting protein kinase regulates yorkie activity to promote tissue growth”. *Current Biology* 22(17), pp. 1582–1586.
- Chénard, C. A. and Richard, S. (2008). “New implications for the QUAKE RNA binding protein in human disease”. *Journal of Neuroscience Research* 86(2), pp. 233–242.
- Cheng, J., Tiyaboonchai, A., Yamashita, Y. M., and Hunt, A. J. (2011). “Asymmetric division of cyst stem cells in *Drosophila* testis is ensured by anaphase spindle repositioning”. *Development* 138(5), pp. 831–837.
- Cheng, J., Türkel, N., Hemati, N., Fuller, M. T., Hunt, A. J., and Yamashita, Y. M. (2008). “Centrosome misorientation reduces stem cell division during ageing”. *Nature* 456(7222), pp. 599–604.
- Cheng, M. H. and Jansen, R. P. (2017). “A jack of all trades: the RNA-binding protein vigilin”. *Wiley Interdisciplinary Reviews: RNA* 8(6).

- Chintapalli, V. R., Wang, J., and Dow, J. A. (2007). “Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease”. *Nature Genetics* 39(6), pp. 715–720.
- Cinquin, O., Crittenden, S. L., Morgan, D. E., and Kimble, J. (2010). “Progression from a stem cell-like state to early differentiation in the *C. elegans* germ line”. *Proceedings of the National Academy of Sciences of the United States of America* 107(5), pp. 2048–2053.
- Collier, B., Gorgoni, B., Loveridge, C., Cooke, H. J., and Gray, N. K. (2005). “The DAZL family proteins are PABP-binding proteins that regulate translation in germ cells”. *EMBO Journal* 24(14), pp. 2656–2666.
- Conn, S. J., Pillman, K. A., Toubia, J., Conn, V. M., Salmanidis, M., Phillips, C. A., Roslan, S., Schreiber, A. W., Gregory, P. A., and Goodall, G. J. (2015). “The RNA binding protein quaking regulates formation of circRNAs”. *Cell* 160(6), pp. 1125–1134.
- Consortium, T. U. (2021). “UniProt: The universal protein knowledgebase in 2021”. *Nucleic Acids Research* 49(D1), pp. D480–D489.
- Corioni, M., Antih, N., Tanackovic, G., Zavolan, M., and Krämer, A. (2011). “Analysis of in situ pre-mRNA targets of human splicing factor SF1 reveals a function in alternative splicing”. *Nucleic Acids Research* 39(5), pp. 1868–1879.
- Corley, M., Burns, M. C., and Yeo, G. W. (2020). “How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms”. *Molecular Cell* 78(1), pp. 9–29.
- Crapster, J. A., Rack, P. G., Hellmann, Z. J., Le, A. D., Adams, C. M., Leib, R. D., Elias, J. E., Perrino, J., Behr, B., Li, Y., et al. (2020). “HIPK4 is essential for murine spermiogenesis”. *eLife* 9, e50209.
- Crittenden, S. L., Bernstein, D. S., Bachorik, J. L., Thompson, B. E., Gallegos, M., Petcherski, A. G., Moulder, G., Barstead, R., Wickens, M., and Kimble, J. (2002). “A conserved RNA-binding protein controls germline stem cells in *Caenorhabditis elegans*”. *Nature* 417, pp. 660–663.
- Davies, E. L. and Fuller, M. T. (2008). “Regulation of self-renewal and differentiation in adult stem cell lineages: lessons from the *Drosophila* male germ line”. *Cold Spring Harbour Symposia on Quantative Biology* 73, pp. 137–145.
- Davies, E. L., Lim, J. G., Joo, W. J., Ho Tam, C., and Fuller, M. T. (2013). “The transcriptional regulator *lola* is required for stem cell maintenance and germ cell differentiation in the *Drosophila* testis”. *Developmental Biology* 373(2), pp. 310–321.
- De Cuevas, M. and Matunis, E. L. (2011). “The stem cell niche: Lessons from the *Drosophila* testis”. *Development* 138(14), pp. 2861–2869.
- Di Fruscio, M., Chen, T., and Richard, S. (1999). “Characterization of Sam68-like mammalian proteins SLM-1 and SLM-2: SLM-1 is a Src substrate during mitosis”. *Proceedings of the National Academy of Sciences of the United States of America* 96(6), pp. 2710–2715.
- Dietzl, G., Chen, D., Schnorrer, F., Su, K. C., Barinova, Y., Fellner, M., Gasser, B., Kinsey, K., Oettel, S., Scheiblauer, S., et al. (2007). “A genome-wide transgenic RNAi library for conditional gene inactivation in *Drosophila*”. *Nature* 448(7150), pp. 151–156.

## References

- Ding, Y., Zhao, L., Yang, S., Jiang, Y., Chen, Y., Zhao, R., Zhang, Y., Zhang, G., Dong, Y., Yu, H., et al. (2010). “A Young *Drosophila* Duplicate Gene Plays Essential Roles in Spermatogenesis by Regulating Several Y-Linked Male Fertility Genes”. *PLoS Genetics* 6(12), e1001255.
- Dinges, N., Morin, V., Kreim, N., Southall, T. D., and Roignant, J. Y. (2017). “Comprehensive Characterization of the Complex *lola* Locus Reveals a Novel Role in the Octopaminergic Pathway via Tyramine Beta-Hydroxylase Regulation”. *Cell Reports* 21(10), pp. 2911–2925.
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). “STAR: ultrafast universal RNA-seq aligner”. *Bioinformatics* 29(1), pp. 15–21.
- Eberhart, C. G., Maines, J. Z., and Wasserman, S. A. (1996). “Meiotic cell cycle requirement for a fly homologue of human deleted in Azoospermia”. *Nature* 381(6585), pp. 783–785.
- Eden, E., Lipson, D., Yogev, S., and Yakhini, Z. (2007). “Discovering Motifs in Ranked Lists of DNA Sequences”. *PLoS Computational Biology* 3(3), e39.
- Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). “GORilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists”. *BMC Bioinformatics* 10(1), p. 48.
- Edenfeld, G., Volohonsky, G., Krukkert, K., Naffin, E., Lammel, U., Grimm, A., Engelen, D., Reuveny, A., Volk, T., and Klämbt, C. (2006). “The splicing factor crooked neck associates with the RNA-binding protein HOW to control glial cell maturation in *Drosophila*”. *Neuron* 52(6), pp. 969–80.
- Ehrmann, I., Crichton, J. H., Gazzara, M. R., James, K., Liu, Y., Grellscheid, S. N., Curk, T., de Rooij, D., Steyn, J. S., Cockell, S., et al. (2019). “An ancient germ cell-specific RNA-binding protein protects the germline from cryptic splice site poisoning”. *eLife* 8, e39304.
- Ehrmann, I., Dalglish, C., Liu, Y., Danilenko, M., Crosier, M., Overman, L., Arthur, H. M., Lindsay, S., Clowry, G. J., Venables, J. P., et al. (2013). “The Tissue-Specific RNA Binding Protein T-STAR Controls Regional Splicing Patterns of Neurexin Pre-mRNAs in the Brain”. *PLoS Genetics* 9(4), e1003474.
- Eisenberg, D., Schwarz, E., Komaromy, M., and Wall, R. (1984). “Analysis of membrane and surface protein sequences with the hydrophobic moment plot”. *Journal of Molecular Biology* 179(1), pp. 125–142.
- Fabian, L. and Brill, J. A. (2012). “*Drosophila* spermiogenesis”. *Spermatogenesis* 2(3), pp. 197–212.
- Fagg, W. S., Liu, N., Fair, J. H., Shiue, L., Katzman, S., Donohue, J. P., and Ares, M. (2017). “Autogenous cross-regulation of Quaking mRNA processing and translation balances Quaking functions in splicing and translation”. *Genes & Development* 31(18), pp. 1894–1909.
- Feracci, M., Foot, J. N., Grellscheid, S. N., Danilenko, M., Stehle, R., Gonchar, O., Kang, H.-S., Dalglish, C., Meyer, N. H., Liu, Y., et al. (2016). “Structural basis of RNA recog-

- nitiation and dimerization by the STAR proteins T-STAR and Sam68”. *Nature Communications* 7, p. 10355.
- Fortier, T. M., Chatterjee, R., Klinedinst, S., Baehrecke, E. H., and Woodard, C. T. (2006). “how functions in leg development during *Drosophila* metamorphosis”. *Developmental Dynamics* 235(8), pp. 2248–59.
- Francis, R., Maine, E., and Schedl, T. (1995a). “Analysis of the multiple roles of *gld-1* in germline development: Interactions with the sex determination cascade and the *glp-1* signaling pathway”. *Genetics* 139(2), pp. 607–630.
- Francis, R., Barton, M. K., Kimble, J., and Schedl, T. (1995b). “*gld-1*, a tumor suppressor gene required for oocyte development in *Caenorhabditis elegans*.” *Genetics* 139(2), pp. 579–606.
- Galarneau, A. and Richard, S. (2005). “Target RNA motif and target mRNAs of the Quaking STAR protein”. *Nature Structural and Molecular Biology* 12(8), pp. 691–698.
- (2009). “The STAR RNA binding proteins GLD-1, QKI, SAM68 and SLM-2 bind bipartite RNA motifs”. *BMC Molecular Biology* 10(47), doi:10.1186/1471-47.
- Garrey, S. M., Voelker, R., and Berglund, J. A. (2006). “An extended RNA binding site for the yeast branch point-binding protein and the role of its zinc knuckle domains in RNA binding”. *Journal of Biological Chemistry* 281(37), pp. 27443–27453.
- Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., and Bairoch, A. (2005). “Protein Identification and Analysis Tools on the ExPASy Server”. *The Proteomics Protocols Handbook*. Humana Press, pp. 571–607.
- Gilboa, L. and Lehmann, R. (2004). “Repression of primordial germ cell differentiation parallels germ line stem cell maintenance”. *Current Biology* 14(11), pp. 981–986.
- Giuliani, G., Giuliani, F., Volk, T., and Rabouille, C. (2014). “The *Drosophila* RNA-binding protein HOW controls the stability of *dgrasp* mRNA in the follicular epithelium.” *Nucleic Acids Research* 42(3), pp. 1970–86.
- Glazar, P., Papavasileiou, P., and Rajewsky, N. (2014). “circBase: A database for circular RNAs”. *RNA* 20(11), pp. 1666–1670.
- Goeke, S., Greene, E. A., Grant, P. K., Gates, M. A., Crowner, D., Aigaki, T., and Giniger, E. (2003). “Alternative splicing of *lola* generates 19 transcription factors controlling axon guidance in *Drosophila*”. *Nature Neuroscience* 6(9), pp. 917–924.
- Gönczy, P., Matunis, E., and DiNardo, S. (1997). “bag-of-marbles and benign genial cell neoplasm act in the germline to restrict proliferation during *Drosophila* spermatogenesis”. *Development* 124(21), pp. 4361–4371.
- Gordon, A. (2010). *FASTQ/A short-reads pre-processing tools*.
- Graindorge, A., Carré, C., and Gebauer, F. (2013). “Sex-lethal promotes nuclear retention of *msl2* mRNA via interactions with the STAR protein HOW”. *Genes & Development* 27(12), pp. 1421–33.
- Greenbaum, M. P., Iwamori, T., Buchold, G. M., and Matzuk, M. M. (2011). “Germ cell intercellular bridges.” *Cold Spring Harbor Perspectives in Biology* 3(8), a005850.
- Grishin, N. V. (2001). “KH domain: one motif, two folds.” *Nucleic Acids Research* 29(3), pp. 638–43.

## References

- Gustafson, E. A. and Wessel, G. M. (2010). *Vasa genes: Emerging roles in the germ line and in multipotent cells*.
- Hafner, M., Katsantoni, M., Köster, T., Marks, J., Mukherjee, J., Staiger, D., Ule, J., and Zavolan, M. (2021). “CLIP and complementary methods”. *Nature Reviews Methods Primers* 1(1), p. 20.
- Hentze, M. W., Castello, A., Schwarzl, T., and Preiss, T. (2018). “A brave new world of RNA-binding proteins”. *Nature Reviews Molecular Cell Biology* 19(5), pp. 327–341.
- Hime, G. R., Brill, J. A., and Fuller, M. T. (1996). “Assembly of ring canals in the male germ line from structural components of the contractile ring”. *Journal of Cell Science* 109(12), pp. 2779–2788.
- Hu, Y., Roesel, C., Flockhart, I., Perkins, L., Perrimon, N., and Mohr, S. E. (2013). “UP-TORR: Online tool for accurate and up-to-date annotation of RNAi reagents”. *Genetics* 195(1), pp. 37–45.
- Huang, H., Du, G., Chen, H., Liang, X., Li, C., Zhu, N., Xue, L., Ma, J., and Jiao, R. (2011). “*Drosophila* Smt3 negatively regulates JNK signaling through sequestering Hipk in the nucleus”. *Development* 138(12), pp. 2477–2485.
- Hundertmark, T., Gärtner, S. M. K., Rathke, C., and Renkawitz-Pohl, R. (2018). “Nejire/dCBP-mediated histone H3 acetylation during spermatogenesis is essential for male fertility in *Drosophila melanogaster*”. *PLoS ONE* 13(9), e0203622.
- Idler, R. K. and Yan, W. (2012). “Control of Messenger RNA Fate by RNA-Binding Proteins: An Emphasis on Mammalian Spermatogenesis”. *Journal of Andrology* 33(3), pp. 309–337.
- Insko, M. L., Leon, A., Tam, C. H., McKearin, D. M., and Fuller, M. T. (2009). “Accumulation of a differentiation regulator specifies transit amplifying division number in an adult stem cell lineage.” *Proceedings of the National Academy of Sciences of the United States of America* 106(52), pp. 22311–6.
- Israeli, D., Nir, R., and Volk, T. (2007). “Dissection of the target specificity of the RNA-binding protein HOW reveals dpp mRNA as a novel HOW target.” *Development* 134(11), pp. 2107–14.
- Jan, E., Motzny, C. K., Graves, L. E., and Goodwin, E. B. (1999). “The STAR protein, GLD-1, is a translational regulator of sexual identity in *Caenorhabditis elegans*”. *The EMBO Journal* 18(1), pp. 258–269.
- Kadyk, L. and Kimble, J. (1998). *Genetic regulation of entry into meiosis in Caenorhabditis elegans*. Tech. rep. 10, pp. 1803–1813.
- Keene, J. D., Komisarow, J. M., and Friedersdorf, M. B. (2006). “RIP-Chip: The isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts”. *Nature Protocols* 1(1), pp. 302–307.
- Kershner, A. M. and Kimble, J. (2010). “Genome-wide analysis of mRNA targets for *Caenorhabditis elegans* FBF, a conserved stem cell regulator”. *Proceedings of the National Academy of Sciences of the United States of America* 107(8), pp. 3936–3941.

- Kiger, A. A., Jones, D. L., Schulz, C., Rogers, M. B., and Fuller, M. T. (2001). “Stem cell self-renewal specified by JAK-STAT activation in response to a support cell cue”. *Science* 294(5551), pp. 2542–2545.
- Kondo, T., Furuta, T., Mitsunaga, K., Ebersole, T. A., Shichiri, M., Wu, J., Artzt, K., Ichi Yamamura, K., and Abe, K. (1999). “Genomic organization and expression analysis of the mouse qkI locus”. *Mammalian Genome* 10(7), pp. 662–669.
- Laktionov, P. P., Maksimov, D. A., Romanov, S. E., Antoshina, P. A., Posukh, O. V., White-Cooper, H., Koryakov, D. E., and Belyakin, S. N. (2018). “Genome-wide analysis of gene regulation mechanisms during *Drosophila* spermatogenesis”. *Epigenetics & Chromatin* 11(1), p. 14.
- Lamont, L. B., Crittenden, S. L., Bernstein, D., Wickens, M., and Kimble, J. (2004). “FBF-1 and FBF-2 regulate the size of the mitotic region in the *C. elegans* germline”. *Developmental Cell* 7(5), pp. 697–707.
- Larkin, A., Marygold, S. J., Antonazzo, G., Attrill, H., dos Santos, G., Garapati, P. V., Goodman, J. L., Sian Gramates, L., Millburn, G., Strelets, V. B., et al. (2021). “Fly-Base: Updates to the *Drosophila melanogaster* knowledge base”. *Nucleic Acids Research* 49(D1), pp. D899–D907.
- Leatherman, J. L. and Di Nardo, S. (2008). “Zfh-1 controls somatic stem cell self-renewal in the *Drosophila* testis and nonautonomously influences germline stem cell self-renewal”. *Cell Stem Cell* 3(1), pp. 44–54.
- Leatherman, J. L. and Dinardo, S. (2010). “Germline self-renewal requires cyst stem cells and stat regulates niche adhesion in *Drosophila* testes”. *Nature Cell Biology* 12(8), pp. 806–811.
- Lee, M. H. and Schedi, T. (2010). “*C. elegans* STAR proteins, GLD-1 and ASD-2, regulate specific RNA targets to control development”. *Advances in Experimental Medicine and Biology* 693, pp. 106–122.
- Lee, W., Andrews, B. C., Faust, M., Walldorf, U., and Verheyen, E. M. (2009). “Hipk is an essential protein that promotes Notch signal transduction in the *Drosophila* eye by inhibition of the global co-repressor Groucho”. *Developmental Biology* 325(1), pp. 263–272.
- Legrand, J. M. and Hobbs, R. M. (2018). “RNA processing in the male germline: Mechanisms and implications for fertility”. *Seminars in Cell and Developmental Biology* 79, pp. 80–91.
- Li, P., Zhou, X., Xu, K., and Zhang, Q. C. (2021). “RASP: An atlas of transcriptome-wide RNA secondary structure probing data”. *Nucleic Acids Research* 49(D1), pp. D183–D191.
- Liao, Y., Smyth, G. K., and Shi, W. (2014). “featureCounts: an efficient general purpose program for assigning sequence reads to genomic features”. *Bioinformatics* 30(7), pp. 923–930.
- Liao, Y., Smyth, G. K., and Shi, W. (2013). “The Subread aligner: Fast, accurate and scalable read mapping by seed-and-vote”. *Nucleic Acids Research* 41(10).

## References

- Licatalosi, D. D., Mele, A., Fak, J. J., Ule, J., Kayikci, M., Chi, S. W., Clark, T. A., Schweitzer, A. C., Blume, J. E., Wang, X., et al. (2008). “HITS-CLIP yields genome-wide insights into brain alternative RNA processing”. *Nature* 456(7221), pp. 464–469.
- Lim, C., Tarayrah, L., and Chen, X. (2012). “Transcriptional regulation during *Drosophila* spermatogenesis”. *Spermatogenesis* 2(3), pp. 158–166.
- Lin, Q., Taylor, S. J., and Shalloway, D. (1997). “Specificity and determinants of Sam68 RNA binding. Implications for the biological function of K homology domains”. *Journal of Biological Chemistry* 272(43), pp. 27274–27280.
- Liu, Z., Luyten, I., Bottomley, M. J., Messias, A. C., Houngninou-Molango, S., Sprangers, R., Zanier, K., Krämer, A., and Sattler, M. (2001). “Structural Basis for Recognition of the Intron Branch Site RNA by Splicing Factor 1”. *Science* 294(5544), pp. 1098–1102.
- Lo, P. C. and Frasch, M. (1997). “A novel KH-domain protein mediates cell adhesion processes in *Drosophila*”. *Developmental Biology* 190(2), pp. 241–56.
- 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), p. 550.
- Luetjens, C. M., Xu, E. Y., Rejo Pera, R. A., Kamischke, A., Nieschlag, E., and Gromoll, J. (2004). “Association of meiotic arrest with lack of BOULE protein expression in infertile men”. *Journal of Clinical Endocrinology and Metabolism* 89(4), pp. 1926–1933.
- Lukong, K. E. and Richard, S. (2003). “Sam68, the KH domain-containing superSTAR”. *Biochimica et Biophysica Acta* 1653(2), pp. 73–86.
- Lunde, B. M., Moore, C., and Varani, G. (2007). “RNA-binding proteins: Modular design for efficient function”. *Nature Reviews Molecular Cell Biology* 8(6), pp. 479–490.
- Ma, W., Baumann, C., and Viveiros, M. M. (2010). “NEDD1 is crucial for meiotic spindle stability and accurate chromosome segregation in mammalian oocytes”. *Developmental Biology* 339(2), pp. 439–450.
- Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A. R., Potter, S. C., Finn, R. D., et al. (2019). “The EMBL-EBI search and sequence analysis tools APIs in 2019”. *Nucleic Acids Research* 47(W1), W636–W641.
- Maines, J. Z. and Wasserman, S. A. (1999). “Post-transcriptional regulation of the meiotic Cdc25 protein Twine by the Dazl orthologue Boule”. *Nature Cell Biology* 1(3), pp. 171–174.
- Marin, V. A. and Evans, T. C. (2003). “Translational repression of a *C. elegans* Notch mRNA by the STAR/KH domain protein GLD-1”. *Development* 130(12), pp. 2623–32.
- Martin, M. (2011). “Cutadapt removes adapter sequences from high-throughput sequencing reads”. *EMBnet.journal* 17(1), p. 10.
- Matter, N., Herrlich, P., and König, H. (2002). “Signal-dependent regulation of splicing via phosphorylation of Sam68”. *Nature* 420(6916), pp. 691–695.
- Matunis, E. L., Stine, R. R., and de Cuevas, M. (2012). “Recent advances in *Drosophila* male germline stem cell biology”. *Spermatogenesis* 2(3), pp. 137–144.
- Mazan-Mamczarz, K., Lal, A., Martindale, J. L., Kawai, T., and Gorospe, M. (2006). “Translational Repression by RNA-Binding Protein TIAR”. *Molecular and Cellular Biology* 26(7), pp. 2716–2727.

- McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). “Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation”. *Nucleic Acids Research* 40(10), pp. 4288–4297.
- McKearin, D. and Ohlstein, B. (1995). *A role for the Drosophila bag-of-marbles protein in the differentiation of cystoblasts from germline stem cells*. Tech. rep. 9, pp. 2937–2947.
- McKee, B. D., Yan, R., and Tsai, J.-H. (2012). “Meiosis in male *Drosophila*”. *Spermatogenesis* 2(3), pp. 167–184.
- McLeod, C. J., Wang, L., Wong, C., and Jones, D. L. (2010). “Stem cell dynamics in response to nutrient availability”. *Current Biology* 20(23), pp. 2100–2105.
- McMahon, A. C., Rahman, R., Jin, H., Shen, J. L., Fieldsend, A., Luo, W., and Rosbash, M. (2016). “TRIBE: Hijacking an RNA-Editing Enzyme to Identify Cell-Specific Targets of RNA-Binding Proteins”. *Cell* 165(3), pp. 742–753.
- Messina, V., Meikar, O., Paronetto, M. P., Calabretta, S., Geremia, R., Kotaja, N., and Sette, C. (2012). “The RNA binding protein SAM68 transiently localizes in the chromatoid body of male germ cells and influences expression of select microRNAs”. *PLoS ONE* 7(6), e39729.
- Meyer, N. H., Tripsianes, K., Vincendeau, M., Madl, T., Kateb, F., Brack-Werner, R., and Sattler, M. (2010). “Structural basis for homodimerization of the Src-associated during mitosis, 68-kDa protein (Sam68) Qua1 domain”. *The Journal of Biological Chemistry* 285(37), pp. 28893–901.
- Mili, S. and Steitz, J. A. (2004). “Evidence for reassociation of RNA-binding proteins after cell lysis: Implications for the interpretation of immunoprecipitation analyses”. *RNA* 10(11), pp. 1692–1694.
- Moerke, N. J. (2009). “Fluorescence Polarization (FP) Assays for Monitoring Peptide-Protein or Nucleic Acid-Protein Binding”. *Current Protocols in Chemical Biology* 1(1), pp. 1–15.
- Monk, A. C., Siddall, N. A., Fraser, B., McLaughlin, E. A., and Hime, G. R. (2011). “Differential roles of HOW in male and female *Drosophila* germline differentiation”. *PLoS ONE* 6(12), e28508.
- Monk, A. C., Siddall, N. A., Volk, T., Fraser, B., Quinn, L. M., McLaughlin, E. A., and Hime, G. R. (2010). “HOW Is Required for Stem Cell Maintenance in the *Drosophila* Testis and for the Onset of Transit-Amplifying Divisions”. *Cell Stem Cell* 6(4), pp. 348–360.
- Moschall, R., Gaik, M., and Medenbach, J. (2017). “Promiscuity in post-transcriptional control of gene expression: *Drosophila* sex-lethal and its regulatory partnerships”. *FEBS Letters* 591(11), pp. 1471–1488.
- Mukherjee, N., Corcoran, D. L., Nusbaum, J. D., Reid, D. W., Georgiev, S., Hafner, M., Ascano, M., Tuschl, T., Ohler, U., and Keene, J. D. (2011). “Integrative Regulatory Mapping Indicates that the RNA-Binding Protein HuR Couples Pre-mRNA Processing and mRNA Stability”. *Molecular Cell* 43(3), pp. 327–339.

## References

- Nabel-Rosen, H., Volohonsky, G., Reuveny, A., Zaidel-Bar, R., and Volk, T. (2002). “Two isoforms of the *Drosophila* RNA binding protein, How, act in opposing directions to regulate tendon cell differentiation”. *Developmental Cell* 2(2), pp. 183–93.
- Nabel-Rosen, H., Dorevitch, N., Reuveny, A., and Volk, T. (1999). “The balance between two isoforms of the *Drosophila* RNA-binding protein How controls tendon cell differentiation”. *Molecular Cell* 4(4), pp. 573–84.
- Nabel-Rosen, H., Toledano-Katchalski, H., Volohonsky, G., and Volk, T. (2005). “Cell divisions in the *Drosophila* embryonic mesoderm are repressed via posttranscriptional regulation of string/cdc25 by HOW”. *Current Biology* 15(4), pp. 295–302.
- Nakagawa, T., Sharma, M., Nabeshima, Y.-i., Braun, R. E., and Yoshida, S. (2010). “Functional hierarchy and reversibility within the murine spermatogenic stem cell compartment.” *Science* 328(5974), pp. 62–7.
- Nakel, K., Hartung, S. A., Bonneau, F., Eckmann, C. R., and Conti, E. (2010). “Four KH domains of the *C. elegans* Bicaudal-C ortholog GLD-3 form a globular structural platform”. *RNA* 16(11), pp. 2058–2067.
- Nelson, J. O., Chen, C., and Yamashita, Y. M. (2019). “Germline stem cell homeostasis”. *Current Topics in Developmental Biology*. Vol. 135. Academic Press Inc., pp. 203–244.
- Nicastro, G., Taylor, I. A., and Ramos, A. (2015). “KH–RNA interactions: back in the groove”. *Current Opinion in Structural Biology* 30, pp. 63–70.
- Nir, R., Grossman, R., Paroush, Z., and Volk, T. (2012). “Phosphorylation of the *Drosophila melanogaster* RNA-binding protein HOW by MAPK/ERK enhances its dimerization and activity”. *PLoS Genetics* 8(3), e1002632.
- O’Bryan, M. K., Clark, B. J., McLaughlin, E. A., D’Sylva, R. J., O’Donnell, L., Wilce, J. A., Sutherland, J., O’Connor, A. E., Whittle, B., Goodnow, C. C., et al. (2013). “RBM5 Is a Male Germ Cell Splicing Factor and Is Required for Spermatid Differentiation and Male Fertility”. *PLoS Genetics* 9(7), e1003628.
- Pangratiou, G. M. (2020). “Structure and Function of RNA Binding Proteins from Negative Sense-RNA Viruses”. PhD thesis. University of Leeds.
- Parker, G. A. (1970). “Sperm competition and its evolutionary consequences in the insects”. *Biological Reviews* 45(4), pp. 525–567.
- Paronetto, M. P., Zalfa, F., Botti, F., Geremia, R., Bagni, C., and Sette, C. (2006). “The Nuclear RNA-binding Protein Sam68 Translocates to the Cytoplasm and Associates with the Polysomes in Mouse Spermatocytes”. *Molecular Biology of the Cell* 17(1), pp. 14–24.
- Paronetto, M. P., Messina, V., Barchi, M., Geremia, R., Richard, S., and Sette, C. (2011). “Sam68 marks the transcriptionally active stages of spermatogenesis and modulates alternative splicing in male germ cells.” *Nucleic Acids Research* 39(12), pp. 4961–74.
- Paronetto, M. P., Messina, V., Bianchi, E., Barchi, M., Vogel, G., Moretti, C., Palombi, F., Stefanini, M., Geremia, R., Richard, S., et al. (2009). “Sam68 regulates translation of target mRNAs in male germ cells, necessary for mouse spermatogenesis”. *The Journal of Cell Biology* 185(2), pp. 235–49.

- Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). “Salmon provides fast and bias-aware quantification of transcript expression”. *Nature Methods* 14(4), pp. 417–419.
- Perkins, L. A., Holderbaum, L., Tao, R., Hu, Y., Sopko, R., McCall, K., Yang-Zhou, D., Flockhart, I., Binari, R., Shim, H. S., et al. (2015). “The transgenic RNAi project at Harvard medical school: Resources and validation”. *Genetics* 201(3), pp. 843–852.
- Poon, C. L., Zhang, X., Lin, J. I., Manning, S. A., and Harvey, K. F. (2012). “Homeodomain-interacting protein kinase regulates Hippo pathway-dependent tissue growth”. *Current Biology* 22(17), pp. 1587–1594.
- Priti, A. and Subramaniam, K. (2015). “PUF-8 Functions Redundantly with GLD-1 to Promote the Meiotic Progression of Spermatocytes in *Caenorhabditis elegans*.” *G3* 5(8), pp. 1675–84.
- Prowse, N. and Partridge, L. (1997). “The effects of reproduction on longevity and fertility in male *Drosophila melanogaster*”. *Journal of Insect Physiology* 43(6), pp. 501–512.
- Qi, H. (2016). “RNA-binding proteins in mouse male germline stem cells: A mammalian perspective”. *Cell Regeneration* 5(1), 5:1.
- Ray, D., Kazan, H., Cook, K. B., Weirauch, M. T., Najafabadi, H. S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., et al. (2013). “A compendium of RNA-binding motifs for decoding gene regulation”. *Nature* 499(7457), pp. 172–177.
- Reijo, R., Lee, T. Y., Salo, P., Alagappan, R., Brown, L. G., Rosenberg, M., Rozen, S., Jaffe, T., Straus, D., Hovatta, O., et al. (1995). “Diverse spermatogenic defects in humans caused by Y chromosome deletions encompassing a novel RNA-binding protein gene”. *Nature Genetics* 10(4), pp. 383–393.
- Reuveny, A., Elhanany, H., and Volk, T. (2009). “Enhanced sensitivity of midline glial cells to apoptosis is achieved by HOW(L)-dependent repression of Diap1”. *Mechanisms of Development* 126(1-2), pp. 30–41.
- Reynolds, N., Collier, B., Maratou, K., Bingham, V., Speed, R. M., Taggart, M., Semple, C. A., Gray, N. K., and Cooke, H. J. (2005). “Dazl binds in vivo to specific transcripts and can regulate the pre-meiotic translation of Mvh in germ cells”. *Human Molecular Genetics* 14(24), pp. 3899–909.
- Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”. *Bioinformatics* 26(1), pp. 139–140.
- Rodrigues, F., Thuma, L., and Klämbt, C. (2012). “The regulation of glial-specific splicing of Neurexin IV requires HOW and Cdk12 activity”. *Development* 139(10), pp. 1765–76.
- Roy, A., Kucukural, A., and Zhang, Y. (2010). “I-TASSER: A unified platform for automated protein structure and function prediction”. *Nature Protocols* 5(4), pp. 725–738.
- RStudio Team (2020). *RStudio: Integrated Development Environment for R*. RStudio, PBC. Boston, MA.
- Rutz, B. and Séraphin, B. (2000). “A dual role for BBP/ScSF1 in nuclear pre-mRNA retention and splicing”. *EMBO Journal* 19(8), pp. 1873–1886.

## References

- Ryder, S. P. and Williamson, J. R. (2004). “Specificity of the STAR/GSG domain protein Qk1: Implications for the regulation of myelination”. *RNA* 10(9), pp. 1449–1458.
- Ryder, S. P., Frater, L. A., Abramovitz, D. L., Goodwin, E. B., and Williamson, J. R. (2004). “RNA target specificity of the STAR/GSG domain post-transcriptional regulatory protein GLD-1”. *Nature Structural and Molecular Biology* 11(1), pp. 20–28.
- Saccomanno, L., Loushin, C., Jan, E., Punkay, E., Artzt, K., and Goodwin, E. B. (1999). “The STAR protein QKI-6 is a translational repressor”. *Proceedings of the National Academy of Sciences of the United States of America* 96(22), pp. 12605–12610.
- Sada, A., Suzuki, A., Suzuki, H., and Saga, Y. (2009). “The RNA-binding protein NANOS2 is required to maintain murine spermatogonia! Stem Cells”. *Science* 325(5946), pp. 1394–1398.
- Schäfer, M., Nayernia, K., Engel, W., and Schäfer, U. (1995). “Translational control in spermatogenesis”. *Developmental Biology* 172(2), pp. 344–352.
- Scheckel, C., Gaidatzis, D., Wright, J. E., and Ciosk, R. (2012). “Genome-wide analysis of GLD-1-mediated mRNA regulation suggests a role in mRNA storage.” *PLoS Genetics* 8(5), e1002742.
- Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al. (2012). “Fiji: An open-source platform for biological-image analysis”. *Nature Methods* 9(7), pp. 676–682.
- Schotman, H., Karhinen, L., and Rabouille, C. (2008). “dGRASP-Mediated Noncanonical Integrin Secretion Is Required for *Drosophila* Epithelial Remodeling”. *Developmental Cell* 14(2), pp. 171–182.
- Schüpbach, T. and Wieschaus, E. (1989). “Female sterile mutations on the second chromosome of *Drosophila melanogaster*. I. Maternal effect mutations.” *Genetics* 121(1), pp. 101–117.
- Sharma, D., Zagore, L. L., Brister, M. M., Ye, X., Crespo-Hernández, C. E., Licatalosi, D. D., and Jankowsky, E. (2021). “The kinetic landscape of an RNA-binding protein in cells”. *Nature*, pp. 1–5.
- Sharpe, R. (1994). “Regulation of spermatogenesis”. *The Physiology of Reproduction*. Ed. by E. Knobil and J. Neill. Raven Press (New York), pp. 1363–34.
- Shitashige, M., Satow, R., Honda, K., Ono, M., Hirohashi, S., and Yamada, T. (2007). “Increased susceptibility of Sf1+/- mice to azoxymethane-induced colon tumorigenesis”. *Cancer Science* 98(12), pp. 1862–1867.
- Sigrist, S., Ried, G., and Lehner, C. F. (1995). “Dmcdc2 kinase is required for both meiotic divisions during”. *Mechanisms of Development* 53, pp. 247–260.
- Silva, A. P., Chechik, M., Byrne, R. T., Waterman, D. G., Ng, C. L., Dodson, E. J., Koonin, E. V., Antson, A. A., and Smits, C. (2011). “Structure and activity of a novel archaeal beta-CASP protein with N-terminal KH domains”. *Structure* 19(5), pp. 622–632.
- Siomi, H., Matunis, M. J., Michael, W. M., and Dreyfuss, G. (1993). “The pre-mRNA binding K protein contains a novel evolutionary conserved motif”. *Nucleic Acids Research* 21(5), pp. 1193–1198.

- Snoke, M. S. and Promislow, D. E. (2003). “Quantitative genetic tests of recent senescence theory: Age-specific mortality and male fertility in *Drosophila melanogaster*”. *Heredity* 91(6), pp. 546–556.
- Soneson, C., Love, M. I., and Robinson, M. D. (2015). “Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences”. *F1000Research* 4, p. 1521.
- Song, Z., Wu, P., Ji, P., Zhang, J., Gong, Q., Wu, J., and Shi, Y. (2012). “Solution structure of the second RRM domain of RBM5 and its unusual binding characters for different RNA targets”. *Biochemistry* 51(33), pp. 6667–6678.
- Soumillon, M., Necsulea, A., Weier, M., Brawand, D., Zhang, X., Gu, H., Barthès, P., Kokkinaki, M., Nef, S., Gnirke, A., et al. (2013). “Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis”. *Cell Reports* 3(6), pp. 2179–2190.
- Spradling, A., Fuller, M. T., Braun, R. E., and Yoshida, S. (2011). “Germline stem cells”. *Cold Spring Harbor Perspectives in Biology* 3(11), a002642.
- Stork, T., Thomas, S., Rodrigues, F., Silies, M., Naffin, E., Wenderdel, S., and Klämbt, C. (2009). “*Drosophila* neurexin IV stabilizes neuron-glia interactions at the CNS midline by binding to wrapper”. *Development* 136(8), pp. 1251–1261.
- Stoss, O., Novoyatleva, T., Gencheva, M., Olbrich, M., Benderska, N., and Stamm, S. (2004). “p59(fyn)-mediated phosphorylation regulates the activity of the tissue-specific splicing factor rSLM-1”. *Molecular and Cellular Neuroscience* 27(1), pp. 8–21.
- Szostak, E., García-Beyaert, M., Guitart, T., Graindorge, A., Coll, O., and Gebauer, F. (2018). “Hrp48 and eIF3d contribute to msl-2 mRNA translational repression”. *Nucleic Acids Research* 46(8), pp. 4099–4113.
- Taliaferro, J. M., Lambert, N. J., Sudmant, P. H., Dominguez, D., Merkin, J. J., Alexis, M. S., Bazile, C. A., Burge, C. B., Agarwal, V., Bell, G., et al. (2016). “RNA Sequence Context Effects Measured In Vitro Predict In Vivo Protein Binding and Regulation”. *Molecular Cell* 64(2), pp. 294–306.
- Tanackovic, G. and Krämer, A. (2005). “Human splicing factor SF3a, but not SF1, is essential for pre-mRNA splicing in vivo”. *Molecular Biology of the Cell* 16(3), pp. 1366–1377.
- Taylor, S. C., Nadeau, K., Abbasi, M., Lachance, C., Nguyen, M., and Fenrich, J. (2019). “The Ultimate qPCR Experiment: Producing Publication Quality, Reproducible Data the First Time”. *Trends in Biotechnology* 37(7), pp. 761–774.
- Tenenbaum, S. A., Carson, C. C., Lager, P. J., and Keene, J. D. (2000). “Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays”. *Proceedings of the National Academy of Sciences of the United States of America* 97(26), pp. 14085–14090.
- Teplova, M., Hafner, M., Teplov, D., Essig, K., Tuschl, T., and Patel, D. J. (2013). “Structure-function studies of STAR family Quaking proteins bound to their in vivo RNA target sites.” *Genes & Development* 27(8), pp. 928–40.

## References

- Tettweiler, G., Blaquiére, J. A., Wray, N. B., and Verheyen, E. M. (2019). “Hipk is required for JAK/STAT activity during development and tumorigenesis”. *PLoS ONE* 14(12), e0226856.
- Theil, K., Herzog, M., and Rajewsky, N. (2018). “Post-transcriptional Regulation by 3’ UTRs Can Be Masked by Regulatory Elements in 5’ UTRs”. *Cell Reports* 22(12), pp. 3217–3226.
- Timakov, B. and Zhang, P. (2000). “Genetic analysis of a Y-chromosome region that induces triplosterile phenotypes and is essential for spermatid individualization in *Drosophila melanogaster*”. *Genetics* 155(1), pp. 179–189.
- Toledano-Katchalski, H., Nir, R., Volohonsky, G., and Volk, T. (2007). “Post-transcriptional repression of the *Drosophila* midkine and pleiotrophin homolog *miple* by HOW is essential for correct mesoderm spreading”. *Development* 134(19), pp. 3473–81.
- Trcek, T. and Lehmann, R. (2019). “Germ granules in *Drosophila*”. *Traffic* 20(9), pp. 650–660.
- Tsuda, M., Sasaoka, Y., Kiso, M., Abe, K., Haraguchi, S., Kobayashi, S., and Saga, Y. (2003). “Conserved role of nanos proteins in germ cell development”. *Science* 301(5637), pp. 1239–1241.
- Tu, Z., Bayazit, M. B., Liu, H., Zhang, J., Busayavalasa, K., Risal, S., Shao, J., Satyanarayana, A., Coppola, V., Tessarollo, L., et al. (2017). “Speedy A-Cdk2 binding mediates initial telomere-nuclear envelope attachment during meiotic prophase I independent of Cdk2 activation”. *Proceedings of the National Academy of Sciences of the United States of America* 114(3), pp. 592–597.
- Tulina, N. and Matunis, E. (2001). “Control of stem cell self-renewal in *Drosophila* spermatogenesis by JAK-STAT signaling”. *Science* 294(5551), pp. 2546–2549.
- Ueishi, S., Shimizu, H., and Inoue, Y. H. (2009). “Male germline stem cell division and spermatocyte growth require insulin signaling in *Drosophila*”. *Cell Structure and Function* 34(1), pp. 61–69.
- Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., et al. (2015). “Tissue-based map of the human proteome”. *Science* 347(6220), p. 1260419.
- Valverde, R., Edwards, L., and Regan, L. (2008). “Structure and function of KH domains”. *FEBS Journal* 275(11), pp. 2712–2726.
- Van Nostrand, E. L., Freese, P., Pratt, G. A., Wang, X., Wei, X., Xiao, R., Blue, S. M., Chen, J.-Y., Cody, N. A. L., Dominguez, D., et al. (2020). “A large-scale binding and functional map of human RNA-binding proteins”. *Nature* 583(7818), pp. 711–719.
- VanGompel, M. J. and Xu, E. Y. (2011). “The roles of the DAZ family in spermatogenesis”. *Spermatogenesis* 1(1), pp. 36–46.
- VanGompel, M. J. and Xu, E. Y. (2010). “A novel requirement in mammalian spermatid differentiation for the DAZ-family protein Boule”. *Human Molecular Genetics* 19(12), pp. 2360–2369.
- Venables, J. P., Vernet, C., Chew, S. L., Elliott, D. J., Cowmeadow, R. B., Wu, J., Cooke, H. J., Artzt, K., and Eperon, I. C. (1999). “T-STAR/ETOILE: A novel relative of

- SAM68 that interacts with an RNA-binding protein implicated in spermatogenesis”. *Human Molecular Genetics* 8(6), pp. 959–969.
- Vernet, C. and Artzt, K. (1997). “STAR, a gene family involved in signal transduction and activation of RNA”. *Trends in Genetics* 13(12), pp. 479–484.
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). “SciPy 1.0: fundamental algorithms for scientific computing in Python”. *Nature Methods* 17(3), pp. 261–272.
- Volk, T., Israeli, D., Nir, R., and Toledano-Katchalski, H. (2008). “Tissue development and RNA control: “HOW” is it coordinated?” *Trends in Genetics* 24(2), pp. 94–101.
- Wang, Z. and Lin, H. (2004). “Nanos Maintains Germline Stem Cell Self-Renewal by Preventing Differentiation”. *Science* 303(5666), pp. 2016–2019.
- Westerveld, G., Gianotten, J., Leschot, N., Van der Veen, F., Repping, S., and Lombardi, M. (2004). “Heterogeneous nuclear ribonucleoprotein G-T (HNRNP G-T) mutations in men with impaired spermatogenesis”. *Molecular Human Reproduction* 10(4), pp. 265–269.
- White-Cooper, H., Schafer, M., Alphey, L., and Fuller, M. (1998). “Transcriptional and post-transcriptional control mechanisms coordinate the onset of spermatid differentiation with meiosis I in *Drosophila*”. *Development* 125(1), pp. 125–134.
- White-Cooper, H. (2010). “Molecular mechanisms of gene regulation during *Drosophila* spermatogenesis”. *Reproduction* 139(1), pp. 11–21.
- (2012). “Tissue, cell type and stage-specific ectopic gene expression and RNAi induction in the *Drosophila* testis”. *Spermatogenesis* 2(1), pp. 11–12.
- Witt, E., Benjamin, S., Svetec, N., and Zhao, L. (2019). “Testis single-cell RNA-seq reveals the dynamics of de novo gene transcription and germline mutational bias in *Drosophila*”. *eLife* 8, e47138.
- Wright, J. E., Gaidatzis, D., Senften, M., Farley, B. M., Westhof, E., Ryder, S. P., and Ciosk, R. (2011). “A quantitative RNA code for mRNA target selection by the germline fate determinant GLD-1”. *EMBO Journal* 30(3), pp. 533–545.
- Xiao, C., Qiu, S., and Robertson, R. M. (2017). “The white gene controls copulation success in *Drosophila melanogaster*”. *Scientific Reports* 7(1), pp. 1–13.
- Xu, E. Y., Lee, D. F., Klebes, A., Turek, P. J., Kornberg, T. B., and Reijo Pera, R. A. (2003). “Human BOULE gene rescues meiotic defects in infertile flies”. *Human Molecular Genetics* 12(2), pp. 169–175.
- Xu, E. Y., Moore, F. L., and Reijo Pera, R. A. (2001). “A gene family required for human germ cell development evolved from an ancient meiotic gene conserved in metazoans”. *Proceedings of the National Academy of Sciences of the United States of America* 98(13), pp. 7414–7419.
- Xu, S., Hafer, N., Agunwamba, B., and Schedl, P. (2012). “The CPEB Protein Orb2 Has Multiple Functions during Spermatogenesis in *Drosophila melanogaster*”. *PLoS Genetics* 8(11), p. 1003079.

## References

- Xu, S., Tyagi, S., and Schedl, P. (2014). “Spermatid Cyst Polarization in *Drosophila* Depends upon *apkc* and the CPEB Family Translational Regulator *orb2*”. *PLoS Genetics* 10(5).
- Yamaji, M., Tanaka, T., Shigeta, M., Chuma, S., Saga, Y., and Saitou, M. (2010). “Functional reconstruction of NANOS3 expression in the germ cell lineage by a novel transgenic reporter reveals distinct subcellular localizations of NANOS3”. *Reproduction* 139(2), pp. 381–393.
- Yamashita, Y. M., Jones, D. L., and Fuller, M. T. (2003). “Orientation of asymmetric stem cell division by the APC tumor suppressor and centrosome”. *Science* 301(5639), pp. 1547–1550.
- Yamashita, Y. M., Mahowald, A. P., Perlin, J. R., and Fuller, M. T. (2007). “Asymmetric inheritance of mother versus daughter centrosome in stem cell division.” *Science* 315(5811), pp. 518–21.
- Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2014). “The I-TASSER suite: Protein structure and function prediction”. *Nature Methods* 12(1), pp. 7–8.
- Yu, J., Lan, X., Chen, X., Yu, C., Xu, Y., Liu, Y., Xu, L., Fan, H.-Y., and Tong, C. (2016). “Protein synthesis and degradation are essential to regulate germline stem cell homeostasis in *Drosophila* testes”. *Development* 143(16), pp. 2930–45.
- Zaffran, S., Astier, M., Gratecos, D., and Sémériva, M. (1997). “The held out wings (how) *Drosophila* gene encodes a putative RNA-binding protein involved in the control of muscular and cardiac activity”. *Development* 124(10), pp. 2087–98.
- Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C. G., et al. (2018). “Ensembl 2018”. *Nucleic Acids Research* 46(D1), pp. D754–D761.
- Zhang, X. O., Dong, R., Zhang, Y., Zhang, J. L., Luo, Z., Zhang, J., Chen, L. L., and Yang, L. (2016). “Diverse alternative back-splicing and alternative splicing landscape of circular RNAs”. *Genome Research* 26(9), pp. 1277–1287.
- Zhang, Y. (2008). “I-TASSER server for protein 3D structure prediction”. *BMC Bioinformatics* 9(1), p. 40.
- Zhang, Z., Kostetskii, I., Tang, W., Haig-Ladewig, L., Sapiro, R., Wei, Z., Patel, A. M., Bennett, J., Gerton, G. L., Moss, S. B., et al. (2006). “Deficiency of SPAG16L Causes Male Infertility Associated with Impaired Sperm Motility”. *Biology of Reproduction* 74(4), pp. 751–759.
- Zhao, J., Ohsumi, T. K., Kung, J. T., Ogawa, Y., Grau, D. J., Sarma, K., Song, J. J., Kingston, R. E., Borowsky, M., and Lee, J. T. (2010a). “Genome-wide Identification of Polycomb-Associated RNAs by RIP-seq”. *Molecular Cell* 40(6), pp. 939–953.
- Zhao, L., Mandler, M. D., Yi, H., and Feng, Y. (2010b). “Quaking I controls a unique cytoplasmic pathway that regulates alternative splicing of myelin-associated glycoprotein”. *Proceedings of the National Academy of Sciences of the United States of America* 107(44), pp. 19061–19066.
- Zhou, Z., Shirakawa, T., Ohbo, K., Sada, A., Wu, Q., Hasegawa, K., Saba, R., and Saga, Y. (2015). “RNA Binding Protein Nanos2 Organizes Post-transcriptional Buffering System

- to Retain Primitive State of Mouse Spermatogonial Stem Cells”. *Developmental Cell* 34(1), pp. 96–107.
- Zhu, R., Heaney, J., Nadeau, J. H., Ali, S., and Matin, A. (2010). “Deficiency of splicing factor 1 suppresses the occurrence of testicular germ cell tumors”. *Cancer Research* 70(18), pp. 7264–7272.
- Zoller, R. and Schulz, C. (2012). “The *Drosophila* cyst stem cell lineage”. *Spermatogenesis* 2(3), pp. 145–157.
- Zuker, M. (2003). “Mfold web server for nucleic acid folding and hybridization prediction”. *Nucleic Acids Research* 31(13), pp. 3406–3415.