

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

# **Beyond BLASTing: Ribonucleoprotein evolution via structural prediction and ancestral sequence reconstruction**

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

Doctor of Philosophy in  
Genetics

at Massey University, Manawatū Campus

Toni K. Daly  
2016

## **Abstract**

Primary homology in DNA and protein sequence has long been used to infer a relationship between similar sequences. However gene sequence, and thus protein sequence, can change over time. In evolutionary biology that time can be millions of years and related sequences may become unrecognisable via primary homology. This is demonstrated most effectively in chapter 4a (figure 10). Conversely the number of possible folds that proteins can adopt is limited by the attractions between residues and therefore the number of possible folds is not infinite. This means that folds may arise via convergence between evolutionarily unrelated DNA sequences.

This thesis aims to look at a process to will wring more information from the primary protein sequence than is usually used and finds other factors that can support or refute the placement of a protein sequence within the family in question. Two quite different proteins; the Major Vault Protein whose monomers make up the enigmatic vault particle and the argonaute family of proteins (AGO and PIWI) that appear to have a major hand in quelling parasitic nucleic acid and control of endogenous gene expression, are used to demonstrate the flexibility of the workflow.

Principally the method relies on prediction of three-dimensional structure. This requires at least a partially solved crystal structure but once one exists this method should be suitable for any protein. Whole genome sequencing is now a routine practice but annotation of the resultant sequence lags behind for lack of skilled personnel. Automated pipeline data does a good job in annotating close homologs but more effort is needed for correct annotation of the exponentially growing data bank of uncharacterised (and wrongly characterised) proteins. Lastly, in deference to budding biologists the world over, I have tried to find free stable software that can be used on an ordinary personal computer and by a researcher with minimal computer literacy to help with this task.

## Acknowledgments

I have been a student since 1988 but now is the time to thank everyone and of course over the years there have been so many people that have inspired and encouraged me and I will miss some people that I shouldn't have.

Trevor Kitson made me realise that it was possible to be smart and funny, Mike Hardman for having the foresight to grab a box of tissues whenever I walked through his office door when I eventually became an internal student. Although Massey could bend the rules in those pre-studylink days a student loan required a minimum of seven papers.....but I only had one semester! So I physically couldn't attend all my classes, I just grabbed the notes, sat the exams and came back the following year to do the lab work. I want to thank Mark Patchett for just being the best tonic one could ever need when things looked glum, Rosie Bradshaw for her kindness, and Kathryn Stowell, a newly minted lecturer when I was an internal student, has encouraged me ever since.

Lesley Collins for allowing a complete unknown to contribute to her book, Austen Ganley from Albany for helping to prepare me for Vienna when Palmerston North was just too far away and lastly Andrew Sutherland-Smith and David Penny who have stoically listened to my troubles academic and personal once a week for six years. In retrospect I do not know how they coped and without them I surely would have quit. I also thank David for organising payment of my fees, and Massey for paying the bill.

It is an extraordinarily lonely thing to attempt a PhD without the camaraderie and the vicarious learning opportunities of watching fellow students give presentations etc. I want to especially thank Bruce White from the library for his help and patience and Tim White for helping me navigate computer-speak.

I set out to show the students of Northland New Zealand, that you can achieve your goals without the financial, geographical and educational advantages that you perceive everyone else to have. Protein annotation in particular needs help and people all over the world with access to a computer and an internet link can join in.

My biggest thanks though must go to my husband Dan Daly for his continual support despite landing him with three children as well as a full time job while I did my undergraduate stints, and I know that I have been a serious drain on finances and I know that we haven't been out in the kayaks (or out anywhere) for a long time and now I promise that our time will be spent together.



## Preface

This thesis is written according to the regulations of the latest version of the Handbook for Doctoral Study (2016), published online by the Doctoral Research Committee. This thesis complies with the format of a thesis based on publication as described in the handbook and includes both published and unpublished chapters. The chapters do not follow the order of publication in order to better demonstrate the flow of the development of the work. Chapters 1 and 2 have been written by Toni Daly as an introduction and literature review and are not intended for publication.

Chapter 1. General Introduction.

Chapter 2. Literature review of the Major Vault Protein.

Chapter 3a. Toni K Daly *et al.* (2013) Beyond BLASTing: Tertiary and quaternary structure analysis helps identify Major Vault Proteins. *Genome Biology and Evolution* 5: 217-232.

Chapter 3b. Toni K Daly *et al.* (2013) In silico resurrection of the Major Vault Protein suggests it is ancestral in modern eukaryotes. *Genome Biology and Evolution* 5 (8): 1567-1583.

Chapter 4. Toni Daly, X. Sylvia Chen and David Penny (2011) How old are RNA networks? (L J Collins ed. *RNA Infrastructure and Networks*) Landes BioScience and Springer Science.

Chapter 4a. Toni Daly, *et al.* (2016) Long Long AGO: The evolutionary history of Argonaute and PIWI in metazoa by ancestral protein inference and structure prediction. (submitted).

Chapter 4b. Toni K Daly *et al.* (2016) Argonaute gain and loss during fungal evolution. (submitted).

Chapter 4c. Toni Daly *et al.* Argonautes origins in eukaryotes. (in preparation).

Chapter 6. Conclusion.

David Penny is co-author on all of the published and prepared papers and Andrew Sutherland-Smith is co-author on five of them. Contributions to each paper are described in Appendix IV.

# Contents

## Chapter one: Introduction

<b>1. Overview .....</b>	<b>1</b>
1.1. Ribonucleoproteins .....	1
1.2. Protein Evolution .....	2
1.2.1. Evolutionary Aspects of the Chosen Proteins .....	3
1.3. Pipeline.....	4
1.3.1. Basic Local Alignment Search Tool (BLAST) .....	5
1.3.2. Protein Annotation and Prediction .....	6
1.3.3. Tree Calculations .....	7
1.3.4. Ancestral Sequence Reconstruction (ASR) .....	8
1.4. Major Vault Protein (MVP) .....	8
1.4.1. MVP Form and Function .....	9
1.5. Vault Function.....	11
1.5.1. Cellular Location .....	12
1.5.2. Vault Cargo.....	13
1.5.3. Developmental / Scavenging Roles .....	13
1.5.4. Association with lipid rafts .....	14
1.5.5. Detoxification roles .....	15
1.5.6. Multi Drug Resistance (MDR) .....	15
1.5.7. Cell signalling.....	17
1.5.8. Possible future biotechnological use of the vault particle .....	18
1.6. VTRNA .....	19
1.6.1. Vault RNA nomenclature .....	20
1.6.2. VTRNA function .....	22
1.7. Summary .....	24
<b>2. a: Beyond BLASTing .....</b>	<b>26</b>
2.1. Sequence similarity identifiers .....	27
Abstract.....	29
Introduction.....	29
Materials and Methods.....	31
Results.....	34
Discussion.....	41
<b>2. b: <i>In silico</i> Resurrection.....</b>	<b>45</b>
Abstract.....	45

Introduction.....	45
Materials and Methods.....	57
Results.....	50
Discussion.....	56
<b>Chapter Three: Introduction to the Argonaute Family</b>	
<b>3. The defence of the Dark Arts .....</b>	<b>62</b>
<b>How Old Are the RNA Networks?.....</b>	<b>67</b>
Abstract.....	67
Introduction.....	67
Regulatory networks of small RNAs.....	68
RNA regulation and defence against the Dark Arts.....	71
Other regulatory RNAs.....	78
How old are the different interactions of RNA?.....	79
Conclusion.....	81
<b>Chapter Four: The Evolution of the Argonautes</b>	
<b>4. An investigation into Argonaute evolution using 3-D structural prediction ....</b>	<b>83</b>
4.1. Abstract .....	87
4.2. Introduction .....	87
4.2.1. Argonaute proteins.....	88
4.2.2. <i>In silico</i> analysis .....	92
4.3. Methods .....	93
4.4. Results .....	96
4.5. Ancestral reconstruction.....	103
4.6. Evolution .....	108
4.7. Discussion .....	111
4.7.1. Annotation issues.....	111
4.7.2. General.....	112
<b>4. b: Argonaute gain and loss during fungal evolution.....</b>	<b>116</b>
4b.1 Abstract.....	116
4b.2 Introduction.....	116
4b.3 Method.....	121
4b.4 Results.....	123
4b.4.1 Yeast and fungi.....	123
4b.4.2 The <i>R. irregularis</i> AGO expansion.....	127
4b.4.3 Microsporidia.....	134

4b.5 Discussion.....	139
<b>4. c: Argonautes in eukaryotes.....</b>	<b>142</b>
4c.1 Abstract.....	142
4c.2 Introduction.....	142
4c.3 Method.....	146
4c.4 Results.....	149
4c.4.1 SAR (Stramenopile, Alveolate and Rhizaria).....	151
4c.4.2 Amoebozoa.....	156
4c.4.3 Excavates.....	157
4c.4.4 Red and green algae.....	160
4c.4.5 Land plants.....	161
4c.5 Ancestral trees.....	163
4c.6 Discussion.....	167
<b>Chapter Five: Conclusion</b>	
<b>5. Conclusion.....</b>	<b>170</b>
5.1. The chosen proteins.....	170
5.2. Links with the past: Major Vault Protein.....	171
5.2.1. Bacteria.....	172
5.2.2. Archaea.....	174
5.3. Links with the past: Argonaute Family Proteins.....	176
5.4. Challenges of the method.....	178
5.4.1. BLASTp.....	178
5.4.2. MSA.....	179
5.4.3. Structural prediction .....	179
5.4.4. RosettaDock (ROSIE) .....	180
5.4.5. Tree calculation .....	181
5.4.6. Ancestral Sequence Reconstruction (ASR).....	181
5.4.7. FATCAT.....	182
5.4.8. Philanthropy.....	183
5.5. The last word.....	184
Glossary.....	185
References.....	189

## List of Figures

### Chapter one:

Fig. 1.1 Pipeline evolution .....	4
Fig. 1.2 Vault ribonucleoprotein structure .....	10
Fig. 1.2 Refinement of the vault structure (2013). .....	11

### Chapter Two (published papers)

#### Preface

Fig. 2.1 Geneious alignment shading .....	28
---	----

#### Chapter 2a:

Fig. 1 Vault ribonucleoprotein structure.....	30
Fig. 2 MVP monomer comparison.....	32
Fig. 3 Structural effect of the 2ZUO*b constraint.....	35
Fig. 4 RosettaDock results from the crystal structure cap-helix.....	36
Fig. 5 RosettaDock results from the rat MVP shoulder region.....	37
Fig. 6 I-TASSER modelling results for the negative control sequences.....	38
Fig. 7 <i>Naegleria gruberi</i> MVP I-TASSER modelling.....	40

#### Chapter 2b:

Fig. 1 Problems with FastML and PAML.....	48
Fig. 2 Vault ribonucleoprotein structure.....	50
Fig. 3 MrBayes tree of unlikely placements.....	51
Fig. 4 Heatmap showing identity between the bacterial homolog pairs.....	53
Fig. 5 How inserts can affect the I-TASSER score.....	54
Fig. 6 A comparison of MVP structures with the stramenopile ancestor.....	55
Fig. 7 Heatmap of ancestors.....	56
Fig. 8 Structural diagrams of I-TASSER predictions.....	57
Fig. 9 Consurf diagram showing conserved and non-conserved residues from multiple sequences.....	57

#### Chapter Three:

Fig. 1 Transcription of endogenous DNA that gives rise to dsRNA.....	70
Fig. 2 A comparison of RNAi networks involved with the defence of the Dark Arts...	71
Fig. 3 Gemini viruses.....	72
Fig. 4 The CRISPR system.....	73
Fig. 5 Working backwards through four stages of the origin of life.....	76

## Chapter Four (submitted papers)

### Preface

Fig. 4.1 Workflow for chapter four. ....	84
--	----

### Chapter 4a:

Fig. 4a.1 The human AGO2 crystal structure PDB:4W5N.....	91
Fig. 4a.2 Similarities and differences in predicted structure between the difficult-to-resolve flatworm sequences. ....	98
Fig. 4a.3 The human AGO2 crystal structure aligned with the low scoring predicted structures identified in table 1. ....	101
Fig. 4a.4 The difference between alignment and predicted structure in <i>X. tropicalis</i> PIWI3. ....	103
Fig. 4a.5. The C terminal signature in PIWI-like ASR. ....	104
Fig. 4a.6 Metazoan predicted structure for PIWI ancestors.....	105
Fig. 4a.7. The C terminal signature in AGO-like ASR. ....	107
Fig. 4a.8 Metazoan predicted structure for AGO ancestors.....	108
Fig. 4a.9 Predicted structure for the putative sequences identified by BLAST in <i>S. rosetta</i> . ....	110
Fig. 4a.10 An example where BLAST searches are ambiguous.....	113
Fig. 4b.1 Solved structures of the argonaute family .....	119
Fig. 4b.2 A simplified Unikont tree showing the proposed relationship between the various phyla that contributed to the work.....	120
Fig. 4b.3 <i>A. gossypii</i> (UniProtKB:M9MXJ8) putative AGO sequence identified by BLAST. ....	124
Fig. 4b.4. Structural predictions of the reconstructed ancestors from the tree of representative fungi species. ....	126
Fig. 4b.5 Fates of the <i>R. irregularis</i> expansion.....	130
Fig. 4b.6 A comparison of UniProtKB:U9SQW1 with the solved structure of the human argonaute. ....	131
Fig. 4b.7. A comparison of the microsporidian ancestor and <i>M. daphniae</i> with solved structures. ....	135
Fig. 4b.8. An unrooted tree of fungi and metazoan sequences re-created by ASR.....	137
Fig. 4c.1 I-TASSER structural prediction for <i>Trypanosoma brucei</i> . ....	145
Fig. 4c.2 I-TASSER 3-D structural prediction of BLAST results with high number of residue changes per site.....	150
Fig. 4c.3 The divergent argonaute Twi12 from <i>T. thermophila</i> . ....	154
Fig. 4c.4 A comparison between the canonical argonaute and PIWI-tryp within <i>Trypanosomes</i> . ....	158
Fig. 4c.5 Mr Bayes tree of annotated argonaute proteins in trypanosomid protozoans. ....	159

Fig. 4c.6 A rooted tree from the calculated ancestral sequences. ....	165
Fig. 4c.7 Tree of all eukaryote ancestral reconstructions.....	166
Fig. 5.1 <i>Escherichia coli</i> TolA (UniProtKB:P19934).....	172
Fig. 5.2 Bacterial MVP monomer. ....	173
Fig. 5.3 Full size bacterial MVP monomers.....	174
Fig. 5.4 Putative archaea homolog sequences.....	175
Fig. 5.5 The highly conserved C terminal from MVP.....	176
Fig. 5.6 I-TASSER and Phyre2 comparison. ....	180
Fig. 5.7 An I-TASSER comparison between raw and trimmed ASR node 1 sequences. .....	182
Fig. 5.8 FATCAT structural alignment of the truncated ancestor with HsAGO2. ....	182
 Appendix I.....	 204
Appendix II.....	209
Appendix III.....	213
Permission and contributions.....	217





# Chapter One: Introduction

## 1. Overview

In this thesis I have studied the predicted tertiary structure of proteins that provide support to putative evolutionary relationships between primary sequence in protein families. There is increasingly reliable evidence for the effectiveness of structural prediction but the main argument within this thesis is that biologists should not rely on one form of evidence alone. Although for proteins with high primary homology especially within closely related species this is a fair assumption. However there are many instances where this breaks down, e.g. where sequence similarity is lost, or where convergence is a possibility, there needs to be more than one line of evidence to support the assumption of relatedness.

Different protein families require different methods of investigation and the focus has been on computational methods that are ‘readily available’, ‘stable’, ‘simple’ and ‘free’. The traditional thesis chapters (chapters 1 and 5), published papers (chapters 2a, 2b, and 3), and submitted papers (chapters 4a, 4b and (in preparation) 4c); represent the journey to find suitable computational methods for different types of protein and to learn something about their evolutionary history.

### 1.1. Ribonucleoproteins

Ribonucleoproteins (RNPs) are associated with important biological roles, e.g. the spliceosome found in all crown group eukaryotes (Collins and Penny, 2005), as well as ribosomes with catalytic capabilities (Kruger *et al.*, 1982) that are conserved throughout all domains. RNPs are of particular interest in terms of evolutionary biology because in many cases the RNA component may date to an ‘RNA world’ (Gilbert, 1986). A scenario for the evolution of RNPs from this RNA world would be the occurrence of short peptides that could chelate metal ions and possibly provide stability for the RNAs. The RNA would then gradually associate with peptides or small proteins in the form of chaperonins. These could act as a scaffold and protect the RNA from degradation. The

longer the RNA could last the more efficiently a task could be accomplished. Efficiency breeds success, and so RNA could become more complex and proteins indispensable.

This thesis principally investigates two protein families; the Major Vault Protein (MVP described in chapters 1 and 2), and the Argonautes which fall into two broad categories; argonaute-like (AGO-L) and PIWI-like (PIWI-L) (described in chapters 3 and 4). What links these proteins is that they both interact with small RNAs, and in fact there is evidence that the RNA found in the vault particle can be processed into small RNAs utilised by the argonaute (described in chapter 1). The question is this: Do these proteins date back (at least) to the Last Eukaryotic Common Ancestor (LECA)? Additionally, can we say that their associated RNAs are ancestral or a recent recruitment?

## **1.2. Protein Evolution**

Once a functional protein is established it changes by both random drift and natural selection. Frequently point mutations either do not change the resultant amino acid because of redundancy in the code (synonymous) or the substituted residue has similar properties to the one displaced. Most neutral (or slightly deleterious) changes simply lead to polymorphisms in the population. Even where the substituted residue has very different properties it matters in some positions more than others. In each of the proteins that I have looked at there are regions that are more evolutionarily constrained than others for different reasons.

Advantageous changes can become fixed under positive selection and deleterious ones eliminated. An advantageous change in the genome, which is subject to positive selection, will spread faster in the population. Without any selective pressure we should expect that mutations would occur randomly throughout the sequence. We do not in fact see this because a lethal mutation leaves no legacy. Some parts of a protein sequence are under positive selection and some are not (Daly *et al.*, 2013b).

With increasing genetic distance the number of nucleotide substitutions increases and so the number of true substitutions might be underestimated (substitution saturation). Since this thesis considers large evolutionary distances I have concentrated on protein sequence, rather than at the nucleotide level, because amino acid conservation corresponds to regions of structural or functional constraint. In some

instances I examined nucleotide sequence and intron placement to assist with sequencing anomalies rather than evolutionary relatedness.

High levels of primary homology are usually a good indicator of evolutionary relatedness, however multiple point mutations and insertions or deletions, that change the primary sequence of the protein beyond the point where sequence similarity is still detectable, may retain homologous folding patterns (Murzin *et al.*, 1995; Orengo *et al.*, 1997). It is known that structure, or ‘fold’ can persist where primary homology is lost (Illergård *et al.*, 2009). The fate of the protein is dependent on the way these evolutionary changes affect the function of the protein. In some cases very large inserts clearly do not affect the function of the protein because they can be found in well-characterised functional proteins, described for example, in chapter four.

Structural homology analysis has been used to identify members of protein super-families with low sequence homology (Holm and Sander, 1997) and folding studies can also be used to predict function (Watson *et al.*, 2005). There are only a limited number of folds that are adopted (Chothia, 1992) and because of this it is possible that structural similarity between two proteins could have arisen by convergence (Fernandez-Fuentes *et al.*, 2010).

Structural constraints arise from folding, packing or interactions with other proteins, nucleic acid, metal ions, etc. Functional constraints are based on catalytic properties as well as interactions even with other molecules. As the proteins studied could have arisen by convergence, evidence in addition to sequence and structure is included. It is for this reason that oligomerisation capability in the vault monomers and conserved catalytic and C terminal residues in the argonautes were added to the pipeline because only by using the sum of available evidence can relationships be more certain. Constraints in terms of amino acid substitution for both structural and functional reasons were found in the RNPs selected.

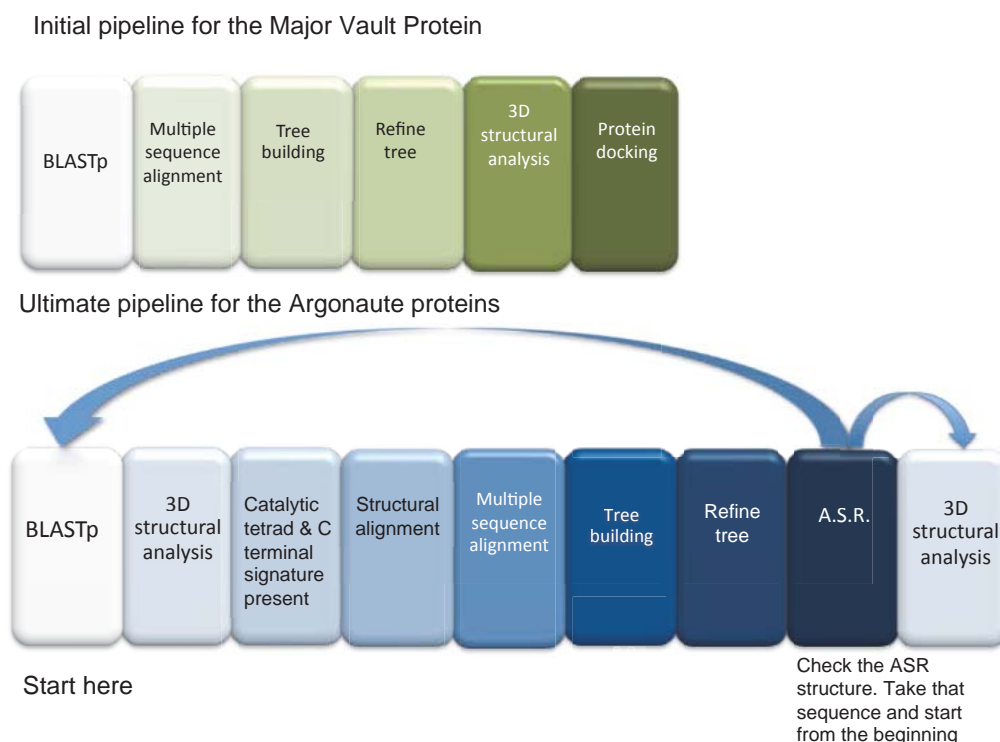
### **1.2.1. Evolutionary Aspects of the Chosen Proteins**

The three ribonucleoproteins are of similar size (~860 residues long) and MVP and AGO-L proteins have full-length solved crystal structures. The reference crystal structures are rat liver major vault protein at 3.5 Å resolution (PDB:2ZUO) (Tanaka *et al.*, 2009), (now superseded by PDB:4HL8 (Casañas *et al.*, 2013)), the human argonaute2 (PDB:4OLA 2.3 Å), and human argonaute2 bound to a defined guide RNA (PDB:4W5O 1.8Å) (Schirle and MacRae, 2012; Schirle *et al.*, 2014). There are no full-

length PIWI solved structures but a number of domains e.g. PDB:4P1Z (the MID domain from MIWI the mouse PIWI 2.3 Å) (Cora *et al.*, 2014) and PDB:3O7V (the PAZ domain human PIWI1 2.1 Å) (Tian *et al.*, 2011b).

### 1.3. Pipeline

The criteria for computer-aided analysis of ‘readily available’, ‘stable’, ‘simple’ and ‘free’ means that an average biologist can utilise the world of computational biology without computing expertise or access to sophisticated equipment. A succession of servers that fulfil this criteria were trialled (described elsewhere) and the pipeline developed (fig. 1.1) for the extraction of more data than is usual from primary protein sequence. This scheme also allows for improved services, new servers and changes of strategy depending on the type of protein selected. The method requires a minimum of a solved homologous crystal structure for part of the protein in order for the structural prediction algorithms (I-TASSER (iterative threading assembly refinement server) (Yang *et al.*, 2015) and Phyre2 (Kelley *et al.*, 2015) described later) to be useful.



**Fig. 1.1 Pipeline evolution**

The first three blocks in the green pipeline shows the method that would be chosen for selecting sequences in order to reconstruct a family tree. Structural prediction and protein docking was added to assess the likelihood that the MVP monomers would align and dock as they do in nature. In chapter 2b ancestral sequence reconstruction (ASR) was also added in order to be able to use a putative ancestral sequence as a seed sequence to search for more remote sequences (Collins *et al.*, 2003a). For the argonaute proteins the pipeline was adapted as docking is not a requirement for the argonaute proteins but the analysis of other identifiers such as the catalytic and C terminal residues was included (described in chapter 4a). The point is that the pipeline can be flexible depending on the protein being studied.

*In silico* research cannot replace the expensive and time-consuming work done by crystallographers. The Protein Data Bank currently holds 37,523 entries, 8,229 of which also contain nucleic acid (January 2016), this indicates that many proteins are amenable to research involving structural analysis. The following section details the basic steps in the pipeline; BLAST, protein prediction, tree construction and ancestral reconstruction and how each relates to the overall goal. A glossary is also included that describes the databases mined for source data plus the servers and algorithms for the steps in general use. Some issues that have arisen have been described in the published work and are also referred to in the discussion section in chapter five.

### **1.3.1. Basic Local Alignment Search Tool (BLAST)**

Although the title of this thesis is ‘Beyond BLASTing’, BLAST (Altschul, 1990) searches are an essential and simple method of mining for sequences (protein or DNA). The search algorithm compares primary biological sequence information from sequence databases using a scoring matrix known as ‘BLOSUM, (BLOcks of amino acid SUBstitution Matrix) (Henikoff and Henikoff, 1992). The sequences studied came principally from UniProtKB (Consortium, 2015) or the National Centre for Biotechnology Information (NCBI) and both enable the user to select a scoring matrix; BLOSUM90 where sequence homology is high and a stringent scoring matrix can be applied through to BLOSUM45 which will be more helpful searching for homologs where the relationship is expected to be weak. BLOSUM62 is the default matrix for protein BLAST. This means that all sequences with more than 62% similarity are merged into one sequence and that will be compared to sequences with less homology. The idea is that there will be a lesser contribution from highly homologous sequences. The default BLOSUM62 or even BLOSUM80 worked well for MVP but even BLOSUM45 will have missed many argonauts because of the very large number of similar sequences.

Pairwise identity scored by the chosen matrix finds the best matching pair of global alignments of two sequences. The databases BLAST tools used in this study use the ‘word’ method (k-tuple method) where short, non-overlapping sequences, known as ‘words’ in the query sequence are matched to sequences in the database. The advantage of this is the speed with which it can be done, particularly with highly conserved sequences. The segment pairs whose scores cannot be improved by extension or trimming are called high-scoring segment pairs or HSPs. The score ‘S’ is the sum of the

matrix score (described above) and gap penalty scores where there are gaps between the sequences in the database and the query sequence.

The BLAST algorithm results give an ‘Expect’ or E value as well as the bit score which indicates the number of alignments between a result (hit) and the seed sequence (or merged sequence from the matrix) with scores equivalent to or better than S that are expected to occur in a database search by chance. An E value of  $1 \times 10^{-4}$  (or less) is considered homologous, although even low E values cannot always be trusted because they can be artificially low if the database is very small. The E value is generated according to the equation:

$$E = Kmn e^{-\lambda S}$$

Where m = the size of the database, n = the length of the query sequence, S = the sum of the matrix score and gap penalty scores. K and lambda ( $\lambda$ ) are parameters related to the maximum value position and to the width of the distribution and depend on the scoring matrices.  $\lambda$  describes how steep is the decay from a high frequency of similar sequences with low S scores and a low frequency of sequences with high S score.

The E value then, is the number of random hits that can be expected to have the score S. It can be seen that the E-value will increase as the database gets bigger and decrease as the database gets smaller.

S is generally normalised:

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

This gives the ‘bit score’ S’ or  $E = mn 2^{-S'}$  (both are resulted). E is used as a cut off in the MVP study but neither score is considered for the argonautes (see chapter five for discussion).

### 1.3.2. Protein Annotation and Prediction

The Genomes OnLine Database (JGI Gold) (Reddy *et al.*, 2014) lists almost 74,000 genome projects as of January 2016. In order to annotate the resultant gene sequences computer aided assignment of protein function has become indispensable. While computational methods offer a fast alternative to expensive experimental studies, automated preliminary pipeline data cannot always be trusted to be correct and the number of ‘uncharacterised’ proteins is growing.

There are many excellent servers and algorithms that deal with sequenced data that work at a genome wide scale to predict protein function. These approaches are well suited for rapid processing of the large amount of uncharacterised sequences, but they are also limited in their specific description of protein function. For instance a similar (or even identical) sequence does not always result in the same structure or function. Most commonly known would be the aberrant folding of the prion protein (Prusiner *et al.*, 1998) but there are also protein homologs with different functions (Kosloff and Kolodny, 2008) that would be grouped in the same family by an automated pipeline. So protein function can evolve separately from sequence or, rapid evolution of both sequence and function could follow once a protein is released from functional constraint by duplication (Ortiz-Rivas *et al.*, 2012).

Pivotal to the body of work are the algorithms that predict structure from primary sequence data. I have mainly used I-TASSER (Yang *et al.*, 2015) and Phyre2 (Kelley *et al.*, 2015). I-TASSER additionally suggests function from sequence and structure via a recent addition of COACH (a function annotation program). These servers are described throughout the work and are listed in the glossary. Structural prediction cannot replace experimental structural determination but are means available to more researchers and fulfil the criteria of readily available, stable, simple to use, and free. Additional evidence to support the determination of relatedness must also be considered.

### **1.3.3. Tree Calculations**

The initial trees are phylograms, in that the branch lengths reflect the number of residue changes (noted on the branches), but only if the trees are rooted by an outgroup. The confidence with which the branches are determined (bootstrap values) are additionally marked as a percentage. Of course relating species by one protein would be foolish, however the trees do generally make ‘sense’ from the view of what would be reasonably anticipated e.g. by comparison to a tree calculated from 18S rRNA. Where a placement has not ‘made sense’ it can usually be attributed to contamination as described in chapters 2b and 4a.

This led to the decision whether to use rooted or un-rooted trees. Choosing an outgroup as a root for the tree made changes to the apparent relatedness of species. Mostly these were minor but chapter 4a describes that in some instances identical MSA produces trees that change radically due to the root. This problem is dealt with on an individual basis. To give a general indication of relationships between sequences and to



provide extra data particularly in terms of numbers of residue changes per site, the initial trees are rooted by an outgroup. However so that only the sequences that truly belong to a particular group or ‘clade’ are included in the calculations for ancestral sequence reconstruction (ASR) the trees need to be unrooted.

#### **1.3.4. Ancestral Sequence Reconstruction (ASR)**

One of the limitations of BLAST is that the search becomes swamped with known sequences and remotely related sequences are harder to find simply because they are outside the maximum number of hits that can be acquired in a single search. This particularly affects the argonaute study and is a fast growing issue due the increasing rate that new sequences are being deposited. The management of large trees and disparate sequences is improved by building ancestors that represent ‘batches’ of protein sequences, whether by species (as in the MVP monomer), or across species (as in the argonautes).

For this reason I use sequences that I am certain are *bona fide* members of a common group or ‘clade’ of sequences (regardless of species) to calculate ancestral sequences via free servers (PAML4, FastML and Mega5). These are therefore sequences that do not appear in any sequence databases and are more remote than the known homologs. These can be used as queries to find sequences that fall outside of the thousand hit maximum. This has already been used successfully to find an RNaseP homolog (Collins *et al.*, 2003b) and was trialled in the search for MVP homologs and proved essential for finding argonautes. The ancestral sequences also represent a much larger number of sequences thus reducing the complexity of the resultant trees.

### **1.4. Major Vault Protein (MVP)**

MVP represented an ideal case for the purposes of demonstrating the utility of three-dimensional studies as a means of enhancing the search for functional sequence homologs. The oligomeric vault structure is very large and so each monomer represents a genuine challenge in terms of structural prediction. The monomers are also capable of independent self-assembly (Stephen *et al.*, 2001) which means that a further check is to see if the monomers will also dock with another in virtual space. Furthermore the monomeric MVP tertiary structure (and hence the sequence) is presumably under strong selective pressure to retain a conformation that forms not only the appropriate monomer structure, but also, the appropriate interface interactions with its neighbours for vault



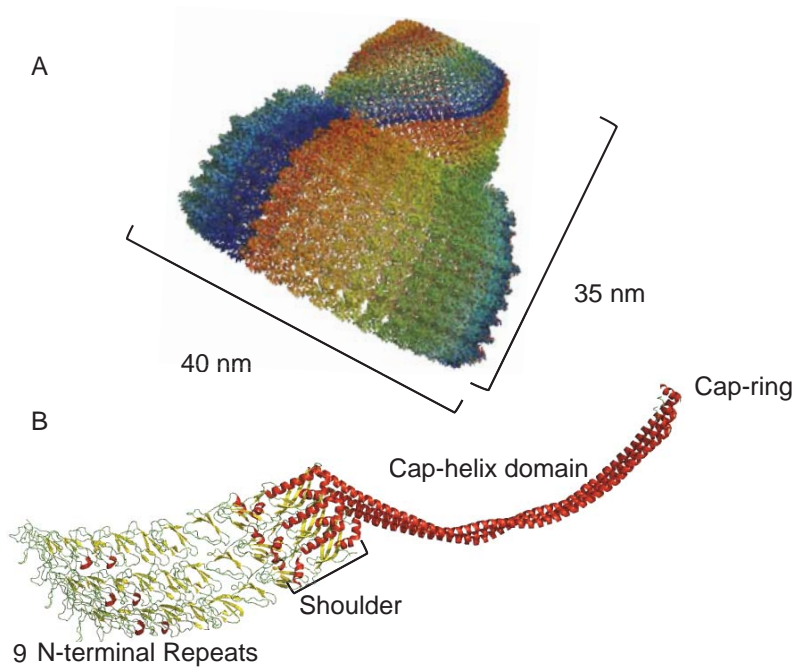
quaternary structure assembly (Qian *et al.*, 2011). The remaining residues can be substituted many times. This is shown most convincingly in fig 9 of chapter 2b.

In most cases of MVP the sequences are orthologs, i.e. they are sequences that are related by a speciation event and so had a common origin. An exception are the MVPs found in the mycetozoa *Dictyostelium* and *Polysphondylium* where MVP monomers appear to have duplicated and vault particles are chimeric made from  $\alpha$  and  $\beta$  MVP. In fact the *D. discoideum* has three version of the MVP gene, the third copy, annotated as an  $\alpha/\beta$  hydrolase (which has some similarity with the C terminal of MVP and has been described as MVPc) can make ovoid vaults in MVP $\alpha^{-/-}$  / MVP $\beta^{-/-}$  cells. These vaults appear similar to those made by either MVP $\alpha$  or MVP $\beta$  alone (Vasu and Rome, 1995). There is no evidence of other species requiring two versions to make the vault.

Vault particles appear to be associated with many pathways but essential to none (yet identified), any underlying basic function has remained elusive (described later in this chapter).

#### **1.4.1. MVP Form and Function**

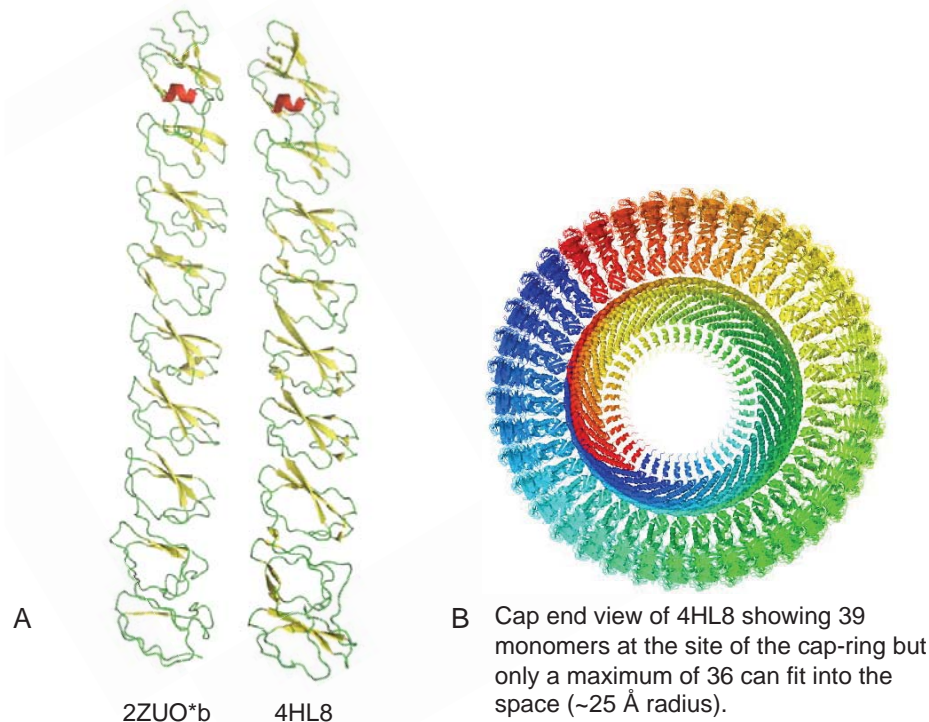
Vault function is not described elsewhere in this thesis and so a summary is provided here. Vaults are ribonucleoprotein particles with a hollow structure, likened to a barrel with cathedral style end caps (see fig. 1.2A for the structure of half a vault (Daly *et al.*, 2013a)). Each half will form spontaneously from 39 Major Vault Protein (MVP) monomers (in rat). Vaults also contain Telomerase Associated Protein 1 (TEP1) (Kickhoefer *et al.*, 1999b) and Vault Associated Poly ADP-Ribosylating Protein (VPARP) (Kickhoefer *et al.*, 1999a), as well as a number of copies of vault RNA (VTRNA) (Kickhoefer *et al.*, 1993; Stadler *et al.*, 2009b), these are not normally essential for vault formation (Stephen *et al.*, 2001), and additionally are found outside of vaults and have other vault independent functions as well (Persson *et al.*, 2009; Amort *et al.*, 2015; Helbo *et al.*, 2015).



**Fig. 1.2 Vault ribonucleoprotein structure**

**A.** Rat MVP quaternary structure showing half a vault coloured by monomer (PDB:2ZUO, 2ZV4 and 2ZV5 now superseded by PDB:4V60). A full vault will have at the lower left, a two fold symmetric second half of the upper half vault. **B.** Three rat MVP monomers coloured by secondary structure. This figure highlights the extensive lateral association required to dock into the vault quaternary structure. Missing from the crystal structure is the very highly conserved C terminal which forms the cap-ring.

Since the article that forms chapter 2 was published there has been a refinement to the vault structure (Casañas *et al.*, 2013). Most of the structure remains the same as the 3.5 Å structure of 2009 (Tanaka *et al.*, 2009) but repeats 1 and 2 of the N terminal of the monomers are observed to have an alternate structure thought to reflect the opening of the vault at the waist where the two halves join (fig. 1.3A). Additionally the later refinement reveals that not all the 39 monomers are physically capable of fitting in the cap-ring space and it is predicted that one in every thirteen subunits (MVP monomer) has a conformation that sits outside of the vault (fig. 1.3 B). Engineered structures with C terminal extensions have shown that tags added to the C terminal are found outside of the vault (Kickhoefer *et al.*, 2009).



**Fig. 1.2 Refinement of the vault structure (2013).**

**A.** Comparison of the N terminal repeats of the original 3.5 Å model (2ZUO\*b i.e. the b strand of the particle) with the same region of PDB:4HL8 which has some extra  $\beta$  sheet in R1 and R2 but otherwise the difference is minimal. **B.** A half vault viewed from the C terminal where the exact structure of the cap-ring remains unsolved but it is believed that all 39 monomers will not fit into the ring.

Vaults particles are widespread amongst eukaryotes and can be directly observed by negative silver stain and cryo-electronmicroscopy or be detected by immuno-blotting with anti-MVP antibodies in diverse species such as sea urchins (Hamill and Suprenant, 1997), cellular slime mould (Vasu *et al.*, 1993), electric ray (Herrmann *et al.*, 1997), and mammals (Kedersha and Rome, 1986). This research finds structural homologs in other eukaryote kingdoms and also in bacteria though they do appear absent from insects (though they have recently been sequenced from chelicerae and myriapoda so the vault particle is found for the ecdysozoa generally) and most likely lost in plants.

## 1.5. Vault Function

In 1999, Kong *et al.*, wrote

“We suspect that the vault has one underlying basic function that may lead to different functional phenotypes in various specialised cells” (Kong *et al.*, 1999).

In contrast to this, as this review chapter will show, vaults do not appear to have an underlying essential task that befits their magnificent and highly conserved structure.

The MVP promoter contains binding sites for a number of transcription factors; STAT, p53, Sp1, E-box, GATA, MyoD and Y-box (Lange *et al.*, 2000). This implies that the vault is involved in many pathways as a particle since the vault self assembles as it is translated via polyribosome templating (Mrazek *et al.*, 2014) and MVP is almost all found in the multimeric form. However, tagged monomer produces a less stable vault particle that is also insoluble. Additionally, free tagged monomer is indistinguishable from whole tagged vault particles if they break apart. The problem can be solved by using pFastBac a dual expression system for the baculovirus that makes tagged and wild type monomer at the same time and results in chimeric vaults with an average of 6-8 copies of the tagged monomer per vault particle. Despite difficulties in making stable tagged vaults they have been used for targeted delivery of cargo within cells (Kickhoefer *et al.*, 2009).

### **1.5.1. Cellular Location**

Vault particles have been shown to move along axons in response to trauma, *e.g.* crush injury (Li *et al.*, 1999), the movement seems too fast and the particles too large to move via diffusion (Luby-Phelps, 2000) and they have been found attached to microtubules (Eichenmüller *et al.*, 2003).

It has been suggested that vaults could be an integral component of the nuclear pore complex (Dickenson *et al.*, 2007), but only 5% of vaults are found in the nucleus. 27% are found to be ‘pore associates’, *i.e.* found within 200 nm of the centre of a nuclear pore on the cytoplasmic rim, and 12% on the nucleoplasmic rim (Paspalas *et al.*, 2009). Nuclear pore plugs are similar in size and shape to vault particle, both (Chugani *et al.*, 1993), and (Suprenant, 2002), have suggested that they could displace them.

Pores will form when the pore-free *Xenopus* egg extract membrane is incubated with either His-tagged MVP or purified MVP, so it is suggested that vaults are involved in the assembly of nuclear pore complexes (Vollmar *et al.*, 2009). There are tryptophan residues at the vault surface that could be involved in membrane binding and bending, but a precise mechanism has yet to be described (Anderson *et al.*, 2007).

Some MVP monomer is also found in the nucleus, and are especially enriched in the nucleus of sea urchins (Stewart *et al.*, 2005) where the distribution of cellular MVP in zygotes is mostly cytoplasmic but becomes predominantly nuclear in adults. This transition occurs following embryogenesis as the sea urchin moves from the

mesenchyme blastular stage to the larval stage, and appears to be specific to sea urchins (Hamill and Suprenant, 1997).

If vaults are unlikely to move by diffusion due to their size (Luby-Phelps, 2000), then how can they reach the nucleus? We could ascribe a targeting capacity to the vault contents or to VTRNA, but Stewart *et al.*, (2005) suggest that sumoylation (Small Ubiquitin-like Modifier or SUMO sites (residues 308-311), very highly conserved among almost all species and residues 707-710 less well conserved), could be involved in nuclear localisation because RanGAP1 in the nuclear pore complex is a sumoylation substrate, modification of this protein leads to its trafficking from the cytosol to the nuclear pore complex (Matunis *et al.*, 1996).

### **1.5.2. Vault Cargo**

The structure of the vault lends itself to carrying cargo, and all manner of cargo has been suggested. The debate revolves around how the cargo can enter the vault. It is certain that it does, because vaults containing cargo have been micrographed and determined by western blot, (Stewart *et al.*, 2005). An SDS-Page comparison of sea urchin and rat vaults purified using an identical method, shows that either sea urchin vaults are considerably more complex than vault particles from the rat, or have complex cargo (Stewart *et al.*, 2005). Paspalas *et al.*, (2009) identified mRNA as possible cargo and observed that the cargo itself could target the vault particle to specific destinations.

VPARP has been shown to enter vaults that don't already have VPARP (Poderycki *et al.*, 2006) and VPARP has a region known as 'INT' that will bind to a protein to guide it into the vault lumen. That the vault must open was shown by anionic polymers (a semi-conducting fluorescent polymer), that would enter vaults that had the capacity to open but not enter a particle with cross-linked monomers unable to open (Ng *et al.*, 2008).

### **1.5.3. Developmental / Scavenging Roles**

Vaults are also enriched in macrophages and amoeboid and ramified (resting) microglia (macrophage-like cells of the brain and spinal cord) (Chugani *et al.*, 1991). The commonality between macrophages and amoeboid microglia is their scavenging ability and plasticity. Using rat brains Chugani *et al.*, (1991) observed the migration of microglia and, serendipitously found that these were enriched with vaults. During embryonic development ramified microglia cross into the rat brain from blood vessels

from embryonic day 15 and subside up to 14 days postnatally. Amoeboid microglia flood into the brain during the first week after birth, a stage where development leaves cellular debris to be cleaned up by phagocytosis. Microglia cannot easily be replaced on a regular basis like macrophages, so amoeboid microglia differentiate into ramified microglia during the two weeks after birth. These constitute a silent pool of microglia able to detect infection via long branching processes and can be activated in times of need.

Amoeboid and ramified microglia remained enriched for vaults above background brain cell levels in the weeks following birth, and although amoeboid microglia become less numerous compared to ramified microglia, their vault immunoreactivity remains higher. MVP also promotes apoptosis in macrophages via SR-A-mediated TNF- $\alpha$  synthesis (Ben *et al.*, 2013) and in knockout mouse peritoneal macrophages TNF- $\alpha$  synthesis is severely reduced, i.e. affecting the immune system. It must be noted that the murine MVP promoter region lacks some elements so may not be useful as a human model (Mossink *et al.*, 2002b).

Vaults can be developmentally differentially expressed, but haven't been found to take any defined role in any kind of degradation mechanism by the microglia. Stewart *et al.*, (2005) suggest that macromolecules are being imported to the nucleus during development, but again, there is no evidence that vaults have a direct role in nuclear trafficking.

#### **1.5.4. Association with lipid rafts**

MVP is highly expressed in human lung and intestinal epithelial cells. Lung epithelial cells expressing either wild type or mutant ( $\Delta$ F508) CF transmembrane conductance regulator (CFTR) have comparable numbers of vaults. Wild type cells infected with *P. aeruginosa* recruit 10 – 15% of cellular MVP to lipid rafts, but  $\Delta$ F508 mutants recruit only ~4% so are less efficient in this aspect. The infection produces an immune response that generates lipid rafts. Epithelial cells associated with the raft then take up the bacteria. This ends with epithelial cell apoptosis and clearance of the bacteria. The failure of the  $\Delta$ F508 mutants to recruit MVP to the raft, and subsequent poor clearance of infection, can be induced by knocking out MVP using siRNA. Bacterial uptake is then reduced by ~50%, but in this instance MVP knockdown does not appear to have any direct effect on the immune response (Kowalski *et al.*, 2007).



The authors suggest that the vault particle contributes to resistance to lung infection by stabilisation of the lipid raft, perhaps recruiting other proteins to the raft. Tanaka *et al.*, (2009) identified the shoulder section of the MVP monomer as having homology with the stomatin core. The SPFH (Stomatin, Prohibitin, Flotillin and HflK/C) domain is known to associate with lipid rafts, it could be that this conserved shoulder domain is involved with recruitment to the raft.

#### **1.5.5. Detoxification roles**

Aggregates of vaults termed 'vaultosomes' form in response to tellurite ( $\text{TeO}_3^{2-}$ ) and other oxyanions, even at low concentrations (Suprenant *et al.*, 2007), but not as the result of heat shock or UV, indicating that vaultosome formation is not a general stress response. In Mediterranean mussels, MVP was found up-regulated in gill tissue in response to nickel stress and was concluded to be part of a multi-xenobiotic resistance pathway (Franzellitti *et al.*, 2011). The ability to encapsulate and eject toxins from cells via the vault particle would be an advantage worthy of the retention, however although implicated in many pathways.

#### **1.5.6. Multi Drug Resistance (MDR)**

MVP has several roles associated with cancer treatment; lung resistance protein (LRP) was identified in treatment resistant lung cancer and linked with the multidrug resistant (MDR) phenotype over two decades ago. However, LRP is MVP and levels are elevated in treatment resistant cancers and refractory epilepsy (Lazarowski and Czornyj, 2011).

Relatively curable cancers e.g., metastatic testicular cancer, childhood acute myelogenous leukaemia have low levels of expression of MVP (Zurita *et al.*, 2003), whereas high levels are seen in metastatic colon, renal and pancreatic carcinoma that generally have a poor outcome (Izquierdo *et al.*, 1996). Although MVP cannot be assigned an unequivocal role it can be a reliable predictive prognostic marker for treatment in some cancers; bladder (Diestra *et al.*, 2003), melanoma (Schadendorf *et al.*, 1995), testicular tumour (Mándoky *et al.*, 2004) and glioblastomas (Tews *et al.*, 2000), but is of contestable value in others. It also appears to have use as a predictable marker for outcomes of radiotherapy (Lara *et al.*, 2011).

In MDR, the membrane permeable anticancer drug doxorubicin is sequestered into low pH compartments (thought to be lysosomes) and eventually removed from the cell rather than accumulating in the nucleus and killing the cell. Herlevsen *et al.*, (2007) used human bladder cancer cells and siRNA to knockdown MVP and thus vault number.

This leads to greater sensitivity of the cells to doxorubicin but does not reduce the concentration in the cells (Herlevsen *et al.*, 2007). Overexpression of MVP engineered so that the siRNA cannot bind, rescues, and in fact increases vault numbers above normal, but does not increase doxorubicin resistance. Microscopy showed that in MVP knockdown cells, the doxorubicin is still removed from the nucleus to the lysosomes. In normal cells lysosomes containing the drug can be found around the nucleus, but when MVP is knocked down, the lysosomes are scattered throughout the cell rather than by the nucleus. It was postulated that the vaults are disrupting the microtubules required by both vaults and lysosomes. So low levels of MVP appear to cause disruption of lysosomal distribution (by unknown means) rather than facilitate transfer of doxorubicin from the nucleus to the lysosome. Comparing the movement of vault particles and daunorubicin (an anthracycline used to treat leukaemia) in drug-sensitive and drug-resistant non-small lung cancer cells and did not find any relocation of vaults as the drug moved (van Zon *et al.*, 2004). One group found that overexpression of MVP in colon carcinoma cells does confer resistance to doxorubicin (Kitazono *et al.*, 1999). This was specifically refuted in a repeat experiment where overexpression of MVP did not confer drug resistance in tumour cells (Scheffer *et al.*, 2000) (Huffman and Corey, 2005). Knocking out MVP does make cells more susceptible to chemotherapy (Hu *et al.*, 2010).

It has been found however, that hypoxia increases MVP expression, angiogenesis, and prevents many drugs from working as they cannot generate free radicals and this is linked with radiation resistance of tumour cells (Lara *et al.*, 2009). Whether the hypoxia or increased levels of MVP is responsible is difficult to untangle but an increased number of vault particles doesn't predict a good outcome.

*MVP*<sup>-/-</sup> mouse mutants did not show altered efflux when GFP-tagged MVP was added to the cells. In fact in the *MVP*<sup>-/-</sup> mice cancer cell lines, the treatments were no better or worse than cells from wild type mice, and the mice appeared perfectly healthy (Mossink *et al.*, 2002a), although as we have already seen, mice may not be representative of the way that vaults behave in primates. Some anti cancer agents, ethidium bromide and ultraviolet light can also induce transcription via the *MVP* promoter (Shimamoto *et al.*, 2006).

It could be argued that perhaps low levels of vaults means that free VPARP and TEPI are more available. Both are involved with DNA and chromosome health, but



reduction in MVP also constrains the levels of free VPARP and TEP1 (Wiemer *et al.*, 2004). Conversely if the levels of MVP are raised, and VPARP and TEP1 levels are raised in concert, it does not appear to have the protective outcome that may be anticipated.

#### **1.5.7. Cell signalling**

Vaults have been suggested as a mechanism for transporting the phosphatase and tensin homologue (PTEN) into the nucleus. PTEN is a tumour suppressor that causes cell cycle arrest and apoptosis and mutations in *PTEN* are implicated in many cancers. PTEN is a target of the highly conserved onco-miR (miR-21), a mammalian microRNA encoded by the *MIR21* gene (Lagos-Quintana *et al.*, 2001) and may demonstrate a link between inflammation and cancer (Musilova and Mraz, 2015). PTEN has been described as having four ‘nuclear-localisation-like’ sequences, two of which are necessary (but not sufficient) for nuclear localisation. The suggestion is that the vault particle, via the formation of a vault tube, penetrates the nuclear membrane and is the mechanism for PTEN nuclear localisation. However vault tubes have only been found to form at low temperatures (about 21°Celsius) so are perhaps unlikely to be relevant *in vivo* (van Zon *et al.*, 2003).

MVP has been linked to the regulation of a number of intracellular phosphorylation cascades all implicated in cancer. MVP / PTEN is linked to the phosphoinositide 3-kinase/Akt signalling pathway which has functions related to cell growth, proliferation, differentiation, motility, survival and intracellular trafficking and is also involved in many cancers (Zhenbao *et al.*, 2002). The MVP promoter region also has a consensus sequences for a gamma-activated site (GAS), which binds STAT1 homodimers that up-regulate MVP expression. Constitutive activation of STAT1 is seen in many forms of malignant transformation by oncogenes, cytokines and tumour viruses and may be responsible for the raised levels of vaults found in some cancers (Steiner *et al.*, 2006). MVP forms a substrate complex with SHP-2 (Src homology 2 - a domain containing a tyrosine phosphatase) resulting in the tyrosyl phosphorylation of the MVP. This modulates the ability of MVP to associate with other signalling molecules (Kolli *et al.*, 2004). Phosphorylated MVP has also been shown to complex with Erk2 (extracellular-regulated kinase 2), both SHP-2 and Erk2 are stimulated by epidermal growth factor (EGF) which also stimulates basal MVP phosphorylation (Kolli *et al.*, 2004). However in fucoidan-treated macrophages, MVP is necessary for the activation

of p38 and JNK kinases but not Erk (Ben *et al.*, 2013). This may reflect a species difference due to the variations in the mouse MVP promoter but regardless it demonstrates that the vault particle has diversified depending on the environment that it is in.

Mammalian COP1 (constitutively photomorphogenic 1) is over-expressed in some cancers (Dornan *et al.*, 2004; Yi *et al.*, 2005). In plants, COP1 (constitutively photomorphogenic 1) functions to suppress the constitutively active light-signalling cascade during periods of darkness (Deng *et al.*, 1991) using E3 ubiquitin ligase to repress light signalling by targeting photoreceptors and downstream transcription factors for ubiquitinylation and degradation. In animals COP1 and MVP interact and negatively regulate c-Jun and AP-1 (activator protein 1) under normal cellular conditions. Under times of cell stress however, such as UV radiation, this interaction is inhibited, so AP-1 and c-Jun are both elevated. *MVP*<sup>-/-</sup> cells also result in increased levels of AP-1 and c-Jun, a phenotype that can be rescued with the introduction of MVP (Yi *et al.*, 2005). It is intriguing that a protein so important in plants associates with vaults in animals. Searches have been made for MVP in plants, particularly following the annotation of MVP in the barley genome (see discussion regarding barley in chapter 3b). Some proteins in plants that could be broken down MVP homologs (e.g., *Petunia integrifolia* UniProtKB:A9XLF3, *Arabidopsis lyrata* UniProtKB:D7MVK4, *Zea mays* UniProtKB:B8A0P4) have been found but not a convincing plant homolog.

#### **1.5.8. Possible future biotechnological use of the vault particle**

Some researchers can see a utility in the vault structure which could be harnessed for function in a number of fields (Rome and Kickhoefer, 2013). As vaults are recognised as ‘self’ by the immune system, and their shape lends them to carriers of cargo, they have been put to use as vectors. Tags can be applied to the C terminal of MVP to direct them to specific locations (Goldsmith *et al.*, 2009). Vaults have been used as carriers of antigens to induce an immune response (Champion *et al.*, 2009), and to courier drugs (Buehler *et al.*, 2011; Kar *et al.*, 2011) toxins and genes (Lai *et al.*, 2009), and proteins (Kickhoefer *et al.*, 2005). Additionally thermally responsive ‘smart’ vaults have been engineered (vault-poly-N-isopropylacrylamide conjugate) (Matsumoto *et al.*, 2013) which could, in theory, be used *in vivo* for mopping up toxins, or as environmental bioreactors containing detoxification enzymes for environmental remediation (Wang *et al.*, 2015).

## 1.6. VTRNA

It is questionable whether the vault ribonucleoprotein is truly a remnant of the RNA world. Although in chapter 2b the vault protein is shown to most likely have been already present at least in the Last Eukaryotic Common Ancestor (LECA), it seems less likely that it had any RNA component at that time because vault particles can form without any RNA component and VTRNA has not been found in some species that do have vault particles (Franzén *et al.*, 2011).

VTRNAs are transcribed by RNA pol III polymerase and are generally in the range of ~80-140 nucleotides in length (Kickhoefer *et al.*, 1993). RNA pol III also synthesises other small RNAs (tRNAs, rRNA 5S, U6, 7SL and also miRNAs) found in the nucleus and cytosol. VTRNAs are similar to Y RNAs in that they have maintained a stable position on the genome although Y RNAs are found as a single cluster of clear paralogs and VTRNAs are either single copy genes or form a cluster of similar sequences which seem more like tandem duplication, either very young, or subject to concerted evolution (i.e. have evolved together within the species, ‘in concert’).

Initial interest in the vault particle was due to the associated RNA – but it is not clear if vault particles in all species have an RNA component. For sea urchin vaults the VTRNA is structural as the vault collapses without it (Stewart *et al.*, 2005) but in most metazoa >90% of the VTRNA is found outside of the vault particle and most vault particles will form without it. VTRNA is present in most deuterostomes (Stadler *et al.*, 2009a) but hasn’t been found in the trematodes *Schistosoma mansoni* and *Schistosoma japonicum* (Copeland *et al.*, 2009) even though *S. mansoni* does have a vault particle (Reis *et al.*, 2014). *S. japonicum* has an almost complete sequence in pieces, which may be an artefact of sequencing rather than lack of MVP because the sequence fragments remain highly conserved.

The RNA component of the vault ribonucleoprotein is abundant in the cytoplasm of most mammalian cell types and is particularly enriched in the spleen, intestine, heart and kidney. In humans there are four functional genes for the non-coding vault RNA on chromosome 5, three are located adjacent to the protocadherin  $\alpha$  (pcdh) locus (VTRNA1) and one a short distance away adjacent to the SMAD5 locus (VTRNA2) plus two pseudogenes on chromosome 2 and on the X chromosome (VTRNA3) (Van Zon *et al.*, 2001).

### 1.6.1. Vault RNA nomenclature

In order to clarify the nomenclature and differentiate vault RNA from viral RNA (vRNA) the HUGO Gene Nomenclature Committee (HGNC) renamed vault RNA VTRNA in 2009. Table 2.1 shows the current nomenclature and that previously used by some authors; the current HGNC nomenclature will be used throughout this discussion.

**Table 1.1 vault RNA nomenclature**

Approved Symbol	Approved Name	Previous Symbols	Synonyms	Chromosome
VTRNA1-1	vault RNA 1-1	VAULTRC1	vtRNA1-1, hvg-1, HVG1, vRNA, VR1	5q31.3
VTRNA1-2	vault RNA 1-2	VAULTRC2	vtRNA1-2, hvg-2, HVG2, VR2	5q31.3
VTRNA1-3	vault RNA 1-3	VAULTRC3	vtRNA1-3, hvg-3, HVG3, VR3	5q31.3
VTRNA2-1	vault RNA 2-1	MIR886, MIRN886, VTRNA2	vtRNA2, hvg-5, CBL-3, hsa-miR886, nc886	5q31.1
VTRNA2-2P	vault RNA 2-2, pseudogene			2p14
VTRNA3-1P	vault RNA 3-1, pseudogene	VAULTRC4, VTRNA3P	vtRNA3P, hvg-4, HVG4	Xp11.22

Peter Stadler's group in Germany have found VTRNA in species where vaults have not yet been described (Stadler *et al.*, 2009b). Some VTRNA from VTRNA2-1 has been misidentified as microRNA (miR-886) (Landgraf *et al.*, 2007) according to Stadler who has listed it as a 'dead miRNA' in miRBase ([http://www.mirbase.org/cgi-bin/mirna\\_entry.pl?acc=MI0005527](http://www.mirbase.org/cgi-bin/mirna_entry.pl?acc=MI0005527)), yet it has been shown that VTRNA can be both (Persson *et al.*, 2009) so this assertion may be premature.

The analysis of VTRNA is complicated by the nature of miRNA processing. The uncapitalised 'mir' refers to the pre-miRNA, which is processed from longer unstable hairpin pri-miRNA (primary RNA) while a capitalised 'miR' refers to the mature form (20-30 nt in length) (though this nomenclature is inconsistently used in the literature). Fragments of mir-886 have been found (miR-886-5p and miR-886-3p) that could be genuine microRNAs, but they are barely detectable in the cell. An active component influencing the RNA dependant kinase-protein (PKR) is mir-866 (~100 nt in length - slightly longer than most pre-miRNA) (Lee *et al.*, 2011). The study found that mir-866 was neither a canonical pre-microRNA, nor a VTRNA and appears to have avoided the usual Drosha processing characteristic of pri-miRNA. It is also too short to be included amongst the group of long non-coding RNAs (lncRNAs) (Quinn and Chang, 2016). Additionally Lee's group did not find that VTRNA2-1 (or mir-886) was vault

associated. In fact there are a plethora of papers detailing the function of mir-886 in cancer (Treppendahl *et al.*, 2012; Kunkeaw *et al.*, 2013; Lee *et al.*, 2014; Yu *et al.*, 2014; Kong *et al.*, 2015) so it would seem that the incompletely processed miRNA is biologically active.

In most circumstances only one VTRNA is transcribed even if there are functional genes at the *pcdh* and *SMAD5* loci. Humans and chimps though, have a clear distinction of the 3 functional VTRNA1 genes transcribed at the *pcdh* locus, and the VTRNA2-1 gene from the *SMAD-5* locus, and all four VTRNAs can be transcribed at the same time. The promoters are different at each locus and the resulting expression pattern of VTRNAs appears linked in some way to multi-drug-resistance (Stadler *et al.*, 2009a)

Small RNAs derived from VTRNA1-1 are observed by Persson *et al.*, (2009). Termed 'svRNAs', a name not covered by the HGNC nomenclature ('sv' is also used for 'small viral' RNAs). The svRNAs are not processed in the nucleus via Drosha but are dependant on Dicer processing in the cytoplasm, thereafter they enter the argonaute RISC complex in the same manner as any other miRNA and have the theoretical capability of interfering with more than a hundred targets. Persson *et al.*, (2009) specifically investigated the interaction with *CYP3A4* as it encodes a cytochrome P450 enzyme important in drug metabolism. They found that svRNA<sub>b</sub> knockdown results in stepwise elevation of *CYP3A4*.

The Stadler group dismissed functional VTRNA derived from pseudogenes yet there is evidence that some pseudogenes are still able to influence translation of mRNA via small RNAs for example, transposons in mouse oocytes have been shown to be regulated by endo-siRNA from pseudogenes (Tam *et al.*, 2008).

In primates different versions of VTRNA can inhabit the same vault, and can be found free, or in association with TEPI or the La RNA binding protein, or outside of the vault (Kickhoefer *et al.*, 1998). In fact TEPI is required for VTRNA to be stably included in the vault (Kickhoefer *et al.*, 2001). Production of MVP and VTRNA are somehow linked because down-regulation of one also affects the other. In *MVP*<sup>-/-</sup> mice, TEPI, VPARP and VTRNA are severely decreased (Wiemer *et al.*, 2004), although svRNA biogenesis appears independent of VTRNA levels (Persson *et al.*, 2009).

### 1.6.2. VTRNA function

VTRNA1-1 is the RNA most likely associated with vault RNPs but the level of expression from all of the loci deviates from normal (up and down-regulated) in different cancer cell lines (Stadler *et al.*, 2009b). This is complicated by increasing evidence that VTRNA has functions outside of the vault particle (Treppendahl *et al.*, 2012; Amort *et al.*, 2015). EBV infection in Burkitt lymphoma (BL) cell lines were found to have VTRNA1-1 (usually associated with the vault particle) expression induced by latent membrane protein 1 (LMP1) resulting in lower levels of apoptosis facilitating the infection. The same is true of HeLa and breast cancer cell lines. BL *MVP*<sup>-/-</sup> cells demonstrate that the VTRNA1-1 acts outside of the vault particle and knockdown of VTRNA1-1 results in increased apoptosis (Amort *et al.*, 2015). Another host non-coding RNA (mir-21) is also known to be manipulated by LMP1 following EBV infection to reduce apoptosis in nasopharyngeal cancer cells (Yang *et al.*, 2013) so this appears to be a general method of EBV warfare.

The VTRNA genes are regulated by DNA methylation; in blood cells from healthy donors VTRNA1-1 is unmethylated, VTRNA 1-2 is heavily methylated in CD34+ cells and methylated at one allele only in lymphocytes and granulocytes. The same pattern appears in the human leukaemia cell line (HL60), however the methylation pattern for VTRNA1-3 is unmethylated in the healthy cells but heavily methylated in the HL60 cells. This indicates that reduced expression of VTRNA1-3 is associated with a poorer outcome in patients with myelodysplastic syndrome, but evident only in patients that score as a lower risk in the International Prognostic Scoring System (IPSS) in terms of survival and risk of leukemic transformation. This suggests that a function of VTRNA1-3 could be as a tumour suppressor (Helbo *et al.*, 2015) and this is a common theme for much cancer research.

The most varied of the VTRNA expression lies with the VTRNA from the VTRNA2-1 locus (the one closest to the SMAD5 locus also described as MIR-886 or mir-886). VTRNA2-1 transcripts are not usually vault associated (hence the debate regarding whether this is VTRNA2-1 or mir886), but in the case of infection with the DNA double stranded Epstein-Barr virus (EBV) VTRNA2-1 was found to be both up-regulated and vault associated (Nandy *et al.*, 2009). If VTRNA2-1 was vault associated and additionally protected by TEPI, rather than easily accessible to RNA extraction from the cytosol the VTRNA would be more difficult to extract. Peter Stadler's group



found that the VTRNA2-1 transcript was severely down-regulated in androgen insensitive prostate cancer cell line Du145 (Stadler *et al.*, 2009a), it is not clear if these had become vault associated and required a more exhaustive extraction process.

It is most likely that the expression does vary as VTRNA2-1 is down-regulated by the single stranded parvovirus (other VTRNA expression is unchanged) (Nandy *et al.*, 2009). VTRNA2-1 expression in acute myeloid leukaemia (AML) was also found to be regulated by methylation of the methylation promoter and when shut down by methylation was prognostic for a poor outcome in AML (Treppendahl *et al.*, 2012).

It is not only viral interference and methylation that affects expression of the VTRNA genes there is also evidence of down-regulation of all host VTRNA in M2 macrophages in response to infection by the parasitic protozoan kinetoplast *Leishmania*, that causes leishmaniasis, a disease that is often fatal (Farrow *et al.*, 2011). Kinetoplasts (*Leishmania* and the related *Trypanosomes*) are interesting in their own right because the various species have four or five copies of *MVP* compared to one copy in most other species. It is not known if any of the kinetoplasts also has a VTRNA homolog, however *Leishmania* has the capacity to control the host VTRNA expression generally via a necessary pol III transcription factor which also results in *Leishmania* controlled down-regulation of small RNAs from type 2 and type 4 pol III promoters *i.e.* miRNA, tRNA and 7SL.

*Trypanosome cruzi* uses fragments of tRNAs as miRNA, accentuated under nutritional stress (Garcia-Silva *et al.*, 2010) so it is possible that processed tRNAs could be used in vaults. It could be argued that *Leishmania* control of the transcription factor is part of a miRNA war undertaken by parasite and host directed at miRNAs generally and that VTRNA is collateral damage. As VTRNA is variously described to be up regulated and down regulated in response to viral infection (Nandy *et al.*, 2009), cancer (Stadler *et al.*, 2009b), and parasites (Farrow *et al.*, 2011), it may just be that *Leishmania* does what is good for its survival. *Leishmania* can down-regulate three types of host RNA that are known to take part in defence in what has to be one of the most unlikely places, the phagolysosome of a macrophage, the very place that the parasite should meet its doom.

It would be very desirable to see whether VTRNA can be identified from the next-gen sequencing data available for some Excavate groups, such as *Giardia* and

*Trichomonas*, *Leishmania*, *Trypanosomes*; and try and decipher VTRNA function. It seems unlikely that it solely exists for whatever purpose a vault serves.

The main point here is that the MVP is a very widespread gene, and VTRNA is rapidly being seen as a key switch in a number of pathways. The two next papers describe work from the start of the project, and it is necessary to give an update in what the protein really does. Vaults are both highly ornate structurally and expensive for cells to produce, so this investment in resources and their widespread distribution implies a biological importance to their roles. A defining function for vault particles remains elusive. It is these factors which drive the interest and analysis here in investigating their evolutionary history.

## 1.7. Summary

The papers presented in this thesis represent a journey to find freely available software that can be used without specialist computational knowledge, and with sequences of limited primary homology in order to perform structure-based evolutionary analysis of proteins. Each paper is a step toward understanding the limitations of the software as well as looking at the evolution of the proteins in question. The intention is to use simple computing techniques to understand the evolution of quite different types of protein using a method that can be adapted to other proteins.

Understanding the evolutionary processes of protein families gives an insight into the development of new protein function. Automatic approaches for function prediction are usually based on sequence information; I seek to extract more information than is usual from the primary sequence. Using these methods I found issues in the assignment of protein annotation, which I have raised for both proteins.

This approach can be likened to trying to determine how our grandparents may have looked by examining the facial features of all of our cousins. If we include cousins that are not blood relatives (by reasons of infidelity or adoption) we will have introduced error into the construction of our grandparent's features.

The way to minimise this risk in terms of reconstructing ancestral sequences (ASR) is to be certain that we have only included *bona fide* family members. In terms of protein sequences we make use of structural prediction via computational analysis but also any additional evidence that we can. Computing time is costly (if we pay for it),



and slow if we don't. However computing time is amenable to a greater number of researchers than bench time.

In chapter 2a we lay down the criteria required for a protein sequence to be included in the family 'MVP'. However by chapter 2b it was becoming apparent that in some cases the criteria could not be applied ruthlessly where there was additional evidence that resulted in reconsideration of individual cases. I have erred on the side of caution but in some cases the evidence from some sources was more compelling than the failure to meet a cut off point in another area. The criterion for sequence inclusion for ancestral reconstruction is redrawn in the preface to chapter four.

Many studies compare the virtues or otherwise of each algorithm (Hall, 2005; Williams *et al.*, 2006; Hanson-Smith *et al.*, 2010; Thompson *et al.*, 2011). Ultimately those that have merit for ease of use and perceived accuracy survive and those that do not will wither and die and so evolution of computer algorithms will occur by preferential selection. We have specifically used servers that provide a free service and excel in terms of their biologist friendly (rather than geek friendly) interface.

Almost exclusively the sequences have come from UniProtKB (Consortium, 2015) or NCBI. UniProt accession numbers have been used to identify sequences wherever possible. In some instances sequences have been given new accession numbers since this work was started, on occasion prompted by an email pointing out anomalies e.g. in the chimpanzee briefly mentioned in chapter 2b. Sequences found in NCBI can often be found later in UniProt. Where this is the case, the accession number has been updated to the UniProt number for consistency.

Our main point is that when we are looking at the evolution of single genes at deep times we cannot rely on any one source of information but must look at proteins and their evolution holistically, that is we must use all the information available and come to a 'most probable' conclusion.

# Chapter Two: The Evolution of the Vault Particle

## 2. a: Beyond BLASTing

‘Beyond BLASTing: Tertiary and quaternary structure analysis helps identify Major Vault Proteins’ was written to lay the foundations of a method of three-dimensional structural prediction that would allow for identification of vault particle monomers (MVP) where the primary homology was low. The solved rat monomer (PDB:2ZUO oligomer b) is known to form a vault particle with the other identical monomers. By comparing structural predictions made by I-TASSER and alignment with the known structure (2ZUO\*b) via FATCAT, I could then use RosettaDock (now renamed ROSIE) to see if sequences with structural similarity with 2ZUO\*b would also dock side-by-side with a low energy score as they would *in vivo*. Supplementary material for this article is available at:

<http://gbe.oxfordjournals.org/content/5/1/217/suppl/DC1>

## 2. b: *In silico* resurrection

‘*In silico* resurrection of the Major Vault Protein suggests it is ancestral in modern eukaryotes’ describes how the method established in ‘Beyond BLASTing’ could be used to identify MVP in other kingdoms and to establish whether MVP was in the Last Common Eukaryotic Ancestor (LECA).

We additionally used Ensembl (Kersey *et al.*, 2015) which found sequences from some extra species that were added to the study and highlighted inconsistencies with regard to the other databases, for example an MVP sequence with one accession for the excavate *Trypanosoma cruzi* in Ensembl (and NCBI), but two different accession numbers in UniProt for the identical sequence. These were from the same gene and have now been amalgamated.

Ensembl was also used to identify and to link gene position, copy number and alternative splice variants in order to provide additional evidence to support the analysis.

Ensembl proved particularly useful in the effort to untangle the complexities of the primate genomes and proteins, (see chapter 2b).

Some alignments weren't identified, or were aligned poorly, by either of the MSA algorithms associated with the databases (UniProt or NCBI), particularly in the case of fragment alignments. In these cases fragments from a variety of databases could be manually added together in 'TextWrangler 5.0.2' (<http://www.barebones.com/products/textwrangler/>) and resubmitted for BLASTing, or aligned using Geneious Pro 8.0.4 (Kearse *et al.*, 2012).

In the case of the MVP, where there is usually only one copy of the gene, it seems more likely that the identified fragments are artefacts of the sequencing process rather than the highly conserved gene broken into bits? An example of this issue is within the trematodes (parasitic flukes) where *Schistosoma mansoni* (UniProtKB:G4V9U9 (previously C4PYV7) (Berriman *et al.*, 2009) aligns with a fragmented copy of MVP from *Schistosoma japonicum* that appears to have the highly conserved shoulder region missing (UniProtKB:Q5DBU4, Q5C1V0, and Q5BSG7). MVP in *S. mansoni* seems to assist in stabilisation of the infection (Reis *et al.*, 2014) and MVP is also found in the *Clonorchis sinensis* (Chinese liver fluke) (UniProtKB:H2KSH8).

In order to find more remote sequences in the sequence databases ancestors were reconstructed from multiple sequence alignments (MSA) of proteins where the confidence was high that the protein was grouped appropriately. The recreated ancestral sequences were used as BLAST queries, in order to find more remote sequences. Some free and simple to use ancestral reconstruction algorithms are described and compared.

## 2.1. Sequence similarity identifiers

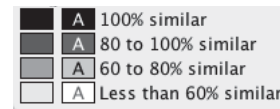
Some alignments were originally directly downloaded from UniProt in CLUSTALW format. The sequence similarity identifiers are as follows:

"\*" = Identical residue, ":" = conserved substitutions, (i.e. in terms of hydrophobicity *etc*), "." = a semi conserved substitution (in terms of size), or a blank space for no similarity.

This does not always give a feel for the percentage of similarity between multiple alignments. An alignment can easily look less significant than it is due to lack of identifiers, which may simply be the result of one sequence being different, having

gaps, or being short. Geneious Pro uses a more complex system that can cope with the problems encountered in the CLUSTAL alignments. All of the alignments are shown with the Geneious sequence similarity colouring described below.

Columns rendered black (100% similar) must have a score that when compared with all other sites must, (according to the specified score matrix), equal or exceed a specified threshold. The score is dependent on the matrix



**Fig. 2.1 Geneious alignment shading**

used, in this case Blosum62, which has a score matrix with a threshold of 1. For instance, Blosum62 has a value of 2 for K (lysine), vs. R (arginine), so a column that contained lysine or arginine residues at that site will be counted as the same and the column will be coloured entirely black. The threshold can be raised so that lysine and arginine would no longer be considered to be the same. In this case, if more than 80% of the column was lysine, then the column would be dark grey, and the arginine would be uncoloured. Figure 7 of '*In silico*' resurrection (chapter 2b) gives a good example.

If 60% – 80% of the column were lysine, and BLOSUM62 considered the other residues dissimilar to lysine, then the column is coloured light grey, (the lettering is now black), and the dissimilar residues are uncoloured. As this gives a superior visual impact of similarity the Geneious colouring has been used throughout. The matrix web address is <http://www.ncbi.nlm.nih.gov/Class/FieldGuide/BLOSUM62.txt>.

I also wish to rephrase the comment that 'Markov models lose information at deep times' (referencing Mossel and Steel 2004 and 2005). It is not the models that lose information rather that the recurrent mutations in aligned sites obscures the relationships and thus the trees are unreliable.

There is necessarily some repetition between these journal articles particularly within the 'methods' section. They are printed as they appear in press, which is a requirement of the Genome Biology and Evolution reprint permission. Supplementary material for this article is available at:

<http://gbe.oxfordjournals.org/content/5/8/1567/suppl/DC1>

# Beyond BLASTing: Tertiary and Quaternary Structure Analysis Helps Identify Major Vault Proteins

Toni K. Daly\*, Andrew J. Sutherland-Smith, and David Penny

Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand

\*Corresponding author: E-mail: t.daly1@massey.ac.nz.

Accepted: December 24, 2012

**Data deposition:** Sequences used in this research — Accession numbers: Q62667, Q5EAJ7, Q4CUM2, Q4QJ7, Q62774, P35240, Q62774, P35240, D2V5B9, D2W0Z9, D2UZ7, D2VSY6, D2VC38, D2VH38 are deposited in UniProtKB. The Rat crystal structure 2ZUO is deposited in the Protein Data Bank. All versions used are current as of January 12, 2013.

## Abstract

We examine the advantages of going beyond sequence similarity and use both protein three-dimensional (3D) structure prediction and then quaternary structure (docking) of inferred 3D structures to help evaluate whether comparable sequences can fold into homologous structures with sufficient lateral associations for quaternary structure formation. Our test case is the major vault protein (MVP) that oligomerizes in multiple copies to form barrel-like vault particles and is relatively widespread among eukaryotes. We used the iterative threading assembly refinement server (I-TASSER) to predict whether putative MVP sequences identified by BLASTp and PSI Basic Local Alignment Search Tool are structurally similar to the experimentally determined rodent MVP tertiary structures. Then two identical predicted quaternary structures from I-TASSER are analyzed by RosettaDock to test whether a pair-wise association occurs, and hence whether the oligomeric vault complex is likely to form for a given MVP sequence. Positive controls for the method are the experimentally determined rat (*Rattus norvegicus*) vault X-ray crystal structure and the purple sea urchin (*Strongylocentrotus purpuratus*) MVP sequence that forms experimentally observed vaults. These and two kinetoplast MVP structural homologs were predicted with high confidence value, and RosettaDock predicted that these MVP sequences would dock laterally and therefore could form oligomeric vaults. As the negative control, I-TASSER did not predict an MVP-like structure from a randomized rat MVP sequence, even when constrained to the rat MVP crystal structure (PDB:2ZUO), thus further validating the method. The protocol identified six putative homologous MVP sequences in the heterobolosean *Naegleria gruberi* within the excavate kingdom. Two of these sequences are predicted to be structurally similar to rat MVP, despite being in excess of 300 residues shorter. The method can be used generally to help test predictions of homology via structural analysis.

**Key words:** homology modeling, BLAST, I-TASSER, RosettaDock, *Naegleria gruberi*.

## Introduction

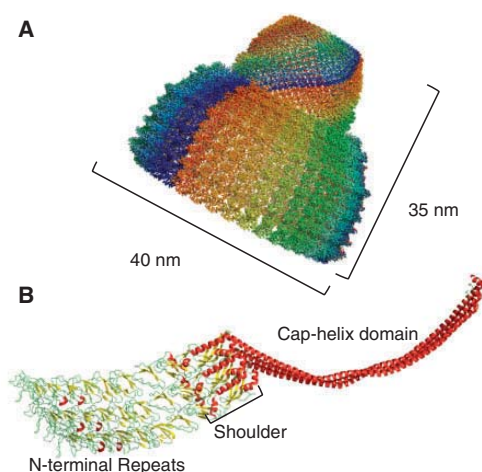
Our interest has included identifying features, proteins, and nontranslated RNAs that may date back at least to the Last Eukaryotic Common Ancestor (LECA). It is increasingly appearing that LECA already had quite a complex cellular and molecular structure (Kurland et al. 2006; Koonin 2010; Neumann et al. 2010). Of particular interest are the smaller untranslated RNAs found in ribonucleoproteins, including the spliceosome (Collins and Penny 2005) and eukaryotic ribosome (Steitz and Moore 2003), where RNA plays a critical catalytic role. Vaults are large oligomeric ribonucleoproteins conserved among a variety of species, many of which contain small untranslated RNAs (vault RNA [vtRNA]) (Stadler et al. 2009). Could the vault RNP date back to similarly early

times? We need to be able to include structural information to test predictions made solely on linear (sequence) information. We first discuss the vaults, then the need for tertiary and quaternary protein structures to help the search for homology.

Vaults can be directly observed by electron microscopy or be detected by immunoblotting with anti-major vault protein (MVP) antibodies, in diverse species such as sea urchins (Hamill and Suprenant 1997), cellular slime mold (Vasu et al. 1993), electric ray (Herrmann et al. 1997), and mammals (Kedersha and Rome 1986). The rat vault RNP structure has been determined to 3.5 Å (Tanaka et al. 2009) defining both the MVP monomeric conformation and how 78 monomers assemble to form a complete vault (a half vault is shown in fig. 1A). The rat MVP monomer consists of four regions: multiple

© The Author(s) 2012. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.



**FIG. 1.**—Vault ribonucleoprotein structure. (A) Rat MVP quaternary structure showing half a vault colored by monomer (PDB: 2ZUO, 2ZU4, and 2ZV5). A full vault will have at the lower left a copy of the upper half vault related by a 2-fold rotation axis. (B) Three rat MVP monomers colored by secondary structure (PDB 2ZUO stripped down to three monomers). This figure highlights the extensive lateral association required to dock into the vault quaternary structure.

N-terminal repeat domains, a shoulder domain, and the cap-helices (fig. 1B); additionally, there is a fourth domain, the capping that is not sufficiently ordered to be observed in the crystal structure and so is not visible in figure 1. The rat monomer begins with nine repeat domains from the N-terminus—these repeats form a “stave-like” structure along the side of the vault barrel. The repeat domains are followed by the shoulder domain that then connects to a 42 turn  $\alpha$ -helical domain known as the cap-helix. The cap-helix represents the top of the vault at a lower diameter than the N-terminal repeats (fig. 1A and B). Interactions between monomers of the long helical cap-helix are key for vault stabilization (Tanaka et al. 2009) and are essential for self-assembly (van Zon et al. 2002).

The equilibrium of monomer to oligomeric vault appears to strongly favor vault formation. For example, in rat liver cell lysate, ultracentrifugation of purified MVP shows 95% of the population as a high molecular weight form (Kedersha et al. 1991); and antibodies fractionate with intact vaults rather than with individual monomers in rat neural cells (Paspalas et al. 2009). Vaults are stable to a wide pH range (4–11), as well as in 1% Triton X-100 and 2 M urea (Kedersha et al. 1991). Extension at the N or C terminal does not prevent vault formation as fusion tagged MVP still assembles into vaults (Kickhoefer et al. 2009). Although vaults have other components, vtRNA, vault poly ADP-ribosylating protein (VPARP), and telomerase associated protein 1 (TEP1), these

are not normally essential for vault formation (Stephen et al. 2001). TEP1 is also found in the telomerase complex; additionally, VPARP and vtRNA are found outside of vaults and so may have other functions as well. Although vault RNPs are linked to many processes (Berger et al. 2009; Vollmar et al. 2009; Lara et al. 2011; Liu et al. 2011) as yet they have no known intrinsic function.

The general issue of homology arises because proteins annotated as MVP via sequence homology, rather than by experimental determination, have been reported in the genome of many species including trypanosomes and paramoecium. Considering that MVP sequences are apparently reasonably widespread, numerous and relatively conserved, it is surprising that convincing homologs (sequences or structures) appear to be missing from nematodes, flies, and fungi. A plant homolog recently reported in domestic barley (*Hordeum vulgare*) (Matsumoto et al. 2011) (UniProtKB: F2E078) has yet to be ascribed to the barley genome, thus could be the result of contamination—an example of contamination has been reported in mosses (Stevens et al. 2007). Thus, we require structural prediction information to help confirm (or not) the presence of vaults in a wider range of eukaryotes.

Traditionally, linear protein sequences have been used to determine homology, with subsequent annotation extrapolated to similar sequences based on a small subset of experimentally characterized proteins. Protein structure may sometimes be minimally affected by amino acid substitutions, and sequences with limited similarity may retain homologous folding patterns (Murzin et al. 1995; Orengo et al. 1997). In addition to sequence comparison, modeling studies have been used to identify members of protein superfamilies with low sequence homology (Holm and Sander 1997) and can also be used to predict function (Watson et al. 2005). Structural prediction studies are especially important for evaluation of the deepest sequence similarities because the Markov models we use for sequence evolution are expected to saturate, and lose information, at the most ancient divergences (Mossel and Steel 2004). Another way of testing structural and functional predictions is to synthesize the inferred ancestral sequences and measure their properties (Finnigan et al. 2012).

To extend one-dimensional sequence homology analysis, we have used a computational approach to help identify putative MVP sequences and to determine whether they are likely to form intact vaults. MVP represents an ideal case for the purposes of demonstrating the utility of three-dimensional (3D) studies as a means of enhancing the search for functional sequence homologs because the oligomeric vault structure is capable of independent self-assembly (Stephen et al. 2001). Furthermore, the monomeric MVP tertiary structure (and hence the sequence) is presumably under strong selective pressure to retain a conformation that forms not only the appropriate monomer structure but also the appropriate interface interactions with its neighbors for vault quaternary



structure/assembly (Qian et al. 2011). Here, we examine previously uncharacterized putative MVP sequences against these structural criteria, enabling us to predict with improved certainty whether the *mvp* gene, or relics of it, is likely to be present in a given species and whether intact vault particles are likely to form. Controls (both positive and negative) are essential to help determine the reliability of the inferred tertiary and quaternary models. It is essential to use tertiary and quaternary information to test homologies suggested in linear (one dimensional) information, and we have used many standard programs that are outlined later.

Electron microscopy of vault particles from a variety of species indicates that the intact vault structure is strikingly conserved. The rat MVP structure (PDB:2ZUO\*b) was chosen as the standard by which we compare the folding of all other models because the whole oligomeric vault is resolved to 3.5 Å (Tanaka et al. 2009). Other structures in the Protein Data Bank (PDB) are fragments limited to the repeat sections of the MVP monomer only: mouse (Querol-Audi et al. 2009) and human (Kozlov et al. 2006), both virtually identical to the rat structure. The rat MVP structure is not necessarily an ideal template for the structure of distantly related MVP sequences, and the amoebozoan *Dictyostelium discoideum* forms a vault from a chimera of two structurally similar MVP paralogs (Vasu and Rome 1995). However, because it is the only full-length oligomeric vault structure, all comparisons have been made to the rat sequence and structure.

## Materials and Methods

### Basic Local Alignment Search Tool Searches

Initial Basic Local Alignment Search Tool (BLAST) searches were undertaken with the rat MVP accession Q62667 and Universal Protein Resource (UniProtKB) using the default BLOSUM62 matrix. All accession numbers refer to UniProt. Later searches used the less stringent BLOSUM45 to identify more remote sequences. "Expect values" (*E*) greater than 0.15 routinely produced BLAST matches that corresponded only to the cap-helix region of MVP (residues 647–802). Similarly, most PSI-BLASTs of the NCBI database identify false positives following the first iteration that align only with the coiled coil and no other MVP region. A PSI-BLAST search should not unduly weight the cap-helix region; however, it appears that there are a limited number of positional homologs involving the repeat areas and an abundance of proteins with the common coiled-coil motif. A search of conserved domains (National Center for Biotechnology Information) shows a very large overlap of conserved domains within the MVP cap-helix region—so the PSI-BLAST search was repeated without the inclusion of the cap-helix and using the kinetoplast sequence from *Leishmania major*. However, the cap-helix was restored following the first iteration as the 1,000 sequences retrieved (default is 500) are

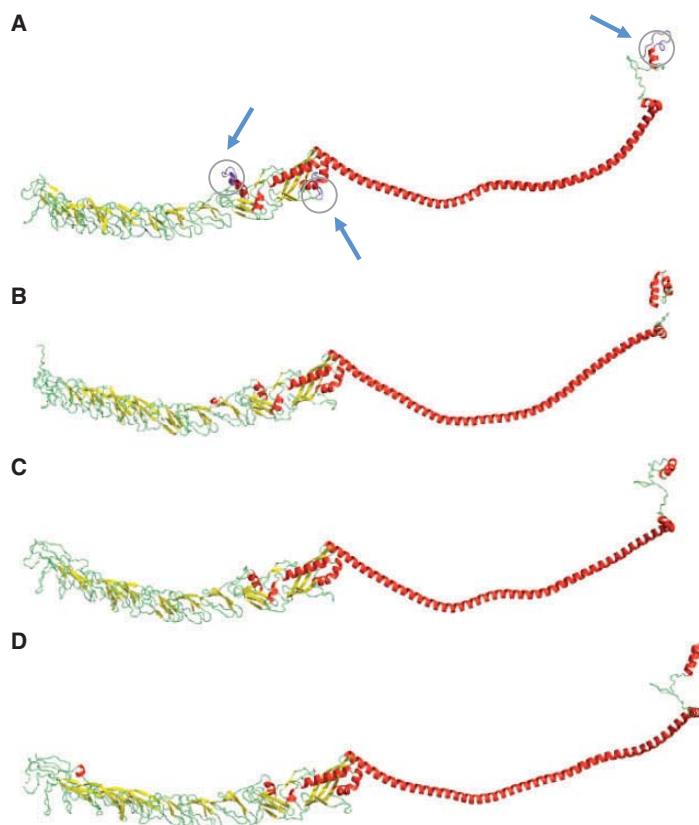
aligned, and a positional matrix is formed and used as the query for the second iteration. Similar searches were also undertaken using ancestral sequences reconstructed from 14 leishmania sequences and 14 trypanosome sequences, but no further sequences were found.

### Iterative Threading Assembly Refinement Server

Iterative threading assembly refinement server (I-TASSER) inputs a query sequence and generates 3D structural models from multiple threading alignments using LOMETS (LOcal MEta Threading Server), a combination of eight threading programs (FUGUE, HHsearch, MUSTER, PROSPECT2, PPA, SP3, SAM-T02, and SPARKS) (Zhang 2008). The submitted sequence initially undergoes a PS-BLAST search to identify possible evolutionary relatives. I-TASSER then uses this BLAST result to generate a position-specific scoring matrix (PSSM or profile) using sequences with an *E* value lower than the threshold (0.005 is the default). The server uses this information to generate a PSI-BLAST using the PSSM as the query. It continues in this manner until no new sequences are added. Still within I-TASSER, the resultant profile is submitted to the PSIPRED server for secondary structure prediction, and both are then submitted to LOMETS. The final structure is presented by MODELLER (Sali and Blundell 1993) using a program that creates a probability density function using geometric criteria that satisfies spatial restraints within the query sequence in comparison to solved structures. It additionally has some ability to predict the shape of the loop structures, which, in the case of the vault, is useful for coverage of the sections missing from the experimentally determined structure (fig. 2A).

I-TASSER is benchmarked by Critical Assessment of Techniques for Protein Structure Prediction (CASP) (Moult et al. 1995), a biannual experiment in which servers are tested on their ability to identify correct folds from protein sequences whose structures have been previously determined but held back from publication by the PDB for the experiment. I-TASSER has scored highly since its inception competing as "Zhang Lab," winning best structure prediction and best function prediction in the most recent test in 2010 (Xu et al. 2011).

The most relevant score for the models predicted by I-TASSER is the *C* score with range −5 to +2. This is the confidence score for the estimated quality of the models calculated from the structural threading and refinement. A *C* score > −1.5 is considered to be a correct fold (Roy et al. 2010). The template modeling (TM) score quantifies structural similarity between two superimposed protein structures analogous to the traditional root mean-squared difference (RMSD). A TM score > 0.5 indicates high confidence that the topology of two models, in this case predicted and native, is the same, and a TM score < 0.17 indicates that the comparison is between random structures. The *C* score is correlated to the TM score (correlation coefficient 0.91) (Zhang 2008). TM weighs small



**FIG. 2.**—MVP monomer comparison. (A) I-TASSER-modeled structure for the full-length rat MVP sequence (Q62667). Residues not observed in the crystal structure (PDB:2ZUO\*b) are circled (shown by arrows). (B) I-TASSER-modeled structure for the sea urchin MVP monomer (Q5EAI7). (C) I-TASSER-modeled structure for the kinetoplast *Trypanosome cruzi* (Q4CUM2) and (D) *Leishmania major* (Q4QJJ7) MVPs.

distance differences greater than large ones and has a length-dependent normalization scale. In contrast, RMSD weighs the pair-wise differences between residues equally meaning that a local difference can have a large impact on the RMSD score, particularly if the protein is large. Because MVP is approximately 850 residues, the RMSD is likely to be of less value. The final control for model quality before submission to RosettaDock was visual comparison to the rat structure, because the "I-TASSER best model" was not necessarily the one that looked most closely like a vault monomer.

Although we considered each output from I-TASSER on a case-by-case basis, some general criteria were applied. For example, to choose a model visually rather than because it is the result with the highest C score, the C scores of the models concerned must be similar. If the C scores are similar, as may occur for targets described by LOMETS as "hard," the first model presented by I-TASSER is not necessarily the best (Roy et al. 2011). Additionally, the C score information lists the

number of decoys and cluster density for each output. If these are also similar for the models being compared, then the model is chosen that is visually closest to the known structure.

If the target is described as "easy," then the first model generally has a significantly higher C score than the rest. LOMETS produced a variety of structures from the *Naegleria gruberi* MVP sequences found via the PSI-BLASTs described as "medium" targets. Visually they were all different, none looked like MVP, and although LOMETS alone does not give a score for confidence, the probability that the models showed the correct folds was described as "medium." They were then submitted to I-TASSER using the rat crystal structure (2ZUO\*b) as a constraint. When a constraint is used, it can be applied with or without a specified alignment. If an alignment is not specified, then the MUSTER (Multi-Source ThreadER) algorithm is used (Wu and Zhang 2008). The initial full-length rat MVP sequence shown in figure 2A could have been used



as a constraint that would have resulted in greater uniformity, particularly with respect to the C terminal and amorphous loop on repeat eight. However, by using 2ZUO\*b, it means that I-TASSER has repeatedly modeled the missing residues ab initio. In fact, the amorphous loop makes a shelf on the inside of the vault that is consistently modeled by RosettaDock. Because the most C-terminal region was not visible in the crystal structure, yet is very highly conserved, future evaluation of the models should highlight any consensus folds in this area. We additionally confirmed the predicted folds using Phyre<sup>2</sup>.

### Phyre<sup>2</sup>

Phyre<sup>2</sup> is an upgrade to the original Protein Homology/analogy Recognition Engine (Phyre) (Kelley and Sternberg 2009). Phyre takes a sequence, builds a profile using PSI-BLASTs, and compares it to templates deposited in the Structural Classification of Proteins database and PDB. Phyre uses three secondary structure prediction programs: PSIPRED, SSPro, and JNet. Each program gives a confidence value for each of three structures: alpha helix, beta sheet, and coil. The confidence values are averaged, and a final, consensus prediction is displayed for each individual prediction. This is computationally less expensive than the multiple alignments used by I-TASSER generating much quicker results and has the advantage that multiple, or even "batch," submissions can be made. Additionally, 20 results can be displayed in full and many more suggested, which means that individual folds can be identified. Phyre and Phyre<sup>2</sup> have been similarly successful in the CASP experiments.

### Flexible Structure Alignment by Chaining Aligned Fragment Pairs Allowing Twists

Flexible structure AlignmentT by Chaining Aligned fragment pairs allowing Twists (FATCAT) (Ye and Godzik 2003) gives a measure of similarity of one structure to another. Structural models predicted by I-TASSER from query sequences were compared with the rat MVP monomer. FATCAT breaks the proteins to be aligned into fragments eight residues long (aligned fragment pairs [AFPs]). These AFPs can be matched, and a twist, gap, or extension can be introduced to match the next AFP if it results in a substantially better superposition. Extensions, gaps, and twists are all scored using a dynamic programming algorithm, so that long AFPs are rewarded and large RMSDs are penalized. This gives the lowest possible chaining score at each juncture. The total chaining score is then combined with the probability of obtaining a greater score, the RMSD of the final superposition, the number of equivalent positions, and the number of twists (with a maximum of five), to give a measure of the structure's significance. This is displayed both as a *P* value and as a raw score. When comparing MVP models, the *P* value is most often reported as "zero," so the raw score gives a sense of

"more" or "less" similar to the rat structure—a high raw score indicates greater similarity to the rat crystal structure that it is being compared with (data not shown).

In this instance, FATCAT was used to space the MVP monomer models for RosettaDock analysis by aligning the query structures with 2ZUO\*b and with 2ZUO\*d (i.e., one monomer width apart) of the rat crystal structure. In some cases, FATCAT will introduce chain breaks to undo twists in the aligned models making them unsuitable for docking analysis; FATCAT can be forced to run a "rigid" alignment that will prevent breaks, and this is a simple and almost instantaneous way of suitably spacing the monomers. Another approach used was to manually position the molecules a monomer width apart in PyMOL (The PyMOL Molecular Graphics System, Version 1, DeLano scientific LLC, 2008), although the advantage of using FATCAT was that the RMSD could be predicted and thus help identify possible docked models where scores were similar across the majority of the models.

### RosettaDock

For vault formation, the MVP monomers dock laterally along the length of both sides to make the barrel shape. RosettaDock is a server that uses a low-resolution Monte Carlo search and backbone optimization algorithm to position the submitted chain pair, followed by a refinement to relax the backbone and accommodate the side chains (Gray et al. 2003).

RosettaDock has very specific requirements; two monomers, side by side, are submitted to see whether they will dock laterally. If the pair of monomers input for docking are initially placed too far apart, then the first local docking search performed may fail to locate them. However, if they are placed too close together (<5 Å), the file is rejected. Additionally, the RosettaDock file cannot total more than 600 residues for submission to the online server as such calculations are computationally too expensive. RosettaDock can be downloaded as a package and thereby the number of residues can be increased. For the online server, the MVP monomers were docked in three sections. The cap-ring domain (C terminal ~60 residues) has not been submitted to RosettaDock because, although it is highly conserved, there is no suitable experimentally determined control structure. As a final complication, in some instances, RosettaDock docks MVP monomers with a large energy score skewed by internal residues that are not involved with the oligomerization interface.

To benchmark RosettaDock, other servers have been tried; ClusPro (Kozakov et al. 2006) is unable to take such large regions of MVP due to a 24-h job limit. GrammX (Tovchigrechko and Vakser 2006) is considerably quicker than Rosetta, but in some instances, it docked the N terminal

of the vault proteins in an antiparallel orientation, which is not consistent with the oligomeric vault crystal structure.

## Results

### Positive Control Study for Method Optimization: Tertiary Structure

The first control used the rat MVP sequence (Q62667) (Kickhoefer and Rome 1994) to model the MVP monomer structure via the I-TASSER server (Roy et al. 2010), initially unconstrained, then constrained by the rat crystal structure (PDB:2ZUO\*b) (Tanaka et al. 2009). This confirmed that I-TASSER identified correctly the crystal structure from the full-length rat sequence. The rat MVP crystal structure shows only 812 residues of the total 861 amino acid sequence. Three regions not observed in the crystal structure are residues 429–448 (a presumed disordered loop on repeat 8), 608–620 (part of the shoulder domain), and amino acids 846–861 (the very C terminus, beyond that described as the cap-ring domain). Nevertheless, the I-TASSER prediction for these regions is important because I-TASSER will be modeling full-length homologous MVP

sequences of unknown structure (fig. 2A). FATCAT structural alignment showed generally that the predicted model is very close to the experimental crystal structure regardless of whether the I-TASSER input sequence was constrained to the known rat structure.

As an additional control, the MVP sequence from the purple sea urchin (an echinoderm), *Strongylocentrotus purpuratus* (Q5EAI7) was analyzed. This urchin MVP has 64% sequence identity with the rat, and intact vaults have been seen via cryo-electron microscopy (Stewart et al. 2005), but the urchin MVP does not have a crystal structure determined. The urchin MVP sequence was submitted to I-TASSER without 2ZUO\*b constraint, and the resulting fold (fig. 2B) is very similar to that of the rat (fig. 2A). MVP sequences from the kinetoplasts *Trypanosoma cruzi* (Q4CUM2) (fig. 2C) and *L. major* (Q4QJJ7) (fig. 2D) were also analyzed (unconstrained) to model the structure that could be anticipated for excavate MVPs. Results are reported in table 1.

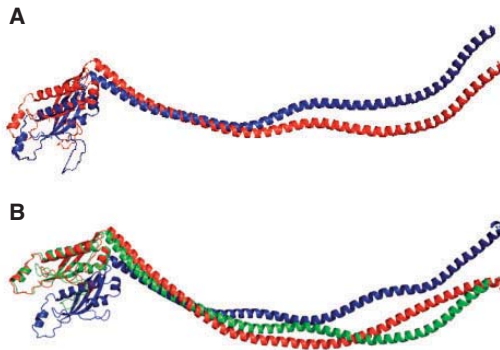
All sequences were also submitted to I-TASSER using 2ZUO\*b (from the rat crystal structure) as a constraint to determine the influence a structural constraint has on the modeling. The use of this constraint has no discernable effect on

**Table 1**

I-TASSER and RosettaDock Results for Positive and Negative Controls

UniProtKB Accession Number	Organism	Length	% Identical Sites versus Q62667	I-TASSER C Score	I-TASSER TM Score	RosettaDock Score for Cap-Helix	RosettaDock Score for Shoulder and Cap-Helix
<b>Positive controls, unconstrained</b>							
2ZUO*b	<i>Rattus norvegicus</i>	812				−261	−435
Q62667	<i>R. norvegicus</i>	861	100	0.42	0.77 ± 0.10	−280	−254
Q5EAI7	<i>Strongylocentrotus purpuratus</i>	857	64	1.12	0.87 ± 0.07	−291	−503
Q4CUM2	<i>Trypanosoma cruzi</i>	838	48	1.11	0.87 ± 0.07	−304	−498
Q4QJJ7	<i>Leishmania major</i>	833	48	1.91	0.99 ± 0.04	−302	−504
<b>Positive controls, constrained by 2ZUO*b</b>							
Q62667	<i>R. norvegicus</i>	861	100	1.02	0.85 ± 0.08	−292	−508
Q5EAI7	<i>S. purpuratus</i>	857	64	1.07	0.86 ± 0.07	−255	−492
Q4CUM2	<i>T. cruzi</i>	838	48	1.18	0.88 ± 0.07	−247	None docked
Q4QJJ7	<i>L. major</i>	833	48	1.33	0.90 ± 0.06	−266	None docked
<b>Negative controls, unconstrained</b>							
Randomized rat MVP	<i>R. norvegicus</i>	861	16	−1.76	0.50 ± 0.15	No cap-helix	—
Q62774	<i>R. norvegicus</i> (myosin 1A)	842	16	0.96	0.84 ± 0.08	−258	No shoulder
P35240	<i>Homo sapiens</i> (merlin)	595	17	−0.76	0.62 ± 0.14	No cap-helix	—
<b>Negative controls, constrained by 2ZUO*b</b>							
Randomized rat MVP	<i>R. norvegicus</i>	861	16	−2.93	0.38 ± 0.13	No cap-helix	—
Q62774	<i>R. norvegicus</i> (myosin 1A)	842	16	0.62	0.80 ± 0.09	−191	No shoulder
P35240	<i>H. sapiens</i> (merlin)	595	17	−1.33	0.55 ± 0.15	Helix does not dock	

NOTE.—The I-TASSER confidence (C) score (>−1.5 is considered a correct fold, range −5 to +2, higher is better). The RosettaDock energy score is lower for the shoulder and cap-helix combined, indicating that the shoulder improves docking. It should be noted that the lateral docking capacity of the cap-helix in the rat MVP was reduced in comparison to the other positive control sequences (fig. 3). This was improved by using the 2ZUO\*b constraint for the I-TASSER rat MVP prediction. In general, using the constraint during I-TASSER modeling reduced the likelihood that RosettaDock would successfully dock the modeled monomers. The other positive control MVP sequences were also submitted to I-TASSER constrained by the rat crystal structure 2ZUO\*b. With the exception of rat and *L. major*, this made very little difference to the I-TASSER score, but it did reduce the possibility of finding docked monomers in the excavates. In the case of the rat, both I-TASSER and RosettaDock scores are considerably improved by using the 2ZUO\*b constraint. The score for *L. major* is reduced by the 2ZUO\*b constraint but still well above the threshold of confidence that the model is correct. (B) Comparison between negative control I-TASSER models with and without the 2ZUO\*b constraint shows that the constraint does not make I-TASSER any more likely to find that the structure matches the rat crystal structure template but does lower the confidence that LOMETS has in the resulted structure. The high C score for the rat myosin (shaded) reflects the myosin V (PDB 2DFS) database structure identified by I-TASSER as most similar. The low score for the randomized rat MVP sequence reflects little similarity to any of the structures in the PDB.



**Fig. 3.**—Structural effect of the 2ZUO\*b constraint. (A) Structural comparison of the shoulder and cap-helix region of two rat MVP models either constrained by 2ZUO\*b (red) or unconstrained (blue). The kink in the unconstrained cap-helix modeled by I-TASSER results in poor docking in RosettaDock. The rat MVP sequence constrained by 2ZUO\*b (red) entirely aligns with 2ZUO\*b (obscured), and this model docks readily in RosettaDock. (B) Urchin MVP shoulder and cap-helix region structural comparison between models either constrained by 2ZUO\*b (red) or unconstrained (blue) relative to 2ZUO\*b (green). In this case, the unconstrained urchin MVP model docks more readily than the constrained model.

the rat, sea urchin, and kinetoplasts sequences in terms of the repeat and shoulder domains. However, the cap-helix structures were altered by the constraint, which had a subsequent effect on the docking performance of the structures (fig. 3 and table 1). All sequences were additionally submitted to the Phyre<sup>2</sup> protein fold recognition server. Phyre<sup>2</sup> confirmed the I-TASSER results with 100% confidence (data not shown).

#### Positive Control Study for Method Optimization: Quaternary Structure

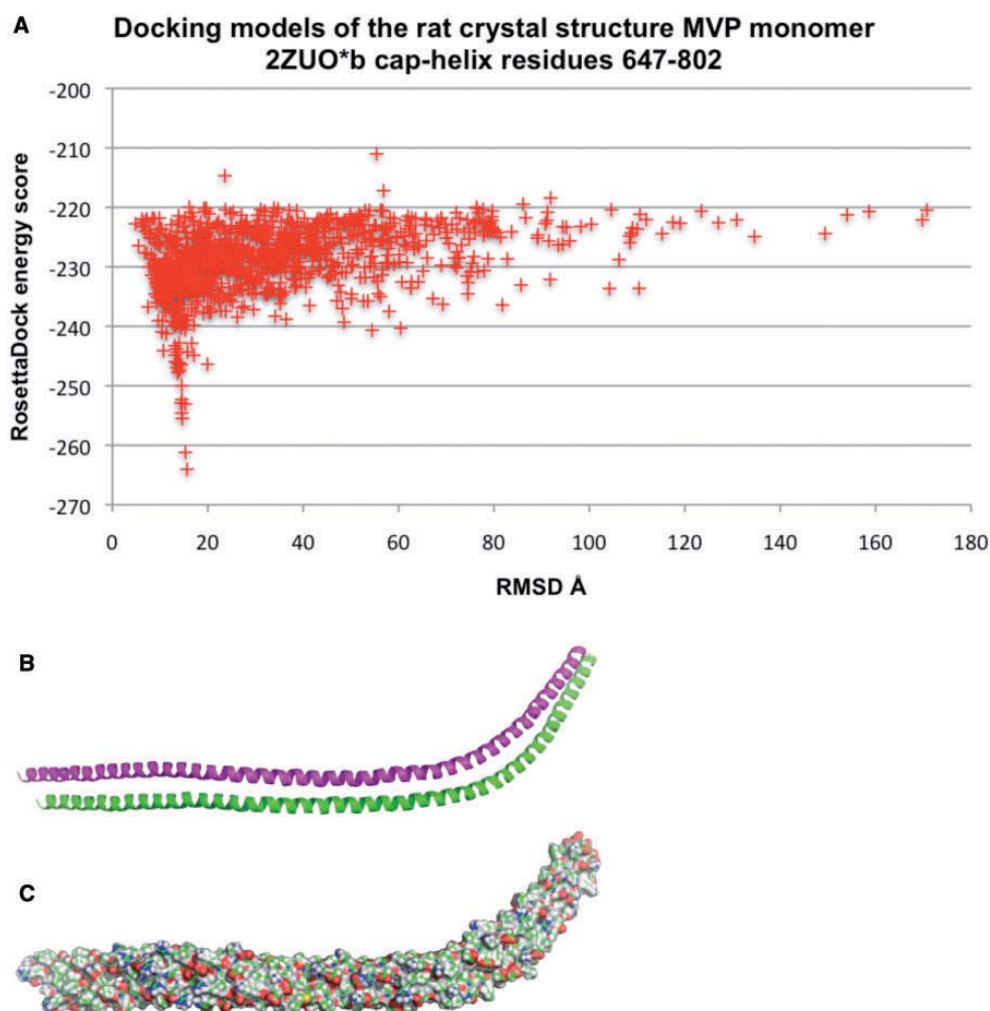
As a further control to determining whether the putative MVP sequences fold in a similar manner to the characterized rat monomer structure, we need to ascertain whether sequences with high structural homology to the MVP monomer are likely to dock with each other and form a vault. As a control, we analyzed rat MVP monomer structures with RosettaDock, with either MVP monomers taken directly from the crystal structure or the full-length rat monomeric MVP structure predicted by I-TASSER. RosettaDock predicts an oligomeric vault structure similar to that of the crystal structure, which can form with good low energy scores (table 1). The MVP C-terminal long  $\alpha$ -helix has been shown to be essential for self-assembly of monomers into oligomeric vaults (van Zon et al. 2002). Therefore, the cap-helix regions (amino acids 647–802) of two separated rat MVP monomers from the crystal structure (PDB:2ZUO\*b and 2ZUO\*d) were submitted to RosettaDock to test how well it would reassemble the lateral associations of docked MVP pairs required for vault assembly.

In most animal species, vaults are homo-oligomeric complexes constructed from identical MVP monomers, so the interactions between monomers are all the same. This means that the docking of one monomer pair can be used to infer vault formation if the appropriate lateral association forms. RosettaDock considers 1,000 structures and searches for the lowest energy conformations of which 10 are output. Each docking solution has an overall energy score (RosettaDock energy score, y axis) that is plotted against the RMSD (x axis) from the starting positions (Å) of the monomers. Score graphs showing a characteristic "funnel" suggest that the 1,000 pairs are clustered in conformation, giving a higher confidence in the lowest energy docked pairs resulted (Lyskov and Gray 2008). A score graph showing the energy scores versus RMSD for residues 647–802 from monomers of the rat crystal structure is shown in figure 4A, together with a cartoon (fig. 4B) and a surface rendered (fig. 4C) representation of the lowest energy docked pair of MVP monomers. Docked monomer surface and MVP ribbon representations were rendered with PyMOL.

A lower energy score can be found when the shoulder region (502–646) is included, indicating that the shoulder area probably contributes to the proper alignment and docking of the monomers (supplementary material S1, Supplementary Material online) consistent with the rat MVP crystal structure. Using the rat shoulder alone indicates a high probability that the shoulders will interact (fig. 5). Oligomerization of a domain homologous to the MVP shoulder has been experimentally demonstrated (Kuwahara et al. 2009).

Using the MVP domains separately, we show that the monomers are likely to dock along their entire length, even when missing the stabilizing effect of the coiled coil (supplementary material S1, Supplementary Material online) again consistent with the interdomain contacts identified from the MVP crystal structure. In each case, the energy score is low and negative, and the RMSD shows that the distance from the starting structure is well clustered. Because the monomers submitted to RosettaDock have been spaced by FATCAT one monomer width apart, the starting distance between the molecules is approximately 15–20 Å and the resulting RMSD for successful docking can be predicted. Thus, we test both that the modeled monomer MVP 3D tertiary structures are consistent with the rat MVP monomer structure and that those modeled monomers are likely to assemble into vaults.

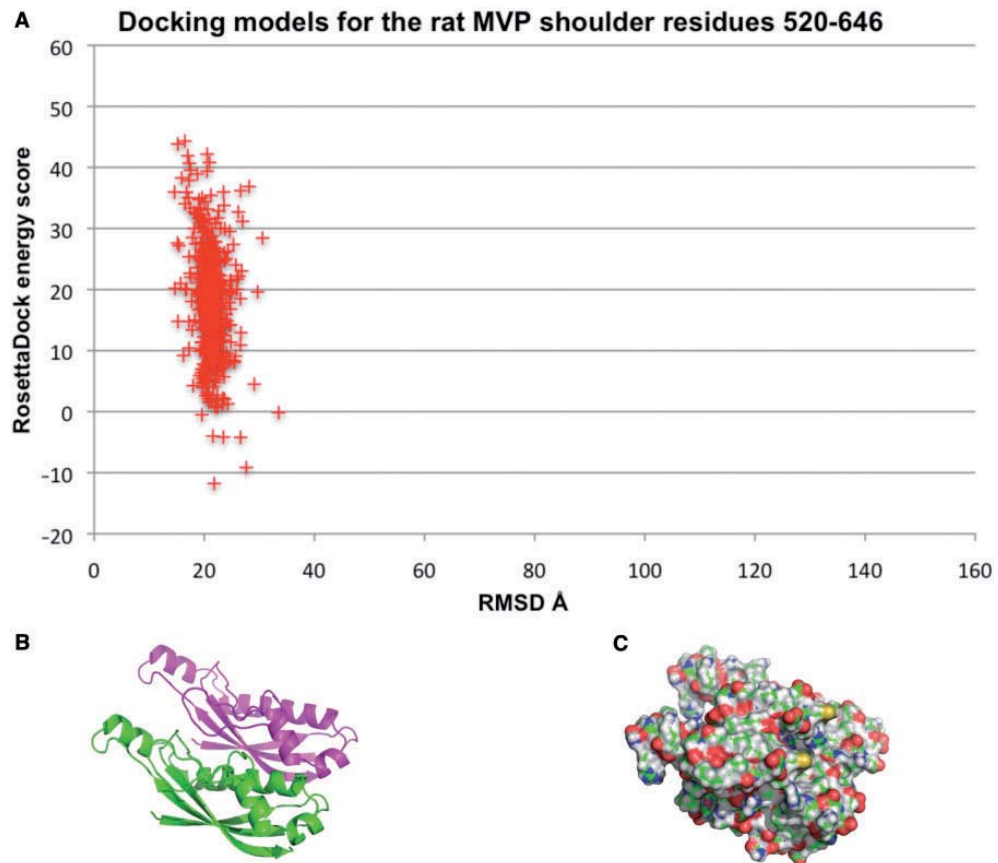
Within the vault, MVP monomers contact their adjacent monomers laterally, but vaults are also able to open in a petal-like fashion from their equator (Kedersha et al. 1991; Yang et al. 2010), potentially complicating the docking analysis. Indeed, less than a third of the lateral noncovalent interactions between MVP monomers in the vault occur between the N-terminus and the shoulder domain (residues 1–519) with oligomerization dominated by interactions between the C-terminal cap-helix regions (van Zon et al. 2002;



**FIG. 4.**—RosettaDock results from the crystal structure cap-helix. (A) Score graph depicting RosettaDock energy score versus RMSD (Å) of the docked monomers compared with their starting positions. The funnel shape of the score graph indicates a high confidence in the structure of the models with lowest energy score. (B) Cartoon of the lowest energy model (energy score -264) shaded by monomer. (C) Surface rendering of the lowest energy model.

Tanaka et al. 2009). This is demonstrated for the rat crystal structure MVP monomer by less favorable RosettaDock energy scores for the docking of the N-terminal sections of the monomer compared with the C-terminal shoulder and cap-helix consistent. In the case of 2ZUO\*b, all 10 top models were docked along the length of the monomer. The RosettaDock output files list the pair energies across the interface; one of the 2ZUO\*b cap-helix models showed residues paired as described for the crystal structure (Tanaka et al. 2009). However, the other RosettaDock output models, even those that included the shoulder—which could be expected to align the helix in position, showed various pairings.

This indicates either some redundancy in the docking arrangements between the monomers in the shoulder and cap-helix or a lack of fine resolution in the RosettaDock prediction—given that the residues that interact across the oligomerization interface (identified in the crystal structure) are well conserved (see MVP sequence alignment marked with known interactions, [supplementary material S2](#), [Supplementary Material](#) online). The remaining MVP positive control sequences were analyzed in the same way (table 1 and [supplementary material S1](#), [Supplementary Material](#) online), predicting the formation of vault particles. For the other positive controls, including full-length I-TASSER-modeled rat MVP, the RosettaDock



**Fig. 5.**—RosettaDock results from the rat MVP shoulder region. (A) Score graph representing the RosettaDock energy scores versus RMSD (Å) for the 1,000 models generated by RosettaDock for the shoulder region of MVP (residues 520–646). The energy score for the shoulder region docking is higher than for the cap-helix (table 1). (B) Cartoon of the shoulder domain from the lowest energy model of the two docked monomers (energy score –12) shaded by chain. (C) Surface rendering of the lowest energy docked monomers.

energy score for the repeat sections was significantly lower, that is, more favorable than the shoulder/cap-helix regions (table 1). Thus, the results are consistent with I-TASSER and RosettaDock being able to detect genuine vaults.

#### Negative Controls

As a negative control, the full-length rat MVP sequence was randomized in three fragments: repeat domains, shoulder domain, and cap-helix. Randomization was confined within each fragment to determine whether the cap-helix was having an undue influence regarding the I-TASSER modeling, because this region strongly influenced the BLAST results (mentioned earlier). Two remote sequences found in BLAST searches were used as additional negative controls: rat myosin 1A (a similar sized protein to MVP) (Q62774) that does not

have an experimentally determined structure and human merlin, (P35240) a neurofibromatosis-2 tumor suppressor that has the structure of its FERM domain determined (PDB 3U8Z). All sequences were subject to the same protocol and submitted to I-TASSER with and without constraints to the rat crystal structure 2ZUO\*b.

As expected, the randomized rat MVP sequence could not be modeled on any existing structural template with confidence. The top scoring models, based on human importin  $\beta$  (PDB 1QGR), were of low confidence (table 1; C score –1.76 and constraint by 2ZUO\*b reduced this to –2.93) and so not considered a "correct fold" by I-TASSER. The rat myosin 1A sequence was identified as most structurally similar to the inhibited state of myosin V (PDB 2DFS) with reasonable confidence regardless of the 2ZUO\*b constraint (C score 0.96, and 0.62 with constraint) (table 1 and fig. 6B). Additionally, Phyre<sup>2</sup>

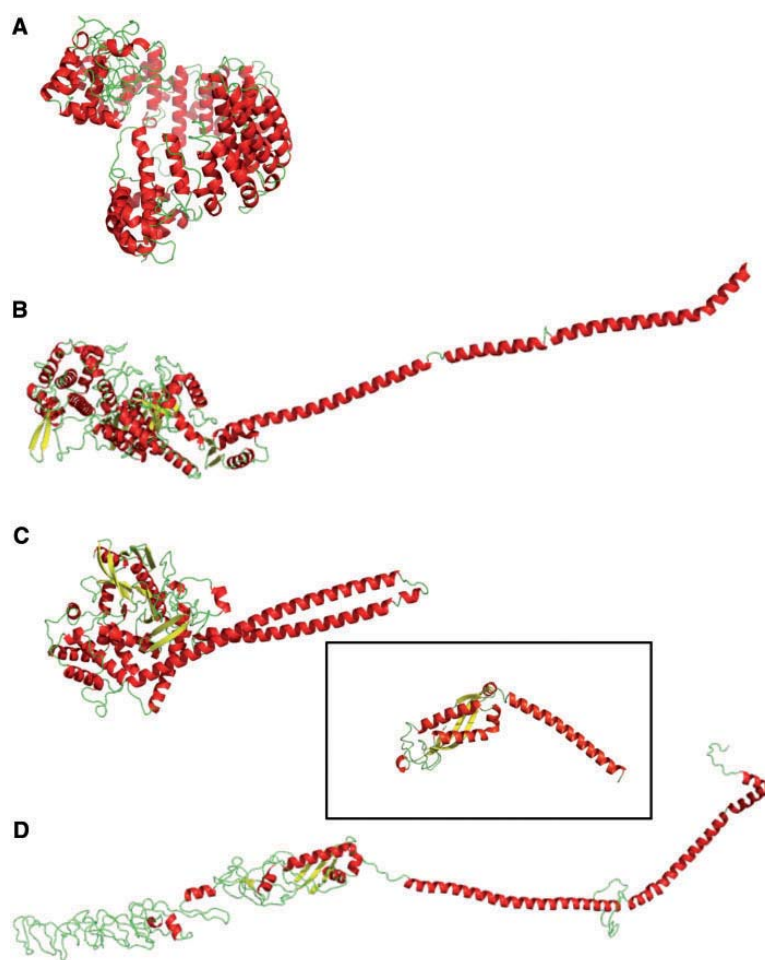


could not report a model for the randomized rat sequence and also identified myosin V as the most similar template for the myosin 1A sequence.

However, using the 2ZUO\*b constraint did influence the structural prediction for the merlin protein sequence (fig. 6C unconstrained, and fig. 6D constrained, by 2ZUO\*b). Although there is a crystal structure for the FERM domain, I-TASSER predicted the unconstrained sequence to be more similar to the merlin homolog in the armyworm caterpillar (PDB 2ILKA) presumably because this is full length rather than the 300 residues of the FERM domain. The shoulder domain in the 2ZUO\*b constrained prediction does look very similar to the MVP shoulder, which was identified as similar to the stomatin core of *Pyrococcus horikoshii* (Tanaka

et al. 2009) (see fig. 6 insert). Phyre<sup>2</sup> identified the merlin sequence specifically as moesin (the fourth part of the FERM domain) from the armyworm (PDB 2ILJA) as their first rated sequence, although the human merlin FERM domain was identified with 100% confidence and 100% coverage but presumably not given the top rating because the sequence was significantly longer than the PDB structure.

Because I-TASSER did not predict that a coil, similar in any way to the cap-helix, would form with the randomized rat MVP sequence, RosettaDock modeling was not carried out. However, the rat myosin 1A was predicted to form a coil structure similar to MVP, so this modeled structure was aligned via FATCAT to monomer positions b and d of the vault complex and submitted to RosettaDock. In this case,



**FIG. 6.**—I-TASSER modeling results for the negative control sequences. (A) Randomized rat MVP. (B) Rat myosin 1A. (C) Human merlin unconstrained. (D) Human merlin constrained by 2ZUO\*b. Insert is the stomatin core from *Pyrococcus horikoshii*.

the lowest energy model did dock along the length of the coiled coil, residues 675–815 (supplementary material S3, Supplementary Material online). The myosin motor domain was submitted to RosettaDock, and this also docked along its length, though with much higher (and positive!) energy scores (lowest energy score +1,020). The putative cap-helix for the 2ZUO\*b constrained human merlin model only partially docked, due to an interruption in the coiled structure (residues 476–513). The shoulder area and truncated cap-helix (residues 315–475) were resubmitted and docked with an energy score of –113. Residues 1–407 representing a combination of the shoulder and the relatively unstructured sequence (in comparison with the repeated  $\beta$  sheets of MVP) were also predicted to dock laterally along its entire length though with a high energy score of +573 (see supplementary material S4, Supplementary Material online). This demonstrates that not all proteins with some homology with MVP (as they were retrieved using BLAST) will be predicted to fold similar to MVP even using the known rat crystal structure as a constraint. In the case of the merlin protein, where the rat constraint did influence the structures output by I-TASSER, it was then very difficult to dock identical monomers in RosettaDock. Thus, it is important that a suite of approaches is used to test structural homology.

#### Investigation of MVP Sequences from *N. gruberi*

Next we used the protocol to find MVP sequences in other genera. Initial BLASTp searches resulted in hundreds of putative MVP sequences, which were reduced to a data set of those with *E* value reported as "zero" and of a similar length (~850 residues) to the complete rat MVP sequence. No sequences matching these criteria were found from the ecdysozoa, or from fungi, but some were from kinetoplasts (excavates), some oomycetes (stramenopiles), and

paramecium (an alveolate). With the criteria relaxed to include sequences with an *E* value up to 10, and any length, then the most remote (compared with rat) excavate sequence that has any kind of MVP annotation was found in *N. gruberi*, an excavate of the clade Heterolobosea, thought to be a very anciently diverged free-living protist. *Naegleria gruberi* has two putative MVP-like protein sequences with an initial PfamA (Finn et al. 2010) annotation of an "MVP shoulder domain" (UniProtKB:D2V5B9, which may not be complete, and D2W0Z9, which is described as "complete"). These two sequences are considerably shorter, 559 and 530 residues, respectively, and contain 17% (148/861) and 19% (166/861) identical sites compared with rat MVP. The size difference is mainly in the body of the vault with *N. gruberi* having fewer repeats domains, suggesting that either repeats have been gained in metazoa since their ancestors diverged from Heterolobosea or that *N. gruberi* has lost a region of the gene within the repeat section compared with the longer characterized MVP sequences. The sequence similarity between these two *N. gruberi* proteins is 35%, indicating that they have been evolving independently for a long time. If the rat MVP repeat region sequence is truncated in an equivalent manner, the percentage of identical sites rises to 25% in both cases (148/588).

The free living *N. gruberi* is often considered to be a representative genome present at a very early stage of eukaryote evolution (Fritz-Laylin et al. 2010). It is predicted to have 15,727 protein coding genes, 3,784 of these are found in at least three other eukaryotic supergroups and a further 349 are found in at least one other supergroup. In contrast, parasitic protists have a reduced genome, relative to their ancestors, owing to their lifestyle. I-TASSER modeled the *N. gruberi* putative MVP sequences into MVP folds with high TM and C scores both unconstrained and constrained by the 2ZUO\*b template (table 2 and fig. 7). In both instances, the models

**Table 2**  
I-TASSER and RosettaDock Results for the *Naegleria gruberi* Sequences

UniProtKB Accession Number	Length	% Identical Sites versus Q62667	I-TASSER C Score	I-TASSER TM Score	RosettaDock Score For Cap-Helix	RosettaDock Score for Shoulder and Cap-Helix
<b>Sequences submitted to I-TASSER without constraint</b>						
D2V5B9	559	17	–0.74	0.62 ± 0.14	–227	–441
D2W0Z9	530	19	0.07	0.70 ± 0.12	–226	–113
<b>Sequences submitted to I-TASSER constrained by 2ZUO*b rat crystal structure</b>						
D2V5B9 <sup>a</sup>	559	17	0.98	0.85 ± 0.08	–209	–438
D2W0Z9	530	19	–0.26	0.68 ± 0.12	–287	–113
D2UZF7	845	13	–1.29	0.55 ± 0.15	No cap-helix	—
D2VSY6	833	13	–2.03	0.56 ± 0.15	None dock	None dock
D2VC38	694	16	–0.24	0.72 ± 0.11	–197	None dock
D2VH38	418	13	–3.15	0.36 ± 0.12	–165	None dock

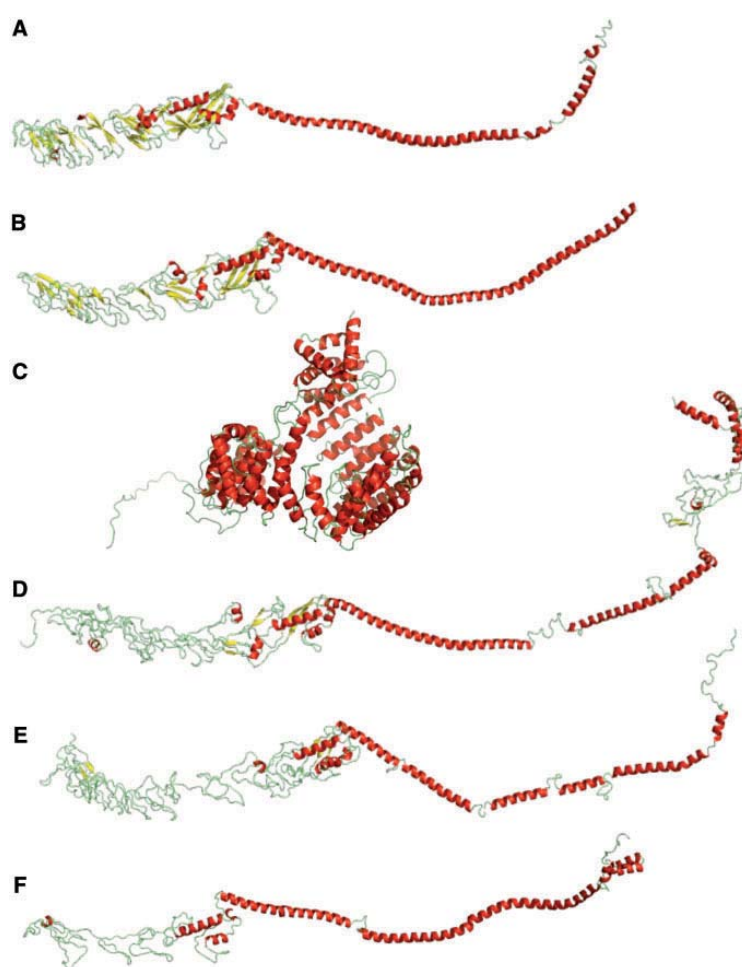
Note.—For D2V5B9<sup>a</sup> constrained by 2ZUO\*b, the lowest 10 energy score models did not dock. Docked models were identified from the expected RMSD and were 47th and 45th lowest energy, respectively. In both cases, the energy scores were all very similar, and there was no compelling consensus model (see supplementary material S5, Supplementary Material online). The highlighted gray cells are scores for I-TASSER predictions that do not resemble the MVP fold being structurally similar to human importin  $\beta$ .

(fig. 7A and B) clearly resembled the MVP structures from figure 2.

Because trypanosomes and leishmania have multiple copies of MVP homolog, it was hypothesized that *N. gruberi* may also have sequences not found by BLASTp, so a PSI-BLAST was conducted (Altschul et al. 1997) (Schäffer et al. 2001) using the first 625 residues of the *L. major* control sequence Q4QJJ7 as the query (see Materials and Methods). Four more *N. gruberi* sequences were retrieved with limited similarity (maximum 16%) to either rat or *L. major* MVP (table 2). All these were first submitted to LOMETS rather than I-TASSER in the interests of speed, but none of the resulting models

predicted a structure that resembled MVP. As a further test, the sequences were submitted to I-TASSER constrained by 2ZUO\*b (fig. 7C–F).

Of these additional sequences, only D2VC38 (694 residues; fig. 7E) is modeled by I-TASSER to resemble MVP with a C score indicating confidence in the model. Although this is the second- "best" model from I-TASSER, the C score is equivalent to the first model, and the cluster density is similar for both models with a similar number of decoys, meaning that the distinction between the two structurally dissimilar models is not certain (detailed in Materials and Methods). Although the model resembles MVP, there are clearly  $\beta$  sheets absent from



**Fig. 7.**—*Naegleria gruberi* MVP I-TASSER structural modeling. (A) D2V5B9, 559 residues. (B) D2W0Z9, 530 residues both identified from a BLASTp search of the UniProtKB database and submitted to I-TASSER without constraint. (C–F) Models derived from sequences retrieved via a PSI-BLAST of the National Center for Biotechnology Information (NCBI) database and submitted to I-TASSER constrained by the rat crystal structure 2ZUO\*b. (C) DZUF7, 845 residues. (D) D2VSY6, 833 residues. (E) D2VC38, 694 residues. (F) D2VH38, 418 residues. UniProt accession numbers are provided for consistency. See also table 2.



the repeat domains. Sequence D2VSY6 (833 residues; fig. 7D) was also modeled resembling MVP. However, in this instance, the C scores of all the models are considerably lower and with a greater difference between the first and second models. The cluster density between these two models is also lower indicating that this prediction is probably no more likely than a random prediction. It could be that the extra C-terminal residues have contributed to the poor C score, even though experimentally a vault can still form with additional C-terminal residues. Interestingly, Phyre<sup>2</sup> identified the shortest sequence (D2VH38) as the bacterial transmembrane protein colicin Ia. A sequence identified as a "colicin uptake transmembrane protein" found in cyanobacteria *Lyngbya majuscula* (F4Y3B4) has 54% sequence homology with rat MVP and is predicted to fold identically to MVP by I-TASSER and Phyre<sup>2</sup>. F4Y3B4 is annotated by family and domain databases Pfam, InterPro, and PROSITE as MVP (see Discussion).

#### Quaternary Structure Prediction for *N. gruberi* Sequences

The *N. gruberi* sequences were submitted to RosettaDock in two fragments. Although it is possible that putative *N. gruberi* vaults are hetero-oligomeric, as found for dictyostelids, the RosettaDock modeling indicates that the shorter D2W0Z9 monomers dock along their entire length more readily than do either D2V5B9 or combinations of both (table 2, combination data not shown). Although the energy score graphs do not demonstrate a clear funnel, and therefore less consensus among the models generated by RosettaDock, the energy scores of the docked models are similar to those of the positive control models.

Constraining the *N. gruberi* sequences (D2V5B9 and D2W0Z9) in I-TASSER by 2ZUO\*b reduced the models propensity to dock in RosettaDock. Constrained D2V5B9 models were identified by their RMSD—which could be predicted as we knew their starting distance apart, rather than by their energy score, as the energy scores were very similar and the consensus poor. We know from the control studies that the constraint can adversely affect the monomer docking depending on the sequence divergence between the query and constraint structure. It may be that if a constraint needs to be used for very remote sequences such as those found via PSI-BLAST, it would be an improvement to use a high confidence I-TASSER output model from a more closely related species as a constraint in preference over a structure from the PDB. The poorer docking of the more remote *N. gruberi* putative MVP sequences is likely due to the greater divergence of sequence and structure resulting in inaccuracies in monomer modeling. For example, in D2VSY6 (fig. 7D), the interruption to the helical structure within the cap-helix section is hindering docking (table 2 and fig. 7). The failure of the rat constraint to improve modeling also reflects this divergence from mammalian MVP sequences.

Given all these results, we propose that *N. gruberi* is capable of making a vault complex with either D2V5B9 or D2W0Z9, both genes have recently been provisionally (and independently of ourselves) reannotated as *mvp* (05/16/12) with the repeat areas additionally annotated as such, thus supporting our results. When used as the query sequence in a UniProt:KB BLAST at default settings, these sequences identify all known MVP sequences. The I-TASSER C scores indicate high confidence that the modeled MVP folds are correct and the predicted structures dock along their entire length in RosettaDock. The more remote sequences from *N. gruberi* (DZUF7, D2VSY6, D2VC38, and D2VH38) appear unlikely to be genuine MVP homologs or have diverged significantly from an ancestral MVP sequence. None of DZUF7, D2VSY6, D2VC38, and D2VH38 retrieves any MVP sequences when used as the query sequence in a BLAST at default settings, and although there is some evidence of lateral docking between monomers, this is most likely due to a natural tendency for coils to interact, and the docking does not extend over the entire length as is required for vault formation.

Excavate databases were searched using PSI-BLASTs independently to retrieve sequences with even the slightest resemblance to MVP. Putative MVP sequences from the parasites *Giardia intestinalis* (UniProtKB:C6LY21) and *Trichomonas vaginalis* (UniProtKB:A2FTW3) were also retrieved, but I-TASSER did not identify any kind of convincing MVP structural homolog (data not shown). Interestingly though, a BLAST search with the *G. intestinalis* putative MVP sequence retrieves MVP from both rat and cow within default parameters (*E* values: 9.3 and 4.2, respectively). Additionally, excavate genome databases were searched using the gene sequences from *L. major* and *T. cruzi* without resulting in any hits other than in trypanosomes and leishmanias where in excess of 50 sequences were retrieved. It has been suggested that the trypanosomes evolved from within the bodonids (euglenozoa) (Deschamps et al. 2011). The *Bodo saltans* annotation is incomplete, but if an MVP homolog exists, we should have expected to retrieve something of it. The lack of any readily identifiable putative MVP homolog in any other excavate, based on currently available sequences, is very intriguing. We therefore conclude that even though some protein sequence homology exists within other excavates, our 3D studies indicate that there is no current evidence that other sequenced excavates are capable of forming a vault particle.

## Discussion

### Three-Dimensional Methodology

The approach described here, using protein structure modeling and docking algorithms, was developed to help answer the question as to the extent that tertiary and quaternary structures will aid the identification of homologous proteins.

The particular application is the question in which species do we find genuine MVP, and if we do, will the MVP monomers form a vault? In this case, BLASTing provides valuable data on the presence of MVP homologs but does not inform directly on the likelihood of any identified MVP monomers assembling into vaults. In general, we need to use more comprehensive methods to demonstrate that limited sequence identity does not preclude vault formation. Here, we show that both tertiary and quaternary structures can be used in addition to information from primary sequences.

It could be argued that the sequence similarity is sufficient for protein prediction servers to be biased toward presenting a structure that is more similar to MVP because there are insufficient alternative templates. However, there are a number of solved structures that could reasonably be ascribed to these sequences, for example, TolA, the stomatin core, band 7 proteins, flotillin, and the colicin membrane spanning protein identified by Phyre<sup>2</sup>. These may hint at possible ancestry for MVP though all are bacterial proteins. Searching for vault specific domains, for example, shoulder or repeats in Pfam (Finn et al. 2010) results in far fewer putative homologs than the BLAST searches. This is undoubtedly because annotation lags far behind sequencing.

It may also be argued that once the structure of the protein is predicted to be MVP-like, then RosettaDock is more likely to find that it does dock. In fact, coiled-coil motifs are likely to dock though usually through twisted supercoiling (Burkhard et al. 2001) rather than lateral association. The I-TASSER-predicted myosin1A coil motifs are docked by RosettaDock, although this example is oversimplified by the absence of the light chains normally present in vivo. However, we have shown for the newly identified MVPs that the lateral docking extends to the shoulder and repeat sections with energy score not dissimilar to the rat and sea urchin where vaults have been observed to form. In the repeat areas in particular, MVP sequence homology is less than 20% versus rat, and our argument is that only those residues that are essential to maintain the shape and lateral docking have been retained.

Although sequence homology of more than 50% is often predictive of structural homology (Clark et al. 2009; Sawyer et al. 2009), there are instances when structure can be dissimilar even with high sequence homology, for example, the prion protein (Pan et al. 1993) and engineered examples (Gronenborn et al. 1991). In this study, we are looking toward the opposite end of the similarity scale, how slim the sequence homology can be and yet structural similarity "sufficient for function" be retained (Holm and Sander 1997). We use MVP as an example to show that structural prediction analysis can extend sequence homology searches. The principles established here could apply to any protein structure. It is more time consuming to check proposed homologies using structural forms but is readily attainable. An important point is that we should not specify too narrow an assumption of the expected structure of a protein. For example, using the rat

tertiary MVP structure as a constraint appears to hinder the detection of related structures in the very distantly related excavates and can disrupt docking by RosettaDock.

Seeking traditional homologous sequences through BLAST searches takes just a matter of minutes, with PSI-BLAST a little longer. This is partly why the simple BLAST solution is so attractive. However, methods that test whether sequence homology implies similarity of function, using structural approaches that can detect more distantly related homologs, are more computationally expensive. In general, a protein the length of MVP (~860 residues) is estimated by I-TASSER to take 50 h and is limited to one job per IP address. Both LOMETS and Phyre<sup>2</sup> are very much quicker taking a matter of hours but do not give quantitative results such as the C score. LOMETS is limited to one job, but Phyre<sup>2</sup> will accept batch jobs. FATCAT is almost instantaneous, but the RosettaDock server also takes up to 50 h for the 600-residue MVP sections depending on server load. In summary, this is a much slower method than simply BLASTing, but as annotation lags far behind sequencing, we need to go beyond BLASTing and be much more rigorous in our determination of protein homology.

#### Informing Evolutionary Studies

The evolutionary history of the vault MVP should help identify possible past functions and illuminate current thoughts on function. The big picture questions are these: are vaults ancestral, having been retained in some species, but fallen into disrepair or lost beyond all recognition in others, or alternatively have they been comprised parts that had other functions and have come together in a fairly remote eukaryote and vaults formed thereafter? If we could be confident which species have functional vaults, and which do not appear to have need for them, or possibly maintain the MVP monomer for another purpose, we should be able to clarify their role. We can suggest that this exquisite example of form, with no known fundamental function, was in LECA and as putative MVP has also been reported in bacterial genomes (H6L4P8 provisional annotation MVP) could conceivably have been present in the last universal common ancestor LUCA. It seems unlikely that vaults would be present in some very diverse groups (such as kinetoplasts, alveolates, amoebozoans, and metazoans) but not be present in others. Finding a link between species that do not appear to have a need to maintain the vault and whether vRNA is associated with it might illuminate an underlying basic function. Equipped with a personal computer, an internet connection, and a means of viewing pdb files, anyone can extend sequence homology analysis to investigation in three dimensions, and we suggest that in silico analysis should routinely be used to check for presumptive structure relationships between potentially ancestrally related proteins. However, that is the work for the future. In all these studies, we require the power from tertiary and quaternary studies to combine with the power of purely sequence-based

studies to enrich the techniques available for molecular evolution.

## Supplementary Material

Supplementary materials S1–S5 are available at *Genome Biology and Evolution* online (<http://www.gbe.oxfordjournals.org/>).

## Acknowledgments

T.K.D. performed the analyses and wrote the drafts, A.J.S.-S. assisted with the analysis programs, and D.P. designed the original research project. All authors regularly discussed the results and contributed to the final manuscript. This work was supported by internal grants from the Institute of Fundamental Sciences, Massey University.

## Literature Cited

- Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389–3402.
- Berger W, Steiner E, Grusch M, Elbling L, Micksche M. 2009. Vaults and the major vault protein: novel roles in signal pathway regulation and immunity. *Cell Mol Life Sci.* 66:43–61.
- Burkhard P, Stetefeld J, Strelkov SV. 2001. Coiled coils: a highly versatile protein folding motif. *Trends Cell Biol.* 11:82–88.
- Clark AR, Sawyer GM, Robertson SP, Sutherland-Smith AJ. 2009. Skeletal dysplasias due to filamin A mutations result from a gain-of-function mechanism distinct from allelic neurological disorders. *Hum Mol Genet.* 18:4791–4800.
- Collins L, Penny D. 2005. Complex spliceosomal organization ancestral to extant eukaryotes. *Mol Biol Evol.* 22:1053–1066.
- Deschamps P, et al. 2011. Phylogenomic analysis of kinetoplastids supports that trypanosomatids arose from within bodonids. *Mol Biol Evol.* 28:53–58.
- Finn RD, et al. 2010. The Pfam protein families database. *Nucleic Acids Res.* 38:D211–D222.
- Finnigan GC, Hanson-Smith V, Stevens TH, Thornton JW. 2012. Evolution of increased complexity in a molecular machine. *Nature* 481:360–364.
- Fritz-Laylin LK, et al. 2010. The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. *Cell* 140:631–642.
- Gray JJ, et al. 2003. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J Mol Biol.* 331:281–299.
- Gronenborn AM, et al. 1991. A novel highly stable fold of the immunoglobulin binding domain of streptococcal protein-G. *Science* 253: 657–661.
- Hamill DR, Suprenant KA. 1997. Characterization of the sea urchin major vault protein: a possible role for vault ribonucleoprotein particles in nucleocytoplasmic transport. *Dev Biol.* 190:117–128.
- Herrmann C, Zimmermann H, Volkandt W. 1997. Analysis of a cDNA encoding the major vault protein from the electric ray *Discopyge ommata*. *Gene* 188:85–90.
- Holm L, Sander C. 1997. An evolutionary treasure: unification of a broad set of amidohydrolases related to urease. *Proteins* 28:72–82.
- Kedersha NL, Heuser JE, Chugani DC, Rome LH. 1991. Vaults. III. Vault ribonucleoprotein particles open into flower-like structures with octagonal symmetry. *J Cell Biol.* 112:225–235.
- Kedersha NL, Rome LH. 1986. Isolation and characterization of a novel ribonucleoprotein particle—large structures contain a single species of small RNA. *J Cell Biol.* 103:699–709.
- Kelley LA, Sternberg MJ. 2009. Protein structure prediction on the web: a case study using the Phyre server. *Nat Protoc.* 4:363–371.
- Kickhoefer VA, Rome LH. 1994. The sequence of a cDNA encoding the major vault protein from *Rattus norvegicus*. *Gene* 151:257–260.
- Kickhoefer VA, et al. 2009. Targeting vault nanoparticles to specific cell surface receptors. *ACS Nano* 3:27–36.
- Koonin EV. 2010. The incredible expanding ancestor of eukaryotes. *Cell* 140:606–608.
- Kozakov D, Brenke R, Comeau SR, Vajda S. 2006. PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins* 65: 392–406.
- Kozlov G, et al. 2006. Solution structure of a two-repeat fragment of major vault protein. *J Mol Biol.* 356:444–452.
- Kurland CG, Collins LJ, Penny D. 2006. Genomics and the irreducible nature of eukaryote cells. *Science* 312:1011–1014.
- Kuwahara Y, et al. 2009. Unusual thermal disassembly of the SPFH domain oligomer from *Pyrococcus horikoshii*. *Biophys J.* 97:2034–2043.
- Lara PC, Pruschy M, Zimmermann M, Henriquez-Hernandez LA. 2011. MVP and vaults: a role in the radiation response. *Radiat Oncol.* 6: 148.
- Liu B, et al. 2011. Up-regulation of major vault protein in the frontal cortex of patients with intractable frontal lobe epilepsy. *J Neurol Sci.* 308: 88–93.
- Lyskov S, Gray JJ. 2008. The RosettaDock server for local protein-protein docking. *Nucleic Acids Res.* 36:233–238.
- Matsumoto T, et al. 2011. Comprehensive sequence analysis of 24,783 barley full-length cDNAs derived from 12 clone libraries. *Plant Physiol.* 156:20–28.
- Mossel E, Steel M. 2004. A phase transition for a random cluster model on phylogenetic trees. *Math Biosci.* 187:189–203.
- Moult J, Pedersen JT, Judson R, Fidelis K. 1995. A large-scale experiment to assess protein structure prediction methods. *Proteins* 23:ii–iv.
- Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP—a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol.* 247:536–540.
- Neumann N, Lundin D, Poole AM. 2010. Comparative genomic evidence for a complete nuclear pore complex in the last eukaryotic common ancestor. *PLoS One* 5:e13241.
- Orengo CA, et al. 1997. CATH—a hierarchic classification of protein domain structures. *Structure* 5:1093–1108.
- Pan KM, et al. 1993. Conversion of alpha-helices into beta-sheets features in the formation of the scrapie prion proteins. *Proc Natl Acad Sci U S A.* 90:10962–10966.
- Paspalas CD, et al. 2009. Major vault protein is expressed along the nucleus-neurite axis and associates with mRNAs in cortical neurons. *Cereb Cortex.* 19:1666–1677.
- Qian W, He X, Chan E, Xu H, Zhang J. 2011. Measuring the evolutionary rate of protein-protein interaction. *Proc Natl Acad Sci U S A.* 108: 8725–8730.
- Querol-Audi J, et al. 2009. The mechanism of vault opening from the high resolution structure of the N-terminal repeats of MVP. *EMBO J.* 28: 3450–3457.
- Roy A, Kucukural A, Zhang Y. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc.* 5: 725–738.
- Roy A, Xu D, Poisson J, Zhang Y. 2011. A protocol for computer-based protein structure and function prediction. *J Vis Exp.* 57:e3259.
- Sali A, Blundell TL. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol.* 234:779–815.
- Sawyer GM, Clark AR, Robertson SP, Sutherland-Smith AJ. 2009. Disease-associated substitutions in the filamin B actin binding domain confer enhanced actin binding affinity in the absence of major structural disturbance: insights from the crystal structures of filamin B actin binding domains. *J Mol Biol.* 390:1030–1047.

- Schäffer AA, et al. 2001. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* 29:2994–3005.
- Stadler PF, et al. 2009. Evolution of vault RNAs. *Mol Biol Evol.* 26: 1975–1991.
- Steitz TA, Moore PB. 2003. RNA, the first macromolecular catalyst: the ribosome is a ribozyme. *Trends Biochem Sci.* 28:411–418.
- Stephen AG, et al. 2001. Assembly of vault-like particles in insect cells expressing only the major vault protein. *J Biol Chem.* 276: 23217–23220.
- Stevens MI, Hunger SA, Hills SFK, Gemmill CEC. 2007. Phantom hitch-hikers mislead estimates of genetic variation in Antarctic mosses. *Plant Systematics Evol.* 263:191–201.
- Stewart PL, et al. 2005. Sea urchin vault structure, composition, and differential localization during development. *BMC Dev Biol.* 5:3.
- Tanaka H, et al. 2009. The structure of rat liver vault at 3.5 Å resolution. *Science* 323:384–388.
- Tovchigrechko A, Vakser IA. 2006. GRAMM-X public web server for protein-protein docking. *Nucleic Acids Res.* 34:W310–W314.
- van Zon A, et al. 2002. Structural domains of vault proteins: a role for the coiled coil domain in vault assembly. *Biochem Biophys Res Commun.* 291:535–541.
- Vasu SK, Kedersha NL, Rome LH. 1993. cDNA cloning and disruption of the major vault protein alpha gene (*mvpA*) in *Dictyostelium discoideum*. *J Biol Chem.* 268:15356–15360.
- Vasu SK, Rome LH. 1995. Dictyostelium vaults: disruption of the major proteins reveals growth and morphological defects and uncovers a new associated protein. *J Biol Chem.* 270:16588–16594.
- Vollmar F, et al. 2009. Assembly of nuclear pore complexes mediated by major vault protein. *J Cell Sci.* 122:780–786.
- Watson JD, Laskowski RA, Thornton JM. 2005. Predicting protein function from sequence and structural data. *Curr Opin Struct Biol.* 15:275–284.
- Wu S, Zhang Y. 2008. MUSTER: improving protein sequence profile-profile alignments by using multiple sources of structure information. *Proteins* 72:547–556.
- Xu D, Zhang J, Roy A, Zhang Y. 2011. Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. *Proteins* 79: 147–160.
- Yang J, et al. 2010. Vaults are dynamically unconstrained cytoplasmic nanoparticles capable of half vault exchange. *ACS Nano* 4: 7229–7240.
- Ye Y, Godzik A. 2003. Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics* 19:ii246–ii255.
- Zhang Y. 2008. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* 9:409.

Associate editor: Dan Graur

# In Silico Resurrection of the Major Vault Protein Suggests It Is Ancestral in Modern Eukaryotes

Toni K. Daly\*, Andrew J. Sutherland-Smith, and David Penny

Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand

\*Corresponding author: E-mail t.daly1@massey.ac.nz; tonidaly@mac.com.

Accepted: July 17, 2013

## Abstract

Vaults are very large oligomeric ribonucleoproteins conserved among a variety of species. The rat vault 3D structure shows an ovoid oligomeric particle, consisting of 78 major vault protein monomers, each of approximately 861 amino acids. Vaults are probably the largest ribonucleoprotein structures in eukaryote cells, being approximately 70 nm in length with a diameter of 40 nm—the size of three ribosomes and with a lumen capacity of 50 million Å<sup>3</sup>. We use both protein sequences and inferred ancestral sequences for in silico virtual resurrection of tertiary and quaternary structures to search for vaults in a wide variety of eukaryotes. We find that the vault's phylogenetic distribution is widespread in eukaryotes, but is apparently absent in some notable model organisms. Our conclusion from the distribution of vaults is that they were present in the last eukaryote common ancestor but they have apparently been lost from a number of groups including fungi, insects, and probably plants. Our approach of inferring ancestral 3D and quaternary structures is expected to be useful generally.

**Key words:** vault ribonucleoprotein, ancestral reconstruction (ASR), BLAST, I-TASSER, RosettaDock, last eukaryotic common ancestor.

## Introduction

Phylogenetic reconstruction of the last eukaryotic common ancestor (LECA) and ultimately the path of life itself is a goal of evolutionary biologists. Molecular phylogenetics has sped up this search and has shown that LECA had many more properties than simply a nucleus and mitochondria. For example, LECA had linear genetic material, essential for meiosis and the advantages that sex and recombination bring (Ishikawa and Naito 1999), but it does lead to the issue of terminal erosion of chromosomes. Although there are a number of fixes, the telomerase complex is the standard caretaker of eukaryote telomeres (Nosek et al. 2006) and is also ancestral. LECA already had introns and a complex spliceosome to process them (Collins and Penny 2005). LECA could synthesize sterols, essential for phagocytosis and cell signaling (Desmond and Gribaldo 2009). If the vault particle were also in LECA, what possible role could it have?

Our interest has been in using in silico methods for inferring 3D structure of proteins (Daly et al. 2013) from tertiary structures determined by standard X-ray crystallography methods. These do not require strict adherence to known 3D structures, they do allow variation, but still based on known structures.

In addition, we have used quaternary structural information, estimating the extent that the tertiary models will assemble into the expected quaternary vault structure. Our approach here is to combine the three methods: searching for proteins that are widespread in eukaryotes by BLAST searches; using I-TASSER to test that the sequences found by BLAST searches (or their inferred ancestral sequences) will really fold into the expected tertiary structures; and using RosettaDock to infer quaternary structure.

In addition, the search for proteins widespread in eukaryotes has recently been extended to allow for some losses in specific lineages (Tabach et al. 2013). These authors reported a loss of some proteins, particularly a loss of homologs to the ciliated sensory ending component (BBS-1) in plants and fungi. However, this loss did not affect the conclusion that BBS-1 proteins were ancestral in eukaryotes. Our primary contribution here is to consider tertiary and quaternary structure in the search for ancestral eukaryote proteins, especially the vaults. Allowing loss of a few ancestral proteins from specific groups is an important advance.

Currently, we do not know the full details of deeper eukaryote phylogeny, probably because the ability of

© The Author(s) 2013. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com



Markov models to reconstruct sequences (or the tree) falls off exponentially at deeper times (Mossel and Steel 2005). This means that the definite relationship of the main groups of eukaryotes is not yet known. There had been hints that the root could be within the excavates in 2007 (Rodríguez-Ezpeleta et al. 2007). More recently, Cavalier-Smith (2010) proposed that the root of the eukaryotic ancestor lay between euglenozoa and the rest of the eukaryotes, that is, within the excavates, breaking euglenozoa away from excavata. However, with equal confidence, he had previously favored the root between the opisthokonts (animals plus fungi = fungamals) and all other eukaryotes (Stechmann and Cavalier-Smith 2003).

Regardless of the placement of the root, the accepted approach is to find features that are in all the major groups of eukaryotes, for example, those defined by Keeling et al. (2005). These are considered to be 1) Opisthokonts (fungi and animals) and the amoebozoa (arguably a supergroup of their own); 2) Plants (Plantae); 3) Excavates (such as *Naegleria*, *Trichomonas*, *Giardia*, and also including Euglenoids); 4) Stramenopiles and Alveolates, together known as chromalveolates; and 5) Rhizaria, these include Radiolarians and Foraminifera. More recently, the chromalveolates have been grouped with Rhizaria forming a supergroup known as SAR (Stramenopile, Alveolate, Rhizaria) (Burki et al. 2007; Elias et al. 2009). Our strategy is to identify vaults in as many of these groups as possible, especially because vaults appear to have been lost in several significant groups of eukaryotes (see later).

With the publication of the *Naegleria gruberi* genome (Fritz-Laylin et al. 2010), the number of genes in the putative LECA increased by 700 additional genes to over 4,000, summarized by Koonin (2010). This is likely to be a conservative estimate, as it does not account for gene loss in a few lineages. Using our in silico protocol of searching for remote sequence homologs of the major vault protein (MVP) monomer, and assessing their putative tertiary and quaternary structure, we found that there were at least two plausible candidate genes in *N. gruberi* encoding proteins predicted to fold as MVP (UniProtKB:D2V5B9 and D2W0Z9) and we predicted that they would ultimately form a vault particle (Daly et al. 2013).

If we are to propose that a particle such as the vault were present in LECA, we would anticipate that many signals from sequence homology would have been largely erased by successive substitutions. We expect that the tertiary and quaternary structural information would persist longer even where other primary sequence becomes randomized (Mossel and Steel 2005). The distribution of amino acid substitution is not random because some residues are essential for tertiary and quaternary structure, so consequently we expect them to be more highly conserved than residues of lesser structural import.

The primary (initial) function of vaults is not known. Extant vaults are associated with signaling pathways (Berger

et al. 2009) they are known to be upregulated in treatment resistant cancers (Herlevsen et al. 2007) and epilepsy (Liu et al. 2011), but these clearly were not their original role in protists. Vaults are enriched in tissue types that are involved with scavenging such as in macrophages (Chugani et al. 1991) and in lipid rafts (Kowalski et al. 2007). They have been observed containing cargo (suggested to be mRNA [Paspalas et al. 2009]), and sea urchin vaults appear to contain many proteins (Stewart et al. 2005). Researchers are now using vaults to deliver cargo such as vaccines (Champion et al. 2009) and drugs (Buehler et al. 2011) to targeted cells. Again, carrying vaccines is clearly not their original function, nevertheless, everything we can learn about their distribution and functions will help understand their original role.

As mentioned earlier, we use a 3-fold protocol for inferring structure using BLASTp, I-TASSER, and RosettaDock. First, a BLASTp search of the UniProt and NCBI protein sequence databases is used to find potential MVP homologs. Second, the tertiary structure of these sequences is predicted by I-TASSER (iterative threading assembly refinement server) (Zhang 2008). This program uses a combination of protein structure prediction techniques to produce potential models of secondary and tertiary structures for a given sequence, based on a structural template. If there is structural similarity to an MVP monomer, we anticipate that I-TASSER will predict the greatest similarity to the rat MVP, as it is the only 3D crystallographic structure in the Protein Data Bank that is almost full-length. For this reason, we have used the rat sequence (UniProtKB:Q62667) as the standard against which others are measured. Third, the output structures from I-TASSER are then submitted to RosettaDock (Lyskov and Gray 2008) to determine whether the predicted monomeric MVP structures are expected to assemble into vaults. Because of this potential for some groups to lose vaults, it is important to test the predicted tertiary and quaternary structure of vault sequences to see that they really do fold and dock in the expected manner (Daly et al. 2013).

Sequences that meet the three criteria of being more or less complete at the primary sequence level, structurally creditable as MVP monomers similar to the rat structure (Tanaka et al. 2009; the only complete crystal structure in the protein data bank), and are predicted by RosettaDock to dock laterally, were grouped phylogenetically and used to infer ancestral MVP sequences using PAML4 (Phylogenetic Analysis by Maximum Likelihood) (Yang 2007) and FastML (Ashkenazy et al. 2012). When we reconstructed the putative ancestor of all eukaryotes, we added Mega5 (Tamura et al. 2011) to try to limit bias shown by both PAML4 and FastML described later. There are more sequences potentially available than have ultimately been used because it is often difficult to decide whether a sequence annotated as complete really is, when there are apparent gaps. There is also the issue of orphan sequences, where there is a cDNA or mRNA sequence that is not ascribed to a gene. In some cases, these are placed

on a branch on a phylogenetic tree consistent with established taxonomy increasing the likelihood of being correct. In other instances, the position on a tree seems so unlikely that RNA contamination must be suspected, instances of these problems are described later.

Because some groups of eukaryotes may lack vaults, we make an additional test and infer ancestral sequences for a broader group, for example, invertebrates—(though insects appear to lack vaults). We then undertake the same tertiary and quaternary structure predictions of the inferred ancestral sequences. Ancestral reconstruction (ASR) takes a multiple sequence alignment (MSA) (nucleotides or proteins), together with a tree representing the sequences in the MSA and calculates the most likely ancestral sequence at each node of the tree. For example, ASR has been used to calculate a putative ancestral RNase P sequence to submit as a BLAST query in the search for evolutionarily distant protein homologs (Collins et al. 2003). Usually, the ASRs retrieved known sequences for proteins associated with RNase P with a higher *E* value than by BLASTing with known sequences; in one instance a protein homolog was found in *Giardia lamblia* using the reconstruction that could not be retrieved using any of the known sequences. Ancestral proteins have also been experimentally resurrected (that is, synthesized) from sequences determined by ASR (Chang et al. 2002; Gullberg et al. 2010). However, here we only resurrect MVP in silico, with a combination of two protocols, first by reconstructing the ancestral sequence for each group (using PAML4 and FastML), and second by inferring the structures using I-TASSER. Explicitly, our test is—will the inferred ancestral MVP sequences be as capable of forming vault particles using our modeling protocol, as the extant MVP?

ASR can use a variety of methods, perhaps the most reliable uses posterior probabilities from known trees in their reconstructions (maximum likelihood [ML] and empirical Bayes). Empirical Bayes may overlook the best guess in terms of most likely substitution resulting in slightly less accurate sequence reconstruction but may better preserve structural and functional properties (Williams et al. 2006). An additional form of ASR involving topological empirical Bayes, which weights the trees differently to other methods, has not been found to alter the resultant sequence (Hanson-Smith et al. 2010). We have found that a combination of two ASR algorithms (PAML4 and FastML) combined with human intervention results in a suitable ancestral sequence to put forward for ancestral protein structural prediction, or in silico resurrection.

## Materials and Methods

Full-length MVP sequences were found by BLASTp and PSI BLASTing of the NCBI and UniProtKB databases using the rat MVP sequence (UniProtKB:Q62667) as the query. Using our established protocol of tertiary and quaternary structural modeling (Daly et al. 2013), we retrieved 116 eukaryote and

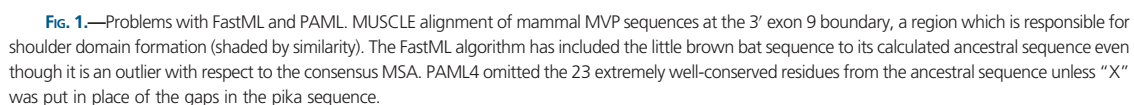
10 bacterial protein sequences that fulfilled the criteria of structural homology with sufficient lateral docking capacity for vault particle formation. Much of the available sequence data had not been subject to detailed analysis, with only few sequences ascribed to a chromosomal position even if they are from genomic DNA. Some MVP sequences are derived from mRNA, for example, cat (*Felis catus*; UniProtKB:Q18PA2), diamondback rattlesnake (*Crotalus adamanteus*; UniProtKB:J3RZY3), and barley (*Hordeum vulgare*; UniProtKB:F2E078). A tree constructed from all protein sequences used in the ancestor of all eukaryotes is available online ([supplementary material S1b, Supplementary Material](#) online).

Each MSA used for ASR was generated by MUSCLE (Edgar 2004). Trees were also calculated for each MSA using MrBayes (Huelsenbeck and Ronquist 2001) with both algorithms run via the Geneious platform (Geneious Pro 5.5.7 Biomatters available from <http://www.geneious.com/>, last accessed August 2, 2013). Most ASR algorithms require a tree formed from the MSA under scrutiny. FastML will calculate a tree from the MSA but we found that MrBayes trees produced the most plausible and reliable trees for submission to both PAML and FastML, although computationally the most expensive. At least four sequences are required for MrBayes tree, which means that any groups with less than four representative sequences could not be used for ASR. Although we do not expect that a tree built from a single gene would necessarily be the same as a tree built using combined gene sequences (Philippe et al. 2011), we have used the method that produces the same tree for MVP for a given species set each time it is calculated to limit systematic error. We also tried calculating the tree using different roots to be certain that the ASR algorithms were being provided with the best initial data. Sequences are continually being added to the databases and the ancestral MVP sequences can be refined. Both PAML4 and FastML ASR methods use ML analysis (empirical Bayes) to estimate the ancestral sequence. However, there are unfortunately significant differences between methods in their handling of sites with missing data. PAML deletes sites in the ancestral sequence where any one sequence contains a gap in the MSA, whereas FastML uses a binary matrix to reconstruct indels and adds them back into regions of more highly conserved sequence. This means that the PAML sequences are shorter, and the FastML sequences much longer. One standard fix with PAML is to use “X” in the position of gaps; this stops the automatic deletion of sites with gaps (PAML FAQ; <http://abacus.gene.ucl.ac.uk/software/paml.html>, last accessed July 2, 2013).

## Limitations of MVP Ancestral Sequence Reconstruction

PAML and FastML generally give identical ancestral sequence for the same input MSA and where there are no gaps. Because gaps result in ambiguity, sequences were checked carefully for completeness to limit the inclusion of sequences

FastML includes insertions in the ancestral sequence even if it is representative of only one species, and so the FastML ancestral sequences are longer. Deleting insertions in the MSA where they are represented in only a single species solved this ambiguity. It is more parsimonious that an insertion has occurred in a single species, than the alternative where deletions have occurred in all other species.





The PAML mammal MVP ancestral sequence, derived from an alignment of 35 mammal MVP sequences, was 740 residues in length while the FastML ancestor from the same MSA had 965 residues. In contrast, the average length of extant MVP sequences is approximately 880 residues. However, replacing missing residues with XXs does not work with large numbers of MVP sequences, or for more diverse sequences such as for invertebrates, as there are gaps in the ASR due to sequence divergence, rather than there being a single sequence region missing.

We took this observation to the extreme in testing how different the PAML and FastML ancestors would be, using 119 MVP sequences to calculate an MVP ancestor of everything. The resulting PAML sequence is 185 residues long in contrast to the FastML ancestral sequence of 1,382 residues in length! It is unrealistic to think that ancestral sequences were generally either significantly shorter or longer. A simpler explanation is that the PAML algorithm has a bias toward reconstructing shorter sequences and FastML to reconstructing longer sequences the more ancient the ancestor becomes. At this point, we added Mega5 (Tamura et al. 2011) to our repertoire of ASR algorithms. Mega5 allows control over the percentage of residues from the MSA that must be considered. When set at use all sites—Mega5 resulted sequences shorter than the total length of the MSA where FastML would have given a results that was as long as the total of the MSA, that is, it was selecting which residue was most likely and not necessarily including residues when they appeared to be restricted to just a few sequences, or deleting all residues where there was a gap in one or few sequences as PAML4 would. This meant that the resultant sequence was a more realistic length. However, it did have a bias of its own, in that it removed some highly conserved sequence that was not present in what could be argued to be more ancient species (discussed later).

FastML additionally has the option of marginal reconstruction (where the residue replaced is based on the posterior probability of the next step at that position from the tree), or joint reconstruction (where the probability is the product of the next two steps, that is, the next most likely substitution is based on two steps rather than one). There were very minor differences at the local level, that is, mammal, invertebrate, and so forth, but the sequence of the ancestor of all eukaryotes then had 17% sequence difference depending on whether marginal or joint reconstruction was used. In practice, the method of reconstruction made little difference to either I-TASSER or RosettaDock. Ultimately, PAML and FastML sequences were combined to generate the final ancestral sequence for each group (described later) for submission to I-TASSER. Inserts unique to just one genus were removed when reconstructing the ancestor of all eukaryotic MVPs—resulting in a more realistic range of ancestor of 678 (PAML)—892 (FastML) residues. A comparison was made

between various methods for reconstructing the overall ancestor (discussed later).

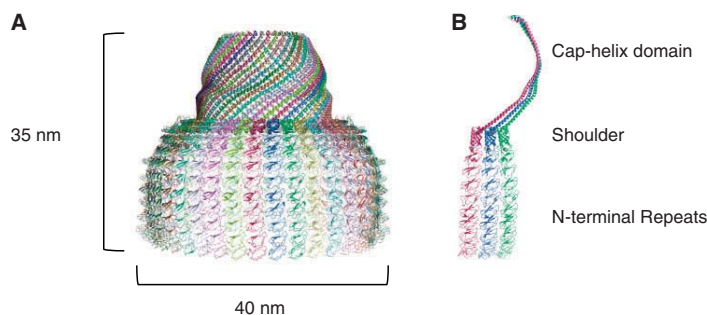
#### Determination of Vault Particle Formation

The completed ASR MVP sequences were analyzed by I-TASSER without constraint (described in Daly et al. 2013) to test that they would be predicted to fold similarly to the rat MVP. Briefly, I-TASSER uses a suite of threading programs known collectively as LOMETS (Wu and Zhang 2007) and outputs up to five structural predictions scored by the confidence in the topology of the model; known as the C score (range is from  $-5$  to  $+2$ ). We have used a C score cut off of greater than  $-1.5$ , which is indicative of a correct fold (Roy et al. 2010). There is an additional score calculated by I-TASSER the template modeling (TM) score that quantifies structural similarity between two superimposed protein structures analogous to the traditional root mean squared difference. A TM score of greater than 0.5 indicates high confidence that the topology of two models, in this case predicted and native (rat) MVP are the same. We have therefore additionally used a TM score of 0.5 or higher as a cut off for inclusion in our analysis.

Then two identical copies of the shoulder and coil domains of each I-TASSER output of ancestral MVP models were submitted to RosettaDock (Gray et al. 2003). For oligomeric vault formation, the crystal structure shows that MVP monomers dock laterally along the length of both sides to make the distinctive barrel shape (Tanaka et al. 2009). RosettaDock uses a low resolution Monte Carlo search and backbone optimization algorithm to optimally position a submitted monomer pair, followed by a refinement to relax the backbone and accommodate the side chains (Gray et al. 2003). Bona fide vault monomers would dock along their entire length with a negative RosettaDock energy score.

Docking a pair of full-length MVP monomers cannot be done via the RosettaDock web server because of a 600 residue limit. It has previously been demonstrated that the coiled coil region is essential for vault formation (van Zon et al. 2002), but we have found that improved in silico docking usually includes the MVP shoulder region as well (Daly et al. 2013). Our test requires that MVP shoulder and coiled coil (known as the cap-helix) would be predicted to dock laterally with an identical monomer, indicating vault particle formation (fig. 2).

Explicitly then, our determination of an MVP monomer is that it will form a vault particle by meeting both our two I-TASSER cut off criteria and then docking with both a negative RosettaDock energy score and the majority of the 1,000 models produced by RosettaDock clustering around this structure. In general, we would expect that such a docked pair of monomers would be in the lowest 10 energy models. All sequences used for the MSAs fulfilled this criteria, as did the ancestors produced by the ASR algorithms (table 1).



**Fig. 2.**—Vault ribonucleoprotein structure. (A) Rat MVP quaternary structure showing half a vault colored by monomer (PDB: 2ZUO, 2ZU4, and 2ZV5). A full vault will have an opposing copy of the upper half vault associated at the N terminii. (B) Three rat MVP monomers (PDB 2ZUO stripped down to three monomers). This figure highlights the extensive lateral association required to dock into the vault quaternary structure. All ribbon diagrams are rendered in PyMol version 1.3.

## Results

### Reconstructing Eukaryote Ancestral MVP Sequences

In the cases of metazoa (63 sequences), amoebzoa (9 sequences), and kinetoplast MVP (29 sequences), complete protein sequences with high homology to the rat crystal structure could be found by simple BLASTp searches using default parameters. There are additionally many more sequences that are fragments of MVP. But in the case of the stramenopiles (e.g., diatoms and oomycetes), there were only just enough sequences from different species to create an ancestor (5 sequences), and although there were five alveolate sequences they came from just two ciliate species: *Paramecium tetraurelia* (3 sequences) and *Oxytricha trifallax* (2 sequences). Sequences that fulfilled the I-TASSER criteria for inclusion (I-TASSER C score of  $>-1.5$  and TM score of  $>0.5$ ) and with a negative RosettaDock energy score (lower is more favorable), but had insufficient representation for ASR, were used as individual sequences. The inferred 3D structures of all of the ancestors are shown later in figure 8 and in the [supplementary material S3a](#) ([Supplementary Material](#) online).

### Metazoa

MSAs were calculated to optimize various phylogenetic groupings. Eventually metazoa were split into; mammals, other sarcopterygii (coelacanth, *Xenopus laevis*, *X. tropicalis*, the Carolina anole, diamondback rattlesnake, chicken, and turkey), fish, and invertebrates ([supplementary material S2a–d](#) [[Supplementary Material](#) online] for these four trees). In the invertebrates, we have representative sequences from sponges (where there are 20 sequences though few are complete), cnidarians, bivalves, annelids, nematodes, and echinoderms. However, we have not found sequences from any arthropods. Although there are a few sequences with limited homology that will fold to resemble parts of MVP, we suggest that the whole vault particle has been lost and the *mvp* gene

degraded beyond recognition in this group. Because the lancelet (an isolated lineage) was difficult to place, it was initially omitted from all of the groups and only added to the final tree of all opisthokonts ([supplementary material S1a](#), [Supplementary Material](#) online).

### Other Opisthokonts

Opisthokonts comprise all metazoa, fungi, plus choanoflagellates, and capsaspora (the latter two are neither animal nor fungi but are closely related and share many gene homologs) (Sebe-Pedros et al. 2011). There were too few MVP sequences to calculate an ancestral sequence for capsaspora or the choanoflagellata. *Capsaspora owczarzaki* is a single cell eukaryote that is neither an animal nor a choanoflagellate but closely related to both and is a symbiont of the freshwater snail *Biomphalaria glabrata*. There are three putative *Cap. owczarzaki* MVP homologs—an insufficient number to infer an ancestor. The choanoflagellates are represented by two species: *Salpingoeca rosetta* and *Monosiga brevicollis*. A single capsaspora MVP sequence and the two choanoflagellate sequences were added to metazoan MVP sequences in the reconstruction of the opisthokont ancestor ([supplementary material S1a](#), [Supplementary Material](#) online). The tree of opisthokonts placed the capsaspora and choanoflagellates within the invertebrates. This is probably a reflection of the increased evolutionary rate of the sponge, *Amphimedon queenslandica*, the parasitic nematodes; *Clonorchis sinensis* and *Schistosoma mansoni* (Tsai et al. 2013), and the tunicate *Oikopleura dioica* (Denoeud et al. 2010). The placement of *Cap. owczarzaki* with hydra is more surprising.

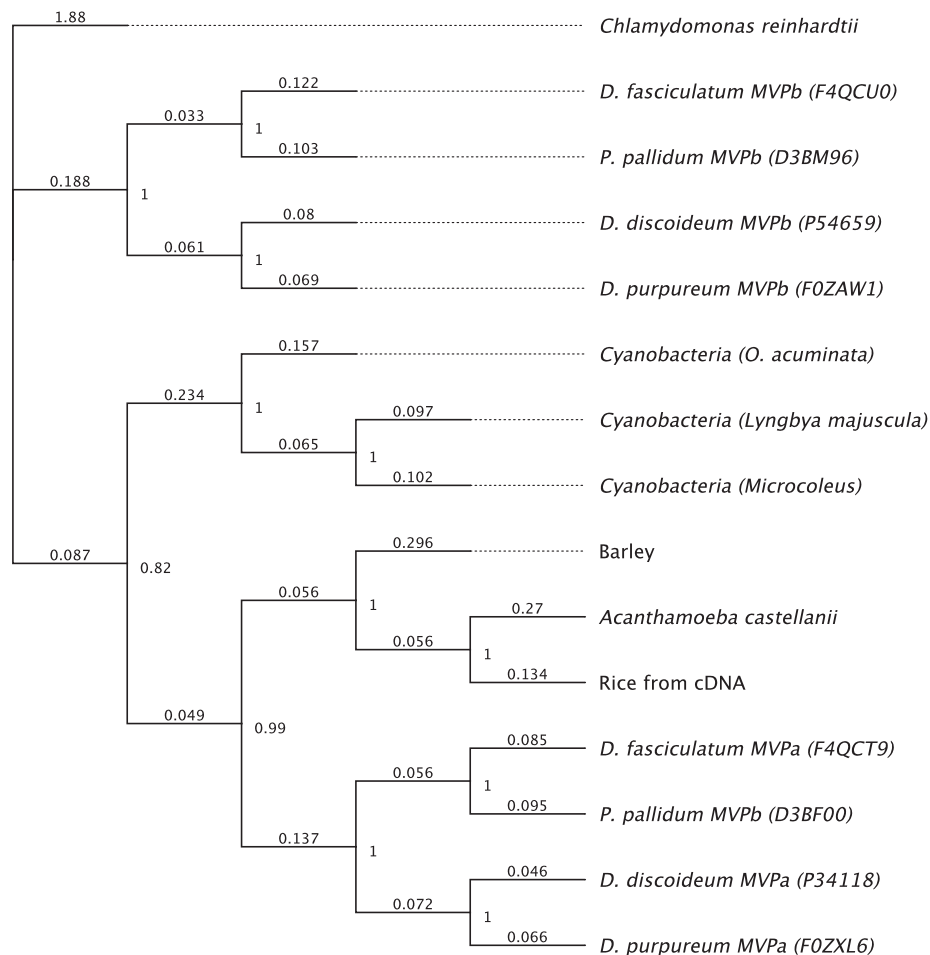
Included within the grouping opisthokont are the fungi. There are proteins that will fold similarly to MVP but these require the constraint of the rat structure when submitted to I-TASSER and do not score within our cut off criteria. Some of these models have been submitted to RosettaDock and show that they will dock although the score is poor in comparison

with metazoa. Additionally, vaults have not been found in fungi and are generally described as missing (Suprenant 2002). The few sequences that we have retrieved are unlikely to form vault particles and we speculate that they may have derived originally from the *mvp* gene and are now significantly diverged, they are annotated as uncharacterized proteins.

### Amoebozoa

An amoebozoan MVP MSA was constructed for MVP $\alpha$  and MVP $\beta$  separately to reconstruct the ancestor of each MVP

form. The dictyostelids form chimeric vaults with proteins from both  $\alpha$  and  $\beta$  genes (Vasu and Rome 1995). Both MVP sequences from *Polysphondylium pallidum* (UniProtKB: P34118 and D3BM96) are annotated MVP $\beta$ ; this is clearly a mis-annotation because P34118 is phylogenetically positioned within the MVP $\alpha$  sequences (fig. 3; [supplementary material S2e](#), [Supplementary Material](#) online). An ancestral MVP sequence was also reconstructed from an MSA of MVP  $\alpha$  and  $\beta$  sequences combined and it is interesting to note that this ancestor docked readily in RosettaDock, indicating that the product from a single original gene could have made a



**FIG. 3.**—MrBayes tree showing the unlikely position of the barley, rice, and, cyanobacteria MVP sequences grouped within the cellular amoebozoa. Because the plant sequences have yet to be attributed to genomic DNA, it seems more likely that they are cDNAs derived from contaminants. The tree shown is rooted by *Chlamydomonas reinhardtii* but other root choices produce the same results. The number by the node represents posterior probability, the number on the branches represents the number of replacements per residue—of course the same residue may have been replaced multiple times. Note that *Polysphondylium pallidum* D3BF00 has been mis-annotated, both *Pol. pallidum* sequences are designated as MVP $\beta$  but D3BF00 has greater sequence similarity with the MVP $\alpha$  sequences of the other amoebozoa.

vault in this group (table 1). In *Dictyostelium discoideum*, knocking out expression of either of MVP $\alpha$  or MVP $\beta$  interferes with vault structure in that the vaults are abnormally ovoid though vaults still form (Vasu and Rome 1995).

### Excavates

There are currently in excess of 30 gene sequences with homology to MVP within the leishmania and trypanosomes (which are grouped within euglenozoa) that fulfill our criteria for vault particle formation and 29 of them are complete. However, neither gene nor protein sequences have been found in any other excavates with the exception of the heterolobosea *N. gruberi*. The situation with excavates has become more complicated in that Cavalier-Smith (2010) has redefined them to be on both sides of his proposed rooting of modern eukaryotes. Our general approach has been to search for MVPs in all major groups of eukaryotes. Nevertheless, we have considered both options, the root being within the traditional excavates, or not. Consequently one of the *Naegleria* sequences has been included in the final tree as an individual but has not contributed to an ancestor other than the all eukaryotic MVP ancestor.

Kinetoplasts fall within euglenozoa and were treated either as two groups, (leishmania and trypanosomes), or as one MSA (and tree) to reconstruct a general kinetoplast ancestor. This was because there is a complex history of gene duplication and the sequences between groups clearly indicate a greater relationship across species in tiers rather than within any individual species ([supplementary material S2f](#), [Supplementary Material](#) online). If the root of the eukaryotic ancestor requires that euglenozoa are removed from the excavates (Cavalier-Smith 2010), the main point is that the vault particle still appears on both sides of the proposed root, even though it may have been lost in a number of lineages.

### Plants

The first land plant sequence identified as an MVP homolog was deposited in databases in March 2011 (Matsumoto et al. 2011) from mRNA of *H. vulgare* (domesticated barley; GenBank:BAK00750 UniProt:F2E078). This was unexpected because up to then no plant had then been shown to have either whole vaults or MVP monomers. Furthermore, the barley MVP sequence has a surprisingly high level of homology (55% identical residues) with rat (UniProtKB:Q62667). A BLAST search of the NCBI plant genomic database using the barley cDNA sequence (NIASHv2093C22) (GenBank:AK369549) resulted in a match to a cDNA sequence in the rice database (*Oryza sativa*) (GenBank:CT836653) with 60% homology to barley, a value considerably lower than expected since both taxa are members of the grass family (Gramineae). However, further BLASTing failed to retrieve either rice genomic DNA or protein sequence. A BLAST of the recently released barley genome (Klaus 2012), using stretches of cDNA

described as MVP from the original find (NIASHv2093C22), has also failed to retrieve any hits.

This could be the result of incomplete genomic sequencing (though to affect the same protein in two species—barley and rice—seems unlikely), or that in both cases the mRNA annotated as MVP, was a contaminant. For example, a DNA extract made from an Antarctic moss using RAPDs (random amplified polymorphic DNA) appeared to be very diverse (Skotnicki et al. 2004); however, it turned out that the DNA of the moss extract came from a mixture of three sources (moss, fungi, and protozoa), and so contamination had occurred from animals living in the clumps of moss (Stevens et al. 2007). Reciprocal BLASTp using the barley MVP protein sequence as the query identifies MVP homologs in UniProt and NCBI with greatest similarity to the cellular slime mold *Pol. pallidum* (UniProtKB:D3BF00), and *D. discoideum* MVP $\alpha$  (UniProtKB:P34118); 60% identical residues shared with each. The translated rice MVP cDNA sequence showed homology to *D. discoideum* MVP $\alpha$  with 68% identical amino acids. I-TASSER predicts that the barley and rice sequences fold into the canonical MVP structure with a greater confidence score than even that of the rat sequence (UniProtKB:Q62667). Additionally, RosettaDock docks identical monomers with a superior (lower) energy score to rat (table 1). So if the barley and rice sequences are truly expressed from the plant genomes, then they are compelling MVP sequences, via linear sequence homology, as well as structure and docking analysis and a functional vault is predicted in both species. However, if they are genuine grass (Graminae) MVPs, their phylogenetic placement within amoebzoa seems unlikely.

Within the amoebzoan tree (fig. 3), three species of Cyanobacteria: *Lyngbya majuscula* (UniProtKB:F4Y3B4), *Oscillatoria acuminata* (UniProtKB: K9TKX8), and *Microcoleus* sp. PCC 7113 (UniProtKB: K9WB38), have homologous sequences described as colicin uptake transmembrane protein that are predicted to fold as MVP. The *Lyngbya* sequence F4Y3B4 is annotated as MVP by InterPro (a membership of 11 protein family databases) (Hunter et al. 2012). The cyanobacterium MVP homologs are more similar in sequence homology to each other (~74%) than to the cellular slime molds (~56%). However, it would not be anticipated that plant or bacterial sequences would group with the amoebzoa. When trees are made of all the individual sequences used in the study, the position of the plant and cyanobacteria sequences remain with the amoebzoa grouping with *Acanthamoeba castellanii* (UniProtKB:L8GQU5), a free living soil protozoa and occasional human pathogen ([supplementary material S1b](#), [Supplementary Material](#) online). In the case of the cyanobacteria, it is possible that horizontal gene transfer (HGT) could be responsible, possibly once from a eukaryote, and then shared between cyanobacteria. Although HGT from eukaryote to prokaryote is rare, there are incidences that have been described (Desmond and Gribaldo 2009; Schönknecht

	F. litoralis...	M. Marina...	O. acumin...	Microcoleu...	L. majuscu...	S. grandis...	P. pacifica...	P. pacifica...	C. coralloi...	H. auranti...
F. litoralis (Bacteroidet...		63.8%	15.4%	15.6%	16.3%	52.6%	52.3%	21.7%	18.1%	10.7%
M. Marina (Bacteroidet...	63.8%		16.4%	15.9%	16.2%	52.8%	52.0%	21.2%	19.0%	10.2%
O. acuminata (Cyanoba...	15.4%	16.4%		74.0%	74.5%	14.8%	16.7%	16.8%	31.0%	12.9%
Microcoleus (Cyanoba...	15.6%	15.9%	74.0%		81.4%	15.7%	16.9%	16.1%	31.9%	13.4%
L. majuscula (cyanoba...	16.3%	16.2%	74.5%	81.4%		15.7%	16.8%	16.5%	31.1%	13.3%
S. grandis (Bacteroidet...	52.6%	52.8%	14.8%	15.7%	15.7%		50.7%	21.1%	17.8%	11.3%
P. pacifica2 (Deltaprot...	52.3%	52.0%	16.7%	16.9%	16.8%	50.7%		23.3%	18.8%	11.3%
P. pacifica (Deltaprote...	21.7%	21.2%	16.8%	16.1%	16.5%	21.1%	23.3%		18.4%	10.6%
C. coralloides (Deltapr...	18.1%	19.0%	31.0%	31.9%	31.1%	17.8%	18.8%	18.4%		14.6%
H. aurantiacus (Chloro...	10.7%	10.2%	12.9%	13.4%	13.3%	11.3%	11.3%	10.6%	14.6%	

**FIG. 4.**—A table produced from an alignment of putative bacterial MVP homologs shaded by the distances as a percentage of identical residues between each pair.

et al. 2013). Clearly, it is necessary to be very careful when attributing a total mRNA extract to just a single species. We suspect that the barley and rice homologs are contaminating sequences from unsequenced amoebozoia. This is our hypothesis until further notice.

We cannot totally discount MVP in land plants, even though the barley and rice sequences appear unlikely. Our prediction is that as additional amoebozoia are sequenced, we will find that one of them has an MVP that is closer to, for example, the barley or the rice sequence. There are a few remote candidate plant MVP sequences: *Petunia integrifolia* (UniProtKB:A9XLF3), *Arabidopsis lyrata* (UniProtKB:D7MVK4), *Zea mays* (UniProtKB:B8A0P4), but all fall far short of our criteria for inclusion as MVP. The only MVP sequence from the super group Plantae that falls within our criteria of folding without constraint in I-TASSER with a C score greater than  $-1.5$  and a TM score greater than  $0.5$ , other than rice and barley is a sequence from the single-celled green algae *Chlamydomonas reinhardtii* (UniProtKB:A8JEL9). Owing to the uncertainty around, and low number of, plant MVP sequences barley, rice, and *Chl. reinhardtii* sequences have been included as individual sequences, but not assigned particularly to plants and were used only in the reconstruction ancestor of all eukaryotic MVP (supplementary material S1b, Supplementary Material online).

Although the validity of the cyanobacterium MVP sequences are uncertain, there are a number of putative MVP homologs found in a variety of bacteria with approximately 16% sequence identity with the cyanobacteria putative MVP, but approximately 25% with all other eukaryotic MVP. These bacterial sequences include the following: *Corallococcus coralloides* (UniProtKB:H8MNI3), *Plesiocystis pacifica* SIR-1 (UniProtKB:A6FXM1), and (UniProtKB:A6FXE2), *Microscilla marina* ATCC 23134 (UniProtKB:A1ZGE7), *Saprospira grandis* (UniProtKB:H6L4P8), *Flexibacter litoralis* (UniProtKB:I4AHY9), and *Herpetosiphon aurantiacus* (UniProtKB:A9AUD4). All sequences are predicted to fold into the shape of MVP according to our I-TASSER criteria and are able to dock in accordance with our RosettaDock criteria. Additionally, *Sap. grandis* has been provisionally annotated as MVP and *Fle. litoralis* as MVP

shoulder domain containing. The matrix (fig. 4) shows the relationship between the bacterial sequences.

*Plesiocystis pacifica* is a fruiting gliding bacterium that has a sterol synthesis pathway related to eukaryotes and the genes are likely to have been acquired by HGT (Desmond and Grihaldo 2009). We now find that it also has two copies of a putative MVP homolog. It is a member of the deltaproteobacteria suggested to be a symbiont with a methanogenic archaea, at the root of eukaryotes (López-García and Moreira 1999) suggesting a possible source of ancestral MVP, though the MVP could also have been acquired from a eukaryote by HGT. We did not include bacterial sequences other than the three cyanobacteria in any ancestral MVP sequence reconstruction.

#### Alveolates

Alveolates fall within the super-group of chromalveolates, or SAR, a group reasoned to be the result of a single endosymbiosis process between a bikont (a protist with two flagella) and a red alga containing a plastid (bestowing the capability of photosynthesis) (Keeling 2004). Although there have been many challenges to the membership of this group; the alveolates and stramenopiles remain core members even though many of the alveolates can no longer photosynthesize (Keeling 2009). So are vault particles also found within the alveolates? *Paramecium tetraurelia* is a well-researched alveolate ciliate that feeds on bacteria, algae, and yeast and has a protein sequence (UniProtKB:A0CI16) containing a domain annotated as MVP shoulder, and is predicted by I-TASSER to adopt the MVP fold with a very high C score of  $+1.13$ . RosettaDock confirms that it is likely to oligomerize, with an energy score of  $-402$  (both scores are more favorable than that of the rat). There are three homologous sequences in *P. tetraurelia*, insufficient to calculate an ancestral sequence; however, sequences added October 31, 2012, from the ciliate *O. trifallax* (UniProtKB:J9IML7 and UniProtKB:J9HVS2) are also predicted to fold as MVP. An ancestral MVP sequence was reconstructed from two *O. trifallax* and three paramecium sequences. Two of the paramecium sequences appear to be fairly recent duplications (UniProtKB:A0CI16 and A0DWW7)



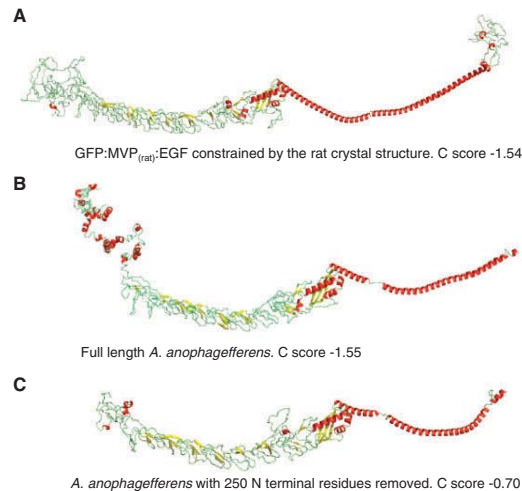
with 95% amino acid identity but the third (UniProtKB: A0EGV2) shows only 42% identity with the other two (supplementary material S2g, Supplementary Material online).

### Stramenopiles

The other main group of the chromalveolates (SAR) are the stramenopiles, with their ancestral sequence reconstructed from five sequences, four from oomycetes: *Phytophthora infestans* (potato blight—annotated as MVP; UniProtKB: D0N745), *Phytophthora sojae* (soybean stem and root rot; UniProtKB: G4Z1M3), and *Phytophthora ramorum* (sudden oak death; UniProtKB: H3G9I8), *Pythium ultimum* (a plant pathogen of many food crops and grasses; UniProtKB: K3X224), and *Aureococcus anophagefferens* (UniProtKB: FOYA32), a harmful algal bloom (supplementary material S2g, Supplementary Material online). The *Aur. anophagefferens* sequence is clearly different to the other stramenopiles, it has approximately 16% similarity with the other stramenopile sequences; however, the fold predicted by I-TASSER is very similar.

The highest scoring I-TASSER model for the complete *Aur. anophagefferens* sequence had a C score of  $-1.55$ , that is, just outside our cut off score of  $-1.5$ . However, one of the reasons for a lower than anticipated C score is the extension of the sequence either at the C or N terminal beyond the rat MVP crystal structure template. Extensions to the core MVP sequence do not necessarily prevent vault formation because vault particles form with Green Fluorescent Protein (GFP) fused to the N terminus of MVP (van Zon et al. 2003), and tags, for example, epidermal growth factor (EGF), added to the C terminal to direct the particle to particular cells (Kickhoefer et al. 2009). Indeed, when the sequence of such an engineered protein (GFP:MVP<sub>(rat)</sub>:EGF) was submitted to I-TASSER the C score was much lower than our cut off score of  $-1.5$ , and the highest C score model predicted did not look convincingly like an MVP. Even when constrained by the rat crystal structure the C score was only  $-1.54$ , still low compared with  $0.42$  for the rat sequence (Q62667) alone (fig. 5A). When the *Aur. anophagefferens* sequence was resubmitted with the N terminal non MVP-like domain truncated it resulted in a score of  $-0.70$ , well above our threshold score (fig. 5B and C).

There are three points to be made here: first, that the C score is affected by the extended sequence presumably because it lacks a template for modeling. Second, the extra sequence may interfere with in silico docking if not predicted to fold correctly. We have confined docking between two monomers to the shoulder and coiled coil regions, as the coil was found to be critical to the vault formation in a yeast two-hybrid system (van Zon et al. 2002) and our previous work shows that pairs of the repeat domain region, in general, will readily dock along their length (Daly et al. 2013). Finally, the *Aur. anophagefferens* sequence has greater homology with the green algal MVP sequence used to root the alveolate

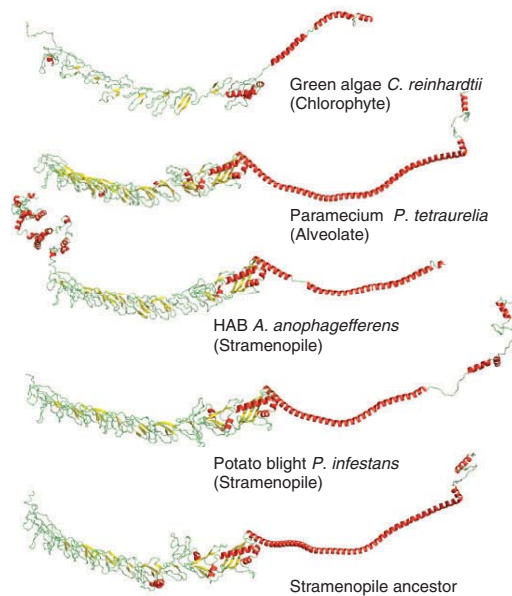


**Fig. 5.**—(A) GFP:MVP<sub>(rat)</sub>:EGF (1,152 residues and constrained by the rat crystal structure 2ZUO\*b). The C score was  $-1.54$  low compared with  $+0.42$  for the rat sequence (UniProtKB: Q62667) alone. (B) *Aureococcus anophagefferens* (UniProtKB: FOYA32) complete sequence 962 residues submitted to I-TASSER without constraint and resulted with a C score of  $-1.55$ . (C) The same sequence with the N terminal 250 residues removed — resulted with a C score of  $-0.70$ .

and stramenopile tree (supplementary material S2g, Supplementary Material online) (26% full length, 35% with the N terminal removed) than with any of the stramenopile or alveolate sequences, full length or truncated. However, structurally I-TASSER predicts *Aur. anophagefferens* MVP to fold more similarly to the other chromalveolate MVPs. The stramenopile ancestral MVP structure is unaffected by the inclusion of the algal bloom sequence with minimal primary sequence homology (fig. 6).

The inclusion of rhizaria as part of the super-group with chromalveolates (known as SAR) is becoming more compelling (Burki et al. 2010; Parfrey et al. 2010). Rhizaria are difficult to culture and consequently underrepresented in sequence databases (Sierra et al. 2013), BLASTing the few genomes sequenced thus far has not retrieved any sequences that resemble MVP.

Finally, a tree was made from both the ancestors (where possible) and the individuals that represented poorly covered families (supplementary material S3a, Supplementary Material online). Initially, the ancestor was comprised of all our eukaryote data set plus the three cyanobacteria—which we had identified as either contamination or gained from eukaryote via HGT. This fulfilled all of our criteria but it could be argued that the number of kinetoplast sequences that were included influenced the resultant ancestral sequence. We therefore limited the number of sequences to one per species. Additionally—because of the issues that affected the output



**FIG. 6.**—Although *Aureococcus anophagefferens* has greater sequence similarity with the green algae, *Chlamydomonas*, I-TASSER predicts that structurally it is more similar to either the ciliate paramecium or to the oomycete *Phytophthora infestans*. The stramenopile ancestor is unaffected structurally by the inclusion of *A. anophagefferens* even though it has very low primary sequence homology with the oomycetes sequences.

from each of the ASR algorithms that we used, making them either unrealistically long (FastML) or unrealistically short (PAML4)—we made ancestors by removing from MSAs, inserts that were present in only one ancestor (columns 2–5), or by deleting inserts represented by just one genus in the MSA of the individuals (columns 6–9). This resulted in sequences that were of a more likely length (number of residues shown) because this removed most of the gaps that the various algorithms dealt with in different ways.

1. Sequences derived from an MSA of the ancestors for the five major groups; amoebozoa, opisthokonts, kinetoplasts, alveolates, and stramenopiles—FastML joint (856 residues) columns 2 and FastML marginal (856 residues) column 3 in figure 7A, Mega5 (853 residues) column 4, PAML (819 residues) column 5. Rat is included in column 1 for comparison.
2. Ancestral sequences reconstructed from individual species using an MSA limited to one sequence per species, with inserts unique to a single genus removed—resulting in FastML joint (892 residues) and marginal (892 residues) sequences columns 6 and 7, Mega5 (770 residues) column 8, and PAML (679 residues) column 9.

Our main point is that regardless of how the ancestor is reconstructed, whether from ancestral sequences from each major

group or from sequences from individual species used all together to make an ancestor, the resultant protein sequence folds and docks within the constraints of our original criteria.

Although the sequences had reduced overall similarity, there were blocks of highly conserved sequence (alignment [supplementary material S3b](#), [Supplementary Material](#) online). Particularly highly conserved is a sequence region close to the C terminus (fig. 7C). The crystal structure for this region has not been resolved but ab initio modeling by MODELLER (Sali and Blundell 1993), part of the I-TASSER suite of programs, consistently predicts the structure depicted in figure 7D. This fold was also found by Phyre<sup>2</sup> (Kelley and Sternberg 2009), which also retrieved known MVP structures.

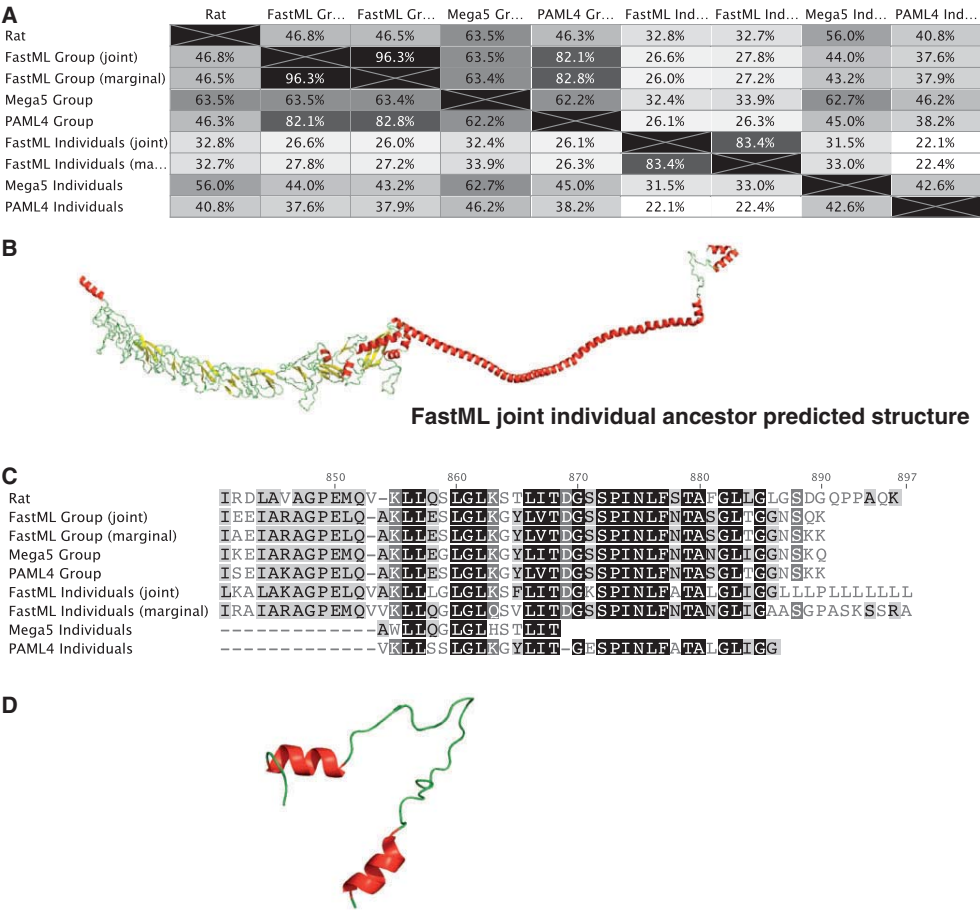
A BLASTp using just these conserved sequences resulted in hits only from known MVP sequences no matter how loose the parameters were. A structural search with the Dali server (Holm and Rosenström 2010) also failed to retrieve any other models with similar folds. This indicates that this sequence is found only in MVP. Could this define MVP? It could be essential for sealing the cap, or to hook the vaults onto cellular structures, vaults have been shown to bind to microtubules via their caps (Eichenmüller et al. 2003). This addition could have expanded the function of the vault from sequestration of ions or molecules to transportation.

This gives us an interesting dilemma. Even though the PAML4 individual ancestor is the shortest, it is Mega5 that has left out the very most highly conserved region of the alignment when asked to include 100% of the MSA. This is because the sequences from the branches (*Chl. reinhardtii*, *N. gruberi*, and *Aur. anophagefferens*) do not have the C terminal folds that appear to be specific to vaults. This could be correct and the extra sequence has arisen more recently. However, this is at odds with Cavalier-Smith's latest version of the root of eukaryotes because extant vault MVP, from both sides of the proposed root, has this structure. It is unlikely to have arisen twice because the primary sequence is so highly conserved across all domains.

A summary of our results is shown in figure 8 based on a eukaryote tree (Keeling et al. 2005).

The Consurf representation (fig. 9A and B) (Ashkenazy et al. 2010) shows the nonconserved residues (blue), and the highly conserved residues (red) from an MSA of MVP sequences from all species discussed. The nonconserved residues are generally either solvent exposed on the exterior surface of the vault or line the interior, but are not those involved with inter molecular contacts docking monomers for vault formation. The conserved residues cluster around the shoulder of the vault, and also along the length of each monomer within the lateral contacts.

In general, the docking is relatively poorer amongst the ancestors (table 1) than docking in the individual sequences that made up the original ASR input. This is to be anticipated; a core MVP fold is conserved with sequence variation existing



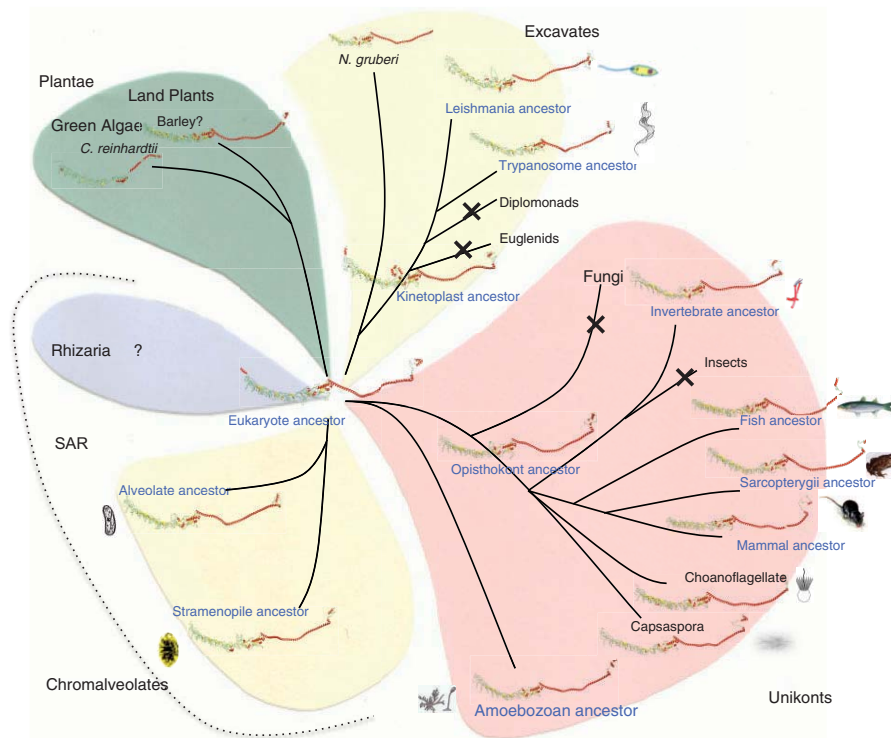
**Fig. 7.**—(A) Table produced from MSA of ancestral sequences of the super-groups identified in the text and from the alignment of the ancestors made by individuals (one per species). Rat has been included for comparison. (B) The I-TASSER structural prediction for the reconstruction of the ancestor from 89 individual sequences, this cartoon depicts the FastML reconstruction that bears least sequence similarity with either the rat or with the other ancestors. (C) The MSA close to the C terminal identifying an area of very high conservation. (D) Cartoon diagram of this region modeled by I-TASSER utilizing the ab initio modeling capacity of MODELLER as this area has not been resolved in the crystal structure.

amongst groups that still allows interface docking, possibly through covariance of sites between monomers within particular species. Our RosettaDock analysis with the known rat structure (2ZUO monomers) indicates redundancy in docking possibilities, that is, the docked rat monomer pairs were not all utilizing the same residues as those found in the solved crystal structure (Daly et al. 2013). If, as expected, the mutation rate was equivalent at all positions within the MVP sequence positions, then the docking of one MVP monomer for another would quickly deteriorate so there must be selective pressure to maintain the residues important for docking even if the vault structure was ancestrally rather simpler.

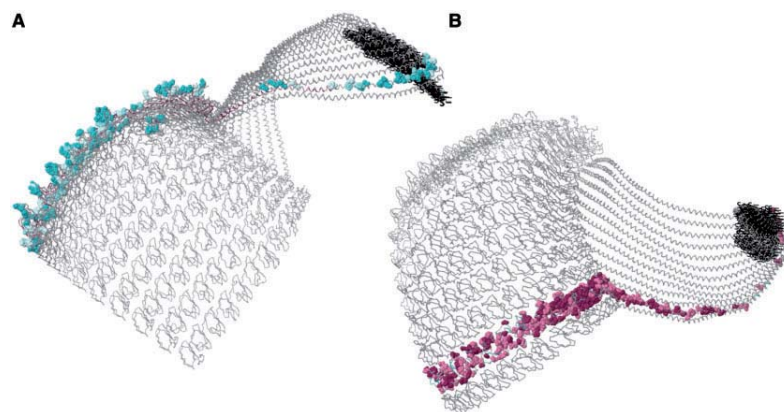
Discussion

From our general approach of reconstructing tertiary structures, and inferring quaternary formations, the MVP gene appears to be ancestral to eukaryotes and it is likely that vault particles were present in LECA. MVP is retained in most eukaryote super-groups; opisthokonts (fungi plus animals), amoebozoa, chromalveolates (though we are not so sure about rhizaria that have been latterly included with the chromalveolates in the SAR supergroup), and excavates distributed in groups both sides of the proposed initial divergence of the last common eukaryote ancestor (Cavalier-Smith 2010). Plantae is rather more controversial, although *Chl. reinhardtii*





**Fig. 8.**—Structural diagrams of I-TASSER predictions from individual extant sequences (black type face) and from reconstructed sequences, derived from a combination of PAML4 and FastML ASR (blue type face).



**Fig. 9.**—Consurf diagrams of the structural back bone of oligomerized rat MVP (PDB:2ZUO) showing one chain with spheres depicting the similarity score of the MSA for all sequences. Thirteen monomers (of a total 39) of a half rat vault are shown for clarity. (A) Nonconserved residues are shown as blue spheres. These nonconserved residues stick out from the surface of the vault (outward or inward), but are not involved in docking. (B) Shows completely conserved residues (red) and highly conserved residues (pink). The most conserved residues are found in the shoulder region and along the length of the monomer within the lateral contacts. The extreme C terminal is also highly conserved but unresolved in the crystal structure (2ZUO), those residues show as black points.

**Table 1**

I-TASSER and RosettaDock Results for Individuals from Poorly Represented Groups and from ASRs

Accession Number	Organism	Length	% Residues Identical to Rat	I-Tasser TM Score, <sup>a</sup> Max Is 1.0	I-Tasser C Score <sup>b</sup> (Range -5 +2)	RosettaDock Energy Score <sup>c</sup>
Q62667 <sup>d</sup>	<i>R. norvegicus</i> (Rat)	861	100	0.77 ± 0.10	0.42	-254
Extant sequences used individually						
A9V809	<i>Monosiga brevicollis</i> (Choanoflagellate)	861	59	0.90 ± 0.06	1.34	-515
F2UN76	Salpingocea (Choanoflagellate)	853	59	0.86 ± 0.07	1.07	-474
E9CE06	<i>Capsaspora owczarzaki</i> (Capsaspora)	860	63	0.82 ± 0.09	0.77	-199
F2E078	<i>Hordeum vulgare</i> (Barley)	843	55	0.86 ± 0.07	1.06	-496
CT836653	<i>Oryza sativa</i> from cDNA (Rice)	831	60	0.90 ± 0.06	1.37	-511
F4Y3B4	<i>Lyngbya majuscula</i> (Cyanobacteria)	879	54	0.77 ± 0.10	0.43	-440
D2V5B9	<i>N. gruberi</i> (Heterolobosea)	559	17	0.62 ± 0.14	-0.74	-441
A8JEL9	<i>Chlamydomonas reinhardtii</i> (Chlorophyte)	529	17	0.62 ± 0.14	-0.86	-17
Ancestors created from combined PAML and FastML ASR						
ASR	All Eukaryotes <sup>e</sup>	892	32	0.79 ± 0.09	0.56	-156
ASR	Stramenopiles	912	45	0.74 ± 0.11	0.2	-208
ASR	Alveolate	871	45	0.88 ± 0.0	1.18	-161
ASR	Leishmania	995	34	0.58 ± 0.14	-1.06	-180
ASR	Trypanosome	916	38	0.81 ± 0.09	0.69	-205
ASR	Kinetoplast	1,025	34	0.53 ± 0.15	-1.46	-148
ASR	Amoebozoa	859	55	0.80 ± 0.09	0.67	-248
ASR	Opisthokont	913	65	0.96 ± 4.6	-0.38	-174
ASR	Invertebrate	853	67	0.90 ± 0.06	1.34	-261
ASR	Fish	887	67	0.88 ± 0.07	1.24	-519
ASR	Sarcopterygii	857	72	0.87 ± 0.07	1.11	-226
ASR	Mammal	945	79	0.78 ± 0.10	0.51	-238

NOTE.—The rat is given at the beginning for comparison; it is the only vault for which its 3D structure is known.

<sup>a</sup>I-TASSER TM score higher is better—cut off is -1.5.<sup>b</sup>I-TASSER C score higher is better—cut off is 0.5.<sup>c</sup>RosettaDock energy score lower is better.<sup>d</sup>Q62667 is included for comparison. Rat MVP is the only complete MVP 3D X-ray crystallographic structure in the Protein Data Bank.<sup>e</sup>All Eukaryotes. These figures refer to the joint ancestor of all individuals (one sequence per species), all final ancestors, scored within our criteria.

seems to be a bona fide inclusion, the grasses look more like contamination. Additionally, *Chl. reinhardtii*, *Aur. anophagefferens*, and *N. gruberi* do not have the highly conserved helices and loop at the C terminal. Although we have concentrated, for obvious reasons, on the MVPs, we see prediction of 3D structure as a general approach that could be used much more frequently for understanding earlier phases of molecular evolution.

MVP is under selective pressure to maintain structure and appropriate residues for docking. In general, the docking scores for the putative ancestors are lower than for the extant sequences though our prediction would still be that these sequences would be capable of self-assembling into a vault particle as they are known to do in metazoa.

Is there any evidence that could support this assertion? If we consider the proteins that are known to associate with vaults there are differences even within extant species. Vault poly ADP-ribose polymerase from the PARP protein family (VPARP) is found only in metazoa and amoebozoa, and therefore seems like a more recent adaptation (Citarelli et al. 2010). These authors found six clades of PARP protein and suggest

that LECA already had at least proteins from clades 1 and 6. The only MVP sequence that we have used that comes from a species where no PARP family member has been found is *Aur. anophagefferens*; although vaults can form without PARP (Stephen et al. 2001).

Similarly, TEP1 (telomerase-associated protein-1) is a component of vault particles in metazoa and amoebozoa. TEP1 contains a TROVE domain (Telomerase, Ro, and Vault), that binds vault associated RNA (vtRNA) (Poderycki et al. 2005). TEP1 is ubiquitous but vaults form without it and without vtRNA. In metazoa, the only characterized group, approximately 80% of the vtRNA is found outside of the vault (Kickhoefer et al. 1998). The sequence homology of vtRNA—even within metazoa, is slim (Stadler et al. 2009). It is a pol III transcribed RNA and outside of the A and B box regions, structural homology would be the best search method to find it in other groups.

If vault particles were formed in LECA—with or without any other association—what functional role could they have played? Extant vault particles open at low pH (Goldsmith et al. 2009) and anions can enter, possibly attracted by positively

charged amino acids facing the vault interior (Ng et al. 2008). Vaults have been associated with detoxification processes (Suprenant et al. 2007), though they have never been proven to be vital (Herlevsen et al. 2007). Some kind of early encapsulation of substances toxic to the cell would be a desirable trait. One possible function could be protection from harmful bacteria that are engulfed by eukaryotes.

Vault particles are probably missing from plants, have not been found in insects and although traces of MVP monomer sequences appear in some fungi, they fall short of having vault forming capability using our criteria (not shown). The loss of vault particles in plants and fungi might be explained because they do not normally consume bacteria? Again, their loss in insects might be explained by their hosting of complex communities of bacterial, fungal and viral symbionts when feeding on plants hosting pathogens and producing toxic chemicals (Frago et al. 2012) that would be protective without the need for vault particles.

Protein compartments that encapsulate and compartmentalize contents are ubiquitous, although a variety of designs are utilized. In many ways vaults are reminiscent of virus particles; they are large assemblies that have a protein shell composed of multiple copies of a single protein and have a large central cavity. However, the geometry of viruses can be classified as; icosahedral, helical or complex (the classification given to the pox virus), but none have the radially symmetrical halves joined together like the vault particle.

Prokaryotes also form compartments, both larger and smaller than the vault (Heinhorst and Cannon 2008) that concentrate linked functional mechanisms; however, these are mostly icosahedral, for example, carboxysomes. Although vault particles were originally thought to be absent from prokaryotes, there are a number of convincing homologs which could have been acquired by HGT from eukaryotes. However, there are other proteins, ubiquitous in prokaryotes that have sequence and structural similarities with MVP in whole or in part. BLASTs with the rat MVP sequence repeatedly result in TolA and band 7 protein homologs being identified within default parameters. In fact, the cyanobacterial MVP homologs have been annotated colicin uptake transmembrane protein, which is a pathway that utilizes TolA. The mechanism for colicin uptake has been mostly studied in *Escherichia coli* and comprises the Tol/Pal system. The function of TolA is not fully understood, it is involved in the structural integrity of *E. coli* and related bacterial cell membranes. It is also involved in active transport across the membrane but can be parasitized by colicins produced by other *E. coli* resulting in the death of the cell (Li et al. 2012). The Tol system also allows uptake of phage DNA, although generally deleterious imported DNA may contain genes that could give the cell an advantage. It seems unlikely that the Tol/Pal system would be retained specifically for the uptake of pathogens, but conservation of an

active, if promiscuous transport system, might have been essential to the early eukaryote.

One of the bacterial sequences, *Her. aurantiacus* (UniProtKB:A9AUD4), that folds as MVP within our criteria is annotated as a band 7 protein. These band 7 sequences are ubiquitous proteins that include stomatins, prohibitin, flotillin HlfK/C, and podicin, known collectively as SPFH domain-containing proteins. Tanaka et al. (2009) identified the shoulder domain of MVP as homologous to the stomatin core from *Pyrococcus horikoshii*. Band 7 proteins have been found to form ring-like oligomeric structures, for example, membrane-bound prohibitin rings in mitochondria (Tatsuta et al. 2005), and free ring structures in cyanobacteria (Boehm et al. 2009). SPFH domain proteins are often linked with lipid rafts (Browman et al. 2007). Extant vault particles are also found in association with lipid rafts (Kowalski et al. 2007). Vault particles are capable of detoxification of anions (Suprenant et al. 2007), and are linked with multi drug resistance in both cancer and epilepsy (Herlevsen et al. 2007; Liu, Mao, et al. 2011). The capacity for sequestration or even ejection of toxins from the early eukaryote would be a reason for the high level of conservation.

## Conclusion

MVP has been identified by our Ancestral Sequence Reconstruction methods in; opisthokonts, amoebozoa, excavates (including euglenids), chromalveolates, bacteria, and possibly plants. We additionally predict that these MVP monomers could dock to form complex oligomeric vaults as they are known to do in opisthokonts and amoebozoa. We propose that vaults in LECA could have functioned in membrane transport, the sequestering of cell toxins, or provide protection from engulfing pathogenic bacteria, but have now diversified into the multitude of roles seen today, to the point where they are being harnessed and utilized for drug and vaccine delivery and possible future bioremediation.

## Supplementary Material

Supplementary materials S1–S3 are available at *Genome Biology and Evolution* online (<http://www.gbe.oxfordjournals.org/>).

## Acknowledgments

T.K.D. performed the analyses and wrote the drafts; A.J.S.-S. assisted with the analysis programs; and D.P. designed the original research project. All authors regularly discussed the results, and contributed to the final manuscript. This work was supported by the Institute of Fundamental Sciences, Massey University.

## Literature Cited

- Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N. 2010. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.* 38:W529–W533.
- Ashkenazy H, et al. 2012. FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Res.* 40:W580–W584.
- Berger W, Steiner E, Grusch M, Elbling L, Micksche M. 2009. Vaults and the major vault protein: novel roles in signal pathway regulation and immunity. *Cell Mol Life Sci.* 66:43–61.
- Boehm M, et al. 2009. Structural and mutational analysis of band 7 proteins in the cyanobacterium *Synechocystis* sp. strain PCC 6803. *J Bacteriol.* 191:6425–6435.
- Browman DT, Hoegg MB, Robbins SM. 2007. The SPFH domain-containing proteins: more than lipid raft markers. *Trends Cell Biol.* 17:394–402.
- Buehler DC, Toso DB, Kickhoefer VA, Zhou ZH, Rome LH. 2011. Vaults engineered for hydrophobic drug delivery. *Small* 7:1432–1439.
- Burki F, et al. 2007. Phylogenomics reshuffles the eukaryotic supergroups. *PLoS One* 2(8):e790.
- Burki F, et al. 2010. Evolution of Rhizaria: new insights from phylogenomic analysis of uncultivated protists. *BMC Evol Biol.* 10:377.
- Cavalier-Smith T. 2010. Kingdoms Protozoa and Chromista and the eozoan root of the eukaryotic tree. *Biol Lett.* 6:342–345.
- Champion CI, et al. 2009. A vault nanoparticle vaccine induces protective mucosal immunity. *PLoS One* 4(4):e5409.
- Chang BSW, Jönsson K, Kazmi MA, Donoghue MJ, Sakmar TP. 2002. Recreating a functional ancestral archosaur visual pigment. *Mol Biol Evol.* 19:1483–1489.
- Chugani DC, Kedersha NL, Rome LH. 1991. Vault immunofluorescence in the brain: new insights regarding the origin of microglia. *J Neurosci.* 11:256–268.
- Citarelli M, Teotia S, Lamb RS. 2010. Evolutionary history of the poly(ADP-ribose) polymerase gene family in eukaryotes. *BMC Evol Biol.* 10:308.
- Collins L, Penny D. 2005. Complex spliceosomal organization ancestral to extant eukaryotes. *Mol Biol Evol.* 22:1053–1066.
- Collins LJ, Poole AM, Penny D. 2003. Using ancestral sequences to uncover potential gene homologues. *Appl Bioinformatics.* 2:S85–S95.
- Daly TK, Sutherland-Smith AJ, Penny D. 2013. Beyond BLASTing: tertiary and quaternary structure analysis helps identify major vault proteins. *Genome Biol Evol.* 5:217–232.
- Denoeud F, et al. 2010. Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science* 330:1381–1385.
- Desmond E, Grihaldo S. 2009. Phylogenomics of sterol synthesis: insights into the origin, evolution, and diversity of a key eukaryotic feature. *Genome Biol Evol.* 1:364–381.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.
- Eichenmüller B, et al. 2003. Vaults bind directly to microtubules via their caps and not their barrels. *Cell Motil Cytoskeleton.* 56:225–236.
- Elias M, Patron NJ, Keeling PJ. 2009. The RAB family GTPase Rab1A from *Plasmodium falciparum* defines a unique paralog shared by chromalveolates and rhizaria. *J Eukaryot Microbiol.* 56:348–356.
- Frago E, Dicke M, Godfray HCJ. 2012. Insect symbionts as hidden players in insect-plant interactions. *Trends Ecol Evol.* 27:705–711.
- Fritz-Laylin LK, et al. 2010. The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. *Cell* 140:631–642.
- Goldsmith LE, Pupols M, Kickhoefer VA, Rome LH, Monbouquette HG. 2009. Utilization of a protein “shuttle” to load vault nanocapsules with gold probes and proteins. *ACS Nano.* 3:3175–3183.
- Gray JJ, et al. 2003. Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J Mol Biol.* 331:281–299.
- Gullberg M, et al. 2010. Characterization of a putative ancestor of coxsackievirus B5. *J Virol.* 84:9695–9708.
- Hanson-Smith V, Kolaczowski B, Thornton JW. 2010. Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. *Mol Biol Evol.* 27:1988–1999.
- Heinhorst S, Cannon GC. 2008. A new, leaner and meaner bacterial organelle. *Nat Struct Mol Biol.* 15:897–898.
- Herlevsen M, Oxford G, Owens CR, Conaway M, Theodorescu D. 2007. Depletion of major vault protein increases doxorubicin sensitivity and nuclear accumulation and disrupts its sequestration in lysosomes. *Mol Cancer Ther.* 6:1804–1813.
- Holm L, Rosenström P. 2010. Dali server: conservation mapping in 3D. *Nucleic Acids Res.* 38:W545–W549.
- Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* 17:754–755.
- Hunter S, et al. 2012. InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.* 40:D306–D312.
- Ishikawa F, Naito T. 1999. Why do we have linear chromosomes? A matter of Adam and Eve. *Mutat Res.* 434:99–107.
- Keeling PJ. 2004. Diversity and evolutionary history of plastids and their hosts. *Am J Bot.* 91:1481–1493.
- Keeling PJ. 2009. Chromalveolates and the evolution of plastids by secondary endosymbiosis. *J Eukaryot Microbiol.* 56:1–8.
- Keeling PJ, et al. 2005. The tree of eukaryotes. *Trends Ecol Evol.* 20:670–676.
- Kelley LA, Sternberg MJ. 2009. Protein structure prediction on the Web: a case study using the Phyre server. *Nat Protoc.* 4:363–371.
- Kickhoefer VA, et al. 1998. Vaults are up-regulated in multidrug-resistant cancer cell lines. *J Biol Chem.* 273:8971–8974.
- Kickhoefer VA, et al. 2009. Targeting vault nanoparticles to specific cell surface receptors. *ACS Nano.* 3:27–36.
- Klaus FX. 2012. A physical, genetic and functional sequence assembly of the barley genome. *Nature* 491:711–716.
- Koonin EV. 2010. The incredible expanding ancestor of eukaryotes. *Cell* 140:606–608.
- Kowalski MP, et al. 2007. Host resistance to lung infection mediated by major vault protein in epithelial cells. *Science* 317:130–132.
- Li C, et al. 2012. Structural evidence that colicin A protein binds to a novel binding site of TolA protein in *Escherichia coli* periplasm. *J Biol Chem.* 287:19048–19057.
- Liu JL, Mao ZY, Gallick GE, Yung WKA. 2011. AMPK/TSC2/mTOR-signaling intermediates are not necessary for LKB1-mediated nuclear retention of PTEN tumor suppressor. *Neuro Oncol.* 13:184–194.
- Liu B, et al. 2011. Up-regulation of major vault protein in the frontal cortex of patients with intractable frontal lobe epilepsy. *J Neurol Sci.* 308:88–93.
- López-García P, Moreira D. 1999. Metabolic symbiosis at the origin of eukaryotes. *Trends Biochem Sci.* 24:88–93.
- Lyskov S, Gray JJ. 2008. The RosettaDock server for local protein-protein docking. *Nucleic Acids Res.* 36:233–238.
- Matsumoto T, et al. 2011. Comprehensive sequence analysis of 24,783 Barley full-length cDNAs derived from 12 clone libraries. *Plant Physiol.* 156:20–28.
- Mossel E, Steel M. 2005. How much can evolved characters tell us about the tree that generated them? Oxford: Oxford University Press.
- Ng BC, et al. 2008. Encapsulation of semiconducting polymers in vault protein cages. *Nano Lett.* 8:3503–3509.
- Nosek J, Kosa P, Tomaska L. 2006. On the origin of telomeres: a glimpse at the pre-telomerase world. *BioEssays* 28:182–190.
- Parfrey LW, et al. 2010. Broadly sampled multigene analyses yield a well-resolved eukaryotic tree of life. *Syst Biol.* 59:518–533.

- Paspalas CD, et al. 2009. Major vault protein is expressed along the nucleus-neurite axis and associates with mRNAs in cortical neurons. *Cereb Cortex*. 19:1666–1677.
- Philippe H, et al. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol*. 9(3):1–10.
- Poderycki MJ, Rome LH, Harrington L, Kickhoefer VA. 2005. The p80 homology region of TEP1 is sufficient for its association with the telomerase and vault RNAs, and the vault particle. *Nucleic Acids Res*. 33: 893–902.
- Rodríguez-Ezpeleta N, et al. 2007. Toward resolving the eukaryotic tree: the phylogenetic positions of Jakobids and Cercozoans. *Curr Biol*. 17: 1420–1425.
- Roy A, Kucukural A, Zhang Y. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc*. 5: 725–738.
- Sali A, Blundell TL. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*. 234:779–815.
- Schönknecht G, et al. 2013. Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science* 339: 1207–1210.
- Sebe-Pedros A, de Mendoza A, Lang BF, Degnan BM, Ruiz-Trillo I. 2011. Unexpected repertoire of metazoan transcription factors in the unicellular holozoan *Capsaspora owczarzaki*. *Mol Biol Evol*. 28:1241–1254.
- Sierra R, et al. 2013. Deep relationships of Rhizaria revealed by phylogenomics: a farewell to Haeckel's Radiolaria. *Mol Phylogenet Evol*. 67: 53–59.
- Skotnicki ML, Mackenzie AM, Ninham JA, Selkirk PM. 2004. High levels of genetic variability in the moss *Ceratodon purpureus* from continental Antarctica, subantarctic Heard and Macquarie Islands, and Australasia. *Polar Biol*. 27:687–698.
- Stadler PF, et al. 2009. Evolution of vault RNAs. *Mol Biol Evol*. 26: 1975–1991.
- Stechmann A, Cavalier-Smith T. 2003. The root of the eukaryote tree pinpointed. *Curr Biol*. 13:R665–R666.
- Stephen AG, et al. 2001. Assembly of vault-like particles in insect cells expressing only the major vault protein. *J Biol Chem*. 276:23217–23220.
- Stevens MI, Hunger SA, Hills SK, Gemmill CEC. 2007. Phantom hitchhikers mislead estimates of genetic variation in Antarctic mosses. *Plant Syst Evol*. 263:191–201.
- Stewart PL, et al. 2005. Sea urchin vault structure, composition, and differential localization during development. *BMC Dev Biol*. 5:3.
- Suprenant KA. 2002. Vault ribonucleoprotein particles: Sarcophagi, gondolas, or safety deposit boxes? *Biochemistry* 41:14447–14454.
- Suprenant KA, Bloom N, Fang JW, Lushington G. 2007. The major vault protein is related to the toxic anion resistance protein (TelA) family. *J Exp Biol*. 210:946–955.
- Tabach Y, et al. 2013. Identification of small RNA pathway genes using patterns of phylogenetic conservation and divergence. *Nature* 493: 694–698.
- Tamura K, et al. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol*. 28:2731–2739.
- Tanaka H, et al. 2009. The structure of rat liver vault at 3.5 Å resolution. *Science* 323:384–388.
- Tatsuta T, Model K, Langer T. 2005. Formation of membrane-bound ring complexes by prohibitins in mitochondria. *Mol Biol Cell*. 16: 248–259.
- Tsai IJ, et al. 2013. The genomes of four tapeworm species reveal adaptations to parasitism. *Nature* 496:57–63.
- van Zon A, et al. 2002. Structural domains of vault proteins: a role for the coiled coil domain in vault assembly. *Biochem Biophys Res Commun*. 291:535–541.
- van Zon A, et al. 2003. The formation of vault-tubes: a dynamic interaction between vaults and vault PARP. *J Cell Sci*. 116:4391–4400.
- Vasu SK, Rome LH. 1995. *Dictyostelium* vaults: disruption of the major proteins reveals growth and morphological defects and uncovers a new associated protein. *J Biol Chem*. 270:16588–16594.
- Williams PD, Pollock DD, Blackburne BP, Goldstein RA. 2006. Assessing the accuracy of ancestral protein reconstruction methods. *PLoS Comput Biol*. 2:0598–0605.
- Wu S, Zhang Y. 2007. LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Res*. 35:3375–3382.
- Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24:1586–1591.
- Zhang Y. 2008. I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* 9:40.

Associate editor: Dan Graur



# Chapter Three: Introduction to the Argonaute Family

## 3. The defence of the Dark Arts

‘How old are RNA networks?’ was written as the final chapter in a book ‘RNA Infrastructure and Networks (Collins, 2011) and also published as a stand-alone journal article (Daly *et al.*, 2011). The work introduces the ‘Argonaute’ family of proteins as well as describing some other proteins that also use RNA in their ‘defence of the Dark Arts’. This phrase comes from J.K. Rowling of Harry Potter fame, but is a perfect term for the arms race that likely exists within every cell (from all domains of life) against every shred of parasitic nucleic acid (collectively termed ‘the Dark Arts’) and evolution within the Dark Arts that can outwit the latest cellular defence tactic. I am indebted to David Penny for thinking of such an apt title.

Some of the ideas in this manuscript have since been enlarged upon and the genesis of small RNAs becomes more varied and complex with each publication. There are many pathways in addition to those described in figure 2. A description of the genesis of microRNA (miRNA) was left out completely as miRNA is principally utilised in the control of endogenous transcripts to fine tune gene expression rather than controlling parasitic nucleic acid, which was the main focus of the chapter. There are also a myriad of ways in which transcripts can avoid Dicer processing e.g., where dsRNA is formed naturally described in figure 1 and PIWI-associating RNA (piRNA) in figure 2 forms via the ping-pong amplification loop thus avoiding Dicer processing. Additionally PIWI family proteins are not confined to germ-line cells as supposed at the time of writing (Rajasethupathy *et al.*, 2012).

Animal small RNAs outside of piRNA additionally require Drosha processing in the nucleus but this too can be avoided. Three classes of mirtron can form even in mammals via a lariat debranching enzyme (Ladewig *et al.*, 2012) and processed VTRNA described in chapter one known to downregulate *CYP3A4* (Persson *et al.*, 2009) avoids Drosha processing. Of particular interest (mentioned in chapter four (4c)

and explored further in chapter five) is the possibility that processed tRNA and rRNA can be used to supply guide RNA to argonautes.

I am specifically responsible for the sections entitled ‘Regulatory networks of small RNAs’, ‘RNA regulations and defence against the Dark Arts’ and for table 1 and figures 1 – 4 as well as contributing generally to the manuscript. Although chronologically this chapter is out of order, it is placed here in order to introduce the argonautes.



## CHAPTER 17

### HOW OLD ARE RNA NETWORKS?

Toni Daly,<sup>1</sup> X. Sylvia Chen<sup>2</sup> and David Penny<sup>\*,1</sup>

<sup>1</sup>*The Allan Wilson Centre, Institute of Molecular BioSciences, Massey University, Palmerston North, New Zealand;*

<sup>2</sup>*Department of Biochemistry, University of Otago, Dunedin, New Zealand*

*\*Corresponding Author: David Penny—Email: d.penny@massey.ac.nz*

**Abstract:** Some major classes of RNAs (such as mRNA, rRNA, tRNA and RNase P) are ubiquitous in all living systems so are inferred to have arisen early during the origin of life. However, the situation is not so clear for the system of RNA regulatory networks that continue to be uncovered, especially in eukaryotes. It is increasingly being recognised that networks of small RNAs are important for regulation in all cells, but it is not certain whether the origin of these networks are as old as rRNAs and tRNA. Another group of ncRNAs, including snoRNAs, occurs mainly in archaea and eukaryotes and their ultimate origin is less certain, although perhaps the simplest hypothesis is that they were present in earlier stages of life and were lost from bacteria. Some RNA networks may trace back to an early stage when there was just RNA and proteins, the RNP-world; before DNA.

#### INTRODUCTION

With new classes of RNA continuing to be discovered, it appears as if many classes of RNA are likely to occur in eukaryotes, bacteria and archaea—but what about the actual RNA-protein networks in which they are involved. For example, small RNA regulation of gene expression has been seen in both eukaryotes and prokaryotes and even viruses use small RNAs. Although the basic mechanism of target recognition and cleavage is similar in all these groups, the proteins and interactions differ. This chapter considers some ideas for the time of origin of some key classes of RNA and their associated networks. We readily accept the idea that some RNA-protein interactions are very old but when we get down to the elaborate pathways of RNA-based regulation there was an early assumption that such regulation only evolved when organisms became more complex (e.g., multicellular). Now we see that the networks of RNA-protein interaction are more general; but can we

infer the presence of particular RNA functions in **Ida** (Initial Darwinian Ancestor), or in the later **Luca** (the Last Universal Common Ancestor), or in the even later **Fred** (Fairly Remote Eukaryotic Daddy).<sup>1,2</sup>

When it comes to understanding the main roles of RNA in modern cell biology there is a potential problem of the ‘alphabet soup’ formed from so many classes of small RNAs. Defining these subgroups is important for identifying subclasses that are reasonably closely related—they have well-defined homologies and clearly related functions as collated and updated in the RNA database Rfam.<sup>3</sup> However, our interest here is at the other end of the spectrum—once we have these many classes and networks of small RNAs, can we put them into higher level evolutionary groups with more general functions. As an example, Boria et al<sup>4</sup> identified sbRNAs (stem bulge RNAs) in nematodes as evolutionary homologues of the more widespread Y-RNAs. Y-RNAs are involved with the protein R<sub>o</sub> in a network of interactions that both assist misfolded RNAs but are also involved with DNA transcription. This equivalence of R<sub>o</sub> and sbRNAs reinforces the message that RNA networks evolve and as we understand more about how RNA and its associated proteins evolve we can begin to surmise how such regulation could have evolved much earlier in the beginning of life.

## REGULATORY NETWORKS OF SMALL RNAs

Networks involving small RNAs can regulate translation with some proteins also used in defence networks. Small (~20-30 nt) RNAs have many functions including, fine-tuning expression of temporal and tissue specific mRNA, destruction of aberrant mRNAs via cleavage, repression, up-regulation and also translational control via DNA methylation and maintenance of histone architecture.

Three basic regulatory classes of ncRNA networks are recognized in eukaryotes, determined by their biogenesis and mode of operation<sup>5</sup> (siRNA-based, miRNA-based and piRNA-based). miRNA sequences are usually found in intergenic and intronic regions of eukaryotes and in some viruses. They have their own promoters and regulatory units. RNA Pol II transcribes long unstable primary RNAs that form hairpin loops (pri-miRNA). Some miRNA genes are found in the UTR regions of coding sequences (indicating that the transcript can be processed as either a miRNA or an mRNA) and some are even found in exons. A complex cellular milieu of miRNAs allows tuning of thousands of genes through combinatorial interactions within 3' untranslated regions and could account for cell specificity and/or temporal control.<sup>6</sup> miRNAs were once thought to be cell-specific but recent work has shown that in plants at least, miRNAs move extensively between cells and can control protein levels in remote tissue.<sup>7</sup> In a viral example, the herpes virus (dsDNA) produces a range of pri-miRNA transcripts which are processed to pre-miRNAs in the host nucleus. They utilise the endogenous miRNA network using the host RNase III endonuclease (Drosha) and a dsRNA binding protein (Pasha). These are then exported via the same pathway as endogenous miRNAs. These viral miRNAs can extend the lytic phase of the cell, by suppression of apoptosis. Other DNA viruses seem to produce just one or two miRNAs.<sup>8</sup> miRNAs have an obligate nuclear processing phase, so viruses that enter the nucleus are more likely to encode them. HIV is an RNA retrovirus that encodes at least three miRNAs.<sup>9,10</sup> Similarly with bacteria, the success of the plant bacterium *Agrobacterium tumefaciens* is enhanced by knocking out the siRNA pathway, but still requires an intact plant miRNA pathway.<sup>11</sup>

In animals, the network of miRNA-based reactions involves exportin-5 mediating nuclear export of pre-miRNAs, a cytoplasmic RNase III endonuclease Dicer, together with the dsRBP Loqs, (*Drosophila*) and trans-activator binding protein (TRBP) (in humans), transforms the pre-miRNA into the mature transcript. Plants lack a Drosha and Pasha homologue and instead use a nuclear Dicer-like endonuclease (DCL1), which makes similar cuts, but they also require a dsRBP (HYL1) to accurately process the miRNA precursors.<sup>12</sup> Such changes over time are what we expect from an evolutionary process and this makes it likely that the plant/animal ancestor had a comparable system.

Animal miRNAs often have limited complementarity, either as a duplex (miRNA:miRNA\*) or with the target mRNA. Usually the strand with the lower 5' end stability enters the Argonaute domain of the RISC complex (RNA induced silencing complex). Argonautes are an ancient protein family known as 'slicers'. Plant and animal miRNAs usually load into Ago1 although miRNAs have been found in all four human Agos. It was believed that the miRNA\* strand is degraded, but recent work in *Drosophila* indicates that the miRNA\* strand can be modified by 2'-O-methylation at the 3' end then loaded into Ago2 in the RISC complex, the domain usually occupied by siRNA.<sup>13</sup>

In animals, introns can be linearized by the lariat debranching enzyme and the resulting RNA folds to a pre-miRNA structure. These spliced lariats are processed by Dicer and have been termed mirtrons. They have the same action as a miRNA, but do not require the cleavage capabilities of Drosha. Initially discovered in *D. melanogaster* and *C. elegans*, mirtrons have now been found in mammals, including primate specific sets.<sup>14</sup> Additionally the HIV virus encodes a miRNA; hiv1-miR-TAR that uses the cellular mirtron pathway.<sup>9</sup> It has been suggested<sup>14</sup> that this pathway could predate Drosha mediated cleavage. An alternate would be that they have evolved independently three times in different lineages and that viruses have taken advantage of both pathways. Most animal Agos now lack cleavage ability<sup>15</sup> and suppress mRNA translation by physical impedence, or by suppression of transcription via heterochromatin and DNA methylation.<sup>16</sup>

Animal miRNAs usually suppress translation by binding to the target mRNA, but the exact mechanism remains under debate. Three models<sup>17</sup> for preventing initiation are; miRNA:RISC complexes compete for ribosome binding, compete for cap binding, or prevent circularization of the mRNA by inducing de-adenylation (circularization enhances mRNA translation). The latter two strategies could also promote the degradation of mRNA unprotected by cap or polyA tail. Another suggestion is that even if translation has begun, the miRNA:RISC complex could cause ribosomes to drop off prematurely. However, miRNAs can also up-regulate translation (particularly when the cell is under stress).<sup>18</sup> Binding position is relevant—an example being if the interaction of mammalian miR10a:RISC is within the 3'UTR of ribosomal proteins then miR10a suppresses translation, but in the 5'UTR it could activate it.<sup>19</sup> Thus there is a wide diversity of miRNA processing and the existence of small RNA networks appears universal in eukaryotes.

Recently small RNAs derived from tRNA were discovered.<sup>20</sup> These load preferentially into Ago3 and Ago4 where one of the classes of tRNA derived small RNAs (tsRNA) has a moderate capability to down-regulate mRNA. Ago3 and Ago4 may act as a buffer by loading unstructured small RNAs preventing Ago2 from becoming overloaded.<sup>21</sup> Intriguingly they are confined to the cytoplasm, though their processing signature indicates that they are processed in the nucleus. Cellular localisation also plays a part in the standard miRNA-processing network. The miRNA:RISC complex in the cytoplasm often relocates to P bodies enriched in mRNA degradation proteins.<sup>22</sup> Smaller complexes composed of just siRNA or miRNA and Ago2 can also be found in the nucleus. Possibly,

a small proportion of the cytoplasmic Ago2:RISC complex is stripped of Dicer and TRBP prior to, or during translocation to the nucleus. In the case of siRNA or perfectly complementary miRNAs, the target is cleaved and the Ago2 is released and exported to the cytoplasm. In *C. elegans*,<sup>23</sup> NRDE3 (an Ago) together with siRNA is necessary and sufficient for location to the nucleus and silencing of nuclear mRNAs. Although there are 27 Agos in nematodes, there appear to be little redundancy in that mutants lacking this protein (*nrde3*<sup>-</sup> nematodes) are defective for all nuclear mRNA silencing.

miRNA-mediated cleavage appears similar to the RNA interference mechanism mediated by siRNAs (see later). However, there are subtle differences in their requirement for complementarity; siRNA and plant miRNAs have perfect complementarity to their target, but the animal miRNA:miRNA\* duplex contains mismatches bulges and GU wobble pairs. Their biogenesis is more strikingly different; most siRNAs silence their encoded targets and are processed from complementary dsRNA transcripts from invasive nucleic acid during infection (exo-siRNA) and so processing is often in the cytoplasm. In contrast, miRNA sequences are encoded in the organism's genome and so the first steps in the processing pathway occur in the nucleus and requires export. In plants and worms, each siRNA precursor gives rise to many siRNA duplexes, but only one miRNA:miRNA\* is generated from each pre-miRNA. Additionally miRNAs are nearly always conserved in related organisms but exo-siRNAs are rarely conserved.<sup>24</sup>

Endogenous-siRNA (endo-siRNA) is involved in a pathway that could evolutionarily link the nuclear and cytoplasmic processing. This process offers protection from genomically integrated parasitic DNA so its transcripts are derived from the genome and processed in the nucleus. The aim is the suppression of parasitic DNA either by guiding histone modification and DNA methylation, or cleavage of mRNA. The endo-siRNA pathway shares some of the miRNA proteins. For instance, *Drosophila* uses the double stranded RNA binding protein (dsRBP) Loquacious (Loqs) and Dicer 2 for the endo-siRNA pathway and another isoform of Loqs is used for the miRNA pathway in partnership with Dicer 1.<sup>25</sup> The commonality is the endogenous root of the RNA, but the need to quell exogenous transcripts would likely have arisen before silencing one's own.

Piwi interacting RNAs (piRNAs) were originally detected in *Drosophila* germline cells and they are slightly larger (~24-31 nt) RNAs with the 2' O-methyl 3' ending reminiscent of siRNAs. They are particularly known for the suppression of transposons and repeat sequences (see later), but flies and vertebrates have an additional type of piRNA expressed in germline cells during the pachytene and prepachytene stages of meiosis (known as primary piRNA). They are derived from the 3'UTR of cellular transcripts involved in diverse cellular processes. They do not exhibit the 'ping-pong' type amplification where secondary piRNAs are produced via an Aub/Ago3 dependent loop as seen in the suppression of transposons (see later) and they are depleted for repeat elements. The mechanism of primary piRNA production is poorly understood but it is thought that they are generated at the same time as the mRNA is being translated in the cytoplasm<sup>26</sup> and may be derived directly from mRNAs selected for piRNA production. There is some crossover of pathways here because in *Drosophila* at least, primary piRNAs that are involved with the suppression of transposable elements have been found in the somatic cells surrounding germ cells. These cells lack Aub and Ago3 and appear to come from one cluster (flamenco) and target one class of transposon (gypsy elements). Gypsy elements mobilise initially by being copied into RNA, but are then processed into virus-like particles in the cytoplasm.<sup>27,28</sup> It is intriguing that the siRNA, pathway which specialises in dealing with viral RNA, appears to have neglected gypsy

elements and that in flies at least, a variant of the piRNA pathway has developed to deal with them. The important conclusion is again that there are many variations in details within eukaryotes, but the same basic mechanism seems to re-occur.

## RNA REGULATION AND DEFENCE AGAINST THE DARK ARTS

All cells and even viruses, are subject to de novo invasion by both exogenous and endogenous nucleic acids including transposons, viruses, pseudogenes—essentially what has been called ‘selfish DNA’. We refer to this potential invasion of DNA as the ‘Dark Arts’ and (with acknowledgement to JK Rowling) ‘Defence against the Dark Arts’ heavily involves RNA networks.<sup>29</sup> Exogenous nucleic acid can occasionally be advantageous for an organism, but in most cases it will be deleterious and possibly lethal.

We first discuss the small RNAs involved in RNA interference (RNAi) pathways in eukaryotes and viruses and the following section the clustered, regularly interspaced, short palindromic repeats (CRISPR arrays) found in bacteria and archaea. We expect that at all stages during the origin of life, the same basic biological principles apply,<sup>30</sup> and that parasites would always have been present.<sup>31</sup> Viruses have also evolved small-RNA based mechanisms which use the conserved RNAi machinery of the host.<sup>8</sup> The co-evolution of these mechanisms is important for understanding molecular epidemiology ‘in action’.

We consider two forms of siRNA networks as described earlier; exogenous siRNAs (exo-siRNA) from invasive transcripts and short RNAs produced from endogenous transcripts (endo-siRNA). Exo-siRNAs are processed from long double stranded RNAs (dsRNA), often transcribed from viral or plasmid sources during infection in animals, fungi, protists and plants. It forms a basis for antiviral defence and requires at least three key proteins, the RNaseIII endonuclease—Dicer, an RNA dependant RNA polymerase (RdRP) and a member of the Argonaute (Ago) family.

There are 3 clades of Argonaute; ‘Argonaute-like’, ‘Piwi-like’ (both found in prokaryotes) and ‘group 3 Argonautes’ found in *C. elegans*.<sup>32</sup> The Ago proteins are found in all three domains, but are quite diverse in sequence. They possibly originated as a structural support for catalytic RNA and eventually took on the catalytic role leaving the RNA as a guide? There is large variation between eukaryotes in their complement of these proteins (see Table 1) but our main question here is whether there are basic mechanisms across all eukaryotes. To be more precise, did the last common ancestor of eukaryotes (Fred) have this defence mechanism, which has subsequently been modified in the different eukaryote lineages? It has been unclear whether mammals use the exo-siR system given the complex immune system available to them. There is some evidence that mammalian viruses encode RNAi suppressors and that implies that they could use exo-siR,<sup>33</sup> and more recently virally encoded siRNAs have been found in mammalian cells. For example, the abundance of virally encoded siRNAs increases in cells defective for the interferon pathway (IFN  $\alpha/\beta$ ) (initially thought to preclude the possibility of exo-siRNA in mammals).<sup>34</sup>

Exo-siRNA deals with invading RNA, but eukaryote genomes are largely comprised of selfish DNA that has accumulated over time and a variant of the siRNA system (endo-siRNA) deals with this. In plants RDR2 copies transcripts from silent loci, principally transposons, thought to be produced by RNA pol IV. This produces long dsRNA which is cleaved to make sets of cis-acting siRNA (casiRNA) that affect the transcripts of the gene loci that produced them. RDR6 can copy miRNAs from the *TAS* locus, which then enters



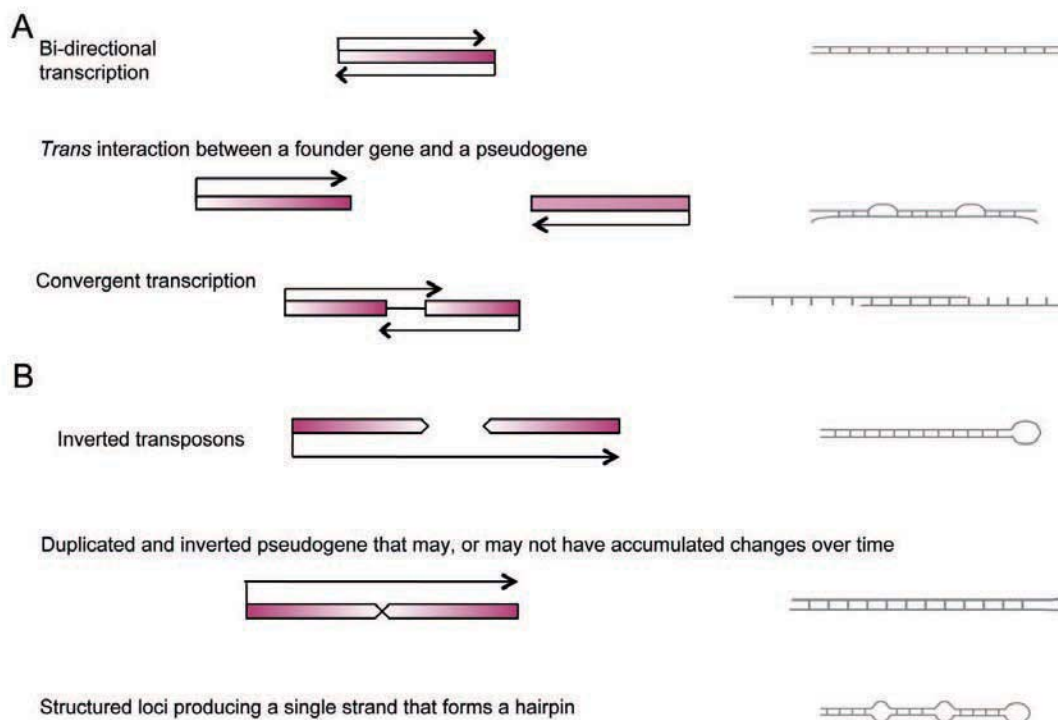
**Table 1.** Some of the key enzymes in the RNAi pathways and variation between species

Species	RNase III Endonuclease	RNA Dependent RNA Polymerase	Argonaute-Like	Piwi-Like
<i>S. pombe</i>	DCR1	RDRP1	Ago1	
<i>A. thaliana</i>	DCL1-4	RDR1, RDR2, RDR6	Ago1-10	
<i>D. melongaster</i>	DCR1-2, Drosha	A subunit of RNA pol II has RdRP capability	Ago1-2	Piwi, Aub, Ago3
<i>C. elegans</i>	DCR1, Drosha	Rrf-1, Rrf-3	Alg1-2, T22B3.2, T23D8.7, ZK757.3	PRG1-2, ERGO1 Plus at least 18 group 3 Argonautes
<i>H. sapiens</i>	DCR1, Drosha	RdRP composed of hTERT in complex with <i>RMRP</i> <sup>37</sup>	Ago1-4	HILI, HIWI, HIWI2 PIWIL3

the RNAi pathway and is trans-acting (tasiRNA); i.e., produced from a discrete locus. Plants also have a third method of producing siRNA via RDR6 by copying overlapping transcripts, one produced constitutively and one produced in times of biotic stress. These are natural antisense transcript-derived siRNAs (natsiRNA).<sup>35</sup> This continued evolution of RNA-protein networks makes it harder to recognise specific networks, but does not obscure the general mechanisms of defensive RNA networks.

Classical RdRP homologs have not been found in mammals, yet endo-siRNAs appear ubiquitous in eukaryotes. This defence is important in flies and an elongation subunit of RNA pol II has been found to have RdRP capability and is involved in the RNAi pathway.<sup>36</sup> An alternate RdRP is also available to mammals; an RdRP composed of hTERT in complex with *RMRP* has been shown to produce siRNAs in humans.<sup>37</sup> Endogenous dsRNA can also derive from DNA transcripts that would result in RNA prone to forming duplexes (Fig. 1A),<sup>38</sup> e.g., bi-directional transcription of a single gene, trans-interaction between transcripts from a founder gene and pseudogene, or overlapping transcription across two genes (convergent transcription). Long single stranded RNA transcripts that fold back on themselves could form from long transcripts of transposons in inverted orientation, or from duplicated inverted pseudogenes, or from structured loci. All of these could result in hairpin duplex structures that can be processed using the siRNA pathway with the addition of a double stranded RNA binding protein (dsRBP) (Fig. 1B).

In relation to the Dark Arts, exo-siRNA defends against exogenous viral and plasmid nucleic acid infection and endo-siRNAs against RNA from endogenous repeating units, transposons, integrated viral sequences and pseudogenes. This defence can occur in any cell, but is especially important in organisms (such as animals) where there is a division into somatic and germ-line cells. The matched endogenous repeat sequences were originally termed repeat associated small interfering RNAs (rasiRNA). This term has been retained for plant endo-siRNAs that target repeat sequences. However, piRNAs aren't simply longer siRNAs—their biogenesis is different, because they are independent of Drosha and Dicer endonucleases and they interact with a specific clade of Ago proteins known as Piwi, AUB (Aubergine) and Ago3 in *Drosophila* (MILI1, MIWI and MIWI2



**Figure 1.** A) Transcription of endogenous DNA that gives rise to dsRNA. Bidirectional transcription resulting in dsRNA can arise when sense and antisense promoters are found in the 5'UTR. Parent and pseudogene transcripts can form endo-siRNAs that regulate the parent gene. Pseudogenes are essentially unsuccessful duplications, 'faulty' in some way, yet many are retained and don't appear to accumulate as many mutations as would be expected. siRNAs can also derive from overlapping transcripts, this occurs naturally in plants giving rise to natsiRNAs produced at times of stress. Convergent transcription in the geminivirus gives rise to a very short overlap but is sufficient to produce the dsRNA needed for exo-siRNA. B) Transcripts prone to forming hairpin loops form from the 'read through' of transposons or pseudogenes that have duplicated and inverted.

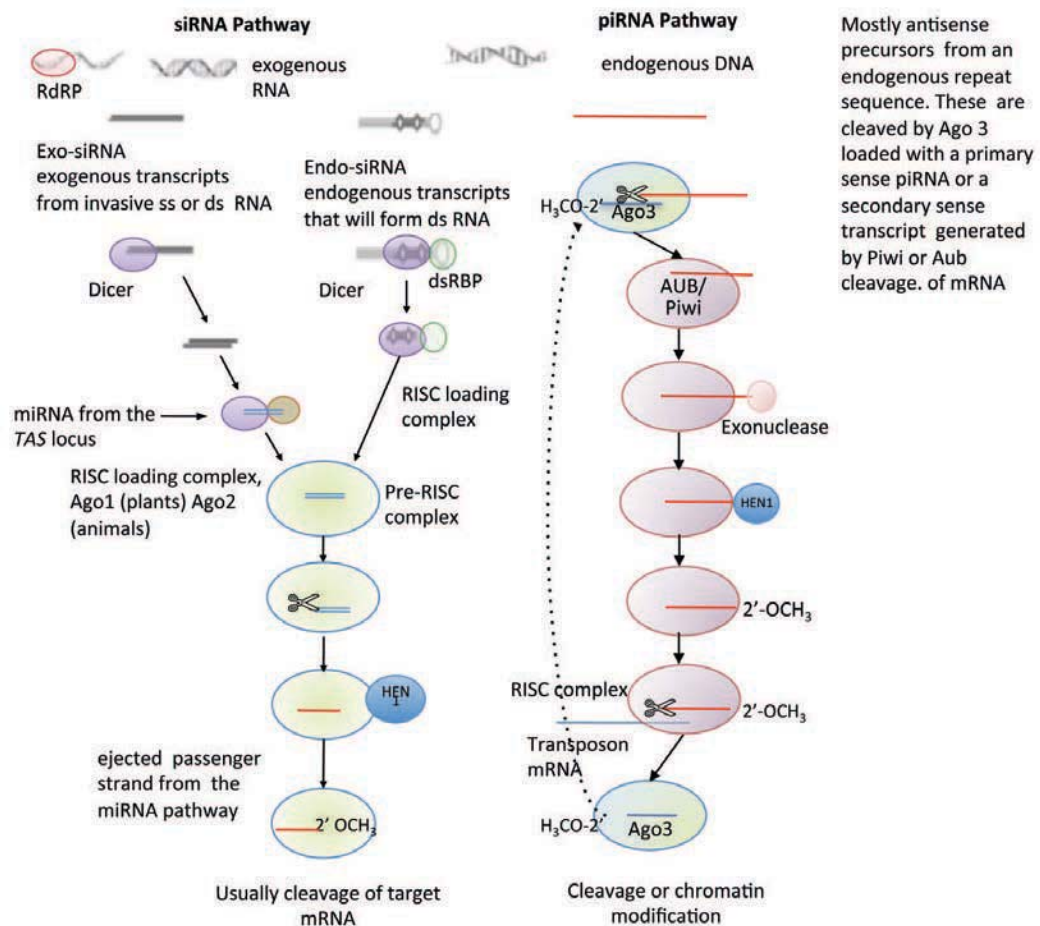
in mice, HILI, HIWI1, HIWI2 and HIWI3 in humans). The evolution of RNA networks continues; it is not a 'once and for all' setup.

Transposons make up the bulk of most eukaryote DNA and are classed as either Type I transposons (that have an RNA intermediate and can be thought of as 'copy and paste'), or Type II elements (where the segment of DNA can move, 'cut and paste'). piRNAs are processed from long endogenous predominantly antisense transposon or repeat sequence transcripts often found concentrated in hotspots on the genome.<sup>39</sup> Their biogenesis is different from the other pathways made possible because the Piwi clade of Argonautes have retained their catalytic capability. Although precursor piRNAs are mostly antisense, a small number of sense piRNAs are produced. These load into Ago3 and cleave antisense retrotransposon transcripts or the antisense precursor piRNA leaving a 5' antisense product or secondary piRNA. These load preferentially into Piwi or Aub and bind and cleave transposon mRNA forming a 5' sense piRNA that loads into Ago3 and guides cleavage of an antisense mRNA or the antisense precursor. This is known as a 'forward feed' or 'ping pong' amplification loop and can generate large numbers of piRNAs without the need for Dicer processing. Plants and worms have also amplification steps in their siRNA pathways, generating secondary siRNAs though the mechanisms are quite different. Silencing is



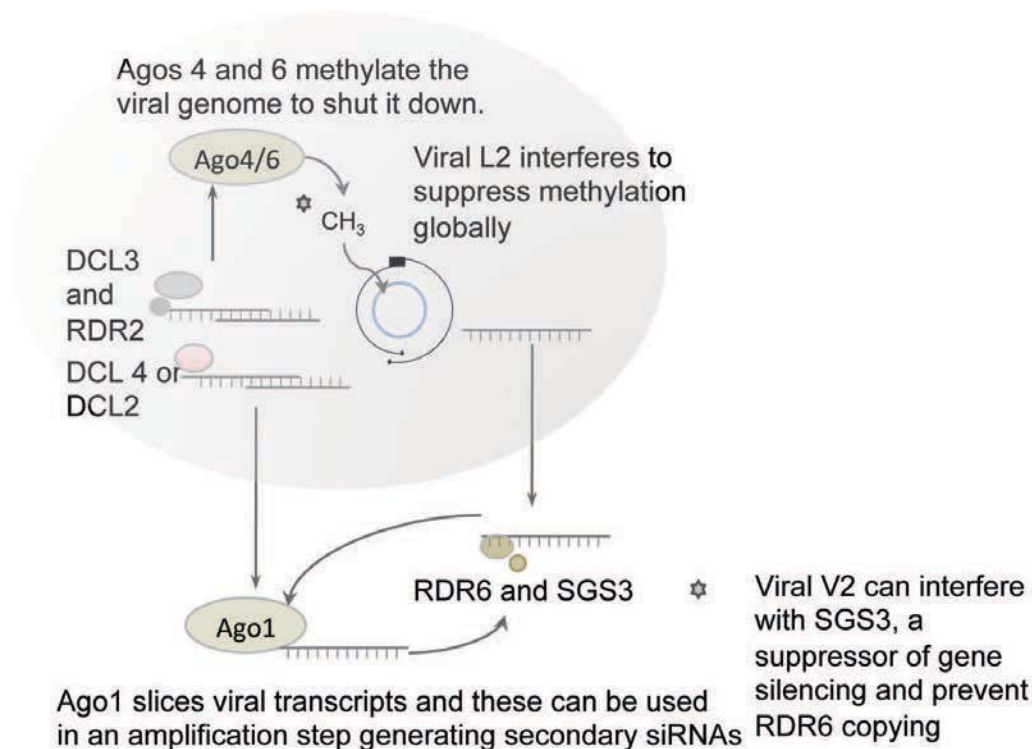
achieved principally by DNA methylation via the RITS (RNA-induced Initiation of Transcriptional Gene Silencing) complex containing Piwi and AUB. In mammals this epigenetic regulatory role is heritable through maternal transmission.<sup>40</sup> A comparison of some features of RNAi defence networks is shown in Figure 2.

Some genomic regions give rise to both endo-siRNA and piRNA and target the same transcripts, indicating cooperation/interaction between networks. Indeed, integration across the entire RNAi networks seems necessary, otherwise a viral infection might overwhelm the machinery required to suppress transposons to the mutual benefit of virus and transposon. Alternatively, there could be some redundancy among the molecular machinery that keeps all pathways working. For instance, it is known that the plant DCL2 can take over the functions of DCL4. Is this redundancy driven by fortuitous gene duplication, or in response to the plant dependence on the RNAi networks?



**Figure 2.** A comparison of RNAi networks involved with the defence of the Dark Arts. Current invasions can be quashed by exo-siRNAs utilising the RNA derived from the invader, either conveniently provided, for example by the geminivirus, or formed into double stranded RNA by RdRP. Endogenous invasive DNA is dealt with by the endo-siRNA pathway, sharing many proteins with the exo-siRNA and also the miRNA networks. The piRNA pathway is also adapted to prevent expression of endogenous invasive DNA, usually by modification of DNA architecture and specifically in the germline. Although the piRNA pathway is independent of Dicer it utilises a variant of the Argonaut protein confined to germline cells (see Table 1).

Viral particles evolve rapidly and even miRNA genes within related viruses may have little homology though they maintain sequence similarity with their mRNA targets for host mRNAs and with some cellular miRNAs. Viruses can suppress host miRNA and siRNA pathways via RNA silencing suppressors (RSS) at many steps. Some DNA viruses score 'own goals' by producing dsRNA, an example being the cassava mosaic virus, a geminivirus that produces dsRNA via convergent transcription from opposite promoters. The plant cytoplasmic DCL4 slices these into siRNAs that suppress the viral transcripts. However, the Dark Arts are never idle; viruses have evolved RSS to turn off the host defence, by inactivating DCL4. Nevertheless, the plant can retaliate and use DCL2 as a substitute for DCL4.<sup>41</sup> The geminiviruses that give rise to convergent transcripts can inhibit the RdRP needed for the miRNA and siRNA pathways in plants,<sup>42</sup> and prevent RNA-directed DNA methylation (Fig. 3). Plants have counter-evolved a method of methylating and silencing geminivirus minichromosomes using RNA-directed methylation via Ago4.<sup>43</sup> Geminivirus counteracts by suppressing methylation globally. A similar arms race occurs wherever viral nucleic acid is found, that is, everywhere. From an evolutionary perspective the RNAi networks must be very old. Not only because RNAi is ubiquitous, but because it makes use of ancient proteins and is involved with

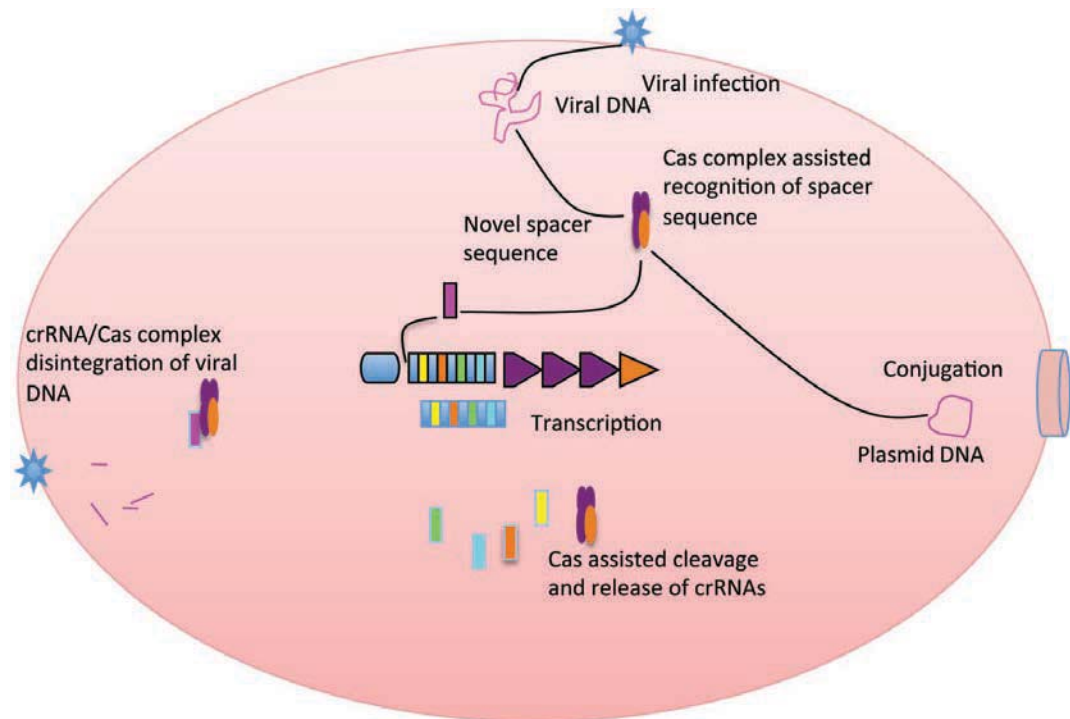


**Figure 3.** Geminiviruses are ssDNA viruses that replicate via rolling replication forming dsDNA without a dsRNA stage. Transcription is bidirectional and the transcript overlaps result in a dsRNA of a small segment of the *AC1* gene, which is indispensable for replication. DCL4 preferentially dices this into small (22nt) virus-derived RNAs (exo-siRNA) which load into Ago1 and slice mRNA transcripts arising from the *AC1* gene. The cleaved mRNA can be converted by plant RDR6 into secondary siRNAs in an amplification step, which also converts other viral ss transcripts into siRNAs. An alternate fate of the viral overlapping transcript is to be extended by RDR2 and diced by DCL3 in the nucleus, where it enters Ago4 or 6 and shuts down viral replication. However, the viruses can interfere with this shutdown at different stages.

the maintenance of other very old proteins such as the highly conserved histones which serve to enhance copying fidelity.

So what of bacteria and archaea, how does their arsenal stack up against parasitic nucleic acids? “Clustered, regularly interspaced, short palindromic repeats” (CRISPRs)<sup>44</sup> occur extensively in bacteria and archaea as a defence against phages, plasmids and transposons, using small RNAs to interfere with invasion. CRISPRs appear ubiquitous in archaea and the apparent lower incidence (~40%) in bacteria may be an artefact of lab maintained strains. However, an alternate explanation is that some pathogenic bacteria, reliant on plasmids for antibiotic resistance or virulence factors, have mechanisms to avoid the acquisition of CRISPR systems. Some *E. coli* strains have CRISPR arrays containing spacer sequences derived from *cas* genes termed ‘antiCRISPRs’ that can destroy the DNA of invading CRISPR/*cas* systems.<sup>45</sup> These species lack endogenous *cas* genes yet the antiCRISPR is selectively retained presumably to permit horizontal gene transfer, or for an as yet unidentified function. The CRISPR system is dynamic and evolutionary and an outline is given in Figure 4.

Spacers between the CRISPR repeats are usually from sources external to the organism (e.g., from the phage) and possession of a spacer sequence matching a challenging phage gives resistance to that phage.<sup>46</sup> A suite of conserved proteins involved in DNA



**Figure 4.** The CRISPR system is a combined exo and endo-siRNA that results in a dynamic, heritable immune system. crRNAs are produced from an endogenous transcript akin to endo-siRNA and diced using Cas proteins, these are analogous but not homologous to the Dicing proteins of the siRNA system. The spacer sequence is acquired from a current infection (top and right), though it isn't clear how rapidly this can be utilised, it would seem that it would be slower than eukaryote exo-siRNA due to the integration step. A major difference between the prokaryote and eukaryote defence against the Dark Arts is that evidence thus far would indicate that cleavage is directed against invading DNA (left), rather than mRNA.

transactions was identified as typically occurring close to the CRISPR locus and called *cas* genes—(CRISPR associated). They have functional domains typical of nucleases, polymerases, helicases and nucleotide binding proteins<sup>47</sup> and are required for the acquisition of the spacer sequences. Although the mechanism is speculative, studies have shown that the core *cas* genes Cas1 and Cas2 occur at all CRISPR loci and are involved with acquisition of the spacer sequences. Disruption to these genes results in loss of ability to acquire new spacer sequences, but does not prevent the function of existing ones. Entire CRISPR/*cas* systems can also be found in plasmids,<sup>48</sup> and the same subset can be found in distantly related organisms. This indicates that they can be acquired from plasmids—by the very mechanism that they normally prevent.

Constitutive transcription of the entire repeat/spacer array gives long precursors which are processed into short crRNAs. Each crRNA corresponds to one spacer sequence flanked by two partial repeats.<sup>47</sup> The significance of the flanking sequences is that the resulting crRNAs can recognise ‘self’ DNA because the ends of the crRNAs will be able to attach by Watson-Crick pairing along their entire length, whereas predatory DNA will result in the crRNA being free at both ends.<sup>49</sup> Disruption of invading target DNA with a self-splicing intron that restores the mRNA, results in loss of resistance in *Staphylococcus epidermidis*, despite having the correct spacer sequence. This indicates that mRNA is not the target.<sup>50</sup> Bidirectional processing was found in the archaeon *Sulfolobus*, suggesting the possibility that RNA duplexes could form and given that archaea contain Argonautes, could this possibly represent a form of siRNA pathway akin to eukaryotes.

RNA cleavage has been demonstrated in vitro in another archaeon *Pyrococcus* via an RNP complex comprised of crRNA and Cas proteins (Cmr1-Cmr6) encoded from a Cas associated module known as RAMP (repeat associated mysterious protein, mainly found in thermophilic archaea).<sup>51</sup> RAMP proteins can be located distally from the CRISPR and do not appear to move with the CRISPR array. DNA cleavage wasn’t detected in *P. furiosus*, this may mean that the *Pyrococcus* CRISPR system is directed at mRNA, or directed at the genome of an RNA phage. The spacer sequences in *P. furiosus* have not been identified. If they are aimed at RNA phage, this is not surprising given that so few RNA phage have been sequenced and spacer sequences matching any RNA phage have not (yet) been found. It is possible that species with RAMP encoded Cas proteins have an alternate mechanism of silencing invasive nucleic acid.

For prokaryotes, with both a single center for DNA replication (Ori) and a short life span, it is vital that the genome is small so that replication can be carried out quickly.<sup>52</sup> It would be costly to allow CRISPR arrays to grow unchecked and there is evidence of spacer deletion at the 3’ end of the array.<sup>47</sup> This would allow historic invasions to be forgotten if the cost of maintaining the surveillance was high. There are also higher levels of expression of mature crRNA from the leader end of the transcript (from more recently acquired spacers) and lower levels of distal sequences.<sup>47</sup> Such a strategy ensures protection against current threats, whilst keeping a low level of surveillance for the re-emergence of an old foe. The Dark Arts cannot be so easily subdued. Mutations in a 3-5 proto-spacer adjacent motif (PAM) can prevent proto-spacer sequence recognition, or target sequence mutation can evade crRNA target recognition, phage from *Leptospirillum* can reshuffle 25 nt blocks<sup>53</sup> confounding the slightly larger spacer sequences of the crRNA. And so it goes on. For the convenience of molecular evolutionists, CRISPRs give a historical insight into past host and pathogen encounters!

Thus it seems that all living cells can use some type of small RNA to defend themselves. Furthermore, we expect that there always would have been viruses, so the



simplest hypothesis at present is that small RNAs have always been used in Defence Against the Dark Arts. Eukaryotes, bacteria and archaea have all evolved slightly different methods of dealing with parasitic DNA, but the use of RNA is a common thread. The diversity in mechanisms shows evolution has continued, but reinforces that the whole network system is very old.

## OTHER REGULATORY RNAs

Our knowledge of noncoding RNA regulated pathways has expanded markedly in the past decade.<sup>54-56</sup> The roles of networks of regulatory and catalytic ribonucleoprotein particles (RNPs) in eukaryotes have been well studied, including classical examples such as the small-nuclear RNPs (snRNPs) in mRNA splicing,<sup>57,58</sup> small-nucleolar RNPs (snoRNPs) in rRNA processing—either the 2' hydroxyl of ribose is to be methylated, or a uracil converted to pseudouracil,<sup>59</sup> and the RNase P in tRNA processing.<sup>60</sup> Genome wide expression studies of eukaryotic model organisms have shown that over 90% of the genome is transcribed, which suggests that the richness of noncoding regulatory elements extends well beyond what is currently known.<sup>61</sup>

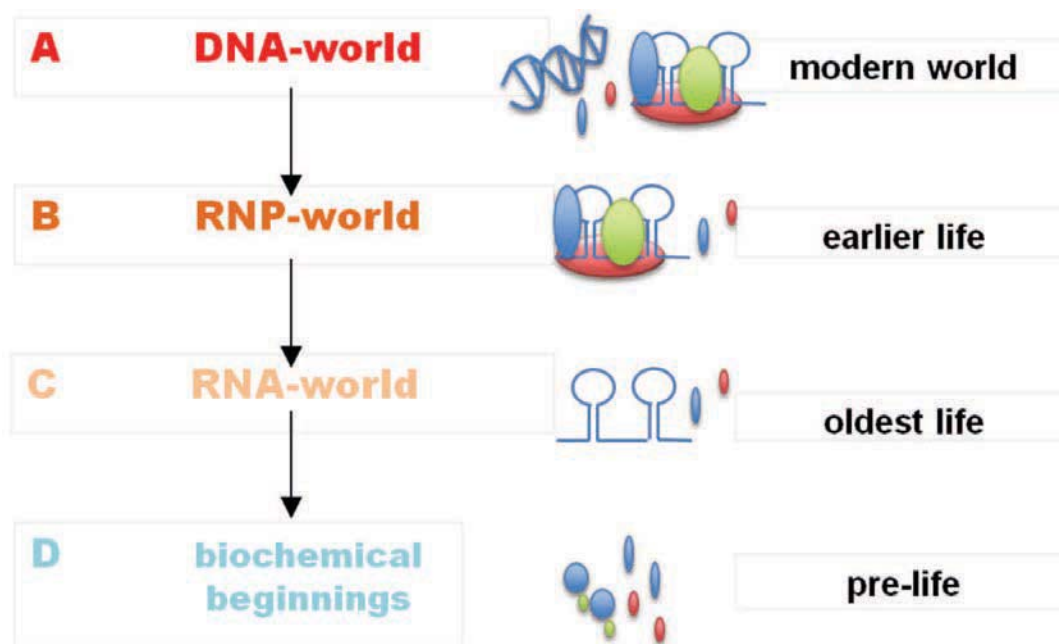
The conserved infrastructural network of RNP-mediated networks in eukaryotic cells indicates a pre-eukaryotic origin for many RNP functions.<sup>62</sup> Both experimental and computational genome-wide mining of noncoding RNAs have revealed that many regulatory RNAs have highly flexible expression patterns.<sup>63,64</sup> An example of such RNAs are snoRNAs that have been found in all eukaryotes and form one of the largest families of noncoding RNAs.<sup>3</sup> They are not restricted to rRNA biogenesis, snoRNAs in the Cajal body (“small Cajal body RNAs”—scaRNAs)<sup>65</sup> modify snRNAs that function in the spliceosome. They have been identified as precursor for smaller RNAs;<sup>66</sup> a human HBII-52 snoRNA (SNORD 115) is processed into smaller RNAs, which regulate alternative splicing of the serotonin-receptor gene,<sup>67</sup> and alternative splicing itself appears very ancient in eukaryotes.<sup>68</sup> Similarly, a C/D box snoRNA from the Epstein-Barr virus is processed into smaller RNAs.<sup>69</sup> Many snoRNAs in modern genomes appear to arise by duplications.<sup>70,71</sup> Large numbers of new snoRNAs are continuously being identified through experimental and bioinformatic screens.<sup>72</sup> All these findings, together with the existence of many orphan snoRNAs without known targets<sup>73</sup> suggest additional network interactions. Identification of snoRNA-derived small RNAs from such a wide range of organisms also implies that some alternative regulatory roles of snoRNAs have an ancient origin. In some eukaryotic lineages with smaller genomes (such as diplomonads and microsporidia) and in archaea, there are fewer annotated snoRNAs, but those that are there appear functional.<sup>74</sup>

Cis-regulatory RNAs have been identified in both eukaryotes and prokaryotes,<sup>3</sup> and classic examples include riboswitches in bacteria and plants,<sup>75</sup> iron-response elements (IREs),<sup>76</sup> and the eukaryotic histone 3'UTR stem-loop.<sup>77</sup> Cis-regulatory RNAs are usually located in the untranslated regions (UTRs) in mRNAs, though some are in coding regions.<sup>78</sup> In association with RNA-binding proteins they regulate translation, splicing, stability and localization. For proteins, there are hundreds of identified RNA-binding protein domains,<sup>79</sup> but the identification of complementary protein-binding motifs in RNA is still at an early stage. An important role of cis-regulatory RNA elements is mRNA localization in eukaryotes. Expression of many eukaryotic RNAs involves packaging of mRNA-protein complexes into transport particles, trafficking of the particles within the cytoplasm and finally translated at

their target destination. A few examples are localization of budding yeast ASH1 mRNA to the bud tip,<sup>80</sup> localization of  $\beta$ -actin mRNA to the leading edge of migrating fibroblasts,<sup>81</sup> localization of Nonos mRNA to the posterior end of the *Drosophila* embryo,<sup>82</sup> and localization of human Vimentin mRNA to the perinuclear region of the cytoplasm.<sup>3,83</sup> The noncoding motifs on the 3'-UTRs of the above mRNAs function to bind specific proteins for the transport of the mRNAs. Existence of these RNAs in all three kingdoms of life also hints at this network arising prior to the universal common ancestor (Luca).

## HOW OLD ARE THE DIFFERENT INTERACTIONS OF RNA?

In general a good case can be made for many of the general classes of RNA networks to arise relatively early in evolution,<sup>84,85</sup> but we need to outline the options carefully. Current approaches attribute importance to several overlapping earlier stages<sup>86</sup> during the origin of life (see Fig. 5). We start with the familiar world of DNA, proteins and RNA and work backwards to earlier stages. The biological world that we are familiar with has DNA as the main coding molecule and this allows several steps of error checking (against the complementary strand as just one example). The replication accuracy of DNA-based system is much higher than for RNA-coding systems and as far as we can



**Figure 5.** Working backwards through four stages for the origin of life. (A) Our modern world has DNA as the main information storage molecule and its double-stranded structure allows a higher replication accuracy. The standard model suggests an earlier RNA-protein (RNP) world (B) where RNA was the main molecule for genetic information and proteins had the main catalytic role. Many regulatory networks may date back to this time. At a still earlier stage (C), the standard model is an RNA-world where both catalysis and information storage was carried out by RNA—presumably aided by short (noncoded) peptides. The error rate is assumed to be relatively high and thus only relatively short RNAs could be coded. Earlier still there must have been a chemical stage (D) where perhaps autocatalytic cycles were important. It is not specified when membranes (allowing protocells) first arose.



tell, the size of even the larger eukaryote genomes is no longer limited by the accuracy of replication of DNA.

Our modern world (Fig. 5A) has the three familiar groups of macromolecules—proteins, RNA and DNA. On any evolutionary scenario there will have been earlier and simpler worlds and under the standard theory it is predicted that our modern world was preceded (Fig. 5B) by a ribonuclear protein world (RNP) with RNA having the coding roles and proteins doing most of the catalysis. Part of the evidence that DNA evolves last comes from the necessity for protein enzymes (ribonucleotide reductases) that use free radical mechanisms to reduce an OH of ribose (in a ribonucleotide RNA precursor) to a deoxy-ribonucleotide.<sup>87</sup> As far as we know, only proteins (not RNA) can catalyse such complex free radical reactions. In the proposed ribonucleoprotein-world (RNP of Fig. 5B) protein is doing most catalysis (especially of small molecules), leaving RNA with the main coding function and RNA-protein networks would probably have had many regulatory roles.

From our knowledge of modern biochemistry, protein enzymes copying single-stranded RNA are much less accurate and have higher error rates than copying the double stranded DNA molecule. This lower accuracy means that RNA genomes would be much shorter than DNA-based organisms. Indeed, we see today that RNA viruses have much shorter genomes than double-stranded DNA viruses.<sup>88</sup> Effectively there is a positive feedback cycle of increasing fidelity of replication allowing longer coding sequences, which allows a further increase in replication fidelity—we call this the Darwin-Eigen cycle.<sup>89</sup> But the important point here is that from this stage we will have many RNA-protein interactions and so, in principle, some of the RNA-protein networks could have been established at this stage. Or were they lost in Luca and then later re-established—but only in eukaryotes? Losing them and later re-establishing them, seems a less likely hypothesis.

At this point (ribonucleic acids and proteins) we are still on reasonably firm ground; but what comes earlier? Currently the best hypothesis for the earlier stage (Fig. 4C) is that RNA is the first polymer that had the ability to reproduce itself (either directly, or indirectly through a cycle and with or without assistance of short noncoded peptides). It has long been considered<sup>90</sup> that enzyme cofactors with a ribodinucleotide structure (such as NAD, FAD, NADP) were remnants of an early RNA-world. The evidence for this hypothesis of the early role of RNA is increasing.<sup>91</sup> In this proposed RNA-world, RNA would have had coding, catalytic and regulatory roles. An interesting point is that some RNAs and their interactions, appear to date from this proposed RNA-world. RNA is not as an effective catalyst as proteins<sup>85</sup> (probably because proteins form more specific and stable 3D structures) and so this leaves space for improved protein catalysts in going from the second to the third stage of Figure 5. There had to be earlier stages before organised polymers formed (whether RNA, protein, or DNA) and we summarize these stages as ‘biochemical beginnings’ (Fig. 5D).<sup>1</sup> Basically by definition, we do not expect any macromolecules to have persisted from this stage.

Some classes of functional and regulatory RNA are so widespread that they are accepted as ‘universal’ and are inferred to have been present in the last common ancestor, Luca. These include rRNA, tRNA, mRNA, RNase P, SRP RNA. A striking discovery was that the catalytic core of the ribosome (forming the amino-acyl bonds linking amino acids into proteins) was composed of RNA. In other words, the ribosome was a ribozyme.<sup>92</sup> Thus the three core aspects of protein synthesis—the messenger, transfer RNA and ribosome—are all RNA molecules that have a network of interactions. Evolutionarily this makes sense if an RNA-world preceded an RNP-world. However there are few convincing theories

for the origin of protein synthesis because we need these three classes of RNA before proteins can be synthesized. In an evolutionary context, we cannot evolve something ‘because it will be useful in the distant future’—it must have a function ‘here and now’. Certainly rRNA, tRNA and mRNA must date back to an RNA-world (before proteins). Sometimes we can define useful subquestions—such as whether the intron/exon structure, with its associated RNAs involved in splicing, is ancestral to all modern eukaryotes.<sup>93</sup> Some RNAs and their interactions may yield progress relatively easily; others will be more difficult for now.

A recent discovery is that tRNAs are part of the proof-reading that increases accuracy,<sup>94</sup> the tRNA(Leu) has catalytic activity in removing a mischarged tRNA(Leu), even though protein does help stabilise some of the intermediates along the catalytic pathway. This means that we need to consider that tRNAs are not passive in amino acid recognition; they are also part of the regulatory network. Although not so well known as the RNAs involved in protein synthesis, the signal recognition particle, 7S in eukaryotes,<sup>95</sup> also appears to occur in all living systems, so this RNA-protein network also appears ancient.

Were these RNA interactions all present in Luca? We will not follow this question here, but have written on it<sup>89</sup> under the names introns ‘first’, ‘early’ or ‘late’. Because the position of the root of the eukaryote tree is not certain,<sup>96</sup> our strategy<sup>93</sup> has been to search for RNAs that occur in all of the five (or six) deepest lineages of eukaryotes. For example, the intron/exon structure of eukaryote genomes, with its requirement for splicing, appears universal in all modern eukaryotes, so it is expected to have been in the last common eukaryote ancestor (Fred). But that only puts the problem back further. For example, did snRNAs arise before, or after, another milestone—the endosymbiotic event that led to the origin of mitochondria? Details on the relative ages of classes will depend on the models of genome reduction.<sup>97</sup> There are plenty of interesting and fundamental questions for the next decade.

## CONCLUSION

Every eukaryote group appears to have networks that handle small RNAs a little differently. However, this is likely a normal evolutionary process creating variants arising from an early ancestral system in Fred. Perhaps it is still premature to decide how the ancestral eukaryote functioned in this respect, but the discovery of different forms of RNA networks in so many groups of eukaryotes has turned the focus back to earlier stages.

Parasitic RNA would have been present in the RNA and RNP-worlds; that is the nature of biology. Thus we expect that at least from the time of Luca there would have been mechanisms to protect the early genome from parasites of various kinds—whether viral or transposon-like. We expect that the Defence Against the Dark Arts idea would have always been a problem for cells since they all require some type of defence mechanism. Although the machinery differs, prokaryotes and eukaryotes have networks with roughly the same function using small RNAs to keep the Dark Arts at bay. It seems likely that the siRNA system was present in Luca as the Argonautes are in all three domains and are almost universally conserved, indicating their age and importance. The endo and piRNA pathways may have evolved to suppress parasitic nucleic acid that escaped or overrun the exo-siRNA pathway.

Increasingly, more RNAs seem to be universal. Bacteria and viruses have many cis-regulatory elements which bind proteins or metabolites (riboswitches), plants and fungi have riboswitches and now it looks like that cis-regulatory elements (in 5'- and 3'-UTRs) are very common in eukaryotes too. Large scale transcription studies in human and mouse at least have uncovered many UTR-derived RNAs, many of which are likely to contain cis-regulatory elements and they may work in a similar way as those in bacteria and virus (binding to specific proteins and having regulatory roles in transcription).

At present we must keep an open mind about how old the RNA and proteins systems are that form the RNA networks. At present, we favour the simplest hypothesis that the RNA networks are, in their basic form, very old. But we must keep testing this hypothesis, looking for small RNA networks involved in regulation in as wide a range of groups as possible. It is a stimulating time for research on RNA networks.

## ACKNOWLEDGEMENTS

We thank Lesley Collins for important editing and comments for this chapter.

## REFERENCES

1. Yarus M. Getting past the RNA world: The initial Darwinian ancestor. Cold Spring Harbor Perspectives in Biology 2010. doi: 10.1101/cshperspect.a003590.
2. Penny D, Collins LJ. Evolutionary genomics leads the way. In: Caetano-Anolles, ed. Evolutionary Genomics and Systems Biology. Hoboken: Wiley-Blackwell, 2010:3-16.
3. Gardner PP, Daub J, Tate JG et al. Rfam: updates to the RNA families database. Nucl Acids Res 2009; 37:D136-D140.
4. Boria I, Gruber AR, Tanzer A et al. Nematode sbRNAs: homologs of vertebrate Y RNAs. J Mol Evol 2010; 70:346-358.
5. Ghildiyal M, Zamore PD. Small silencing RNAs: an expanding universe. Nat Genet 2009; 10:102-108.
6. Bartel DP, Chen CZ. Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. Nat Rev 2004; 5:396-400.
7. Carlsbecker A, Lee JY, Roberts CJ et al. Cell signalling by microRNA165/6 directs gene dose-dependent root cell fate. Nature 2010; 465:316-321.
8. Cullen BR. Viral and cellular messenger RNA targets of viral microRNAs. Nature 2009; 457:421-425.
9. Ouellet DL, Plante I, Landry P et al. Identification of functional microRNAs released through asymmetrical processing of HIV-1 TAR element. Nucl Acids Res 2008; 36:2353-2365.
10. Lin J, Cullen BR. Analysis of the interaction of primate retroviruses with the human RNA interference machinery. J Virol 2007; 81:12218-12226.
11. Dunoyer P, Himber C, Voinnet O. Induction, suppression and requirement of RNA silencing pathways in virulent agrobacterium tumefaciens infections. Nat Genet 2006; 38:258-263.
12. Dong Z, Han MS, Federoff N. The RNA-binding proteins HYL1 and SE promote accurate in vitro processing of pri-miRNA by DCL1. PNAS 2008; 105:9970-9975.
13. Okamura K, Liu N, Lai EC. Distinct mechanisms for microRNA strand selection by *Drosophila* Argonautes. Mol Cell 2009; 36:431-444.
14. Berezikov E, Chung W, Willis J et al. Mammalian mirtron genes. Mol Cell 2007; 28:328-333.
15. Zamore PD, Haley B. Ribo-gnome: the big world of small RNAs. Science 2005; 309:1519-1524.
16. Moazed D. Small RNAs in transcriptional gene silencing and genome defence. Nature 2009; 457:413-420.
17. Carthew RW, Sontheimer EJ. Origins and mechanisms of miRNAs and siRNAs. Cell 2009; 136:642-655.
18. Leung AKL, Sharp PA. MicroRNAs: A safeguard against turmoil? Cell 2007; 130:581-585.
19. Orom UA, Nielsen FC, Lund AH. MicroRNA-10a binds the 5'UTR of ribosomal protein mRNAs and enhances their translation. Mol Cell 2008; 30:460-471.
20. Lee YS, Shibata Y, Malhotra A et al. A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). Genes and Dev 2009; 23:2639-2649.

21. Haussecker D, Huang Y, Lau A et al. Human tRNA-derived small RNAs in the global regulation of RNA silencing. *RNA* 2010; 16:673-695.
22. Kulkarni M, Ozgur S, Stoecklin G. On track with P-bodies. *Biochem Soc Trans* 2010; 38:242.
23. Guang S, Bochner AF, Pavelec DM et al. An Argonaute transports siRNAs from the cytoplasm to the nucleus. *Science* 2008; 321:537-541.
24. Bartel B. MicroRNAs directing siRNA biogenesis. *Nat Struct Mol Biol* 2005; 12:569-571.
25. Miyoshi K, Miyoshi T, Hartig JV et al. Molecular mechanisms that funnel RNA precursors into endogenous small-interfering RNA and microRNA biogenesis pathways in *Drosophila*. *RNA*. 2010; 16:506-515.
26. Robine N, Lau NC, Balla S et al. A broadly conserved pathway generates 3'UTR-directed primary piRNAs. *Curr Biol* 2009; 19:2066-2076.
27. Malone CD, Hannon GJ. Small RNAs as guardians of the genome. *Cell* 2009; 136:656-668.
28. Li C, Vagin VV, Lee S et al. Collapse of germline piRNAs in the absence of Argonaute3 reveals somatic piRNAs in flies. *Cell* 2009; 137:509-521.
29. Gottesman S. Micros for microbes: noncoding regulatory RNAs in bacteria. *Trends Genet.* 2005; 21:399-404.
30. de Nooijer S, Holland BR, Penny D. Eukaryote origins: there was no Garden of Eden? *PLoS One* 2009; 4:e5507.
31. Boerlijst MC, Hogeweg P. Spiral wave structure in prebiotic evolution—hypercycles stable against parasites. *Physica D* 1991; 48:17-28.
32. Yigit E, Batista PJ, Bei Y et al. Analysis of the *C. elegans* Argonaute family reveals that distinct Argonautes act sequentially during RNAi. *Cell* 2006; 127:747-757.
33. de Vries W, Berkhout B. RNAi suppressors encoded by pathogenic human viruses. *Int J Biochem Cell Biol* 2008; 40:2007-2012.
34. Parameswaran P, Sklan E, Wilkins C et al. Six RNA viruses and forty-one hosts: viral small RNAs and modulation of small RNA repertoires in vertebrate and invertebrate systems. *PLoS Pathogens* 2010; 6:e1000764.
35. Ruiz-Ferrer V, Voinnet O. Roles of plant small RNAs in biotic stress responses. *Annu Rev Plant Biol* 2009; 60:485-510.
36. Lipardi C, Paterson BM. Identification of an RNA-dependent RNA polymerase in *Drosophila* involved in RNAi and transposon suppression. *PNAS* 2009; 106:15645-15650.
37. Maida Y, Yasukawa M, Furuuchi M et al. An RNA dependent RNA polymerase formed by hTERT and the RNase MRP RNA. *Nature* 2009; 461:230-235.
38. Watanabe T, Totoki Y, Toyoda A et al. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* 2008; 453:539-543.
39. Hartig JV, Tomari Y, Forstemann K. piRNAs—the ancient hunters of genome invaders. *Genes Dev* 2007; 21:1707-1713.
40. Brennecke J, Malone CD, Aravin AA et al. An epigenetic role for maternally inherited piRNAs in transposon silencing. *Science* 2008; 322:1387-1392.
41. Ding S, Voinnet O. Antiviral immunity directed by small RNAs. *Cell* 2007; 130:413-426.
42. Wang H, Buckley KJ, Yang X et al. Adenosine kinase inhibition and suppression of RNA silencing by geminivirus AL2 and L2 proteins. *J Virol* 2005; 79:7410-7418.
43. Raja P, Sanville BC, Buchmann RC et al. Viral genome methylation as an epigenetic defence against geminiviruses. *J Virol* 2008; 82:8997-9007.
44. Jansen R, Embden JD, Gastra W et al. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* 2002; 43:1565-1575.
45. Touchan M, Rocha EPC. The small, slow and specialized CRISPR and anti-CRISPR of *Escherichia* and *Salmonella*. *PLoS One* 2010; 5(6):e11126.
46. Bolotin A, Quinquis B, Sorokin A et al. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* 2005; 151:2551-2561.
47. Wiedenheft B, Zhou K, Jinek M et al. Structural basis for DNase activity of a conserved protein implicated in CRISPR-mediated genome defense. *Structure* 2009; 17:904-912.
48. Andersson AF, Banfield JF. Virus population dynamics and acquired virus resistance in natural microbial communities. *Science* 2008; 320:1047-1050.
49. Marraffini LA, Sontheimer EJ. Self versus nonself discrimination during CRISPR RNA-directed immunity. *Nature* 2010; 463:568-571.
50. Marraffini LA, Sontheimer EJ. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* 2008; 322:1843-1845.
51. Hale C, Kleppe K, Terns RM et al. Prokaryotic silencing (psi)RNAs in *Pyrococcus furiosus*. *RNA* 2008; 14:2572-2579.
52. Penny D, Poole AM. Lateral gene transfer: some theoretical aspects. *NZ BioScience* 2003; 12:32-35.
53. Andersson AF, Banfield JF. Virus population dynamics and acquired virus resistance in natural microbial communities. *Science* 2008; 320:1047-1050.



54. Licatalosi DD, Darnell RB. RNA processing and its regulation: global insights into biological networks. *Nat Rev Genet* 2010; 11:75-87.
55. Lioliou E, Romilly C, Romby P et al. RNA-mediated regulation in bacteria: from natural to artificial systems. *New Biotech* 2010; 27:222-235.
56. Sashital DG, Doudna JA. Structural insights into RNA interference. *Curr Opin Struct Biol* 2010; 20:90-97.
57. Newman AJ, Nagai K. Structural studies of the spliceosome: blind men and an elephant. *Curr Opin Struct Biol* 2010; 20:82-89.
58. Staley JP, Woolford JL Jr. Assembly of ribosomes and spliceosomes: complex ribonucleoprotein machines. *Curr Opin Cell Biol* 2009; 21:109-118.
59. Bachellerie JP, Cavaillé J, Huttenhofer A. The expanding snoRNA world. *Biochimie* 2002; 84:775-790.
60. Lai LB, Vioque A, Kirsebom et al. Unexpected diversity of RNase P, an ancient tRNA processing enzyme: challenges and prospects. *FEBS Lett* 2010; 584:287-296.
61. Carninci P. RNA dust: where are the genes? *DNA Res* 2010; 17:51-59.
62. Collins LJ, Penny D. The RNA infrastructure: dark matter of the eukaryotic cell? *Trends Genet* 2009; 25:120-128.
63. Dieci G, Preti M, Montanini B. Eukaryotic snoRNAs: a paradigm for gene expression flexibility. *Genomics* 2009; 94:83-88.
64. Slezak-Prochazka I, Durmus S, Kroesen BJ et al. MicroRNAs, macrocontrol: regulation of miRNA processing. *RNA* 2010; 16:1087-1095.
65. Kiss AM, Jady BE, Darzacq X et al. A Cajal body-specific pseudouridylation guide RNA is composed of two box H/ACA snoRNA-like domains. *Nucl Acids Res* 2002; 30:4643-4649.
66. Taft RJ, Glazov EA, Lassmann T et al. Small RNAs derived from snoRNAs. *RNA* 2009; 15:1233-1240.
67. Kishore S, Khanna A, Zhang et al. The snoRNA MBII-52 (SNORD 115) is processed into smaller RNAs and regulates alternative splicing. *Hum Mol Genet* 2010; 19:1153-1164.
68. Irimia M, Rukov JL, Penny D et al. Functional and evolutionary analysis of alternatively spliced genes is consistent with an early eukaryotic origin of alternative splicing. *BMC Evol Biol* 2007; 7:188.
69. Huttinger R, Feederle R, Mrazek J et al. Expression and processing of a small nucleolar RNA from the Epstein-Barr virus genome. *PLoS Pathog* 2009; 5:e1000547.
70. Chen CL, Liang D, Zhou et al. The high diversity of snoRNAs in plants: identification and comparative study of 120 snoRNA genes from *Oryza sativa*. *Nucl Acids Res* 2003; 31:2601-2613.
71. Zemmann A, op de Bekke A, Kieffmann M et al. Evolution of small nucleolar RNAs in nematodes. *Nucl Acids Res* 2006; 34:2676-2685.
72. Raabe CA, Sanchez CP, Randau et al. A global view of the nonprotein-coding transcriptome in *Plasmodium falciparum*. *Nucl Acids Res* 2010; 38:608-617.
73. Hertel J, Hofacker IL, Stadler PF. SnoReport: computational identification of snoRNAs with unknown targets. *Bioinformatics* 2008; 24:158-164.
74. Gardner PP, Bateman A, Poole AM. SnoPatrol: how many snoRNA genes are there? *J Biol* 2010; 9:4.
75. Henkin TM. Riboswitch RNAs: using RNA to sense cellular metabolism. *Genes Dev* 2008; 22:3383-3390.
76. Pantopoulos K. Iron metabolism and the IRE/IRP regulatory system: an update. *Ann N Y Acad Sci* 2004; 1012:1-13.
77. Dominski Z, Marzluff WF. Formation of the 3' end of histone mRNA: getting closer to the end. *Gene* 2007; 396:373-390.
78. Nguyen MQ, Zhou Z, Marks CA et al. Prominent roles for odorant receptor coding sequences in allelic exclusion. *Cell* 2007; 131:1009-1017.
79. Finn RD, Mistry J, Tate J et al. The Pfam protein families database. *Nucl Acids Res* 2010; 38:D211-D222.
80. Gonsalvez GB, Urbinati CR, Long RM. RNA localization in yeast: moving towards a mechanism. *Biol Cell* 2005; 97:75-86.
81. Condeelis J, Singer RH. How and why does beta-actin mRNA target? *Biol Cell* 2005; 97:97-110.
82. Kugler JM, Lasko P. Localization, anchoring and translational control of oskar, gurken, bicoid and nanos mRNA during *Drosophila* oogenesis. *Fly* 2009; 3:15-28.
83. Bermanno G, Shepherd RK, Zehner ZE et al. Perinuclear mRNA localisation by vimentin 3'-untranslated region requires a 100 nucleotide sequence and intermediate filaments. *FEBS Lett* 2001; 497:77-81.
84. Collins LJ, Chen XS. Ancestral RNA: the RNA biology of the eukaryotic ancestor. *RNA Biol* 2009; 6:1-8.
85. Jeffares DC, Poole AM, Penny D. Relics from the RNA world. *J Mol Evol* 1998; 46:18-36.
86. Penny D. An interpretive review of the origin of life research. *Biol Philos* 2005; 20:633-671.
87. Poole A, Penny D, Sjöberg BM. Confounded Cytosine! Tinkering and the evolution of DNA. *Nat Rev Mol Cell Biol* 2001; 12:147-151.
88. Duffy S, Shackelton LA, Holmes EC. Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet* 2008; 9:267-276.
89. Penny D, Hoepfner MP, Poole AM et al. An overview of the introns-first theory. *J Mol Evol* 2009; 69:527-540.

90. White HB. Coenzymes as fossils of an earlier metabolic state. *J Mol Evol* 1976; 7:101-104.
91. Yarus M. *Life from an RNA World: The Ancestor Within*. Cambridge:Harvard University Press, 2010.
92. Steitz TA, Moore PB. RNA, the first macromolecular catalyst: the ribosome is a ribozyme. *Trends Biochem Sci* 2003; 28:411-418.
93. Collins LJ, Penny D. Complex spliceosomal organization ancestral to extant eukaryotes. *Mol Biol Evol* 2005; 22:1053-1066.
94. Hagiwara Y, Field MJ, Nureki O et al. Editing mechanism of aminoacyl-tRNA synthetases operates by a hybrid ribozyme/protein catalyst. *J Am Chem Soc* 2010; 132:2751-2758.
95. Rosenblad MA, Gorodkin J, Knudsen B et al. SRPDB: signal recognition particle database. *Nucl Acids Res* 2003; 31:363-364.
96. Keeling PJ, Burger G et al. The tree of eukaryotes. *Trends Ecol Evol* 2005; 20:670-676.
97. Poole AM, Penny D. Lateral gene transfer, some theoretical aspects. *NZ BioScience* 2003:32-35.



# Chapter Four: The Evolution of the Argonautes

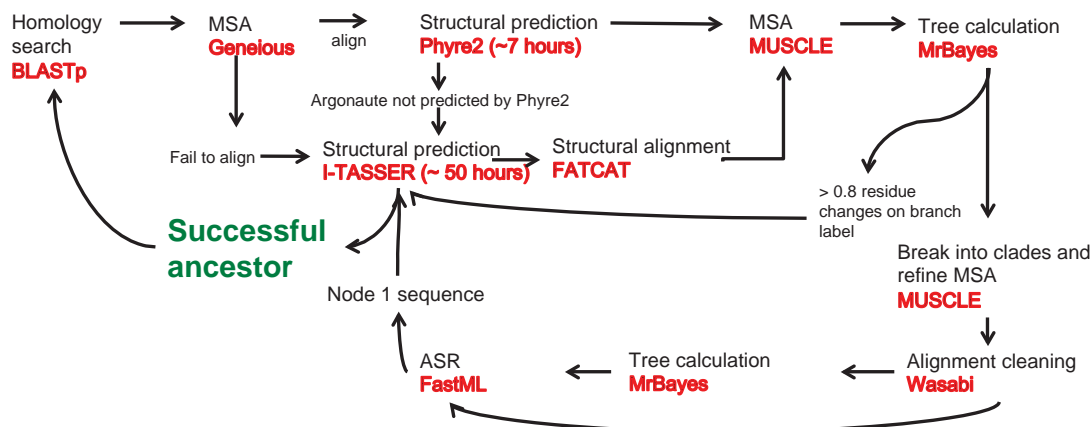
## 4. An investigation into Argonaute evolution using 3-D structural prediction

To determine how useful the method is more generally, and in line with the preference for ribonucleoproteins, the argonaute family was chosen. In order to show that the principle of using three-dimensional structural prediction plus additional evidence of relatedness at the level of protein folds is flexible, the chosen protein is quite different. Although argonaute domains are well described, argonautes are not constrained by the very precise shape and docking requirements of the vault protein, but are under a functional constraint. The argonaute family of proteins have literally thousands of sequences in the databases and number of publications describing new functions has reached dizzying heights.

Argonautes are a family of proteins and may be inferred to have the same (or very similar functions) as new species have evolved, but they are also complicated by the huge number of paralogous sequences arising from intra-genome duplications. Following the intra-genomic duplication process, the redundant gene is freed from selective pressure and can acquire new functions or change its expression pattern or specificity. In addition to this there are literally hundreds of splice variants in some species adding another layer of complexity.

There is also a veritable alphabet soup of small non-coding RNAs that guide them to their target sequence. That argonautes are very old is not in dispute and the nucleic acid component (bacteria and archaea can use DNA as well as RNA) is likely to have been crucial then as it is now. The argonautes are similar in length to the Major Vault Protein but the structure could not be more different and different kinds of evidence are required before they can be assigned to a group to create an ancestor. This story tells of problems with trees and alignments as well as unexpected but useful observations.

The workflow has been adapted for the argonaute study as shown in the figure below. A list of criteria for the inclusion of a sequence in the reconstruction of an ancestor is listed in the table. The criteria cannot be stringently applied because well-characterised argonautes are shown to fall outside of the I-TASSER C score ( $> -1.5$ ) for confidence that the fold is correct (described in detail in chapter 4a). Some sequences fail on a number of counts and can be rejected quite simply. Conversely most sequences pass through the pipeline quite quickly as many are clear homologs and require little consideration.



**Fig. 4.1 Workflow for chapter four.**

The MSA and tree building are carried out on a mini Mac (2.7 GHz Intel core i7). The passage of time for each process will vary considerably according to complexity. Of the servers; the I-TASSER time of ~ 50 hours is computing time, actual time will vary depending on queue length. FATCAT and Wasabi are quite reliably less than three minutes with FastML being more variable. Following the structural prediction step(s) the structures are scanned for the catalytic tetrad and C terminal sequence (described in chapter 4a) as these can add evidence particularly if the C (confidence) score for the structural prediction is poor.

The argonaute story has been split into three parts, chapter 4a metazoa, chapter 4b fungi and chapter 4c the remaining eukaryotes. Metazoa (4a) and fungi (4b) have both been submitted for publishing and 4c (the remainder of the eukaryotes) is still in process. There is then necessarily some repetition as each part of the tale is designed to stand alone as an article. As these have not yet been published there is no requirement to keep to a particular formatting style so these are presented with numbered headings and the referencing as part of the total referencing section for the thesis. Essential supplementary material is supplied in the appendix with exceptions as described where trees for instance are too large to be read even at A3 size and can be found in wikispaces. Posterior probability logors for all of the ancestral reconstructions can also be found in wikispaces.

**Table: Cut off criteria used for the search for distant argonaute sequences**

BLASTp	If the bulk of the length of the query is covered a sequence will be considered regardless of a poor E or bit score. Additionally because the size of the search space will vary between searches the numerical values are not comparable.
MSA Geneious	Geneious alignments are stringent and if a sequence will not align with the majority of the sequences then it is flagged for I-TASSER structural prediction. The sequences that will align are checked via Phyre2
Phyre 2	Phyre 2 will result a structure that is matched to folds found in the PDB but will miss out stretches of sequence that are not matched. A level of confidence is displayed as a percentage and also the percentage of residues resulted in the matched folds. If this is matched to a known argonaute with >90% coverage then the sequence will proceed to a MUSCLE alignment, if not it will be sent to I-TASSER.
I-TASSER	I-TASSER has the capacity for <i>ab initio</i> modeling as well as fold matching and results a maximum of five models. A confidence score greater than $-1.5$ is indicative of a correct fold (scoring is from $-5$ to $+2$ ). However many well documented argonautes (AGO and PIWI) score less than $-1.5$ due to inserts that are modelled well away from the catalytic area. The practical difference between the utility of Phyre2 and I-TASSER is described in chapter five.
Catalytic tetrad	From the I-TASSER predicted structure the catalytic residues (DEDH) in the anticipated positions, or known alternatives (e.g. DEDD/R/K) adds additional evidence when considering a poor C score.
C terminal signature	The C terminal last four residues tuck inside the argonaute close to the 5' RNA binding pocket. Met (two aromatic residues) and a final more variable hydrophobic residue makes it most likely to be an AGO-L argonaute and Leu (two aromatic residues) and usually Leu make this more likely to be a PIWI-L argonaute. The C terminal signature is taken into consideration to support other structural evidence.
FATCAT	If in excess of 70% of residues are in equivalent spacial positions, the P score is zero and the raw score is greater than 1500, then this increases the evidence for structural homology.
MrBayes tree	The branch label describes the number of residue changes per site in the initial trees. If this is greater than 0.8 then the sequences is flagged for I-TASSER prediction if it hasn't already been assessed.
Literature search	Supporting literature may provide added evidence where only some of the above criteria are met.



## **Chapter 4a: The evolutionary history of Argonaute and PIWI in metazoa by ancestral protein inference and structure prediction**

### **4.1. Abstract**

Traditionally, molecular evolution studies analyse the primary structure of genes and proteins. However, with more three-dimensional structures of proteins being determined, or predicted, and because evolutionary divergences lose information at very old times when analysed via Markov models, it is important to include 3-D (three dimensional) structural information for deeper evolutionary analysis. Our test data set uses argonaute proteins, both AGO-like and PIWI-like. BLAST searches may identify superficially similar sequences that are not predicted to fold consistent with being a functional PIWI domain, and may also miss some sequences with limited similarity that are predicted to fold to resemble AGO/PIWI structures. In addition, multiple sequence alignments (MSA) do not always correctly reflect the position of the residues when sequences are analysed by structural prediction. We also investigate how the sequence at the root can affect our trees by demonstrating the effect of ten different root sequences on otherwise identical trees.

We investigate the ancestry of metazoan argonautes, demonstrating that our analysis methodology uses more information than is available directly from primary sequences alone. Our approach models the 3-D structure of proteins and their putative homologs with over 180 protein structures predicted. We use Ancestral Sequence Reconstruction (ASR) and tertiary structure analysis to clarify the evolutionary history of both the AGO and PIWI proteins. We find that there is conservation of catalytic residues in the PIWI domain (common to both AGO-like and PIWI-like proteins), and conservation of C terminal residues that differ between AGO-like and PIWI-like proteins. These methods and observations can be used generally to understand more about protein evolution.

### **4.2. Introduction**

There is a potential problem for the annotation of proteins as whole genome sequencing becomes more routine, and the difficulty arises from several sources. Firstly there is the problem arising from the ability of proteins to evolve into new structures (Alva *et al.*,

2015). Secondly there is the difficulty from the use of Markov models because it has been shown mathematically that recurrent mutations in aligned sites will obscure deeper divergences (Mossel and Steel, 2004). Thirdly there are some problems with existing annotations formulated before the full data was available. As more 3-D protein structures are determined, and as the prediction of 3-D structures improves, it is important to use the full range of techniques to maximise the evolutionary information obtained from sequence data.

We do not infer phylogeny on the basis of one gene, rather we use the argonautes to demonstrate methodology issues that are already known; we study poorly-supported trees, incorrect annotation and alignment errors. However, there are several reasons for optimism, we use proven software that is freely available and simple to use in order to improve confidence and reveal anomalies that can then be assessed. We apply both ancestral sequence reconstruction (ASR) (Ashkenazy *et al.*, 2012) and the prediction of tertiary structure, to the argonaute family of proteins in metazoa to demonstrate the wealth of information attainable from primary sequence but usually not considered. Our interest in the argonaute family of proteins is because of the unusual pattern of loss and duplication of the argonaute genes across all eukaryotes.

#### **4.2.1. Argonaute proteins**

The metazoan argonaute family is divided into three main sub-families: 1, argonautes ‘AGO-like’ (that are ubiquitous in metazoa), 2, ‘PIWI-like’ (they are often thought to be gonad specific in metazoa) and 3, WAGO (Worm AGO) or ‘group 3 argonautes’ (found in *Caenorhabditis elegans*) (Vastenhouw *et al.*, 2003). This latter group 3 appears to have been the result of a burst of gene duplication specific to nematodes (Yigit *et al.*, 2006; Farazi *et al.*, 2008), and retained in some *Caenorhabditis* species (Dalzell *et al.*, 2011) but lost in parasitic nematodes. *Trichinella* species appear to have some unusually long sequences annotated as argonautes in addition to one ordinary AGO-like protein, but this looks like a recent occurrence. Group 3 argonautes are not considered ancestral to metazoa and for this reason we have not included them in this study.

AGO and PIWI sub-families are important regulatory proteins in the metazoan cell, and are related by their use of small RNAs to guide them to RNA or DNA to regulate gene expression. There are other proteins involved in the pathways that create these small RNAs (most notably the Dicer endonuclease), and the RNAs differ by their size and genesis. Endogenous genes can be regulated through post-transcriptional



silencing via microRNAs (miRNAs) produced from longer dsRNA (and processed via Dicer). Similarly, parasitic nucleic acids such as viruses can be targeted via RNA interference (RNAi) using RNA copied from the parasitic sequence via RNA dependent RNA Polymerase (RdRP).

Less well described are the PIWI interacting RNAs (piRNA – those that interact with the PIWI-type argonaute including repeat associated RNA). Metazoa have two classes of PIWI-like proteins that partner up in the so-called ‘ping-pong’ amplification of secondary piRNAs (Czech and Hannon, 2016), large numbers of piRNAs can be copied from existing piRNAs rather than via longer genomic transcripts, thus avoiding Dicer processing. These piRNAs typically target transposon and repeat sequences. This mechanism has been well described in *Drosophila* germline cells (Brennecke *et al.*, 2007) and has also been reported in *Hydra* (Lim *et al.*, 2014). However piRNAs, with the hallmarks of the ping-pong facility, are found in the sponge (*Amphimedon queenslandica*) and the sea anemone (*Nematostella vectensis*) even though they do not have specific germline cells (Grimson *et al.*, 2008). The retention of the two types of PIWI protein suggests that this facility is widespread in metazoa at least. piRNAs are also found in plants and some yeast that do not have PIWI type proteins (Aravin and Tuschl, 2005). Indeed a large majority of the piRNAs in mammals derive from non-transposon, gene-rich areas of the genome, and are processed by Dicer and act outside of the germline cells (Ross *et al.*, 2014). So PIWI and piRNAs are more general than just in germline cells.

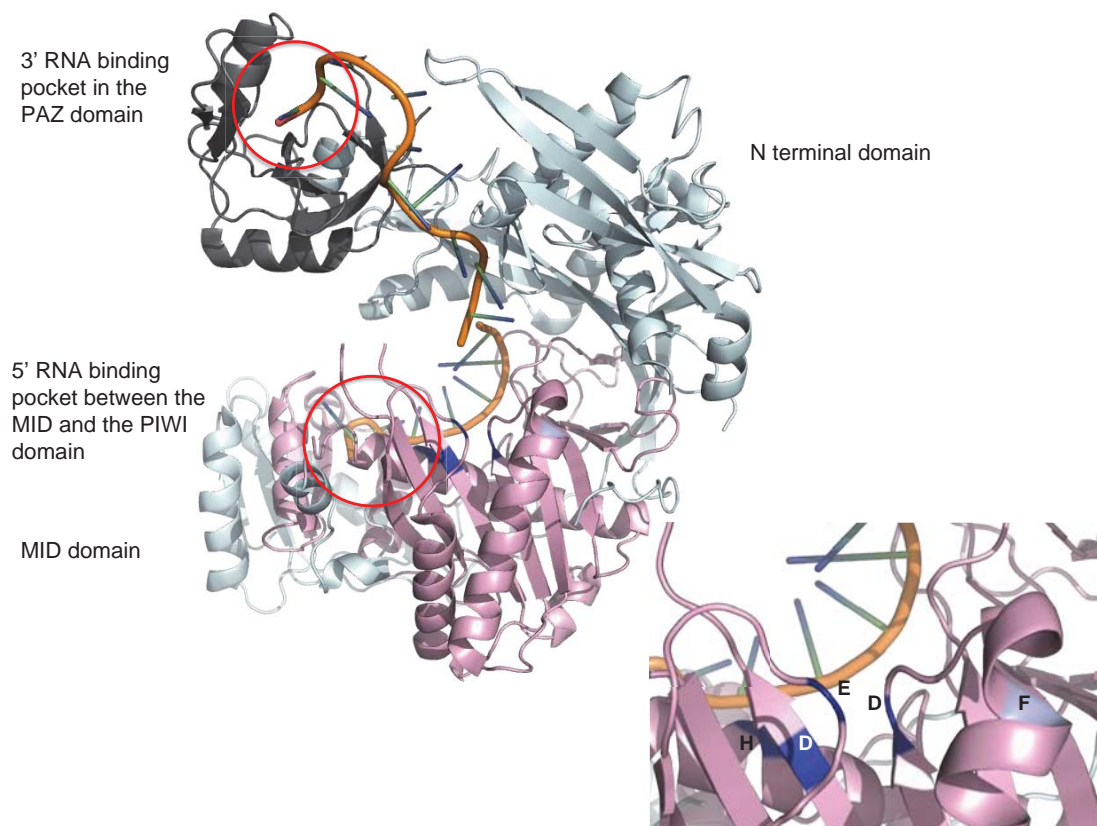
Importantly, although PIWI are often thought to be more recent than AGO proteins from an evolutionary point of view (because of their role in the germline), our results show that they appear to be much older than the development of specific germline cells. This is supported by an in depth review of the PIWI protein (Juliano *et al.*, 2011) that links the PIWI protein with the capacity of stem cells to generate copies of themselves; this would have been essential long before specialised germlines.

AGO and PIWI-type proteins also regulate gene expression via mechanisms other than RNA cleavage. For example, argonautes with bound guide-RNA and associated proteins (known as the RISC complex) can sterically impede translation by binding to the mRNA 3' UTR, first observed in *C. elegans* (Olsen and Ambros, 1999). In addition miRNA:RISC complexes can compete for binding to ribosomes, or to RNA 7-methylguanosine caps, or prevent circularisation of the mRNA by inducing de-

adenylation (circularisation enhances mRNA translation, and is reviewed in (Fabian *et al.*, 2010)). These latter two strategies would also promote the degradation of mRNA unprotected by cap or polyA tail structures. Alternatively, even if translation has begun the miRNA:RISC complex causes ribosomal subunits to either fail to associate competently, or to release the mRNA prematurely by recruiting ribosome anti-association factor eIF6 (Chendrimada *et al.*, 2007). Linearising the mRNA, or interfering with the cap and tail structures, is likely to result in its hydrolysis by nucleases - we term this 'secondary degradation' because it doesn't require a slicing competent argonaute to result in mRNA degradation by other nucleases.

Argonautes (AGO and PIWI) retain a catalytic nuclease tetrad (Nowotny *et al.*, 2005), although most vertebrate AGOs lack oligonucleotide cleavage ability (Zamore and Haley, 2005) and suppress mRNA translation by steric impedence, or by suppression of transcription via argonaute guided heterochromatin and DNA methylation guided by small RNAs. These alternate mechanisms of regulating gene expression impose less functional restraint on the conservation of the catalytic residues.

The machinery required for the biogenesis of the small RNAs that functionally combine with AGO and PIWI does vary, and not all proteins have been found in all species. It is sufficient for our analysis to infer that if an argonaute-like protein is retained it is because of some functional utility. To describe a typical argonaute we take the structure of *Homo sapiens* AGO2, bound to a cleaved RNA (Schirle *et al.*, 2014), which possesses a functional enzymatic RNA slicing capability. The HsAGO2 structure is representative of the four domains: the N terminal domain, the PAZ domain, PIWI and MID domains, (thus there are both PIWI proteins and PIWI domains) common to metazoan argonautes generally. The PAZ and MID domains form the binding pockets for the guide RNA and the catalytic residues are found in the PIWI domain (fig. 4a.1).



**Fig. 4a.1 The human AGO2 crystal structure PDB:4W5N.**

N terminal and MID domains (pale grey), PAZ domain (dark grey) and PIWI domain (pink) are shown together with a cleaved RNA. The RNA binding pockets are circled in red. The positions of canonical catalytic residues (Asp-Glu-Asp-His DEDH sequentially in the primary sequence) are shown in blue and labelled in the enlargement. Phe (F) is labelled in pale blue on the coil adjacent to the catalytic site where it forms hydrophobic interactions (Nakanishi *et al.*, 2013) and is conserved in AGO-like but not generally PIWI-like argonautes. All ribbon diagrams are rendered in PyMol version 1.3.

The PIWI domain contains a ribonuclease H-like domain which is also conserved in other proteins including some involved in processing the guide RNAs (Majorek *et al.*, 2014). So AGO or PIWI BLAST searches may identify domains other than those in the AGO and PIWI proteins. Important to our assessment is the presence of the canonical catalytic residues in any potential AGO/PIWI sequence. Sequentially these residues (abbreviated DEDH) are Asp (D597), Glu (E637), Asp (D669) and His (H807) (numbering from HsAGO2) (Nowotny *et al.*, 2005). Asp and His bind divalent cations  $Mg^{2+}$  or  $Mn^{2+}$ . Phe (F676) is also considered to be important and forms strong hydrophobic interactions with the PIWI domain (Nakanishi *et al.*, 2013). However these residues are not strictly conserved even in characterised functional argonautes e.g. Arg may substitute for Asp or His, and Tyr is often found in place of the Phe (which is generally conserved in AGOs but not usually in PIWI proteins). Nuclease activity is sensitive to substitution of the catalytic residues e.g. H807R abolishes HsAGO2 activity and F676Y impairs it (neither substitution impairs RNA binding) (Faehnle *et al.*, 2013).

Even where the catalytic tetrad is conserved, other regions of the structure may prevent the catalytic residues from cleaving the RNA target (Schürmann *et al.*, 2013) but do not prevent RNA binding which directs the argonaute to its target.

#### **4.2.2. *In silico* analysis**

Biologists know that recurrent mutations at aligned sites will obscure deep relationships in phylogenetic trees (Mossel and Steel, 2004) but build them anyway. These authors showed that the Markov models we use do lose information exponentially at the deepest divergences, even though additional sequences do help in a linear manner. So we have a linear increase in information against an exponential loss of information at deeper divergences! In order to bring more rigour to our tree building we have turned to predicted structural analysis using mainly I-TASSER (Iterative Threading Assembly Refinement Server) suite (Roy *et al.*, 2010; Yang *et al.*, 2015). This is currently the best available set of programmes for predicting 3-D (tertiary structure) structure, being a consistent winner in the CASP (Critical Assessment of protein Structure Prediction) competition (Roy *et al.*, 2010). We also use Phyre2 (Kelley *et al.*, 2015) for fast analysis and FATCAT (Flexible structure Alignment by Chaining Aligned fragment pairs allowing Twists) (Ye and Godzik, 2003) to align structures with known functional argonautes so that we can assess if large inserts appear likely to affect functionality.

Ancestral sequences have been recombinantly expressed e.g. capsid proteins of the coxsackievirus (Gullberg *et al.*, 2010) and the LeuB enzyme from the putative ancestor of the thermophilic bacteria *Bacillus* (Hobbs *et al.*, 2012). ASR *in silico* is relatively easily achieved through free servers and in the past we have used a number of different systems (Daly *et al.*, 2013a; Daly *et al.*, 2013b). Our reconstruction is to test if we can look back in time to the ancestor of the metazoa, rather than to recreate putative ancestral functionality. This requires more sequences and therefore much less primary homology in both the MSA that is used to recreate the ancestor, and also between ancestors themselves than would usually be found where the physical protein was recreated. We developed a pipeline to assist with our search for argonaute family proteins where primary sequence similarity is so low that they may not be found by BLASTing alone. Our pipeline brings increased confidence to the BLAST result by using the evidence provided by structural prediction, retention of catalytic sites, and additional features that we found during our investigation to support or refute the inclusion of protein sequences into MSA.

To this end we report the 3-D structural analysis of ~180 argonaute proteins, and their relatives to understand the evolutionary history of the AGO-like and PIWI-like proteins in metazoa. During this process we find that a small number of argonautes (in this case) have been misidentified/misclassified, and we hope to clarify such issues. We have additionally found that AGO-like and PIWI-like proteins can be differentiated from one another on the basis of the last four residues of the primary sequence and this holds for almost all PIWI-like proteins across all eukaryote kingdoms as well as the vast majority of AGO-like proteins

### **4.3. Methods**

BLASTp and PSI BLASTing (BLOSUM45) of the NCBI and UniProtKB databases was carried out using argonaute sequences from taxa with well described experimentally determined argonaute crystal structures. More remote sequences thus found were then used as further BLAST seeds. Ancestral sequences reconstructed by FastML (Ashkenazy *et al.*, 2012) from MSA were used as further queries. The number of BLAST ‘hits’ exceeded the thousand maximum parameter in UniProt even when the database was constrained to groups such as ‘vertebrates’ or ‘fungi’. This meant that there were a number of potential argonaute sequences not identified this way. Anticipated sequences could be found via NCBI where there is the provision for searching specific groups of taxa or even species but still some sequences could be missed. The resulting overabundance of sequences meant that further BLAST analysis with the ancestral sequence was also likely to retrieve sequences containing a PIWI domain, OB fold (oligonucleotide/oligosaccharide binding fold) or RNaseH-like folds, common to all domains of life and found in other proteins associated with the argonaute processing pathway e.g. Dicer, and so further structural analysis was required. Pfam (Finn *et al.*, 2010) is a protein families database and has the advantage of being able to search for complete domains, i.e. PIWI which can help to narrow down the search. To include the sequence in our phylogenetic analysis the PIWI domain needs to be recognisable, and so it is likely that Pfam will find all proteins containing that domain. Pfam currently (April 2016) lists 1,207 PIWI domain-containing sequences in metazoa. The problem is then an easier one of discarding sequences that are incomplete, duplicated submissions or PIWI domains alone.

The three dimensional structural predictions were carried out by the I-TASSER suite (Zhang, 2008) of algorithms where Phyre2 was inconclusive (including our

inferred reconstructed ancestral proteins). Each sequence submitted for structural prediction takes a minimum of 50 hours to complete (multiple sequences cannot be submitted) so it is important to prioritise the sequences that add most to the knowledge base. To assist in prioritizing sequences for I-TASSER, and in order to exclude non-AGO/PIWI proteins, we calculated MSAs using Geneious (Geneious Pro 8.0.4 <http://www.geneious.com/>). Non-aligned sequences from Geneious alert us to minimal primary homology that can then be checked via structural prediction prior to inclusion in MSA calculated by MUSCLE (Edgar, 2004b) which will align sequences rejected by the Geneious algorithm. We then created large trees using MrBayes (Huelsenbeck and Ronquist, 2001) all algorithms run via the Geneious platform.

Trees were initially unrooted but then a root could be selected, permitting both the posterior probability and number of amino acid changes per site to be displayed. We were alerted when the number of predicted changes per site was noticeably large (i.e. in excess of 0.8 changes per site), given our argument that primary sequence homology can be at the noise level, yet structural homology retained we were expecting this number to be high. We also noted when the posterior probability was particularly low. If the sequence of a particular species seemed phylogenetically ‘not sensible’ (e.g. the ‘barley’ sequence UniProtKB:F2DNY6 (coloured green in all our trees) grouping with rotifers) then we would put those sequences through I-TASSER as a priority to test if this was a biologically interesting outlier, or a sequence misidentification – the latter is much more likely in the supposed barley case. In our previous work (Daly *et al.*, 2013b) we used the I-TASSER confidence score (C score) for fold recognition of Major Vault Proteins (MVP). In the present case with AGO and PIWI proteins, I-TASSER routinely resulted in poorer C scores for sequences that grouped amongst known PIWI-like proteins, even those that we knew to be properly identified and had been functionally characterized. A C score of  $>-1.5$  (where scores range from  $-5$  to  $+2$ ) is indicative of a correct fold (Roy *et al.*, 2010) but human PIWIL2 (a positive control PIWI sequence with a determined crystal structure for part of the protein (residues 386-533 PDB:3QIR) resulted in a score of  $-1.84$ , so we needed an additional method of identifying putative PIWI proteins. As a test for model quality we considered the retention of the DEDH catalytic tetrad in the 3-D structures. Although possession of the catalytic tetrad does not guarantee catalytic ability, having those residues spatially arranged in a manner



similar to experimentally-determined structures, plus an overall recognizable AGO/PIWI fold determines the protein's ancestry, if not its functionality.

For our ancestral reconstructions we made MSA using sequences confidently predicted as genuine AGO or PIWI-like proteins by protein sub-family rather than by phylogeny. ASR calculations are dependent on accurate MSA and tree calculation (Hanson-Smith *et al.*, 2010). We used FastML (Ashkenazy *et al.*, 2012) for ASR which uses maximum likelihood analysis (empirical Bayes) to estimate the ancestral sequence. This method serves to minimise noise resulting from MSA that have some uncertainty. Previously we compared a number of reconstruction methods and found that as long as we removed as many gaps as possible, FastML would produce sequences of a reasonable length (Daly *et al.*, 2013b). Excessive gaps left in the MSA resulted in unreasonably long ancestral sequences. For this reason our MUSCLE (Edgar, 2004b) alignments were visually checked and where gaps in the vast majority of sequences introduced by inserts present in only one species (or even a few species in MSA with a large number of sequences), those inserts were removed. We manually removed gaps where less than 10% of sequences displayed residues. Wasabi (web-based and free to use) will do the same thing automatically (Löytynoja and Goldman, 2010; Veidenberg *et al.*, 2015).

We calculated trees from the MSAs before and after gap removal to control that we do not affect the tree by the removal of these inserts. We could anticipate that the number of substitutions per site would be altered but the basic tree structure needed to remain the same. Supplementary material (S1 Fig.) is shown in appendix I and gives an example of a tree calculated before and after gap removal. There was no difference between the tree structure prior to gap removal and that following either manual or Wasabi removal thus validating our method. Sequences included in the BLAST results that are not AGO/PIWI had now been eliminated otherwise they could introduce noise into the reconstruction. FastML will create trees from this MSA but in order to maintain consistency we submitted our own trees calculated by MrBayes for all of the ancestral calculations.

FastML results in a putative ancestral sequence for each node, but we concentrated on the sequence predicted for the deepest node of each of our groups. The inferred ancestral sequences produced by FastML were submitted to I-TASSER for

structural prediction and were also used as BLAST queries to search for more remote sequences in the UniProt and NCBI databases (Collins *et al.*, 2003a).

In summary we created a pipeline for the sequence of methods that would result in an ancestral reconstruction calculation. We start with a BLASTp search of the UniProt and NCBI databases, but the basis of the work is the 3-D structural prediction. For the argonaute protein family we added an additional analysis module, i.e. checking for the retention of catalytic residues. We used I-TASSER for structural prediction, FATCAT for structural alignment, MUSCLE for multiple sequence alignment, MrBayes for tree building, and FastML for ancestral sequence reconstruction (ASR). We have adapted the work-flow for different proteins in the past by adding modules e.g. docking algorithms for oligomeric proteins, it can be changed to incorporate newer or more accessible servers as needed.

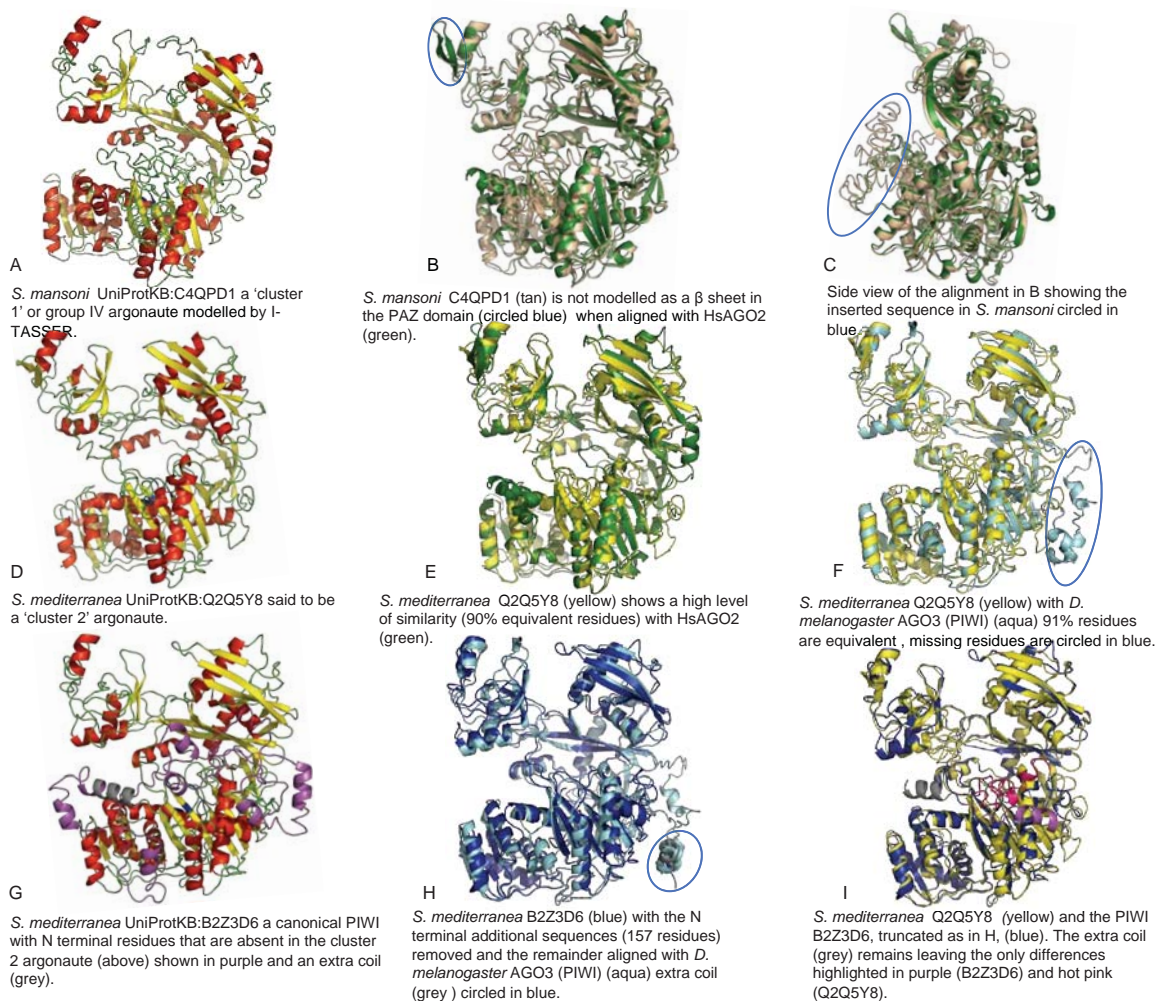
#### 4.4. Results

Most vertebrates have three or four genes for AGOs and a similar number for PIWI. The number of transcripts may be much larger because of many splice variants. For such reasons there was potential ambiguity about precisely which sequence to include in an ancestral sequence. We have therefore erred on the side of caution and included only sequences that we were convinced were the result of different genes.

From over 500 metazoan sequences we made a representative sample of 184 proteins that we were confident would fold correctly as either AGO-like or PIWI-like proteins. In our preliminary trees we had noticed that in some cases the posterior probability was quite poor and some of the Platyhelminthes sequences regularly swapped between being AGO-like or PIWI-like in the PIWI/AGO tree, highlighting the limitations of tree building of deep divergences (Mossel and Steel, 2004). Sequences affected were the parasitic fluke (*Schistosoma mansoni*) AGO2C UniProtKB:C4QPD0, C4QPD1 (and C4QPD2 virtually identical to C4QPD1 – and which was left out of subsequent trees), the free living triclad planarian *Schmidtea mediterranea* PIWIL-1 and PIWIL-2 (UniProtKB:Q2Q5Y9 and Q2Q5Y8 and not found in Pfam) plus the nematode *C. elegans* ERGO1 (UniProtKB:O61931). We investigated the effect of the outgroup used to root the trees by calculating 10 trees using an identical MSA of 184 sequences and changing the root sequence. This meant that the MSA with roots from outside metazoa had 185 sequences and the final four trees were from identical MSAs of 184 sequences, each calculated with a different outgroup from within metazoa.

The results of the tree construction exercise demonstrates tree variability when just one sequence is changed in each tree given identical MSAs (except for the outgroup). We need to confidently determine phylogenetic relationships as they form the basis for subset MSA used for ASR. We found that we cannot rely on an outgroup that is too close to metazoa and we have used the ten trees to form a consensus for our groupings. The similarities between the trees show that all of the vertebrate AGOs are from one duplication. The individual trees are too large to show within the thesis (available in wikispaces) a summary is given in table S2 in appendix I. Vertebrate PIWI2 plus representatives of all lower species and ecdysozoa AGO3 (an insect PIWI confusingly annotated as AGO3) are in the same ‘clade’ in all trees, implying that they all originate from the same ancestor. The insect AUB-PIWI ‘clade’ is isolated in all trees (AUB is also an insect PIWI-like protein). In insects AUB and AGO3 are partners in the ping-pong amplification of piRNAs, however vertebrates partner PIWI2 (in the same ‘clade’ as insect AGO3) with PIWI1. We find no evidence that PIWI1 evolved from AUB and it appears more likely that PIWI1 evolved from a duplication of PIWI2 (the animal version of insect AGO3).

Of the sequences that ‘flip’ between AGO-like or PIWI-like, those of the parasitic fluke *S. mansoni* (UniProtKB:C4QPD0, C4QPD1 and C4QPD2) are considered to represent a different argonaute subfamily described as ‘cluster 1’ argonautes (Zheng, 2013) and later renamed group IV argonautes (Skinner *et al.*, 2014). Zheng (2013) also identified the PIWI-like sequences from the free living planarian *S. mediterranea* (UniProtKB:Q2Q5Y9 and Q2Q5Y8) as a subfamily named ‘cluster 2’. We looked at the differences and similarities between the structural predictions of examples of the cluster 1 and 2 argonautes (fig. 4a.2), and although there are differences between *S. mansoni* and HsAGO2 they are outside of the protein core and we could find no structural difference between the *S. mediterranea* cluster 2 structural predictions and HsAGO2. However, the cluster 2 PIWI are different from the *S. mediterranea* canonical PIWI UniProtKB:B2Z3-D6 (which groups with the insect AGO3 in our trees – one of the PIWI proteins involved in the ping-pong copying). UniProtKB:B2Z3-D6 (the canonical PIWI) and UniProtKB:Q2Q5Y8 (one of the cluster 2 PIWI) are essential for stem cell function (Palakodeti *et al.*, 2008) so it is likely that the cluster 2 PIWI are functioning in a similar manner to AUB (insects) or PIWI1 (vertebrates) as the other partner for the ping-pong system and are genuinely derived from PIWI - as described in Zheng (2013).



**Fig. 4a.2 Similarities and differences in predicted structure between the difficult-to-resolve flatworm sequences.**

**A**, **D** and **G** show the difficult to resolve flatworms coloured by structure. We have used HsAGO2 (PDB:4OLA) a solved crystal structure in **B**, **C** and **E** (shown in green) and the predicted structure for *D. melanogaster* AGO3 (a well described insect PIWI-like argonaute) in **F** and **H** (shown in aqua). **A**. The predicted structure of *S. mansoni* (UniProtKB:C4QPD1), a 'cluster 1' argonaute. **B** 87% of residues in C4QPD1 (shown in tan) are in equivalent positions, (i.e. 87% of the residues of Q4QPD1 have the same geometric position as a residue in the solved structure) when aligned with HsAGO2. There are minor differences in structure, e.g. some loss of  $\beta$  sheet, circled. **C**. Side view of the same alignment showing the inserted sequence away from the core of the argonaute. **D**. The predicted structure for one of the *S. mediterranea* 'cluster 2' argonautes (UniProtKB:Q2Q5Y8). **E**. The *S. mediterranea* sequence (yellow) shows little difference when compared with HsAGO2, 90% of residues are in equivalent positions. **F**. Since the atypical argonaute ultimately fell on the PIWI side of our tree we aligned it with the I-TASSER predicted structure for *D. melanogaster* AGO3 (PIWI). The difference is circled but the equivalent residues were still very high at 91%. **G**. The predicted structure for the canonical *S. mediterranea* PIWI protein (UniProtKB:B2Z3D6). This groups with the PIWI proteins that include the insect AGO3 (PIWI) and PIWI2. The difference between B2Z3D6 and the 'cluster 2' argonaute Q2Q5Y8 is mostly in the N terminal insertion shown in purple with an additional coil that is a structural difference shown in grey. **H**. B2Z3D6 with the N terminal 157 residues removed (dark blue) and aligned with *D. melanogaster* AGO3. 96% of the truncated canonical PIWI residues are in equivalent positions to the well-studied *Drosophila* protein. FATCAT moves the grey helix not shared by the two *S. mediterranea* sequences (circled). **I**. *S. mediterranea* cluster 2 argonaute (Q2Q5Y8) (yellow) aligned with the truncated *S. mediterranea* PIWI protein (B2Z3D6) (dark blue). The equivalent residues are 95%, the grey helix is still a structural difference and the only other minor differences are highlighted in hot pink for Q2Q5Y8 and purple for B2Z3D6. All sequences have the conserved catalytic tetrad Asp-Glu-Asp-His (DEDH). All structural alignments are by FATCAT (Ye and Godzik, 2003).



It is interesting is that the parasitic flatworms including flukes and tapeworms do appear to have lost genuine PIWI-like proteins. UniProtKB:C4QPD0 is necessary for germline maintenance (Wang *et al.*, 2013) so it seems that the AGO-like cluster 1 (or group IV argonautes) are acting in place of PIWI-like proteins. Our results support this given these proteins from *S. mansoni* do flip from side to side of our trees i.e. they are not clearly AGO-like. Zheng (2013) depicts the cluster 1 argonautes as duplications of classical AGO-like proteins but Skinner *et al* (2014) shows them as a duplication of the canonical PIWI proteins. Although we cannot show this conclusively, we find nothing to suggest that the group IV argonautes have evolved from PIWI rather that they have evolved from AGO-like proteins (possibly due to the loss of PIWI). The main point is that there is limited primary homology between any of these proteins (20-25% in all cases), yet the structural predictions are very similar. In the FATCAT structural alignments with well-described argonautes the percentage of equivalently positioned residues is very high, with differences outside the core centre of the protein. In terms of the ‘cluster 2’ argonautes described by Zheng (2013), we find that the inserted sequence at the N terminal gives rise to almost all of the structural difference and if that is removed the first residue is methionine which is often the sequence start. However removing 157 amino acid residues from the N terminal of the canonical *S. mediterranea* PIWI protein (UniProtKB:B2Z3-D6) and placing them at the N terminal of the cluster 2 *S. mediterranea* (UniProtKB:Q2Q5Y8) AGO and recalculating the tree of all metazoa does not make any difference to the position of either protein. This indicates that the differences between them are not as simple as just the N terminal residues.

By checking anomalies, similarities and differences in this way we were able to identify argonaute sub-family sequences that were more similar across species than either PIWI-like or AGO-like proteins are from within a single species. In some instances it was clear that the sub-family groups had all duplicated from an original ancestor. Where enough information could be found this could often be interpreted mechanistically, e.g. human AGOs 1, 3 and 4 are all very closely located on chromosome one (1p34.35, 1p34.3, 1p34) and all vertebrate AGOs 1, 3 and 4 grouped in the same clades on the tree indicating that this duplication occurred prior to vertebrate evolution. Additionally there are more recent duplications within some species. Some of these were left out of the ancestral reconstructions because they are

recent rather than ancestral and do not add information about distant evolutionary relationships, an important point for our analysis of protein structures.

PIWI-like sequences tend to have a longer N terminal (which is predicted to form a disordered region by I-TASSER) and also generally have a poorer C score (Zhang, 2008; Roy *et al.*, 2010; Yang *et al.*, 2015) compared with AGO. This region is sometimes annotated as DUF (domain of unknown function) but is not found only in PIWI proteins. Many sequences have extra residues in this area including those known to be functional. It has previously been reported that disordered regions are likely to acquire more inserts over time (Light *et al.*, 2013). In contrast AGO-like sequences have a longer C terminal (resulting in a similar overall length to the PIWI-like proteins) and a higher C score which is most likely because the full-length solved structures in the Protein Data Base are mostly AGO-like proteins, however we have used I-TASSER specifically because it has some *ab initio* protein folding capability and is not totally reliant on simply aligning sequence to known protein structure (Table 4a.1.).

**Table 4a.1** I-TASSER results – a correct fold is considered  $> -1.5$  (range is  $-5$  to  $+2$ )

Accession number	Organism	Annotation	Catalytic site	I-TASSER C score
Q9UL18	<i>H. sapiens</i>	AGO1	DEDR	1.67
Q9UKV8	<i>H. sapiens</i>	AGO2	DEDH	1.14
Q9H9G7	<i>H. sapiens</i>	AGO3	DEDH	1.90
Q9HCK5	<i>H. sapiens</i>	AGO4	DEGR	1.94
Q96J94	<i>H. sapiens</i>	PIWIL1	DEDH	-0.82
Q8TC59	<i>H. sapiens</i>	PIWIL2	DEDH	-1.84 <sup>a</sup>
Q7Z373	<i>H. sapiens</i>	PIWIL3	DEDH	-1.26
Q7Z3Z4	<i>H. sapiens</i>	PIWIL4	DDAH	-0.98
F6P5N5	<i>X. tropicalis</i> (toad)	AGO1	DEDR	1.55
F7CWA8	<i>X. tropicalis</i>	AGO2	DEDH	0.70
F7D3A7	<i>X. tropicalis</i>	AGO3	DEDH	1.24
F7ALP3	<i>X. tropicalis</i>	AGO	DEGR	1.05
A8KBF3/F6UQE9 <sup>b</sup>	<i>X. tropicalis</i>	PIWIL2	DEDH	-1.47
F6VZI1	<i>X. tropicalis</i>	PIWIL3	DEGH <sup>c</sup>	-1.18
F7BXD6	<i>X. tropicalis</i>	PIWIL4	NDDH	0.83
Q7KY08	<i>D. melanogaster</i>	AGO1	DEDH	-0.63
Q9VUQ5	<i>D. melanogaster</i>	AGO2	DEDH	-1.79 <sup>a</sup>
Q7PLK0	<i>D. melanogaster</i>	AGO3(PIWI)	DEDH	-0.95
O76922	<i>D. melanogaster</i>	AUB (PIWI)	DEDH	-0.91
Q9VKM1	<i>D. melanogaster</i>	PIWI1	DDDK	-0.59
B3VCG6	<i>S. purpuratus</i> (sea urchin)	AGO1	DEDH	-0.96
W4YU94	<i>S. purpuratus</i>	AGO	DEDH	-0.89
17BC22	<i>S. purpuratus</i>	AGO1B	DEDH	-0.57
Q9GPA8	<i>S. purpuratus</i>	Seawi (PIWI)	DEDH	-0.31
B0FLQ9	<i>S. purpuratus</i>	Seali (PIWI)	DEDH	-2.04 <sup>a</sup>

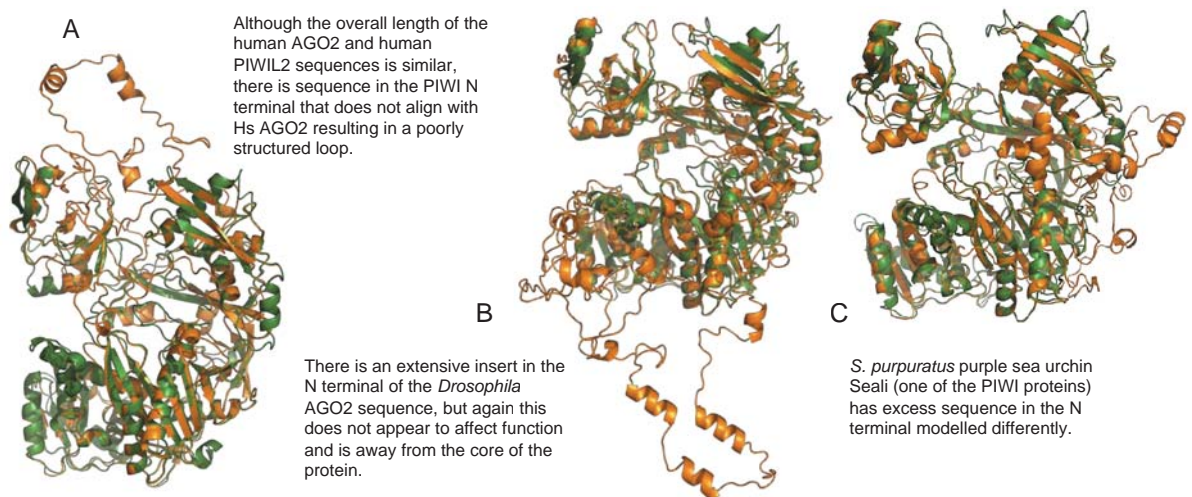


<sup>a</sup> A correct fold is considered  $> -1.5$  (range is  $-5$  to  $+2$ ) so this score alone would have resulted in exclusion from consideration in our previous work. However they are known to be functional proteins and are described further in fig. 4a.3.

<sup>b</sup> These are two accession numbers for sequences with one residue difference (427 V/I).

<sup>c</sup> The sequence alignment shows Ser (DESH) at the catalytic site rather than Gly which is the residue at the catalytic site in the structural model in fig. 4a.4.

Sequences with poor confidence score for their predicted structure (table 1) are *H. sapiens* PIWIL2 (UniProtKB:Q8TC59), *Drosophila melanogaster* AGO2 (UniProtKB:Q9VUQ5) and *S. purpuratus* ‘Seali’ (purple sea urchin PIWI UniProtKB:B0FLQ9) that groups with basal metazoan and insect AGO3 type PIWI. There are insertions in these sequences that explain the poor scores. The human and *Drosophila* sequences have large inserts not found in any of the solved argonaute structures. The mouse PIWI2 homolog (UniProtKB:Q8CDG1) has the same insert and is known to be functional (Itou *et al.*, 2015), as is the *Drosophila* sequence (Abramov *et al.*, 2016) (see fig. 4a.3). The *Drosophila* AGO2 sequence has a long glutamine-rich N terminal insertion also found in honeybee (UniprotKB:A0A088AHH6 – replacing H9KEG2) and carpenter ant (UniProtKB:E2ACD7 which is annotated PIWI-L but groups with AGO-like proteins). This N terminal insert is neither conserved nor widespread throughout ecdysozoa, so doesn’t appear to be ancestral. The general observation though is that PIWI proteins often have poorer C scores than AGO type proteins.

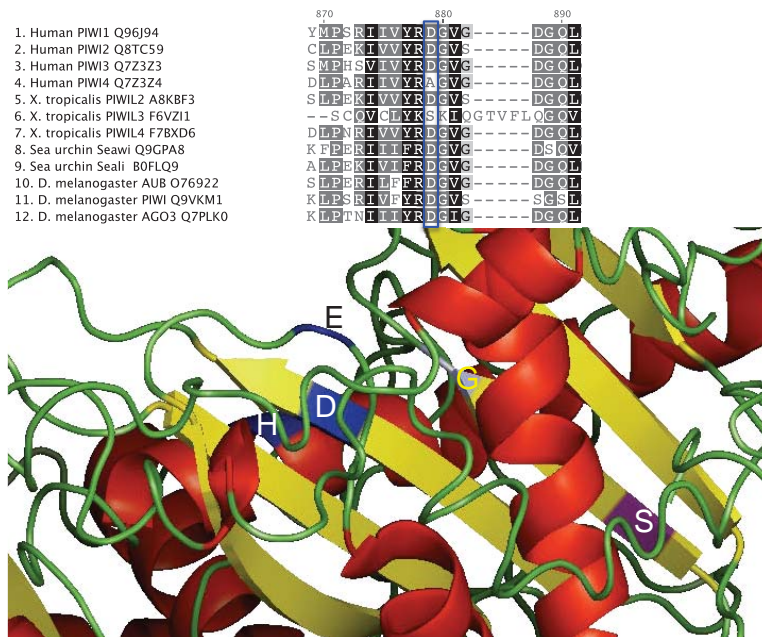


**Fig. 4a.3 The human AGO2 crystal structure aligned with the low scoring predicted structures identified in table 1.**

In each case the human AGO2 crystal structure (PDB:4OLA 859 residues) is shown in green, and in each case the low scoring I-TASSER predictions identified in table 1 are shown in orange. **A.** HsAGO2 is aligned with human PIWIL2 (861 residues). The majority of the two proteins overlap (88% of residues are equivalently placed) but the extension of the PIWIL2 N terminus is clear. This is balanced by extra sequence in the C terminus of HsAGO2. **B.** Shows a similar situation for the low scoring *Drosophila* AGO2 (1214 residues) aligned with HsAGO2, in this case 67% of the residues are equivalently placed

and the excess N terminal is modelled as a long loop but outside of the core of the protein. **C.** The purple sea urchin Seali (PIWI 960 residues) also has extra sequence in the N terminal. None of the sequence with disordered structure has any similarity between the sequences and is likely the result of disordered sequence acquiring further inserts (Light *et al.*, 2013).

Our work has already alerted us to a simple problem that could easily be missed with alignments of multiple sequences; the catalytic residues predicted by the MSA do not always match those found in the 3-D predicted structure. Of the catalytic tetrad (DEDH), Glu (E), appears under less selective restraint probably because it is on a loop in the structure and it seems likely that a Glu close enough to contribute to catalysis would do and so does not need to be conserved in an identical position in a homologous sequence. The sequentially first Asp lies structurally adjacent to the final residue of the tetrad, His, the remaining Asp is usually found at the tip of the adjacent  $\beta$  strand (as shown in the enlargement in fig. 4a.1). In the case of the toad (*X. tropicalis*) PIWI3 (UniProtKB:F6VZI1) the residue in the third position (sequentially), anticipated to be Asp (equivalent to HsAGO2 D669), is Gly in the predicted structure. The alignment predicts Ser but the predicted structure places Ser quite some distance away from the catalytic area (see fig. 4a.4). This looks to be a consequence of a unique five-residue insertion, compared to the rest of the MSA, and then a five-residue deletion the other side of this region. We tried the same alignment with MUSCLE, CLUSTALW, MAFFT, and Geneious with varying alignment parameters, but none of them modelled Gly as being in the catalytic site found in the 3-D predicted structure. CLUSTALW and Geneious differed in that the adjacent Thr was predicted. The predicted structure has a score that infers a correct fold  $-1.18$ . Gly is predicted in this position in other argonautes (table 4a.1) and by functional analysis (HsAGO4) (Schürmann *et al.*, 2013), so we must consider all the evidence to determine that the 3-D structural prediction is, on balance, more likely to be correct than is the MSA. This example illustrates the importance of relying on additional evidence, rather than just BLASTing and subsequent MSA and is not an isolated incidence of this occurring. Ultimately it will be important to test the protein activity based on the suggestions made here.






**Fig. 4a.4 The difference between alignment and predicted structure in *X. tropicalis* PIWI3.**

The MSA (truncated) shows Ser (S) is aligned where Asp (D) would be anticipated for *X. tropicalis* PIWI3 (UniProtKB:F6VZI1) (column outlined in blue). However the structural prediction of the sequence (showing only the catalytic region), places Gly at the tip of the  $\beta$  strand where Asp would be anticipated (grey shading, yellow text). Ser is some distance from the catalytic site, lower on the same  $\beta$  strand (marked in purple). Our point is that if we relied on the MSA alone, we would not recognise that Gly was predicted to be at the catalytic site.

## 4.5. Ancestral reconstruction

When we looked at the ancestral sequences we found that the simplest way to differentiate AGO-like and PIWI-like proteins (at least in metazoa), is that PIWI have a highly conserved C terminal signature of two aromatic residues (Phe or Tyr), between two Leu residues; namely LFFL (occasionally LFYL or LYFL). In contrast, AGO-like argonautes have a moderately conserved C terminal four residue signature of the same aromatic amino acids between Met and Ala/Val e.g. (MYFA). In fact, we found this is increasingly useful for classifying sequences in all species and in all kingdoms.

The PIWI side of the unrooted tree (S3 Fig.) shown in appendix I has three clear branches. The first group (i) (S3) that we looked at included the insect PIWI (annotated AGO3), and the vertebrate PIWI2, (vertebrate PIWI are annotated as either PIWI 1, 2 or 3 or PIWIL 1, 2 or 3 in different vertebrates). We named this ancestor AGO3\_PIWIL2 (Box 1, fig. 4a.5). The ancestor retains LFFL at the C-terminus, as well as an intact catalytic DEDH tetrad. AGO3\_PIWIL2 is the ancestor of the partner of the ping-pong partnership that predominantly binds sense piRNAs, generating antisense piRNAs to be picked up by PIWI or AUB (in *Drosophila*) (Brennecke *et al.*, 2007), or can be used to silence their target.

Box 1. AGO3_PIWI2	Box 2. PIWI_AUB	Box 3. 'Other' PIWI
		
PIWI of this type are annotated as AGO3 within the insects, but this group comprises of all metazoa including vertebrates (PIWI2). These are an essential partner of the ping-pong system.	Insects and basal metazoa <u>but not vertebrates (or deuterostomes other than the sea squirt)</u> have a pair of similar proteins: it seems that either one can partake in the 'ping-pong' autocopy of piRNAs. AGO3 is the other member of the pair.	Comprises of all groups <u>other than ecdysozoa</u> and includes vertebrate PIWI1, which is the other ping-pong partner with PIWI2 in vertebrates. Vertebrate PIWI1 tends to LYYL at the C terminal.

**Fig. 4a.5. The C terminal signature in PIWI-like ASR.**

Sequence logo of the last four C-terminal residues of the ancestral sequences reconstructed for node 1 (in each case) for each group visually representing (by size) the relative probabilities of each residue based on the final trees (S3 Fig.) of all PIWI-like amino acid sequences from the group 'metazoa'.

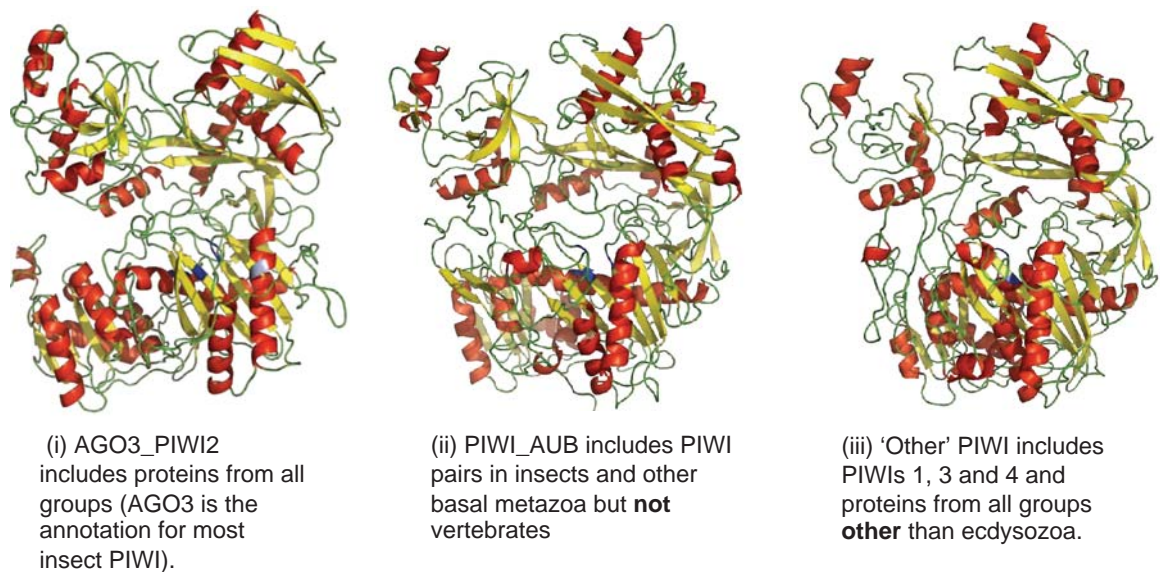
The next branch (ii) (S3 Fig.) is comprised of sequences from the PIWI\_AUB like proteins but these are not specific to insects (or even to ecdysozoa) because homologs of the PIWI\_AUB pairs are found in comb jellyfish, leech, planaria and trematoda as well as the deuterostome *Ciona savignyi* (sea squirt). This branch does not contain any vertebrate sequences, and was probably the result of a duplication either in the earliest metazoa or prior to the common ancestor of metazoa. Either one of these pairs can provide the antisense piRNA binding partner for the ping-pong 'autocopy' activity together with AGO3, and so are functioning akin to vertebrate ping-pong partner PIWI1.

The ancestor of the PIWI\_AUB group (Box 2 of fig. 4a.5) has the C terminal signature of LYYL. These two branches of the metazoan PIWI-like protein tree contain sequences from all of our groups implying that there were already two PIWI proteins in the earliest metazoa. It could be argued that the ping-pong mechanism of making piRNAs could pre-date the more usual Dicer processing pathway for piRNAs. However Dicer is found in basal metazoa, as well as in insects and so two different pathways for the genesis of piRNAs appear to have been already in operation at the dawn of metazoa.

The 'other' PIWI group (Box 3, (iii) in S3 Fig.) consists of the vertebrate specific PIWI (PIWIL1, 3 and 4) and representatives of all species apart from ecdysozoa. PIWI1 appears to function as the vertebrate ping-pong partner for PIWI2 (found in the clade that includes the unfortunately annotated AGO3). PIWI4 would seem to be the more



obvious partner for a system known to be old at least in terms of metazoa because PIWI4 specifically groups with basal metazoa, however it may be that vertebrate PIWI1 has become best adapted to the task. The ancestors in Boxes 2 and 3 (PIWI\_AUB and ‘Other’) could have derived from one protein supporting the hypothesis of two different PIWI at the base of metazoa. Unexpectedly high posterior probability scores for the residues in each of the ancestors demonstrate a high level of conservation. The ancestral sequence predictions were submitted to I-TASSER, all of the ancestors retain the DEDH tetrad structurally located as anticipated and marked in dark blue in fig. 4a.6.



**Fig. 4a.6 Metazoan predicted structure for PIWI ancestors.**

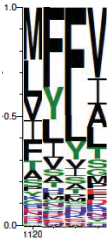


(i) AGO3\_PIWI2 the common substitution Tyr appears in the ancestor where Phe is often observed in AGO-like sequences (marked in pale blue). Note that we have used the nomenclature generally given for insect PIWI (AGO3), this is most definitely a regular PIWI protein and the nomenclature is confusing, (ii) PIWI\_AUB. (iii) ‘Other’ PIWI. It is most likely that ancestor (ii) named PIWI\_AUB and ‘other’ PIWI (iii) are derived from the same original (duplicated) gene. The intact catalytic tetrad Asp-Glu-Asp-His (DEDH) is retained in each ancestor and marked in dark blue.

In some cases Arg-Gly (RG) repeats were found at the N terminal of individual PIWI proteins (and also occasionally AGO sequences) and were shown as a conserved feature in all of the ancestral sequence PIWI. RG/RA (Arg-Gly/Arg-Ala) repeats were noticeable in the ‘other’ PIWI ancestor as well as the RG motif. Conserved RG/RA motifs have been documented in PIWI proteins recognised by the Tudor domains (a protein motif that specifically recognises dimethylated arginines), one of many proteins that interact most notably with PIWI1 (Vagin *et al.*, 2009; Liu *et al.*, 2010).

Of the sequences that would flip from the AGO-like side to the PIWI-like side of the tree until the root was far enough removed *S. mediterranea* PIWIL-1

(UniProtKB:Q2Q5Y9) and PIWIL-2 (UniProtKB:Q2Q5Y8, both described as cluster 2 argonautes), had a PIWI-like signature and ultimately rested on the PIWI side of the tree. The high level of conservation seen in the C terminal signature is not applied over the whole of the PIWI-like ASR sequences but the PIWI ASR are strikingly more conserved than the AGO-like ASR sequences. Ultimately our investigation infers that PIWI-like argonautes of at least two types are found in basal metazoa.

The final metazoan tree of all AGO-like proteins is shown in appendix I (S4 Fig.). This reveals that one branch has only ecdysozoa and basal metazoan sequences (i) (S4 Fig.). We named this group ecdysozoa\_basal AGO2, (Box 1 of fig. 4a.7). The second branch we named AGO1\_‘other’ both following the AGO1 nomenclature observed for insect AGO proteins (S4 Fig. (ii)). This group included lophotrochozoa and ecdysozoa but not any clear representatives of deuterostomes (Box 2 of fig. 4a.7). The nomenclature follows the insect annotation (that is not strictly adhered to) but doesn’t equate to a vertebrate homolog of the same annotation. The vertebrate branch is ancestral almost exclusively to vertebrates (Box 3 of fig. 4a.7 (iii) (S4 Fig.)). It is most likely that all of the four vertebrate AGO-like proteins originated from one ancestor (AGO1\_‘other’ Box 2) and that vertebrates have lost one basal AGO-like protein, with basal metazoa and ecdysozoa having retained two distinct AGO-like proteins.

Box 1. Ecdysozoa_basal AGO2	Box 2. AGO1_‘other’ AGOs	Box 3. Mainly vertebrate AGOs
		
Composed of ecdysozoan AGO2 and basal metazoa, the majority of sequences that went into this ancestor <b>did</b> retain a recognizable AGO signature (including the sponge) but clearly the posterior probability is much lower than for the other ancestors.	This group is almost entirely ecdysozoa (AGO1) and lophotrochozoa, there are no vertebrate sequences.	This group is almost entirely vertebrate AGOs, this is apparently a duplication of ‘other’ AGOs.

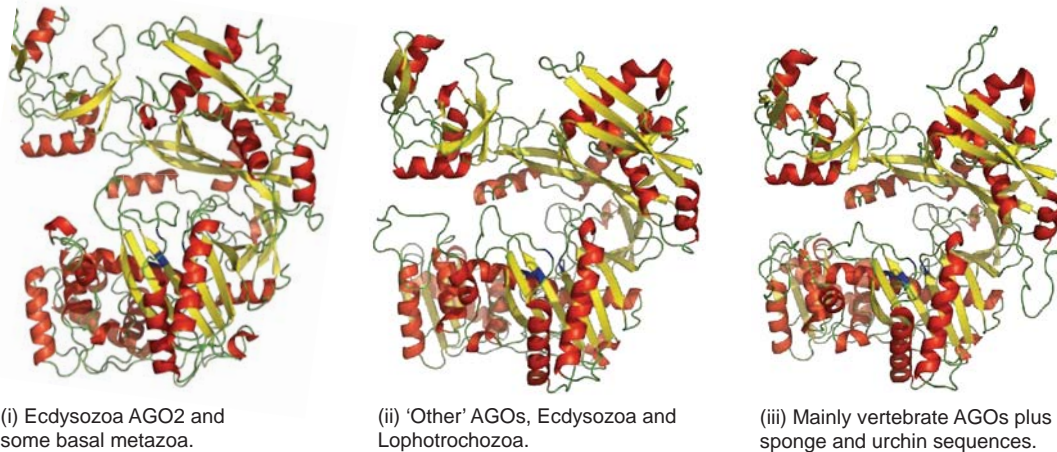


**Fig. 4a.7. The C terminal signature in AGO-like ASR.**

Sequence logo of the last four C-terminal residues of the ancestral sequences reconstructed for node 1 (in each case) for each group visually representing (by size) the relative probabilities of each residue based on the final trees (S4 Fig.) of all AGO-like amino acid sequences from the group 'metazoa'.

We initially treated vertebrate AGO2 separately in terms of generating an ancestor because in our original vertebrate trees it was different at the sequence level from other AGOs, as well as being the only cleavage competent argonaute. However, from our analysis vertebrate AGO2 appears to be the original AGO-like argonaute that was retained into the bilaterians with the other AGO-like argonaute being lost. Vertebrate AGO3 is then a duplication of AGO2, with vertebrate AGOs 1 and 4 arising later from a separate duplication of AGO2. In our reconstruction the catalytic residues of the putative ancestor of AGO 1, 3 and 4 were calculated as DEDR, so although Asp (replaced by Gly in AGO4) was retained, H→R appeared to predate the duplications. The ancestor of all of the vertebrate AGOs (and indeed of all metazoan AGOs) retains the DEDH catalytic tetrad as anticipated. This means that AGO3 (which has H as the final residue of the catalytic tetrad) would have to have reverted to His following the duplication. We checked the FastML reconstruction with ancestral sequences calculated by PAML4 (Yang, 2007) and MEGA6 (Tamura *et al.*, 2013) and the results were concordant with FastML. Ultimately we added the vertebrate AGO sequences (Box 3) and the AGO1 of ecdysozoa and lophotrochozoa (Box 2) to make a single ancestor with only very minor differences from 'other' AGOs (structural prediction of combined ancestor not shown).

Apart from the catalytic tetrad and the C terminal signature the posterior probability of most of the residues, but especially those in the N domain was extremely low. This supports our conclusion that residues can be remarkably varied and yet structural homology remains (fig. 4a.8). Posterior probability logos of all of the ancestors are available in wikispaces.



**Fig. 4a.8 Metazoan predicted structure for AGO ancestors.**

Left (i) Box 1 AGO ancestor from the ancestral sequence reconstruction from MSA of ecdysozoan AGO2 and basal metazoan AGO-like sequences. (ii) 'Other AGOs' from Box 2, which did not have any vertebrate sequences but comprised mainly of ecdysozoa and lophotrochozoa. (iii) Box 3, mainly vertebrate AGOs but includes sponge and sea urchin sequences. The vertebrate duplications are likely to have arisen from one protein, we did combine (ii) and (iii) with very little impact. The catalytic tetrad is marked in dark blue in each structural prediction, all retain the intact catalytic tetrad Asp-Glu-Asp-His (DEDH).

Of the ambiguous cluster 1 or group IV *S. mansoni* argonautes, AGO2C (UniProtKB:C4QPD0 and C4QPD1), plus *C. elegans* ERGO1 (UniProtKB:O61931), retain Met at the fourth from last position. This is conserved in AGO-like proteins from all plants, stramenopiles and fungi (except microsporidia) and ultimately these argonaute variants group on the AGO-like side of the tree. The simplest way to distinguish between PIWI-like and AGO-like proteins is to examine the last four residues.

## 4.6. Evolution

Basal metazoa; lophotrochozoans, cnidaria and sponges in general have three argonautes, one AGO-like and two PIWI-like. It could be that there were originally two distinct AGO-like proteins in metazoa and one has been lost in different species because there are two distinct AGO-like forms in ecdysozoa, plus two different forms in the Queensland sponge (*Amphimedon queenslandica*). It could be that over time the sponge argonautes have become so diverged from one another that they could have arisen from the same protein, however pairwise sequence alignment between UniProtKB:I1FXQ6 (*A. queenslandica*) and mouse AGO2 (UniProtKB:QBCJG0) has 60% sequence identity but *A. queenslandica* (UniProtKB:I1GIS7) aligned with the same mouse AGO2 sequence has just 39% sequence identity. This suggests that the vertebrate AGO-like proteins (AGO1-4) derive from only one of two original proteins. The loss of one of the

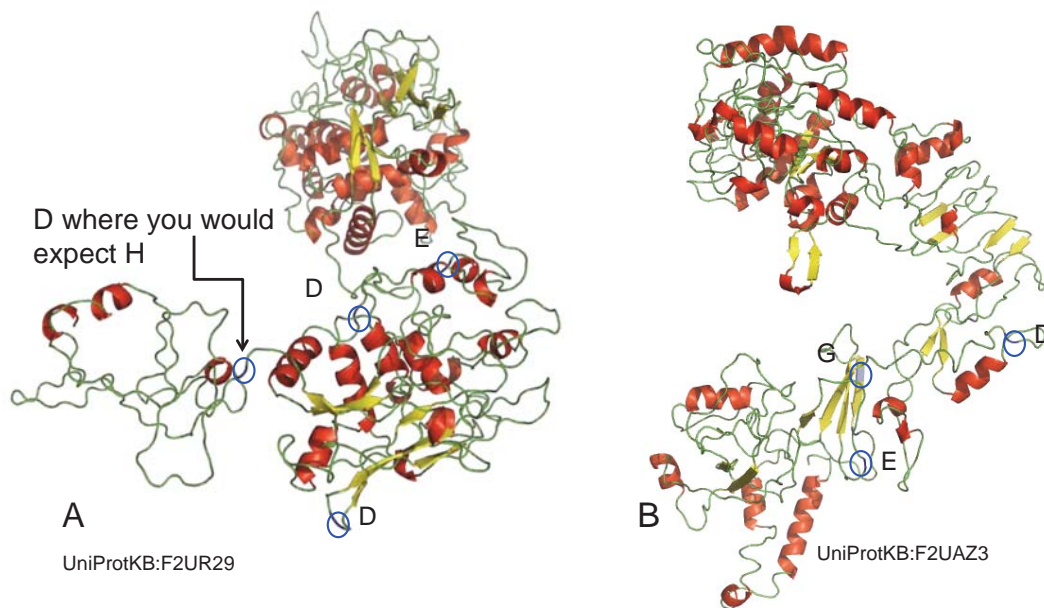
AGO-like argonautes appears to have happened quite early in metazoa because *Strongylocentrotus purpuratus*, as well all of the lophotrochozoans has lost one of the proteins and the sequences that they have retained are more similar to the vertebrate sequences. This is consistent with the proposal that all vertebrate AGOs derive from the same ancestral AGO; the only one retained in the lophotrochozoa.

There is some debate regarding the earliest member of the metazoan family (Jekely *et al.*, 2015). We have found evidence for three AGO-like proteins in the comb jellyfish *Pleurobrachia bachei* (one full length) and three in *Mnemiopsis leidyi* (two full length) and three full-length PIWI type proteins in *P. bachei*. Should the debate fall back to the sponges at the root of the metazoan tree we find two full-length PIWI-like proteins and one AGO-like protein in the freshwater sponge *Ephydatia fluviatilis* and one PIWI-like plus the two AGO-like proteins in the Queensland sponge *A. queenslandica* mentioned above.

The placazoan *Trichoplax adhaerens* has one AGO-like argonaute and we have found no trace of a PIWI-like protein. The *Trichoplax* AGO (UniProtKB:B3SEP3) (Srivastava *et al.*, 2008) appears structurally consistent with orthologues containing the anticipated catalytic residues. Although this organism appears to reproduce asexually there are markers for sexual reproduction and they certainly have the capacity to exchange genetic material between cells; though there doesn't appear to be any differentiation between cells that could be described as male (Signorovitch *et al.*, 2005). *T. adhaerens* also seems to be depleted in terms of small RNAs other than those known to be associated with ribosomes, RNaseP and snoRNAs. The loss of PIWI in the parasitic platyhelminthes (fluke and tapeworm) could have been rescued by an adaptation of an AGO-like protein to function in the maintenance of the germline or the capacity for stem cell generation. *Trichoplax* doesn't appear to have taken this route. In other work we look at plants and fungi that have also lost the PIWI-like argonautes and are still extremely successful.

*Salpingoeca rosetta* is a model choanoflagellate which is a sister group to metazoa. We submitted all of our ancestors (from unikonts, plants, chromalveolates, trypanosomes as well as some chosen individual sequences) as query sequences BLASTing just the *S. rosetta* species held in the NCBI database. Of the retrieved sequences there were only two putative argonaute sequences (UniProtKB:F2UA73 and F2UR29) but when submitted to I-TASSER it became clear that although there were

some hallmarks of the argonaute family, the predicted protein folds were not similar enough to the known fold of the domains required for a functional argonaute (fig. 4a.9).



**Fig. 4a.9 Predicted structure for the putative sequences identified by BLAST in *S. rosetta*.**

A. Catalytic residues can be roughly positioned from the alignment for F2UR29 – this serves only to show that BLAST and alignment are insufficient to predict any kind of functional capability. B. By primary sequence alignment we see Gly (marked in pale blue) where we would anticipate the second Asp. This is not an unusual substitution, see fig. 4a.4, but these sequences have limited credibility even as AGO-like relics.

No proteins of the small RNA processing pathways were found in the choanoflagellate *Monosiga brevicollis* (Grimson *et al.*, 2008). We did our own search and the metazoan AGO ancestor retrieved one sequence via BLAST but none of the predicted folds resembled any of the argonaute domains and so it is likely that the entire system was long ago lost in choanoflagellates. An extensive search using all of the ancestors in the metazoan sister *Capsaspora owczarzaki* also failed to find any sequences resembling argonautes.

So it seems that retention and expansion of both AGO and PIWI proteins is specific to metazoa and both have been, or are in the process of being lost in close relatives. We do know that fungi have lost PIWI-like proteins and that there are some budding yeasts that have lost AGO-like proteins as well (Drinnenberg *et al.*, 2011). Given the extensive and ever expanding repertoire of functions in metazoa it is very interesting that even the closest relatives to metazoa have no apparent need for them (or that the genes are lost for good reason).

## 4.7. Discussion

### 4.7.1. Annotation issues

Annotation is going to be an important task as more genomes are available, and where previously we had less information to be able to classify all proteins correctly. Work such as this argonaute study can contribute to more confident annotation. The exponentially expanding databases of gene sequences and the lag between sequencing and annotation can lead to potential confusion.

One of the problems that we had was determining between different splice variants and different gene copies that complicates homologue identification. There are reports that different argonaute isoforms via splice variation have arisen in response to viral mechanisms to suppress the argonaute defence system (Huang and Zhang, 2012). In the refined trees (S3-4 Fig.) it can be seen that in some cases we have allowed two copies of a sequence identically annotated, e.g. the stickleback (*Gasterosteus aculeatus*) has two copies annotated as AGO3 with different accession numbers. They both fall within the AGO3 ‘clade’ but they are on different chromosomes and so are genuinely different genes that translate into an AGO3 type isoform (91% identical residues) and so they look to have arisen from a relatively recent duplication event. It is possible that AGO3 has recently duplicated in some vertebrates, but in many cases the extra transcripts overlap and could also be splice variants. Where we found that the sequences were splice variants, or we could not prove that they were separate genes (different accession numbers but overlapping transcripts) we removed the extra copies from our analysis. In the case of the stickleback AGO3 is clearly annotated by a separate accession number and (1 of 2, 2 of 2) and identified as resulting from two different genes. Additionally the stickleback (*G. aculeatus*) PIWI1 is more likely to be PIWI3 which is from a different duplication, and the king cobra (*Ophiophagus hanna*) AGO1 (UniProtKB:V8NK59) should be annotated as AGO3. The annotation of insect PIWI2 as AGO3 is also confusing and inconsistent even between insects.

In some cases sequences were clearly incorrectly annotated; the barley sequence (UniProtKB:F2DNY6) (Matsumoto *et al.*, 2011) is annotated as ‘PIWI-like’, which would be the first PIWI protein found in a plant. However it groups convincingly within metazoan PIWI-like proteins, closest to rotifers. A BLAST search with this sequence does not bring up another plant sequence within a thousand ‘hits’. We have suggested contamination within the barley sequences before where the major vault protein barley



(*Hordeum vulgare*) sequence (UniProtKB:F2E078) grouped with MVP sequences from slime moulds (Daly *et al.*, 2013b). In both cases the sequences derive from mRNA and contamination from other species is likely more common than anticipated. Certainly, annotation issues are going to be ongoing.

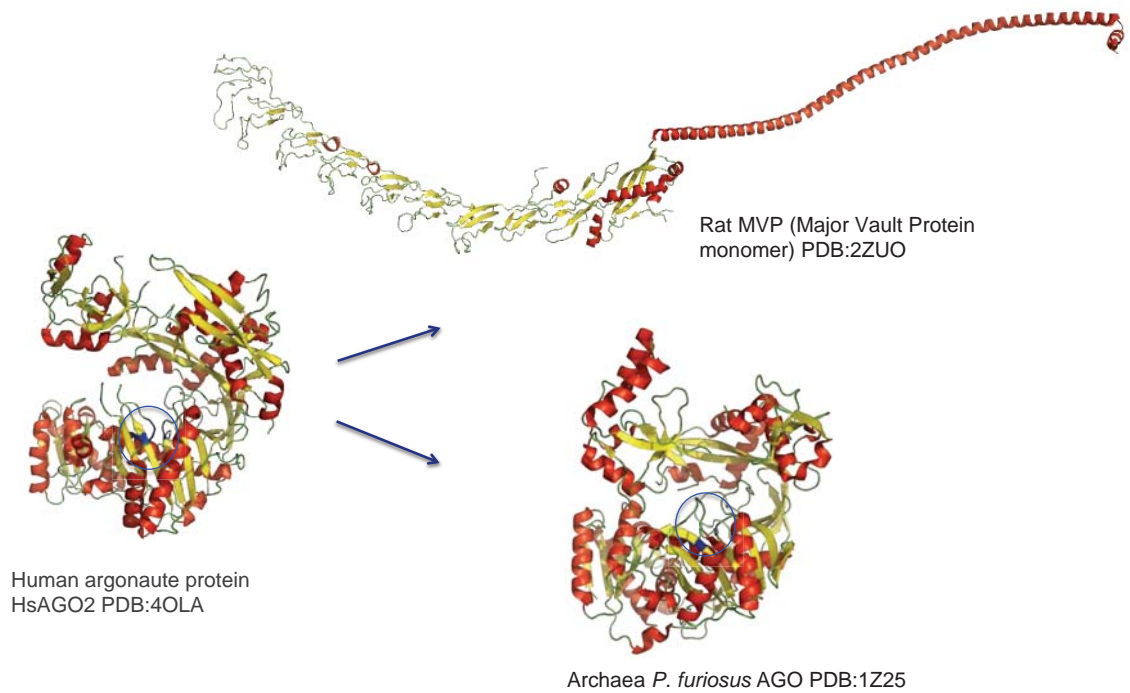
#### 4.7.2. General

Our study highlights the use of three-dimensional structural analysis, together with further evidence (in this case retention of the catalytic residues and also the C terminal signature residues) in order to support our BLAST results. One of the issues is that many of our results are not ‘reproducible’ because the ever-increasing number of deposited sequences quickly puts the more remote sequences out of the thousand hit maximum for UniProtKB.

An extreme example of the need for 3-D structural analysis is the similarity between the following three experimental structures: The human argonaute (HsAGO2 PDB:4OLA, UniProtKB:Q9UKV8) has 14% pairwise sequence identity with the rat MVP monomer (PDB:4V60, UniProtKB:Q62667), and 14% identity with the argonaute from *Pyrococcus furiosus* (PDB:1Z25, UniProtKB:Q8U3-D2). There is 13% pairwise identity between the rat MVP and the *P. furiosus* argonaute. Most of the identical residues are maintained across all three sequences when all are aligned. Because of the limitation of ‘1000’ hits, none of them would be found by using any one of the other two as query sequence. Even BLASTing just the archaea database with HsAGO2 does not retrieve the *Pyrococcus* sequence.

However HsAGO2 and rat MVP are clearly not homologs of one another (fig. 4a.10). The predicted structure of the archaean *P. furiosus* AGO (identical to the solved crystal structure) and retention of the catalytic sites (marked in blue and circled) are extra evidence of homology between the two argonautes. The crucial point is that structural homology is not always predictable from BLAST results, 3-D structural analysis can be essential.





**Fig. 4a.10 An example where BLAST searches are ambiguous.**

There is 14% primary sequence identity between the human AGO protein compared to both the rat MVP and *P. furiosus* AGO sequences. These are all experimentally-determined crystal structures with obvious structural homology between the AGO proteins contrasting with a very different fold for MVP.

Using the archaea sequence (UniProtKB:Q8U3-D2) as a query sequence for a BLAST search of the human database does retrieve human PIWI-like protein 1 (UniProtKB:Q96J94) with an E value of 900 (a usual acceptable E value is  $1 \times 10^{-4}$ ). There are in excess of 600 more similar hits within the human database than PIWI-L1. *P. furiosus* retains the C terminal signature of PIWI type proteins so it would be anticipated that if we could find a sequence in the human database then it would be PIWI-like. A reciprocal BLAST of the archaea database using the human PIWI-L1 protein still fails to find the archaea argonaute.

Almost certainly we are going to have to consider the 3-D structure of proteins (which demonstrates relatedness where primary sequence homology is ambiguous), in order to fully understand the mechanisms of deeper protein evolution. No longer is it sufficient to simply use BLAST results and sequence alignment as evidence for the deepest homologies, we have seen that alignment algorithms do not always predict where a particular residue will likely be found in 3-D space compared with even a very similar sequence, and that trees cannot be relied upon either at deeper divergences because some of our sequences flipped from AGO-like to PIWI-like and back again with every tree.

Some of our initial trees had poor posterior probability. Trees of all metazoan argonautes rooted with basal metazoa skewed the tree because whatever we chose was either a PIWI-like or an AGO-like sequence, which gave the erroneous indication that either there was a single source of argonautes that had duplicated into all PIWI and AGO type of proteins, or moved the sequences that ‘flipped’ from one side of the tree to the other. We found that trees rooted by the trypanosome *T. brucei* (either PIWI-like or AGO-like) separated the other sequences more clearly into either AGO-like or PIWI-like ‘clades’ than closer outgroups were able to.

There are also some branches showing the number of changes in residue per site as greater than one. This implies that more than 100% of the residues have changed, but each site can have a residue change on more than one occasion, so in these cases we need to look at the supporting evidence as a whole. Had we looked at each process and applied a ‘value’ of confidence as a ‘cut off’, e.g., E values, C score, posterior probability, an unlikely number of residue changes, we would have deleted well described and functional argonautes. We have decided not to publish the I-TASSER confidence score for all of our sequences because we have shown in table 1 and fig. 4a.3 that it can be misleading. Similarly posterior probabilities in terms of tree nodes and also of ancestral reconstruction should alert us to possible sources of error that we can check, rather than an arbitrary line resulting in rejection. Our determinations rely on the sum of the evidence which is why we reject the notion that one source alone can be used to determine relationships. This is not an automated process, rather it is a process that can be completed cheaply with a modest computer but with more rigour than an automated pipeline.

Experimental determination of protein structure remains essential for the success of the algorithms used for structural prediction. Once there is a structure determined of your favourite protein then it becomes amenable to structural prediction, as do more distant homologs. This can be used to aid annotation at deeper divergences. This also enhances hypotheses regarding whether or not an apparent homolog could feasibly be a functional protein. For example where we have found that the sequence in *S. rosetta* could be identified by BLAST, and has some structural resemblance to an argonaute, we could not claim that it was a functional argonaute without experimental analysis. Our results imply that it is not an argonaute.

The number of sequenced species is increasing rapidly (though it barely

scratches the surface of the number of species) and this provides great opportunities for evolutionary biologists, but there needs to be caution too. We have shown how easily it is to manipulate trees to support a story and so we need to understand the limitations of the technology that we have. Using structural prediction to turn 1D sequence information into a 3-D structure reveals additional evidence of homology, and by reconstructing ancestral sequence we have been able to see patterns that we could have missed otherwise. From these analyses we have shown that we can extract a much greater value from each sequence which will inform bioinformatics and annotation more cheaply and rapidly than current procedures.

## Acknowledgements

We thank I-TASSER for the provision of free computer hours for structural prediction via their University of Michigan server <http://zhanglab.ccmb.med.umich.edu/I-TASSER/>.

## Supporting Information

**S1 Fig. Gap removal.** Demonstration that conservative removal of gaps in the MSA does not alter the structure of the tree.

**S2 Table. Tree root summary.** A summary of the study to see how the root affects the tree.

**S2 Fig. Ten trees.** Circular trees of the same multiple sequence alignment rooted by ten different sequences. Not provided in appendix due to size but is available in wikispaces please contact [tonidaly@mac.com](mailto:tonidaly@mac.com) for access.

**S3 Fig. PIWI tree.** Line tree of the sequences that went to make the MSA used for ancestral reconstruction of PIWI proteins.

**S4 Fig. AGO tree.** Line tree of the sequences that went to make the MSA used for ancestral reconstruction of AGO proteins.

Posterior probability logos of all of the ancestors are available in wikispaces please contact [tonidaly@mac.com](mailto:tonidaly@mac.com) for access.

**Supplementary material (with the exception of S2 Fig.) can be found in Appendix I.**

## Chapter 4b: Argonaute gain and loss during fungal evolution.

### 4b.1. Abstract

Here we use three dimensional structure prediction of proteins and reconstructed ancestral argonaute sequences of yeast, fungi, and microsporidia to help infer their evolutionary history and their relationship to metazoan sequences. Most phyla retain multiple copies of an AGO-like or PIWI-like protein (or both) but fungi typically retain only two AGO-like argonaute proteins and microsporidia have just one, but it is dissimilar to either of the fungal proteins. Independent loss of AGO appears scattered across diverse fungal species, leaving some species without any argonautes at all. In addition there is a remarkable expansion of argonaute proteins found in the glomeromycotan fungus *Rhizophagus irregularis*, we find that we can predict that 17 of the sequences are likely to form functional argonautes and that these have duplicated from two different early AGO-like proteins.

Prediction of tertiary structures is becoming increasingly essential for studying deeper evolutionary divergences and more information can be obtained from the primary protein sequence to add support to tree calculations. With the increase of genomic information it is easier to determine whether an argonaute-like protein is genuinely missing for a given group or species, or has not yet been identified. This is part of a larger study using ancestral sequence reconstruction (ASR) and tertiary structure analysis to clarify the evolutionary history of the argonaute and PIWI proteins in eukaryotes.

### 4b.2. Introduction

Biologists regularly use the three-fold approach of BLAST searches, multiple sequence alignments (MSA) and tree building (of gene and/or protein primary sequences) to infer evolutionary relationships. These can additionally be used to aid gene annotation which struggles to keep up with the exponentially increasing data banks of sequenced genomes. It has long been known that this is not totally satisfactory and that recurrent mutations in aligned sites will obscure deep relationships (Mossel and Steel, 2004). Our aim here is to increase the amount of information obtained from one dimensional primary sequences retrieved by BLAST, with the use of *in silico* prediction of tertiary structure,

together with ancestral sequence reconstruction (ASR), in order to reduce the information loss at deeper divergences. We will not infer just a single gene phylogeny, but accept the current phylogeny based on multiple sequences.

First we will outline some properties of the argonaute proteins in the fungi, and consider those yeast that do not have argonautes. We then investigate the major expansion of argonautes in the glomeromycote (arbuscular) fungi and the next step considers microsporidia. Finally we analyse the placement of the reconstructed fungal ancestors compared with ancestral sequences from metazoa. The evolution of the different families of argonaute proteins has been very dynamic.

Argonaute proteins are important regulatory proteins with significant roles in unicellular eukaryotes as well as in multicellular plants and animals. They have a number of sub-families; firstly AGO-like which are the only type found in plants (Bohmert *et al.*, 1998), fungi (Aravin and Tuschl, 2005) and stramenopiles (Fahlgren *et al.*, 2013). Plants and stramenopiles frequently retain multiple copies of AGO-like argonautes. These are also found in metazoa and excavates together with the second subfamily of PIWI-like proteins (Cox *et al.*, 1998). Metazoa retain as many as four AGO-like proteins plus four PIWI-like argonautes. Ciliates retain multiple AGO/PIWI-like proteins that group closely with other PIWI-like proteins and load principally scnRNAs that control replication rather than gene expression (Mochizuki and Gorovsky, 2004). The third recognised form of argonaute are found in *Caenorhabditis elegans* (Vastenhouw *et al.*, 2003). This species has five AGO-like proteins (Alg1 and 2, T22B3.2, T23-D8.7, ZK757.3), three that are PIWI-like (PRG1 and 2, plus ERGO1 although the latter annotation is less clear) plus at least 18 WAGO ('worm' nematode argonautes) or type 3 argonautes. Some of these duplications are also conserved in other *Caenorhabditis* species but lost in parasitic nematodes (Dalzell *et al.*, 2011). Other variants have been described in Platyhelminthes (Skinner *et al.*, 2014) (Zheng, 2013), but they are all clearly related by structure.

AGO-like argonautes are guided by small RNAs to control the expression of endogenous and exogenous RNA via cleavage or steric impediment, mostly in the cytoplasm. PIWI-like proteins (guided by PIWI-interacting RNA, piRNA) are capable of suppression (and less commonly up-regulation) of transcription via heterochromatin changes and DNA methylation (Rajasethupathy *et al.*, 2012). They are largely responsible for suppression of transposons in germ-line cells, but are also found in

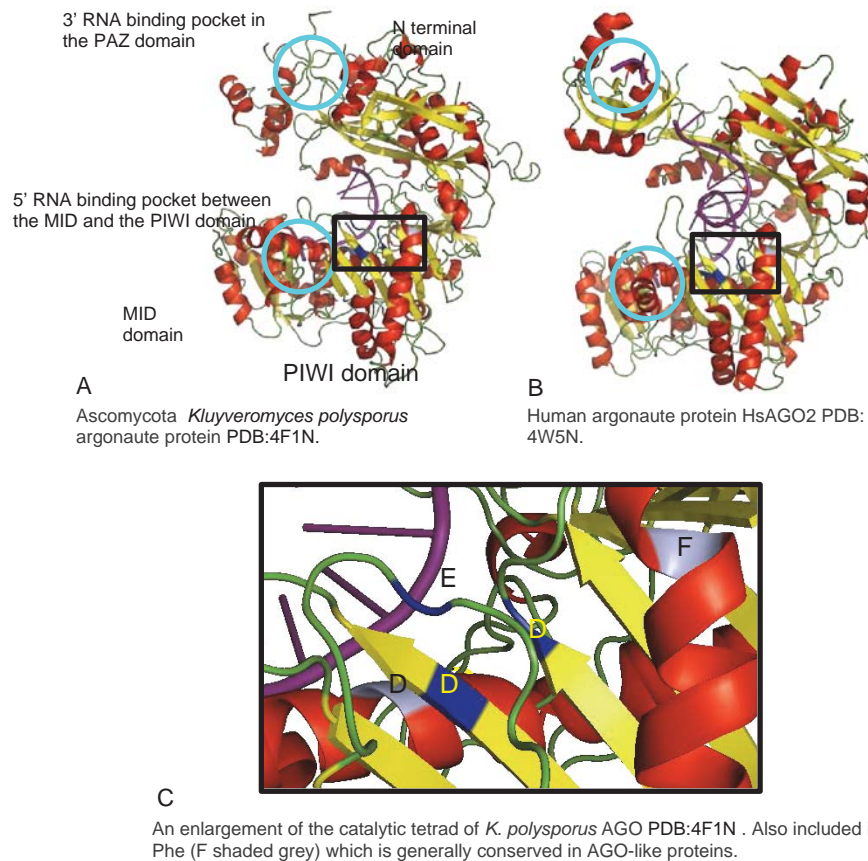
somatic cells (Ross *et al.*, 2014). Many of the fungal sequences are annotated as eukaryote initiation factor (eIF) or transcription initiation factor (TIF) proteins. Some are annotated QDE which refers to the original RNA-induced silencing complex (RISC) posttranscriptional gene-silencing pathway known in fungi as ‘quelling’. There is additionally an RNA-induced transcriptional silencing (RITS) complex responsible for a chromatin-based silencing pathway which reversibly silences unpaired genes and any homologs during meiosis. The latter function being more similar to that of the metazoan PIWI type argonaute.

Some plants have large numbers of argonautes from whole genome duplications but the expansions of one (or just a few genes) in *C. elegans* (and now in *R. irregularis*) may be different. Study has been carried out on the worm (nematode) expansion (Yigit *et al.*, 2006) which suggests that the duplications have evolved to fulfill different functions. Here we use bioinformatics to confirm that the *R. irregularis* argonautes are likely to be functional.

AGO-like and PIWI-like proteins both consist of four domains; the N terminal domain, the PAZ domain, the PIWI and the MID domains. (Note that there are both PIWI domains, and PIWI proteins on this classification). The PAZ and MID domains form binding pockets for RNA, and the catalytic residues Asp-Glu-Asp-His (DEDH) are found in the PIWI domain (Nowotny *et al.*, 2005). Here we use the presence of the catalytic tetrad of the PIWI domain as additional evidence to support structural homology. While we were reconstructing metazoan ancestors we noticed that in addition to the catalytic residues the PIWI-like proteins had retained a signature tetrad of residues at the C terminal that was remarkably conserved across PIWI-like proteins from all of the groups that had not lost the PIWI-like protein. Marginally less well conserved, but still remarkably useful, was the retention of the C terminal residues in the AGO-like proteins that showed the fourth from last residue at the C terminal being Met. So we are looking for two additional pieces of evidence to support the structural prediction; retention of the catalytic tetrad of residues (DEDH), and retention of a typical AGO-like C terminal signature of ‘Met and two aromatic residues followed by a non-polar hydrophobic residue’ given that PIWI-like proteins have reportedly been lost in fungi. We refer to this as the AGO ‘signature’ to avoid confusion with the ‘catalytic tetrad’.



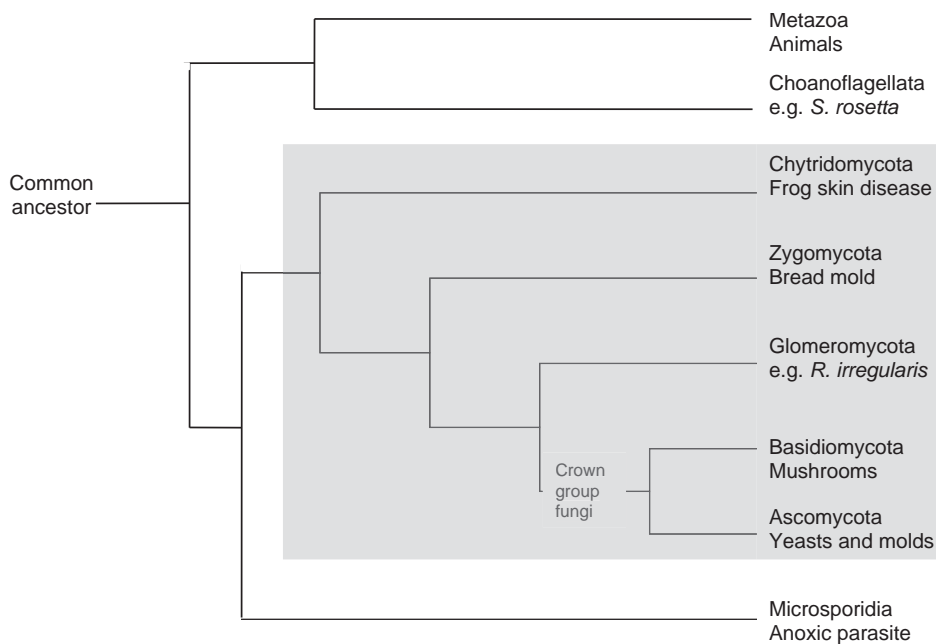
Figure 4b.1A shows the complete crystal structure of the ascomycetes AGO-like protein (PDB:4F1N 1046 residues) (Berman *et al.*, 2000) from *Cluyveromyces polysporus* which has 1251 amino acid residues in the primary sequence. It has a catalytic tetrad of Asp-Glu-Asp-Asp (DEDD), common in Basidiomycota but unusual in Ascomycota. The fungal sequences are typically (though not always) longer than those found in metazoa. Fungal-AGO is shown together with the human AGO2 structure (PDB:4OLA 859 residues, fig. 4b.1B) as a comparison.



#### Fig. 4b.1 Solved structures of the argonaute family

**A.** The solved structure of the fungal AGO-like protein from *K. polysporus* with one half of a cleaved RNA bound (purple). **B.** Human AGO-like argonaute AGO2 with full-length cleaved RNA bound. HsAGO2 is the only human AGO that retains slicing capability. Both cartoons have the catalytic residues marked in dark blue (within the black square), RNA binding pockets are circled in aqua. **C.** The catalytic site of *K. polysporus* is shown in the enlargement. Most argonautes (AGO-like and PIWI-like) retain Asp-Glu-Asp-His (sequentially in the primary sequence DEDH). *K. polysporus* has at its catalytic core Asp-Glu-Asp-Asp (DEDD) but is otherwise identical. DEDD is frequently found at the catalytic site in Basidiomycota but is unusual in Ascomycota. Phenylalanine (F) is also marked in grey, this is usually conserved in AGO-like proteins and appears not essential for catalysis but replacement of this residue can impair or abolish catalysis. All ribbon diagrams are rendered in PyMol version 1.3.

We need to give a phylogeny of the fungamals (fungi and animals) as currently understood (fig. 4b.2). This is much simplified and does not account for timing of divergences but is given in order that the reader may visualise the scheme and the placement of fungi and microsporidia within the group (Capella-Gutiérrez *et al.*, 2012).



**Fig. 4b.2 A simplified Unikont tree showing the proposed relationship between the various phyla that contributed to the work.**

Fungi are shaded and microsporidia are shown as a sister group. The timing of divergences is not indicated.

Arbuscular mycorrhizal fungus (AMF) of the phylum Glomeromycota, are obligate symbionts with coenocytic hyphae, that is, the cells contain multiple nuclei. *R. irregularis* is described as containing 26 AGO-like proteins in its genome (Tisserant *et al.*, 2013). Differences between the genomes in the multiple nuclei are only 0.43 single nucleotide polymorphisms per kb (Tisserant *et al.*, 2013), which is far less than between different strains of the same species. Gene synteny also suggests that the genomes in each nuclei are uniform, so heterokaryosis is not the reason for the multiple gene copies of the argonaute proteins.

AMF have an intricate symbiosis with most vascular plants; the arbuscles are branched structures that facilitate the exchange of nutrients. At this time *R. irregularis* is the only sequenced Glomeromycota fungus, and until recently these organisms have been difficult to culture in the lab. This means that we don't have any other species to compare them with to test if this expansion is specific to *R. irregularis*, or is general to Glomeromycota. We have used other early branching fungal sequences from Chytridomycota and the Zygomycota to build trees that set a background to demonstrate the diversity of the *R. irregularis* sequences.

We also studied the evolution of the argonaute proteins generally and have created ancestors from representative fungal and microsporidian sequences (which

appear to have evolved from different duplications). By using structural information rather than sequence alone we can infer a single ancestral sequence to represent a group of sequences and calculate trees with far fewer branches as long as we know that each ancestor is truly representative of that group. We can then add the fungi and microsporidia ancestral sequences to metazoan ancestors that we have already inferred to help unravel the evolutionary story of both groups.

### **4b.3. Methods**

A BLAST search with human AGO2 as a query sequence against the UniProtKB eukaryote database resulted in 1,000 hits that comprised mostly of plant and metazoan argonaute sequences. Additionally we used the ancestors of the metazoan sequences (unpublished) that we had created as query sequences and searched specifically the UniProt fungal database. This retrieved sequences from *R. irregularis* and other early branching fungi. There are 76 putative AGO sequences for *R. irregularis* but many of these are duplicated database entries from different strains of the same species. We checked NCBI and added a few more sequences from early branching fungi and used these sequences in a MSA using MUSCLE (Edgar, 2004b) to create an early fungal tree using MrBayes (Huelsenbeck and Ronquist, 2001), both run in the Geneious platform (Geneious R8 Biomatters available from <http://www.geneious.com/>, last accessed 9<sup>th</sup> August 2015). Because there are so many fungal argonautes we confined our trees to selected representatives but undertook extensive searches for argonautes where they were reported as absent.

The I-TASSER (iterative threading assembly refinement server) suite of algorithms (Zhang, 2008; Roy *et al.*, 2010; Yang *et al.*, 2015) was used to predict the three-dimensional (tertiary) structures of proteins from their primary sequence to analyse their likely functions. I-TASSER is benchmarked by CASP (Critical Assessment of Techniques for Protein Structure Prediction) (Moult *et al.*, 1995) a biannual experiment in which servers are tested on their ability to identify correct folds from protein sequences whose structures have been previously-determined but held back from publication by the PDB for the experiment. I-TASSER has scored highly since its inception competing as ‘Zhang Lab’ (Xu *et al.*, 2011). I-TASSER was run via the NeSI high-performance computing facility at the University of Auckland (New Zealand). Reasons for prioritising a sequence for analysis could be because they group unusually, for example the microsporidian *Mitosporidium daphniae* sequence

(UniProtKB:A0A098VRC6) groups with fungi away from other microsporidia. In this case this is consistent with other observations (Haag *et al.*, 2014). Another factor would be a very high number of residue changes per site or a very low posterior probability at a node.

Our previous work with the Major Vault Protein (Daly *et al.*, 2013b; Daly *et al.*, 2013a) used the I-TASSER confidence (C score) score in aiding our analysis where a score of greater than  $-1.5$  (out of a range  $-5$  to  $+2$ ) predicts a homologous fold. However we found that even where the annotation is correct, and there is a solved crystal structure, the score could be very low due to sequence inserts that most likely do not affect overall AGO structure and function. For example, we have previously found large inserts in both human PIWI2 protein (UniProtKB:Q8TC59) and *Drosophila* AGO2 (UniProtKB:Q9VUQ5), sequences which give both a low I-TASSER C score and a high score in terms of sequence changes per site and yet these are verified as functional proteins (Itou *et al.*, 2015; Abramov *et al.*, 2016).

We have previously found that we could ‘improve’ the C score by removing some or all of the inserts (Daly *et al.*, 2013b). By checking the I-TASSER prediction, compared to a known argonaute structure such as the human AGO2 (PDB:4OLA) using FATCAT (Ye and Godzik, 2003) we can align the proteins and get a value of how similar they are in terms of equivalent residue positions revealing where inserted sequences in the 3-D structure prediction lie. Removing sequence insertions additionally lowers the magnitude of the residue changes per site on the tree branch which we needed to do for ancestral sequence reconstruction (ASR).

Once we could be confident that the sequences that we had were *bona fide* argonautes we could use them for ASR. We have investigated a variety of algorithms for doing this (Daly *et al.*, 2013b) and are convinced that FastML gives good reconstructions. In MSAs the alignment algorithms insert gaps to produce the best alignment. The longer the MSA the greater the ‘noise’ from the gaps which results in excessively long reconstructed sequences because FastML will simply fill them all in. I-TASSER then predicts unstructured loops that are artifacts of the process. We therefore need to remove gaps that occur as the result of insertions in one or a few sequences (where it is more parsimonious that they are likely to be more recent insertions in a few species, rather than ancestral loss in the majority). This has been done manually by removing gaps where less than 10% of the sequences have residues resulting in gaps in

all the other sequences. This can also be done automatically using Wasabi (Veidenberg *et al.*, 2015). In practice it doesn't make any difference to the resultant tree. FastML will produce an ancestor for each node of the tree. We then check that the ancestor from the deepest node would be predicted by I-TASSER to fold as an argonaute, and then use that sequence as a new BLAST query to search for more remote homologs.

## 4b.4. Results

### 4b.4.1. Yeast and fungi

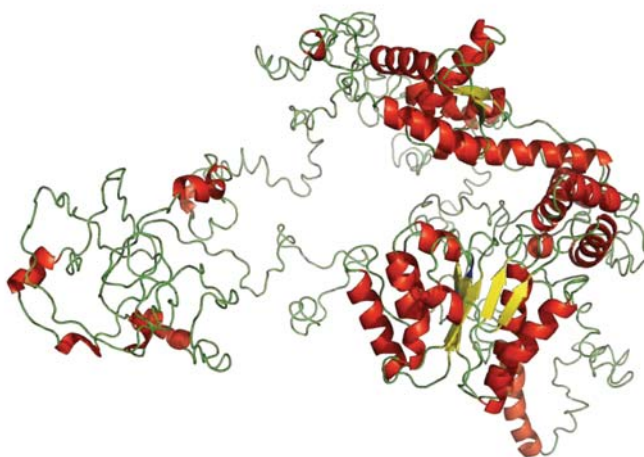
The term 'piRNA' implies interaction with the PIWI type argonaute, rasiRNAs (repeat associated small interfering RNAs) are well described in plants, and are now incorporated as a sub-species within the piRNA family (Klattenhoff and Theurkauf, 2008). rasiRNAs are also found in species such as the fission yeast *Schizosaccharomyces pombe* (Sigova *et al.*, 2004) but neither plants nor fungi have been observed to contain the PIWI type protein. However it is known that siRNAs found in *S. pombe* can induce heterochromatin changes and force transcriptional silencing of the genome in a similar manner to the PIWI-type argonautes of metazoa (Lippman and Martienssen, 2004). This means that the loss of PIWI has not resulted in fungi having to rely solely on post-transcriptional silencing.

Although most fungi have two types of AGO-like argonaute protein, some have none at all. In some cases there appears to be an obvious explanation for gene loss, for example, some fungi harbour the 'killer' virus which appears incompatible with RNAi. The 'killer' virus confers immunity on cells that maintain it from a killer-associated toxin-producing satellite. Destroying 'killer' makes the fungus vulnerable to the toxin (Drinnenberg *et al.*, 2011). However some yeast species; *Candida glabrata*, *Saccharomyces kudriavzevii*, *S. kluyveri*, *Kluyveromyces waltii*, *K. lactis*, *Ashbya gossypii* and *Debaryomyces hansenii* appear to have neither the capability for RNAi, nor harbour the 'killer' virus. Do species harbouring 'killer' survive in the evolutionary long term? Fungi that have lost RNAi appear to be relatively recent events so it could be proposed that fungi which lost RNAi in the past did not survive (Drinnenberg *et al.*, 2011). What is it that ensures survival of those species with neither RNAi nor 'killer'? There are interesting and important questions here.

In our search of novel AGO sequences we found examples where the protein sequence had diverged from the consensus sequence. This commonly occurs following



a duplication event that frees the duplicated gene from functional constraint. However in *A. gossypii* (a species where RNAi is lost) the sequence found by BLAST (UniProtKB:M9MXJ8) was submitted to I-TASSER and the resulting models analysed for evidence of a catalytic site. Although one residue of the conserved catalytic tetrad was in place, and synonymous substitutions raised the sequence identity from 12% to 27% compared with *K. polysporus*, we couldn't be sure if this sequence represents a relic of a broken down argonaute, or an unrelated protein (fig. 4b.3). We put the sequence into Phyre2 where 20% or 200 residues were modelled as an SH3-like barrel from the BAH superfamily but no folds resembled those of the argonaute. So from our evidence; lack of identifiable domain structure, lack of catalytic residues we can exclude this sequence from consideration.



**Fig. 4b.3 *A. gossypii* (UniProtKB:M9MXJ8) putative AGO sequence identified by BLAST.**

The fungal ancestral sequence poorly conserved compared with other fungal sequences. The predicted structure shows one Asp residue is in the catalytic position that would be anticipated but there is no indication for a complete catalytic site.

We explored an alternate possibility; small RNAs deriving from the Ascomycota *Botryotinia fuckeliana* (also known as *B. cinerea*) genome have been found in plant cells, utilizing plant AGOs (Weiberg *et al.*, 2013) even though *B. fuckeliana* has two argonautes of its own. So there is a possibility that a parasitic or symbiont lifestyle could mean that harbouring the AGO/PIWI machinery is unnecessary when plant, or potentially other host proteins could be used instead. However the majority of the yeasts that lack argonautes are not parasitic. We have also found total argonaute loss in the choanoflagellate, *Monosiga brevicollis*. *M. brevicollis* does not harbour any proteins of the argonaute or small RNA pathways (Grimson *et al.*, 2008) and although BLASTing retrieved two putative argonaute sequences in a recently sequenced choanoflagellate, *Salpingoeca rosetta*, neither were predicted to fold at all like any kind of functional



argonaute. We were also unable to find any putative sequences in *Capsaspora owczarzaki*, a sister group to metazoa, so although argonautes are ubiquitous and essential to most species, they do not appear to be indispensable.

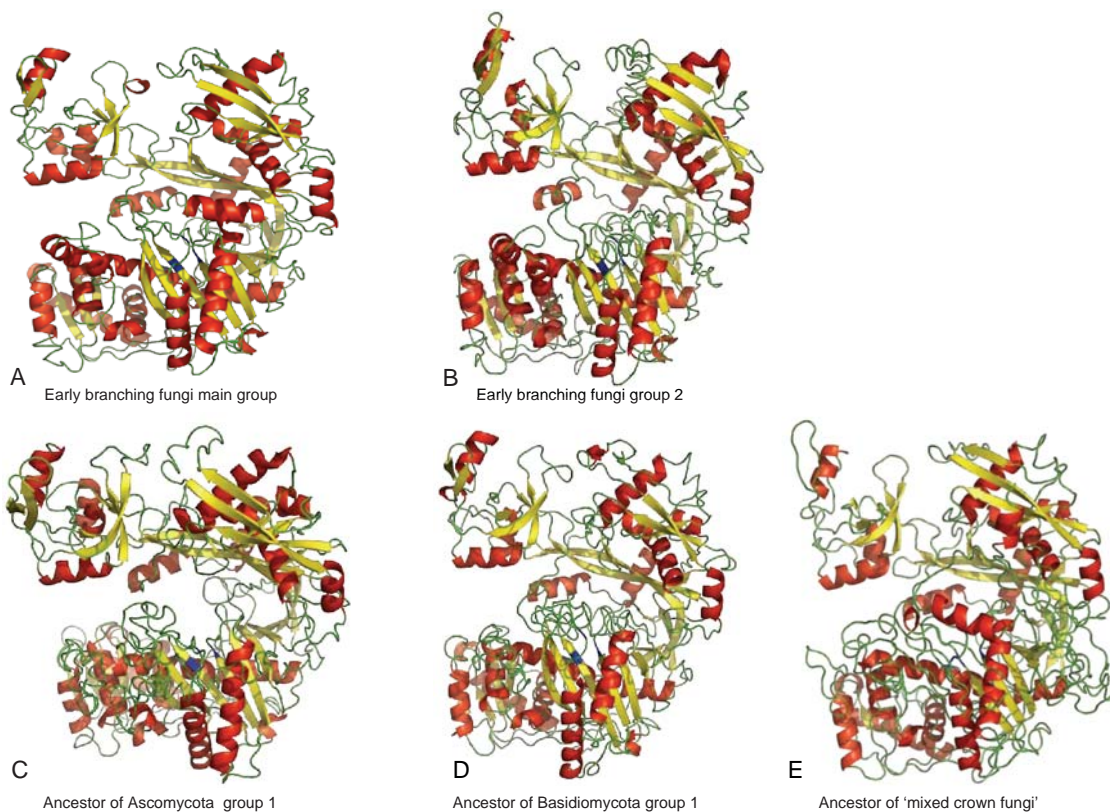
From our representative sample of species we found greater similarity between proteins between species than between the argonautes within a single species, so the simplest explanation is that the duplications occurred earlier in an ancestor. We have found that since basal metazoa harbour two AGO-like argonautes (as well as PIWI), the most parsimonious explanation is that AGO- and PIWI-like proteins were both present prior to the animal / fungal split. Other work supports the idea of two types of argonaute within Basidiomycota termed ‘group A and B’. Also in agreement with our own work ‘group A’ contains argonautes from Ascomycota as well as Basidiomycota (Hu *et al.*, 2013). We have termed this group ‘mixed crown fungi’ and inferred a single ancestor although it is most likely that the ‘mixed crown fungi’ and ‘group one Basidiomycota’ derive from the same original gene.

We also found five sequences that appear in different ‘clades’ in different trees. Eventually *Cryptococcus gattii* (UniProtKB:E6R522 and E6R506), *Pseudozyma aphidis* (UniProtKB:W3VSF6), *Sporsorrium reilani* (UniProtKB:E6ZNC3) and *Trichosporon asahii* (UniProtKB:J5TUC3) were left out of any ancestral sequence calculation. What these sequences have in common is that their C terminal signature shows Ala-Trp-Phe-Met (AWFM), Gly-Trp-Phe-Met (GWFM), Leu-Trp-Phe-Met (LWFM), Leu-Trp-Tyr-Met (LWYM), and Ala-Trp-Phe-Met (AWFM) respectively which are unlike other fungal sequences. However when modelled in I-TASSER these sequences folded as any other argonaute with an intact catalytic tetrad, so although they are argonautes, we are unable to assign them to a particular ancestor with confidence.

There are two ancestors from the early branching fungi. The sequences that made up the early fungal group 2 ancestral sequence are too few for a high quality ancestor so the high posterior probability of the residues at node 1 was expected (posterior probability logos for all ancestors are available in wikispaces). The unrooted tree shows that they are clearly different from the sequences that make up the main group early fungal ancestor and they have more in common with microsporidia. The unrooted tree is available as supplementary material in wikispaces. It was already known that *M. daphniae* was closer to the cryptospora *Rozella allomyces*, however it was surprising that *R. allomyces* would group closer to the microsporidians than did *M.*

*daphniae*. Fortunately we are not assigning phylogeny on the basis of one protein sequence, particularly as there was an unexpected resolution for *R. allomyces* when we calculated a tree of all fungi (including microsporidia) and animals (fungamals) because they grouped with metazoan AGO-like sequences! (S3 available in wikispaces).

The posterior probability at the catalytic sites and at the C terminal AGO signature residues for all of the inferred ancestors is in excess of 0.9 (i.e. 90% confidence) with the exception of the final residue (~0.5). The N terminal probability of the four remaining ancestors is generally poor and this is anticipated since prokaryotes do well without the N terminal at all and this area frequently gains inserts. The phenomena where inserts are gained and retained in already poorly structured regions has been previously documented (Light *et al.*, 2013). The PIWI domain has higher conservation generally than does the remainder of the protein, which was also anticipated. The resultant inferred ancestors are shown in fig. 4b.4 (an unrooted tree of all sequences that contributed to the ancestors is available as supplementary material S1 in wikispaces).



**Fig. 4b.4. Structural predictions of the reconstructed ancestors from the tree of representative fungi species.**

Each of the reconstructed ancestors retains a canonical catalytic tetrad (DEDH) with the exception of ancestor D (Basidiomycota group 1) that retains DEDD.

There are a number of species with paralogs that have occurred after speciation in addition to the *R. irregularis* expansion, e.g. *Mortierella verticillata* harbours six copies of the argonaute protein. Four are grouped with the main group but two have much less similarity; there is ~60% sequence homology within each group, but only ~35% between them. However there were already two different AGO-like argonaute proteins at the time of the split from metazoa as we had predicted from our work on metazoan AGO and PIWI evolution.

#### **4b.4.2. The *R. irregularis* AGO expansion**

We investigated the *R. irregularis* AGO expansion using 3-D structural prediction and retention of catalytic sites. One comparison is the well-documented *C. elegans* (nematode) argonaute expansion (Yigit *et al.*, 2006) in which the *C. elegans* specific argonautes (WAGOs) had arisen from a single protein (RDE-1), determined to be close to the divide between AGO-like and PIWI-like proteins. However, in the case of *R. irregularis* it can only be from an AGO-like ancestor. But do all the duplicated proteins derive from the same original AGO because it appears that there were two AGO-like proteins at the base of the fungal family?

A BLAST search using an annotated *R. irregularis* argonaute (UniProtKB:U9SUD1) returned 36 AGO-like protein sequences from a single strain of *R. irregularis*, several of which are short fragments. Short sequences that contain only a MID and PIWI domain have been characterised from bacteria and archaea (Makarova *et al.*, 2009). Eukaryote argonautes are most likely to contain a PAZ domain to provide for RNA binding pockets, although it is possible that this is for protection rather than a requirement for binding (Hur *et al.*, 2013). It is difficult to distinguish between fragments deriving from short fragments of coding DNA, fragments due to incomplete sequencing, and genuinely short argonautes. We additionally checked our findings against the Pfam database that gives a graphical outline of the domains in sequences from UniProt and NCBI (Finn *et al.*, 2010). Supplementary material S2 in appendix II gives greater detail of the Pfam findings.

An initial tree comprising of early branching fungi and including the *R. irregularis* expansion gave an example of the need for ‘post tree’ structural analysis. Our original BLAST of the fungus *R. irregularis* included the protein UniProtKB:U9UV71. The posterior probability at the node was high and the sequence identity compared to the other sequences in the early branching fungal tree was between

20-40% (well within the range that we would consider for a putative homolog), but the sequence had a score showing 1.033 predicted changes in residue per site. In fact a number of branches in other trees show similarly high substitution rates. Although this implies that more than 100% of the residues have changed – we can show that this is not the case. Some residues may have changed several times but there are core residues that are essential to retain correct structure and catalytic capability. However all sequences that show in excess of 0.8 predicted changes per site are automatically referred to I-TASSER for structural prediction. UniProtKB:U9UV71 failed to be modelled with AGO-like domain structure (although it was superficially similar). Additionally it lacked any sign of the catalytic tetrad, demonstrating that not all sequences identified by a BLAST search necessarily have structural homology. Our conclusion is that UniProtKB:U9UV71 is not currently a functional AGO-like protein, and we do not yet know what it does. Details of the structural predictions for all of the *R. irregularis* sequences and additional information including I-TASSER C scores can be found in appendix II (supplementary material S2).

We do need to make a caveat here; although this sequence (UniProtKB:U9UV71) failed to fold convincingly as a functional argonaute and we were alerted to this by the high number of residue changes per site, the ancestors that we created had similarly high predicted changes in residue per site yet are predicted to fold as competently structured argonautes. Our point is that the residues at some, or even most sites, may have changed many times, but the residues essential for structure and function have much greater constraint and have not been subject to very much change which is important. It is also important that we point out that without 3-D structural prediction we would not be able to easily distinguish the difference.

Twenty-seven of the remaining sequences were long enough to represent full-length proteins rather than fragments and were submitted to I-TASSER. Three atypically short sequences (but longer than obvious fragments) were also submitted to I-TASSER since we know that bacteria have a family of short argonautes; UniProtKB:U9UIX7 (fragment), U9SXZ8 and U9UAG7 (fragment)) of 481, 403 and 544 residues respectively retain the PIWI domain complete with the catalytic tetrad DEDH (and also a Phe that is semi-conserved, noted in fig. 4b.1C), but lack the RNA binding pocket found in the PAZ domain which may not preclude binding (Hur *et al.*, 2013). These are in shaded boxes on the tree in fig. 4b.5 with the caveat that these do

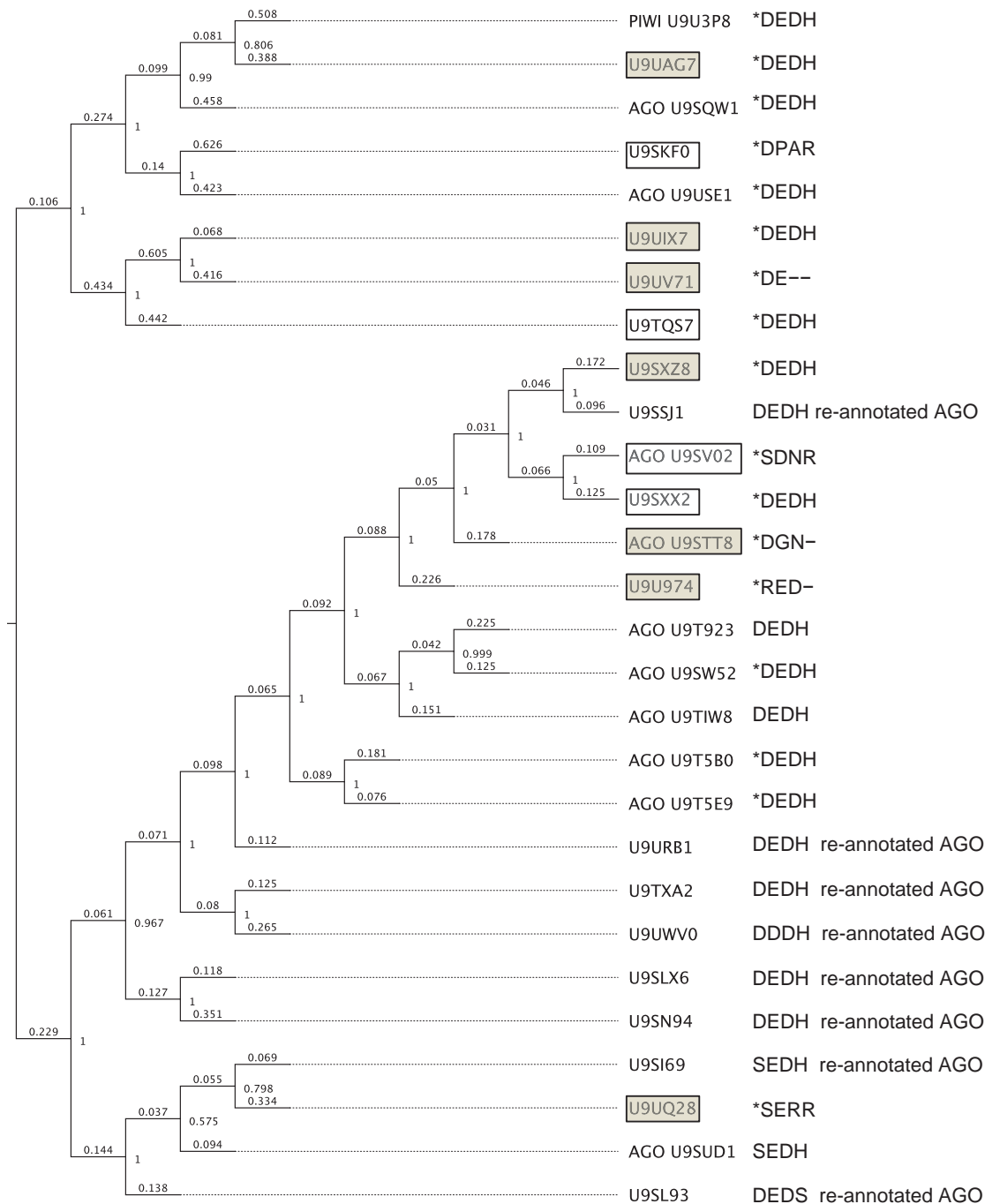
resemble bacterial argonautes. UniProtKB:U9UQ28 is 580 residues but the structural model is poor and the catalytic residues are also unconvincing (also marked in shaded boxes on the tree in fig. 4b.5). Supplementary material S2 in appendix II shows the predicted structure for all of the *R. irregularis* sequences including C scores for the I-TASSER modeling.

Another of the shorter sequences (described as a fragment, UniProtKB:U9T4B0 at 692 residues) does appear to fold with enough structure to possibly be functional (S2 appendix II). Sequences between 600 and 700 residues tend to lack a complete N domain but it is not clear that this will prevent the protein from functioning. The N domain is poorly conserved across all species and may also be dispensable since the short bacterial argonautes previously described lack the N domain altogether (Makarova *et al.*, 2009). We know that slicing is impaired in human AGO proteins when the N domain is compromised but we know that this doesn't affect RNA binding and functionality other than cleavage (Faehnle *et al.*, 2013). So, we have accepted that U9U3P8, U9SW52, and U9T5E9 could function despite a poor N domain but we lack confidence in two similarly medium length sequences. These are U9SVO2 which lacks an N domain but additionally has an unlikely catalytic tetrad (SDNR), (though the absence of a complete catalytic tetrad doesn't necessarily preclude RNA binding, nor binding to a target site), so our uncertainty is marked by boxes without shading in fig. 4b.5. U9U974 lacks much of the PIWI domain including some catalytic residues (S2 appendix II) and so we cannot support a functional argonaute for this structure.

Of the longer sequences UniProtKB:U9STT8 has essential parts of the C terminal missing and is marked by a shaded box as 'unlikely to function' in fig. 4b.5. Since we undertook this work there has been an update in annotation of the sequences, which we show in added text 're-annotated AGO'. It can be seen that one of the sequences (UniProtKB:U9STT8) was already annotated as an AGO type argonaute contrary to our analysis and we have already described our uncertainty over UniProtKB:U9SVO2. Of the sequences that have been 'upgraded' we are in agreement, but UniProtKB:U9TQS7 and U9SXX2 remain 'uncharacterised', the former has some small loss of  $\alpha$ -helix in the 5' RNA binding pocket and the latter has loss of  $\alpha$ -helix in the 3' RNA binding pocket (S2 appendix II). Whether this is sufficient to preclude RNA binding we cannot be sure and so these are boxed in fig. 4b.5. Additionally there seems no justification at all for retaining the PIWI annotation for UniProtKB:U9U3P8,



regardless of its functionality the annotation is anomalous as PIWI proteins have not been found in fungi and the C terminal signature Met-Phe-Phe-Val (MFFV) is typical of AGO-like proteins. In trees containing both AGO and PIWI proteins this groups with AGO.



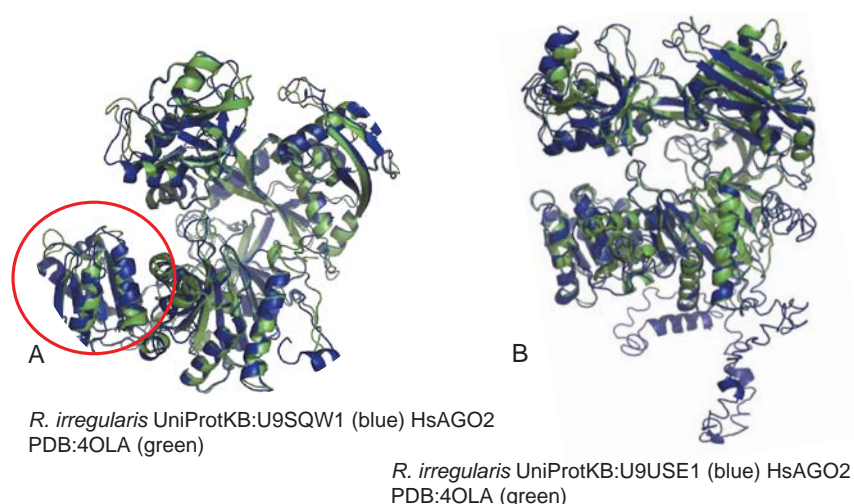
**Fig. 4b.5 Fates of the *R. irregularis* expansion.**

Sequences marked in shaded boxes are those that we cannot support as fully functional argonautes, the boxes without shading show the sequences that we are not certain about. The remainder would appear to be capable of functioning as argonautes. Protein sequences recently annotated as AGO (previously ‘uncharacterised’) are marked. Sequences referred to in the text are marked with an asterisk\* and further information can be found S2 appendix II.



UniProtKB:U9SQW1 is said by Pfam to lack the MID domain. If the MID domain were missing it would affect the 5' RNA binding pocket. However in this case the MID domain structure appears intact in the predicted model (S2 appendix II) so we aligned this sequence via FATCAT (Ye and Godzik, 2003). The I-TASSER prediction shows that the MID domain is complete and structurally is no different than the MID domain of human AGO2 (PDB:4OLA) (fig. 4b.6A). Thus it appears that there are some aspects of Pfam that need to be updated.

Of those remaining, the sequence UniProtKB:U9USE1 was particularly long (1094 residues) and had a relatively poor C score (−1.53) in I-TASSER. We aligned this with the characterised human AGO2 structure (PDB:4OLA) and believe that it would be functional despite this large insert that is located outside the core AGO structure (fig. 4b.6B). Additionally *R. irregularis* UniProtKB:U9SKF0 is predicted to fold as a homolog with a catalytic tetrad of Asp-Pro-Ala-Arg (DPAR). However a nearby glutamate is on a loop in a position where it could extend into the catalytic site.



**Fig. 4b.6 A comparison of UniProtKB:U9SQW1 with the solved structure of the human argonaute.**

**A.** Pfam failed to find evidence of a MID domain in *R. irregularis* UniProtKB:U9SQW1. The FATCAT alignment of HsAGO2 PDB:4OLA (green) and U9SQW1 (blue) is tilted to show the MID domain (circled) and we find that the domains entirely overlap. **B.** Extra sequence may not impair functionality. HsAGO2 (green) is aligned with *R. irregularis* (UniProtKB:U9USE1) (blue) and the inserted sequence doesn't appear to impact the structural integrity of the argonaute. *R. irregularis* (UniProtKB:U9USE1) has a greater percentage of equivalent residues with HsAGO2 than there are between HsAGO2 and *Drosophila* AGO2. Structural alignment is by FATCAT (Ye and Godzik, 2003).

Ultimately we find that 17 of the sequences have a good likelihood of being functional AGO but they do not seem to have arisen from one protein similar to the worm expansion (Yigit *et al.*, 2006) but rather from duplications from two different original sequences, (fig. 4b.5). Removal of the outlier sequences and recalculating the

tree from the remaining 17 protein sequences doesn't make any difference to the tree so it seems that in this case both original genes have duplicated a number of times. We note that it does not make these sequences any less relevant in terms of the rate of duplication of the argonaute gene.

So we can confirm that the argonaute expansion in *R. irregularis* is real but why has it occurred? This is difficult to answer. Perhaps the expansion of argonaute proteins is associated with the comparatively large genome ( $\sim 154.8 \pm 6.2$  Mb) (Sędziewska *et al.*, 2011), supported by Tisserant *et al.*, (2013), rather than as a result of the lifestyle of the fungus, or the nature of its nuclei? The genome is bloated by repeat and transposon DNA and in metazoa it is the PIWI-like protein that is responsible for limiting the impact of transposons so perhaps the loss of the PIWI-L protein could be responsible for these oversized genomes. However, if that were the case then a runaway explosion of transposons should enlarge all fungal genomes!

The powdery mildew *Blumeria graminis* and the black truffle *Tuber melanosporum* (both Ascomycota) also have genomes massively enlarged due to transposons (Spanu *et al.*, 2010), (Wicker *et al.*, 2013) (Martin *et al.*, 2010) but neither species demonstrate the expansion of argonaute genes seen in *R. irregularis*. In fact the large number of transposons present in the genome, including retrotransposons, would argue that the argonaute system that is supposed to suppress them is not working well in *B. graminis* or in *T. melanosporum*, and even with the increased number of argonautes in *R. irregularis* it could be said that the argonaute suppression system wasn't working there either! However Tisserant, *et al.*, (2013) report that transposon families are in decline so it may be that the expansion has occurred in response to the number of transposons and is working to repress them.

The occurrence of coenocytic cells is relatively widespread – not just amongst fungi, e.g., green algae (stoneworts), plasmodial slime moulds (myxogastria), and even early embryo development in *Drosophila* has a multinuclear stage yet there is not a notable expansion of argonaute proteins in *Drosophila*. We do need to be careful here that we are not insisting that the only reason to maintain an argonaute-like protein is in order to have RNAi capability because other functions have been identified (Juliano *et al.*, 2011) (Wei *et al.*, 2012).

One possibility is that the *R. irregularis* expansion is the result of a previous infection of an ancestor with some type of parasitic nucleic acid (possibly viral) that

was the driver for an escalating number of argonautes to deal with it. A study of the prawn AGO-like protein AGO1 noted that different isoforms had different functions in antiviral immunity (Huang and Zhang, 2012), so it may be that a general lower rate of alternate splicing in fungi consistent with a general lower rate of introns (Irimia *et al.*, 2007) has meant that gene duplication was an effective option. However the argonaute sequences in *R. irregularis* have a surprisingly high number of small introns (EnsemblFungi <http://fungi.ensembl.org/index.html>), though each show only one transcript. In any case the main type of alternate splicing in fungi generally is intron retention with skipped exons being quite a rare event (Grützmann *et al.*, 2014).

Tisserant *et al.*, (2013) proposed that ancient whole genome expansion but with slow loss in some genes, notably the kinases and proteins with MATA-related HMG domains would also explain the argonaute expansion. The problem with this is that the sequence similarity between the *R. irregularis* argonautes is much greater than the primary homology between the *C. elegans* argonaute expansion, implying that the expansion is relatively new compared to the expansion in *C. elegans*. It was argued that in nematodes an ancient whole genome duplication had resulted in the expansion seen in *C. elegans* and that the retention of the argonaute genes was the result of their gain of function in this species and less so in other *Caenorhabditis* species and not at all in parasitic nematodes (Dalzell *et al.*, 2011). Nematodes are known to be susceptible to massive gene loss (Aboobaker and Blaxter, 2003) and there is evidence that the parasitic nematode *Pristionchus pacificus* does still have 23 relics of the argonaute gene, yet only one is complete (UniProtKB:H3-DS31) so the possibility of previous genome duplications and slow loss of some proteins does need to be considered in *R. irregularis* as well.

Alternately there could be something in particular about the *R. irregularis* argonaute sequence that causes them to duplicate. The RNase H fold in the PIWI domain is also found in HIV integrase and reverse transcriptase as well as in the Tn5 transposase (Nowotny *et al.*, 2005), so it is possible there is gene sequence that duplicates and diverges in a functional domain-based mechanism. The argonautes in general do seem to be able to duplicate quite readily since many species have a number of paralogs. We do need genome sequences of additional arbuscular fungi to see if this is unique to *R. irregularis*.

#### 4b.4.3. Microsporidia

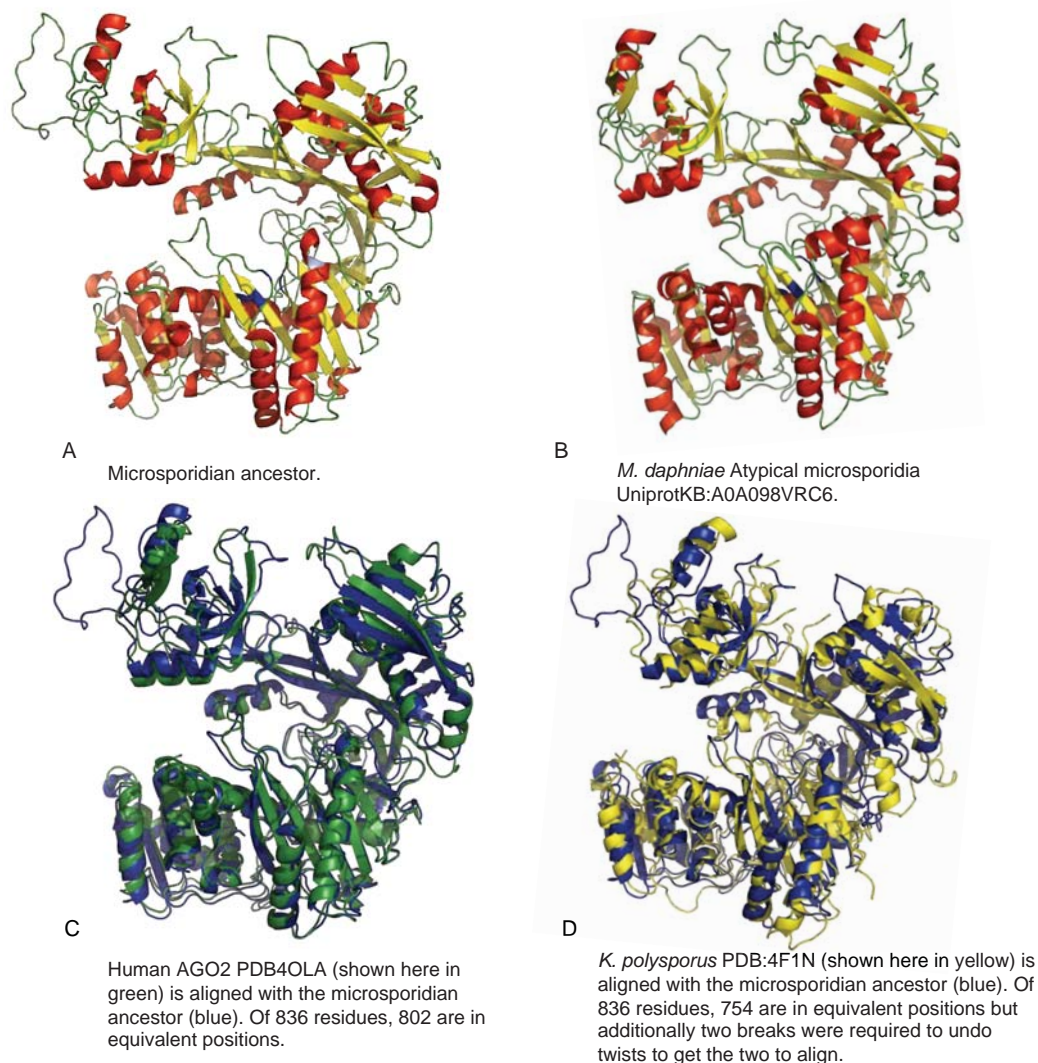
Microsporidia were once considered to be protists but in 1999 Weiss *et al.* placed them within the fungal clade (Weiss *et al.*, 1999). They have since been reclassified as a sister group to fungi (Liu *et al.*, 2006) due to evidence of a common ancestor, but outside of fungi. They are parasites (living in anoxic environments) characterised by the development of a unique polar tube that facilitates infection (Franzén, 2005) and a loss of the mitochondrial genome, reducing the mitochondria to a mitosome, requiring then that ATP is transported into the parasite from the host (Tsaousis *et al.*, 2008).

With the exception of *M. daphniae* all of the microsporidia argonautes have the PIWI-like C terminal signature Leu-Phe-Tyr-Val (LFYV) that we have previously observed only in PIWI-like argonautes. With the exception of *M. daphniae*, the microsporidia grouped either as an isolated clade close to the AGO-like sequences or with the PIWI-like proteins which depended on other sequences in the tree and the tree root (S3 available in wikispaces). *M. daphniae* grouped with the AGO-like early branching fungal sequences and this correlates with recent findings that *M. daphniae* represents a link between microsporidia and fungi and retaining features more similar to fungi (including a mitochondrial genome) but also features unique to microsporidia (Haag *et al.*, 2014). BLASTing with the *M. daphniae* sequence resulted in the closest sequences identified belonging to *R. allomyces* and within the top 1,000 hits were plant and metazoan argonautes but none of the sequences from other microsporidia. For such reasons the microsporidian ancestor does not include UniProtKB:A0A098VRC6 from *M. daphniae*.

The microsporidian ancestor retains the catalytic residues and I-TASSER resulted only two models (indicative of high confidence) for this ancestor with a high C score of 1.49 (fig. 4b.7A.). Because microsporidia is a rapidly evolving parasite group that have undergone extreme genome shrinkage it seems remarkable that this protein should be so well conserved. We checked it against both the fungal solved AGO structure (*K. polysporus* PDB:4F1N) and the human AGO2 structure (PDB:4OLA) to see which it might be closer to. Surprisingly the structural prediction for the microsporidian ancestor aligned much more closely with the human argonaute with a massive 96% of residues in equivalent positions, i.e. of the 836 residues in the microsporidian ancestor, 802 of them have an equivalent geometric position in the human AGO2 (fig. 4b.7C). The microsporidian ancestor and human AGO2 primary



sequences are much closer in length (836 and 859 residues respectively with sequence identity of 29%) whereas *K. polysporus* AGO is 1046 residues which will lower the score (the sequence identity is 21% - also lower due to the extra number of residues), the difference is that there are many small inserts in the *K. polysporus* sequence compared to the microsporidian ancestor. However even if all of these are removed the sequence identity is still less than that between the microsporidian ancestor and the human argonaute. It can be seen in fig. 4b.7D that the 'fit' is simply slightly poorer all over.



**Fig. 4b.7. A comparison of the microsporidian ancestor and *M. daphniae* with solved structures.**

**A.** I-TASSER structural prediction of the ancestor of all microsporidia except *M. daphniae*. **B.** *M. daphniae* grouped with fungi rather than with microsporidia and so was left out of the MSA that contributed to the microsporidian ancestor but is shown here for comparison. **C.** Shows the FATCAT alignments of the microsporidian ancestor (blue) with Human AGO2 (PDB:4OLA green) where 96% of the residues in the ancestor find equivalent structural position with the human protein. **D.** The microsporidian ancestor (blue) aligned with *K. polysporus* (PDB:4F1N yellow) where 90% of its residues have an equivalent position with those of the solved fungal structure. It can be seen that the ancestor of Microsporidia almost entirely aligns with the human argonaute so well that the two can barely be separated.

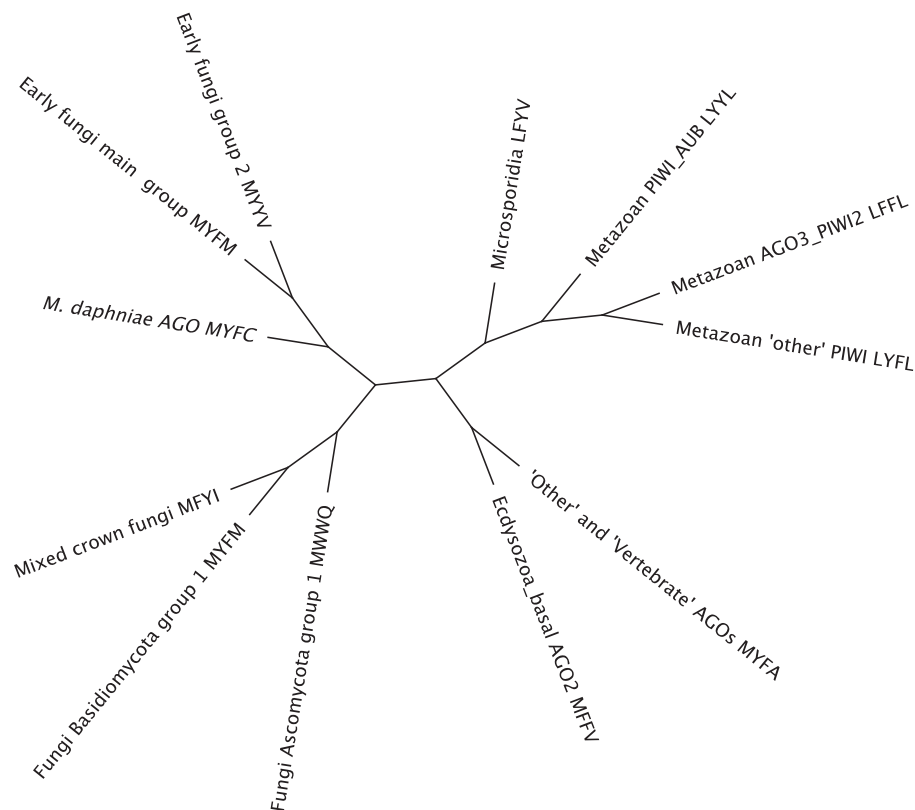
Because the microsporidian sequences sometimes aligned with the PIWI side of a tree that included fungi and metazoas sequences and because the microsporidian ancestor had the C terminal signature of the PIWI-like proteins we tried aligning the microsporidian ancestor with the human PIWI I-TASSER structural predictions (because there are currently no complete solved PIWI-like structures in the PDB). The result was that of all four human PIWI proteins, none aligned any better than the fungal AGO (*K. polysporus* PDB:4F1N) with all requiring breaks to allow for two conformational twists for a FATCAT alignment. In terms of predicted structural homology the microsporidian ancestor is closer to the canonical human AGO2 than to the solved fungal structure (*K. polysporus* PDB:4F1N) or to any of the other ancestors. Although *M. daphniae* had been left out of the microsporidian ancestor we tried aligning the microsporidian ancestor with the predicted structure for *R. allomyces* (the closest sequence to *M. daphniae*). We found that the alignment showed 98% equivalent positions and when we aligned *R. allomyces* (UniProtKB:A0A074aN09) with HsAGO2 the equivalence was 99% so we are confident that despite the PIWI-like C terminal signature, microsporidia have genuine AGO-like argonautes.

From the microsporidia species sequenced so far, none have been found with more than one argonaute. UniProt reports two identical sequences for *Nosema ceranae* (UniProtKB:C4V9J2 dated 2009 (Cornman *et al.*, 2009), and UniProtKB:A0A0F9YTR9 dated 2015 (Pelín *et al.*, 2015)) plus one full sequence and two fragments for *Nosema apis* (UniProtKB:T0MD46, T0LD06 and T0LD08) annotated as AGO, ‘leaf development protein’ and EIF2c2 respectively (Chen *et al.*, 2013). However the fragments are identical to the N terminal and C terminal of the full-length sequence and are likely to be artifacts. However we suggest that there could have been another subfamily in the lineage that led to microsporidia because of its presence in *M. daphniae*. What we can say is that the argonautes found in microsporidia are surprisingly different from fungi if they arose from the same ancestor. This could imply that the common ancestor had two argonautes one of which could have been PIWI-like and was lost in fungi after the split from microsporidia and that microsporidia lost the AGO-like argonaute. It seems unlikely that the last four residues (the C terminal signature) would always mutate from (or to) Met / Leu at the fourth from last residue and the trees cannot be dispersing the sequences based solely on the last four residues so we tend to reject the idea that PIWI and AGO are derived from the same protein after the



common ancestor with animals. Clearly though, species can flourish without either one or the other protein and in some cases with neither so why are AGO-like and PIWI-like proteins so ubiquitous and conserved if not essential in at least some species?

Figure 4b.8 shows an unrooted tree calculated from the ancestors of the species that we have reviewed here together with the metazoan ancestors that we have inferred previously (chapter 4a). There is evidence in metazoa of two PIWI proteins in an early ancestor which form part of the mechanism of so called ‘ping-pong’ amplification of piRNAs. AGO3 is an insect PIWI-like protein (despite its confusing annotation) but this ancestor also includes the vertebrate PIWI2 and this ancestor would be partnered by either one of the other two PIWI-like ancestors. Likewise there were three clear AGO-like branches in the metazoan trees but it is likely that ‘AGO1\_other AGOs’ and ‘Mainly vertebrate AGOs’ derive from the same gene so we combined them to create a combined ancestor (‘other’ and ‘vertebrate’ AGOs).



**Fig. 4b.8. An unrooted tree of fungi and metazoan sequences re-created by ASR.**

The ancestral sequences that we have recreated from fungi and microsporidia are merged with metazoan ancestral sequences that we have previously inferred and displayed as an unrooted tree. *M. daphniae* is included as an ‘orphan’. All of the sequences are predicted to fold in the manner of argonaute proteins with all domains intact. All ancestors have also retained a viable catalytic tetrad (DEDH and DEDD for Basidiomycota group 1). The C terminal signature is attached to the ancestor names. In each case with the exception of the metazoan Ecdysozoa-basal AGO2, the posterior probability for the residues at the C terminal is in excess of 90% probability for all but the final residue.

This tree clearly shows the microsporidian ancestor closer to the metazoan PIWI even though the structural comparison has more in common with the metazoan AGO. *M. daphniae* is included in the tree as an ‘orphan’ because we could not include it with any family with confidence. The ancestors all retain a functional catalytic tetrad even though it is known that it is not necessary for function in terms of suppressing RNA transcripts. The C terminal signature of each ancestor is added to each label (fig. 4b.8). Our initial tree (500 sequences) of representative fungamal sequences that we had considered showed some fungi as well as microsporidia on the PIWI side of the tree. The posterior probability at many of the nodes was 0.5 or 50:50 likelihood of being correct and this demonstrates the problem with large trees. This is why we need additional information (such as the predicted structures) to resolve such ambiguity. A tree of representative fungal, microsporidian and metazoan sequences is available as supplementary material S3 in wikispaces. We had previously removed sequences picked up by BLAST but that are not predicted to fold properly, or have catalytic and C terminal defects that did not give us sufficient confidence to include them. In this refined tree (345 sequences – with only representatives of *R. irregularis* shown for clarity) we found that *M. daphniae* grouped within early branching fungi specifically with the chytrid *B.dendrobatidis* (UniProtKB:F4PD83) different from the tree comprising only of early branching fungi.

In the original unrooted tree that included fungi and microsporidia the *R. allomyces* sequences (UniProtKB:A0A074aN09 and A0A074aVN0) grouped with microsporidia but in the fungamal tree *R. allomyces* grouped with the metazoan AGO-like argonautes. This is not completely unexpected given their 99% structural similarity to HsAGO2 and it indicates that these argonautes are very different from other early branching fungi.

Of the eleven sequences that registered a score of greater than 0.8 residue changes per site, three are sequences from *C. elegans* and these have already been documented to have gained slightly different specificity in the RNAi pathway (Yigit *et al.*, 2006). Two other metazoan sequences are *Drosophila melanogaster* AGO2 (UniProtKB:Q9VUQ5) and the silk moth (*Bombyx mori*) AGO2 (UniProtKB:Q59HV7), both of which are functional argonautes. Six of the sequences are fungi; two sequences from *B. fuckeliana* (UniProtKB:G2XU62 and G2XTN8) that, as we have previously mentioned, can hijack host argonautes, which may provide relief

from conservation of its own argonautes. Both sequences have inserts (dissimilar) that are predicted to fold away from the core and should not impede function regardless of a putative lack of constraint. Leaving *B. dendrobatidis* (UniProtKB:F4PD83) which grouped with *M. daphniae* and has a large N terminal insert which is predicted to be away from interfering with the core of the argonaute, *Schizophyllum commune* (UniProtKB:D8PVS4) which also has a number of small inserts, *R. allomyces* (UniProtKB:A0A074aVN0) which we have already mentioned groups with metazoa and *Pyrenophora tritici-repentis* (UniProtKB:B2WD35) which has the greatest number of changes in residues per site amongst the fungi at 1.3 but still fulfills all our criteria for inclusion.

## 4b.5. Discussion

Our method combines 3-D structural analysis coupled with analysis of phylogenetic relationships by phylogenetic tree calculations. This furthers our understanding of protein family evolutionary histories as well as clarifying issues regarding gene annotation. Where annotation is available the sequences on all trees show the annotation given by UniProtKB (figures 4b.5, S1 and S3). Those that simply have the accession number are annotated as ‘hypothetical’ or ‘uncharacterised’ protein (although we can anticipate changes with each revision), this is the case with many of the fungal proteins not just those from *R. irregularis*. Annotation will always lag behind sequencing because sequencing is increasing so rapidly. One of the *R. irregularis* AGO sequences (UniProtKB:U9U3P8) is annotated as PIWI-like but we have shown that it groups with the other fungi on our trees and also retains the C-terminal AGO-like signature. This is significant because the PIWI-like subfamily of argonautes are absent from fungi, so mis-annotation like this example can cause considerable confusion. Additionally we found inconsistency between the databases, e.g., UniProtKB:A0A0C9MH75 from *Mucor ambiguous* is annotated as PIWI-Like and the identical sequence in NCBI (GAN01273.1) is annotated as ‘translation initiation factor’, neither of which are correct since this too shows all the signatures of an AGO-like protein.

Regarding the *R. irregularis* sequences that lacked homology as argonautes in our analysis we must remember that functionality can evolve as the sequence evolves. We fully expect duplication and rapid evolution of the non-homologous sequence as well. The nematode expansion in *C. elegans* has resulted in diversity both of primary sequence and of function (Yigit *et al.*, 2006). Proteins that have duplicated but lost the

characteristics of functional argonautes now have the opportunity to evolve novel function(s) (Ortiz-Rivas *et al.*, 2012). Increased experimental analysis of the argonaute family reveals further diversity in functional roles in different species. It is likely that sequences that structurally remain AGO-like have related activities, and those that have diverged from the archetypal AGO-like structure may be undergoing either some gene loss or could have adapted novel functions (at least as argonautes).

Our conclusion is that the *R. irregularis* expansion appears to have arisen from two original *R. irregularis* argonaute proteins and although many of the *R. irregularis* proteins are structurally capable of function, some appear to be diverging from such capability. Whether this is from a gain of function or simply falling into disrepair we cannot say.

In terms of our trees, the tree of the fungamal species that went into reconstructing the ancestral sequences of metazoa, fungi and microsporidia (and including some orphans), we found that the posterior probability was frequently low and that the placement of groups could change significantly. With a greater number of sequences in our unikont tree we found some unlikely placements. This problem simply grows as we work through the remainder of the eukaryotic domain and so calculating ancestral sequences that represent particular groups is a way of simplifying what would otherwise be visually complex and potentially less reliable. We argue therefore that ancestral sequence reconstruction and then MSA and tree calculation from those alignments is more reliable than calculating trees with sequences from a very large numbers of species.

In summary, this study illustrates the utility of incorporating three dimensional structural analysis together with further structural and biochemical evidence, in this case retention of the AGO catalytic residues, and retention of the C terminal PIWI or AGO signature, in order to support BLAST results. It is no longer sufficient at deepest divergences to simply use BLAST results and sequence alignment as evidence for homology. We have found that even with a few highly homologous sequences, the available sequence alignment algorithms (MUSCLE, CLUSTALW, MAFFT, and Geneious) do not always concur with structural alignment to a determined structure. We have found that some putative AGO sequences retrieved by BLAST showed some structural resemblance to AGO but yet we were not convinced that all of the sequences were functional.

The algorithms used for structural prediction are reliant on experimental structure determination. Once there is a solved structure of a protein or a domain then more and more sequences become amenable to structural prediction. This could be used to aid annotation at far lower cost and requiring less specialised expertise than experimental protein structure and/or functional characterisation. This also informs on whether or not an apparent homolog could be a functional protein, allowing some predictive analysis before experimental characterisation. This applies generally to genome annotation and characterisation, and not only to our investigation of the AGO and PIWI protein families. We suggest that studies, such as this, will become essential.

## Acknowledgments

The author(s) wish to acknowledge the contribution of NeSI high-performance computing facilities to the results of this research. NZ's national facilities are provided by the NZ eScience Infrastructure and funded jointly by NeSI's collaborator institutions and through the Ministry of Business, Innovation & Employment's Research Infrastructure programme. URL <https://www.nesi.org.nz>. We also thank I-TASSER for the provision of free computer hours for structural prediction via their University of Michigan server <http://zhanglab.ccmb.med.umich.edu/I-TASSER/> and we thank Massey University for financial support for this project.

## Supporting information

**S1 Fig.** Unrooted tree showing groups for ancestral reconstruction A2 size. Not provided in appendix due to size. This is available in wikispaces please email [tonidaly@mac.com](mailto:tonidaly@mac.com) for access.

**S2 Table.** The 3-D structures for *R. irregularis* and additional information.

**S3 Fig.** All contributing fungi species as a rooted tree A2 size. Not provided in appendix due to size. This is available in wikispaces please email [tonidaly@mac.com](mailto:tonidaly@mac.com) for access.

Posterior probability logos of all of the ancestors are available in wikispaces please contact [tonidaly@mac.com](mailto:tonidaly@mac.com) for access.

**Supplementary material S2 can be found in appendix II.**

## Chapter 4c: Argonautes in eukaryotes

### 4c.1. Abstract

Protein annotation is frequently based on primary sequence homology, however there is a rapidly increasing list of uncharacterised proteins. We speculate that tertiary structure prediction will become essential to identify homologs where primary sequence homology barely exists. In order to move forward from the limitations of primary sequence homology and extract more information than is currently the norm we show how to calculate the tertiary structure of argonaute proteins (both AGO-like and PIWI-like) and include other supporting information in order to identify structural homologs. Once identified, we can recreate ancestral sequences *in silico*. We build on previous work and further refine our bioinformatics pipeline making this a process that can be accomplished with minimal resources. We use the reconstructions to assist with our search for argonaute family proteins where primary sequence similarity is so low that they would not be found by BLASTing alone.

Argonaute-like proteins can be found in all domains of life and the argonaute gene has been duplicated independently many times with large expansions in some species and yet loss in a few. Here we report only within the domain Eukaryota, mainly in plants and single celled eukaryotes including the major protozoan parasites. We find evidence of argonaute proteins, even where typically associated small RNAs haven't been identified. We also demonstrate that the PIWI protein, often proposed to be germline-specific in animals, was most likely an original argonaute, possibly found together with another more AGO-like protein in the Last Eukaryote Common Ancestor (LECA). Finally we amalgamate the putative ancestral sequences with those that we have previously inferred in fungi and metazoa to calculate a simple tree representing ~700 argonaute protein sequences.

### 4c.2. Introduction

It has been proposed that LECA had sophisticated cellular machinery and had to contend with parasitic nucleic acid in the form of plasmids, viruses and transposons that would have been ubiquitous then as they are now. Argonautes are ~850 residue proteins guided by small RNAs, that function in the nucleus or cytoplasm and can control the



expression of many endogenous genes. They can mediate gene expression by cleavage of mRNA, or steric impediment of translation, or by chromatin modification guided by endogenous small RNAs. They can also control parasitic exogenous RNA such as viruses (guided by small RNAs copied from the invasive nucleic acid using RNA dependant polymerases).

There are two main classes of argonaute proteins, AGO-like and PIWI-like (Daly *et al.*, 2011). They also control expression of endogenous parasitic DNA such as transposons and repeat sequences via PIWI-like argonautes guided by small RNAs known as PIWI interacting RNA (piRNA) generated via a mechanism known as ‘ping-pong’ which requires a pair of PIWI-like argonautes. piRNA generated this way is known as secondary piRNA. Primary piRNAs generated directly from endogenous nucleic acid can control gene expression by chromatin modification or DNA methylation suppressing or upregulating expression of gene expression outside of the germ line cells that PIWI proteins were thought to be restricted to (Ross *et al.*, 2014). Small non-coding RNAs guide both sequence specific activities.

Defence would be an important issue for the early cell, and PIWI are often suggested to be more recent due to their role in the animal germline, however our results show that PIWI are much older than the development of specific germline cells. This is supported by an in depth review of the PIWI protein (Juliano *et al.*, 2011) that links the PIWI protein with the capacity of stem cells to generate copies of themselves which could have been essential long before a specialised germline. This raises the issue of the apparent loss of PIWI in some kingdoms.

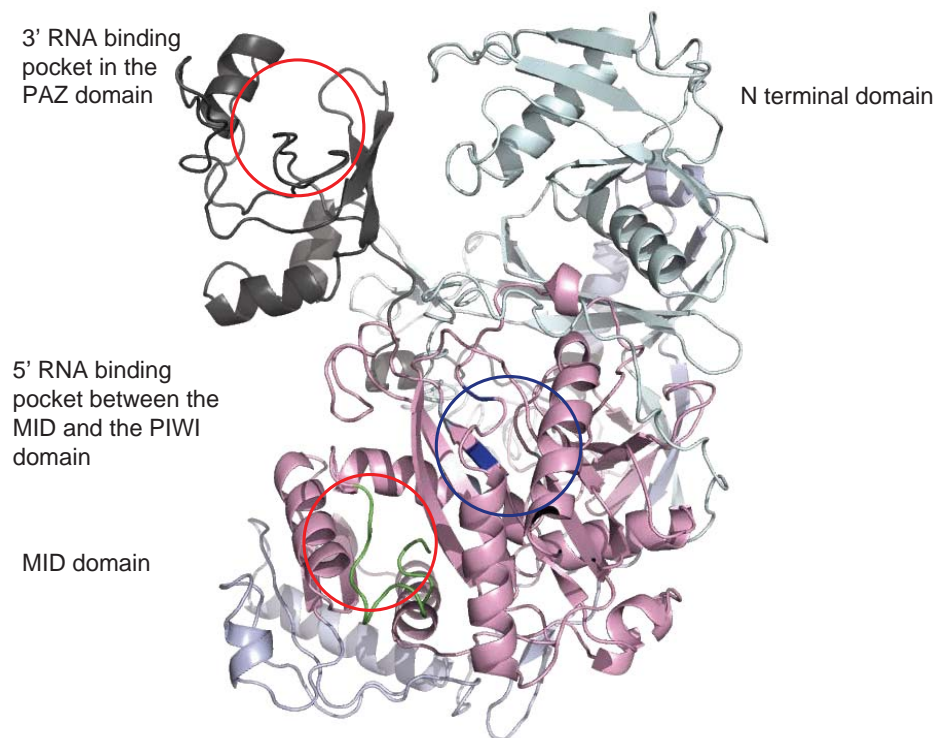
The eukaryote domain has a chequered history of loss of some (or occasionally all) of the proteins required for the RNA interference pathways utilising argonautes (both PIWI-like and AGO-like) and yet their fundamental tasks would appear to be extremely beneficial. In possibly the oldest eukaryote group, the Excavates, the parasitic trypanosomids have at least two different types of protein. These are PIWI-like and AGO-like, the only other kingdom known to retain both types of argonaute are the metazoa. Although it has been argued that the trypanosomid argonautes are monophyletic (Garcia Silva *et al.*, 2010b), we argue in favour of more than one original argonaute with loss of the AGO-like protein in some trypanosome species. There is loss of either PIWI or AGO in different groups, and evidence of loss of function of all argonautes within some yeast species (Drinnenberg *et al.*, 2011) (and also appear to be

absent in choanoflagellates by our own observation), yet massive gain of argonaute sequence in others (Vastenhouw *et al.*, 2003; Tisserant *et al.*, 2013).

The picture is complicated by a variety of pathways for the genesis of small RNAs that enter and guide the argonautes, as well as the flexibility of the Dicer proteins that are usually (but not always) responsible for processing dsRNA to the necessary length. Small RNAs are generally accepted as RNAs less than 200 nt in length, however the RNAs that guide argonautes are typically 18-30 nt. In addition to well described small non-coding RNAs such as microRNAs (miRNA) and PIWI-interacting RNA (piRNA), there are many small RNAs that are derived from tRNAs (Lee *et al.*, 2009; Franzén *et al.*, 2011), rRNAs (Wei *et al.*, 2013), and others (Kawaji *et al.*, 2008) that tend to be excluded from RNA transcriptome screens but are biologically active.

This means that not all the small RNAs that enter the argonaute have the same processing signatures and not all species harbour all processing proteins. It would seem that the minimal requirement would be an argonaute protein and a means of genesis of small RNAs. In the yeast *Saccharomyces cerevisiae* (which is RNAi deficient), the introduction of an argonaute, transactivating response RNA-binding protein (TRBP) and a Dicer was sufficient to enliven the pathway (Suk *et al.*, 2011). However there are many pathways that avoid Dicer processing (Hansen *et al.*, 2016) so it may be that an argonaute protein is the only requirement in some species.

Both AGO-like and PIWI-like argonautes comprise of the same four domains, N, PAZ, MID and PIWI. The PIWI domain comprises of an RNaseH like fold which contains a catalytic tetrad of residues Asp-Glu-Asp-His (DEDH) (Nakanishi *et al.*, 2012). The inclusion of Glu at the catalytic site is more recent, and it has been reported that Asp or Lys could replace the His (Tolia and Joshua-Tor, 2007). These catalytic residues are surprisingly well conserved because, although essential for cleavage, they are not necessary for steric impedance of translation - and the majority of argonautes lack cleavage capability (Chak and Okamura, 2014) (see fig. 4c.1).



**Fig. 4c.1 I-TASSER structural prediction for *Trypanosoma brucei*.**

UniProtKB:Q6T6K0 – this falls on the AGO-like side of our trees. The cartoon shows the domains present in most argonautes (AGO and PIWI) and are coloured as follows; N terminal and MID domains (pale grey and lilac), PAZ domain (dark grey) and PIWI domain (pink). RNA binding pockets are circled in red and the catalytic region is circled in blue with catalytic residues marked in dark blue. Generally slicing competent argonautes have catalytic residues (Asp-Glu-Asp-His DEDH sequentially in the primary sequence) but *T. brucei* has Asp-Glu-Arg-Ala (DERA). The positions of canonical residues (DE) are shown in blue. *T. brucei* (UniProtKB:Q6T6J9 – a PIWI-like argonaute) has the canonical DEDH catalytic tetrad (shown in fig. 4c.4).

A major difference between AGO-L and PIWI-L argonautes is the size and genesis of the small RNAs that guide them. However, we have found that AGO and PIWI proteins harbour different C terminal ‘signatures’. In almost all cases PIWI-L argonautes have a C terminal signature of Leu-two aromatic residues-Leu as the C terminal residues whereas AGO-L argonautes retain Met at the 4<sup>th</sup> from last residue, then two aromatic residues and invariably Ala, Ile, or Val as the final residue. This has been used to successfully identify AGO-L and PIWI-L sequences with very limited primary homology to our query sequences.

There are thousands of argonaute protein sequences in the UniProtKB and NCBI databases. If we use traditional criteria for estimating primary homology in the BLAST search we would miss proteins with structural homology but limited primary sequence homology. We have found that large deep trees can be easily swayed by the outgroup chosen to root them, and also by including ambiguous or non-homologous sequences. For these reasons we need to be able to use a few sequences that represent many more

in order to simplify the trees and also to use as query sequences to find ever more remote homologs.

### 4c.3. Method

Due to the large and increasing number of sequences deposited in NCBI and UniProt it is not always possible to retrieve sequences within 1,000 hits that are at the extremes of primary homology. In order to find really remote sequences our query sequences need to get progressively more remote from well-characterised sequences without losing the structural and catalytic integrity of an argonaute protein (either AGO or PIWI-like). Therefore we become less stringent than most researchers in our approach to BLAST searches over time because we are purposefully looking for sequences with limited homology.

In order to eliminate ambiguous sequences we use three-dimensional prediction of tertiary structure. Once we are able to view the predicted structure in three dimensions we can find additional evidence, for instance the retention of correctly positioned catalytic residues and the C terminal signature before finally including a sequence as a relative. We take the sequences that we confidently report as being part of the same group or ‘clade’ for ancestral sequence reconstruction (ASR). This results in sequences that are a good representation of a number of sequences. By reducing the number of sequences in the multiple sequence alignment (MSA) that informs the final tree we can reduce the ‘noise’. To do this successfully we must be certain that the sequences included in the MSA used to recreate the ancestor are *bona fide* argonautes and that we have placed them in the correct group (AGO-like or PIWI-like).

In this section of our work we have examined almost all the *bona fide* argonaute sequences that we have found. In some cases we have left out strains of species where sequences are highly similar to one another. This is because the ancestor should represent a variety of species rather than be swayed by an over representation of similar sequences. Sometimes it is prudent to take a representative sample where there are so many sequences that the task would be unmanageable (which we had to do with the plants, and with metazoa and fungi in an earlier investigation). For example we used 12 representative plant species and the three plant ancestors are calculated from 197 sequences. Where there are few sequenced species, we try to include as many species as we can find. For instance most of the ciliate ancestors are each calculated from ~20 sequences but the majority of the sequences come from just five species.

Our pipeline starts by taking a well-described argonaute protein sequence and using it as a query to BLAST the UniProt and NCBI databases using BLOSUM 45 and setting the E value at up to 1,000. Where we can, we restrain our searches to a database narrower than ‘eukaryotes’, but this is not always possible. Our pipeline was designed to be flexible, we can add different modules as different web servers become available, or swap modules out for different types of protein. In this case we made greater use of the Pfam database (Finn *et al.*, 2010) by selecting the domain PIWI (common to AGO and PIWI like argonautes) which results in a graphical outline of the recognisable domains from UniProt and NCBI sequences. We have found some sequences that are not included in Pfam as well as some that we reject where the sequence is not modelled well enough to fulfil our criteria for inclusion, though we have erred on the side of caution by not including these sequences where we have low confidence.

From the identified sequences we made a number of multiple sequence alignments (MSA) using different alignment tools (MUSCLE, CLUSTALW, MAFFT, and Geneious) to try to get the best alignment that we could and learn more about the sequences. Geneious global alignment with free ends (Geneious Pro 8.0.4 <http://www.geneious.com/>) either better aligns the more degenerate N terminal, or won’t align the sequences at all. In fact this non-alignment is a useful early alert that a sequence found by BLAST needs to be flagged for 3-D structural prediction. Ultimately we always end up with a MUSCLE (Edgar, 2004b) alignment because although it doesn’t necessarily reflect the best alignment, particularly in terms of the N terminal, it at least forces alignment between disparate sequences where primary homology is more tenuous. This is consistent with our argument that tertiary structure is retained when primary homology is virtually lost between distantly related proteins.

The initial trees created by using MrBayes (Huelsenbeck and Ronquist, 2001) are rooted by the remote outgroup *Pyrococcus furiosus* (UniProtKB:Q8U3-D2 PDB1Z25) (an archaeon) because we have found that outgroups that are closer affect the outcome of the tree. At this stage the large and unwieldy tree is enough to demonstrate the split between AGO-like and PIWI-like proteins and within that split to demonstrate loose ‘clades’ that we can later refine to create putative ancestral sequences.

Sequences that stand out from these ‘pre-trees’ in terms of a very low posterior probability at the node or greater than 0.8 changes of residue per site (although it is consistent with our argument that this number be high) plus those that Geneious cannot

align are, as a priority, analysed with the 3-dimensional structural prediction suite of algorithms that make up I-TASSER (iterative threading assembly refinement server) (Zhang, 2008; Yang *et al.*, 2015).

Computing time for these predictions run at ~50 hours per sequence and only one can be submitted at a time. There is nothing to be gained from submitting sequences that are almost certain to fold correctly though we have additionally used LOMETS (Wu and Zhang, 2007) (part of the I-TASSER suite, again one sequence at a time) and Phyre2 (Kelley *et al.*, 2015) which is faster (but has the limitation of only reporting the closest structural matches rather than a prediction of the structure of the submitted sequence).

Only when we can be sure that the outliers will be predicted to fold with recognisable argonaute domains complete with a reasonable catalytic tetrad and C terminal signature will we then put the sequence into the appropriate ancestral clade. Sequences with large inserts often have very poor confidence scores (C score) as reported by I-TASSER. We have learnt that this does not necessarily preclude function as many well-described functional argonautes have low C scores due to inserts that are modelled away from the core of the protein. We can check this using FATCAT (Flexible structure AlignmenT by Chaining Aligned fragment pairs allowing Twists), (Ye and Godzik, 2003) which aligns our predicted sequence with a well described crystal structure (such as the human argonaute AGO2 PDB:4OLA) and gives the number of residues in equivalent positions. This gives a visual guide to where the inserted sequence lies with respect to the catalytic centre, or indeed where structure that is likely to be essential for function is absent.

We cannot rely on the primary homology alone as the alignments are unreliable over a great number of sequences. Even with smaller MSA problems can occur that can compromise the integrity of the ancestor. Sequences included in the MSA that are not AGO/PIWI need to be eliminated because they introduce noise into the reconstruction. In some cases the alignment does not reflect the sequence that appears at the equivalent place in three dimensions for all of the aligned sequences.

Once sequences are placed into an ancestral clade a MSA can be made without inclusion of an outgroup, this gives us good information for which parts of the sequence to remove. For example, it is more parsimonious to remove an insert present in less than 10% of the sequences that make up the MSA than to assume that the insert has been lost



on the remaining 90% of sequences; we can also use Wasabi (Veidenberg *et al.*, 2015) in an automated way. In practice it makes no difference to the MSA or resultant tree other than residue changes per site or posterior probability at the nodes. The reason for outlier insert removal from the MSA is to prevent FastML (Ashkenazy *et al.*, 2012) from creating unreasonably long ancestors where it includes the inserts in the calculation. We have used FastML previously having compared a number of ASR algorithms and are satisfied that as long as we remove the excess inserted sequence and submit our own trees then a satisfactory ancestor will result for each node of the tree (Daly *et al.*, 2013b).

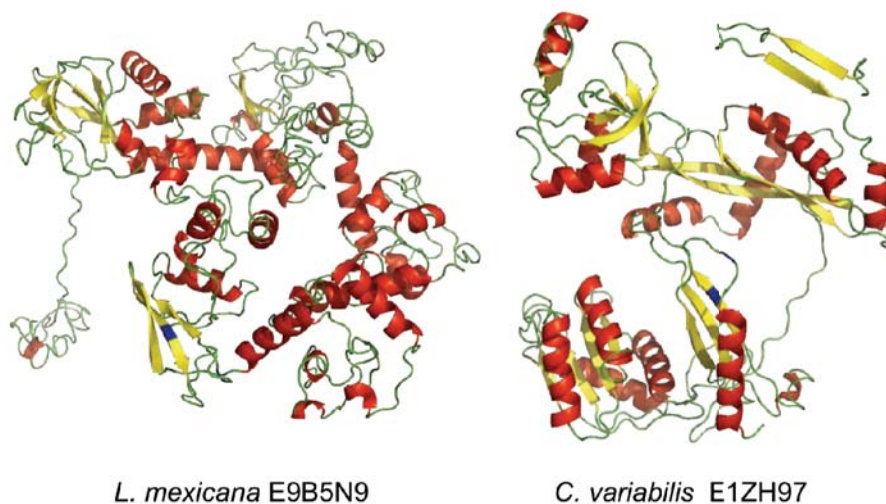
All ancestors from the earliest node in each tree are submitted to I-TASSER for structural prediction and are also used as a BLAST query with the aim of finding more remote sequences. Due to the rapidly increasing number of sequences stored in the databases we usually do find new sequences. Some of these would undoubtedly be found without the use of the ancestral sequence, but some are simply too remote from a well-described sequence to fall within the 1,000 maximum ‘hits’ where the database of argonaute proteins is so large.

#### **4c.4. Results**

Our pre-trees of all eukaryotes (in excess of 500 sequences) split into two halves; closest to the root on one side were PIWI-like sequences, with AGO-like sequences on the other. This initial tree had plants curiously placed on a branch of their own between the metazoan AGO-L and metazoan PIWI-L sequences. We separated out metazoa and fungi and have reported on them previously. The remaining sequences were used to make a smaller tree. Here we found all kinetoplasts, ciliates and amoebozoa on one side, and plants (plus red and green algae) and stramenopiles on the other. We then eliminated sequences that we could not confidently support using 3-D structural prediction as members of either AGO or PIWI families thus refining the trees.

Examples of sequences found by BLAST but highlighted for I-TASSER submission because of an improbably high number of changes per residue are the trypanosomid *Leishmania mexicana* (UniProtKB:E9B5N9) with 2.5 residue changes per site, and the green algae *Chlorella variabilis* (UniProtKB:E1ZH97) with 1.3 residue changes per site. These did not fulfil our criteria (fig. 4c.2) and were rejected. The ‘expect’ or E value is, in many cases, higher than traditionally acceptable (generally  $1 \times 10^{-4}$ ), but because we know that structural homology is retained long after primary

homology has been lost, we will investigate sequences where the E value is much higher depending on the size of the database. Where there are so many sequences available it is more likely that we could miss sequences with structural similarity and limited primary homology. Due to limited primary homology we need more than one line of evidence to support the inclusion of a sequence within an argonaute family. There have been much more difficult decisions to make and so in some cases borderline sequences have been left out, e.g., *Stylonychia lemnae* annotated as PIWI-L (UniProtKB:A0A078AWB2), but modelled by I-TASSER as lacking  $\beta$  sheet in the PIWI domain in three of five models. I-TASSER will result up to five models that may vary in minor details (as in the case of *S. lemnae*), or where the sequence is truly ambiguous the resultant folds can be very different and the scores are generally all poor.



**Fig. 4c.2 I-TASSER 3-D structural prediction of BLAST results with high number of residue changes per site.**

*L. mexicana* (UniProtKB:E9B5N9) and *C. variabilis* (UniProtKB:E1ZH97) both retain some catalytic residues Asp-Glu (marked in blue) in both cases. There is some structural homology in the case of *C. variabilis* but minimal for *L. mexicana*. Neither would be included in the MSA to calculate an ancestral sequence.

In other examples of a high number of residue changes per site we find an acceptable argonaute with a divergent function (Twi12 described later) and this may account for other higher than expected branch numbers.

In other examples of a high number of residue changes per site we find an acceptable argonaute with a divergent function (Twi12 described later) and this may account for other higher than expected changes per site. So from large trees we could allot subgroups of sequences, independently of species, into ‘clades’. Although the sequences are more similar across species than within, they generally fell within their

expected phyla (with the exception of the ‘barley’ sequence that groups with metazoa – rotifers in particular), we suspect contamination here (Daly *et al.*, 2016).

#### **4c.4.1. SAR (Stramenopile, Alveolate and Rhizaria)**

The term SAR covers a supergroup initially comprising of what was known as the chromalveolates proposed to be the result of symbiosis between a red alga (containing a plastid although many of this group have lost photosynthetic capability (Keeling, 2009)) and a protist with two flagella (bikont) (Keeling, 2004). Stramenopiles vary from single-cell glassy diatoms to the large multicellular ‘kelp’. We initially found multiple sequences of AGO type proteins within the oomycetes (water mould/mildews) mainly *Phytophthora* many of which are plant pathogens which is why so many more of these have been sequenced.

Alveolates include dinoflagellates that can result in toxic red tides, apicomplexans (which includes the *Plasmodium* parasite that causes malaria) and the ciliates, such as *Paramecium*. Rhizaria have been more recently added to the supergroup (Burki *et al.*, 2010; Parfrey *et al.*, 2010). They are difficult to grow and underrepresented in sequence databases (Sierra *et al.*, 2013). There is already an interesting pattern of loss and gain of argonautes within the SAR group, with some groups yet to have any species sequenced. Each group is discussed individually in the following pages.

##### **4c.4.1.1. Stramenopiles**

Many of the oomycetes sequences in the databases are from different strains of the same species so we have been careful to include just one ‘set’ of argonaute proteins from each species. As more species have been sequenced the group of argonautes has expanded to include the microalgae *Nannochloropsis gaditana*, and the brown algae *Ectocarpus siliculosus*, two species of diatom *Thalassiosira oceanica* and *Thalassiosira pseudonana*, plus the animal gut parasite *Blastocystis hominis*. None of these appear to have the multiple duplications that have occurred in the oomycetes. All sequences found retain the Met fourth residue from the C terminal that is characteristic of AGO-like proteins. It could be argued that we should have taken the oomycetes as a group and left the remaining algae and diatom sequences as ‘orphans’, or sequences that could not be included in an ancestor, but in our original large trees the diatoms and algae grouped within the oomycetes rather than on the fringes. An exception to this was the coccolith

*Emiliania huxleyi* which is a single cell photosynthetic phytoplankton – phylum haptophyta. This could not be assigned to any ancestor and was included as an orphan.

Ultimately we recreated three ancestors from the 38 sequences retrieved; two reconstructions (stramenopiles group 1 and group 3) have a marked preference for Arg-Gly-Gly-Gly (RGGG) N terminal repeats similar to those that we also noted in the ancestral reconstruction of the metazoan PIWI proteins. Conserved Arg-Gly/Arg-Ala (RG/RA) motifs are known to interact with the Tudor domain of metazoan PIWI1 (Vagin *et al.*, 2009; Liu *et al.*, 2010), but in the stramenopile AGO-like ancestors Ala residues were conspicuous by their absence in the N terminal. The ancestor for stramenopiles group 2 lacks the RGGG repeats, in this respect they are more similar to the metazoan AGO-like argonautes. Each ancestor retains domain structure characteristic of the argonaute family, the catalytic tetrad DEDH, and each retains a variant of the AGO-like C terminal signature (group 1 MYFV, group 2 MYYV and group 3 MFFI).

#### **4c 4.1.2. Alveolates**

Once again most of the original sequences that we found came from just a few ciliates; *Oxytricha trifallax*, *Paramecium tetraurelia*, *Stentor coeruleus*, *Stylonychia lemnae* and *Tetrahymena thermophila*. Ciliates have a very large macronucleus, which has DNA for cellular housekeeping requirements (somatic DNA), and a micronucleus containing germ line DNA (Gao *et al.*, 2015). All of the ciliate argonautes were on the PIWI side of our tree in two distinct groups. The species mentioned above have ~12 PIWI-like proteins in each species. The fish white spot *Ichthyophthirius multifiliis* is an exception, it has only two full-length sequences that grouped in different clades. We also found one full-length sequence in *Pseudourostyla cristata* and *Paramecium caudatum* all are PIWI-like. This implies that ciliates have lost the AGO-like form of the protein and confirms previous work (Cerutti and Casas-Mollano, 2006) and also implies that the duplications may either not be necessary and have been lost or were not present in the common ancestor of all ciliates. Although it seems unlikely that one species of *Paramecium* would have made extensive use of a great number of argonautes and another have only one although only one was described in the original paper (Obara *et al.*, 2000).

We found two photosynthetic species of apicomplexa; *Chromera velia* and *Vitrella brassicaformis*. *C. velia* contains five argonaute sequences but they were too

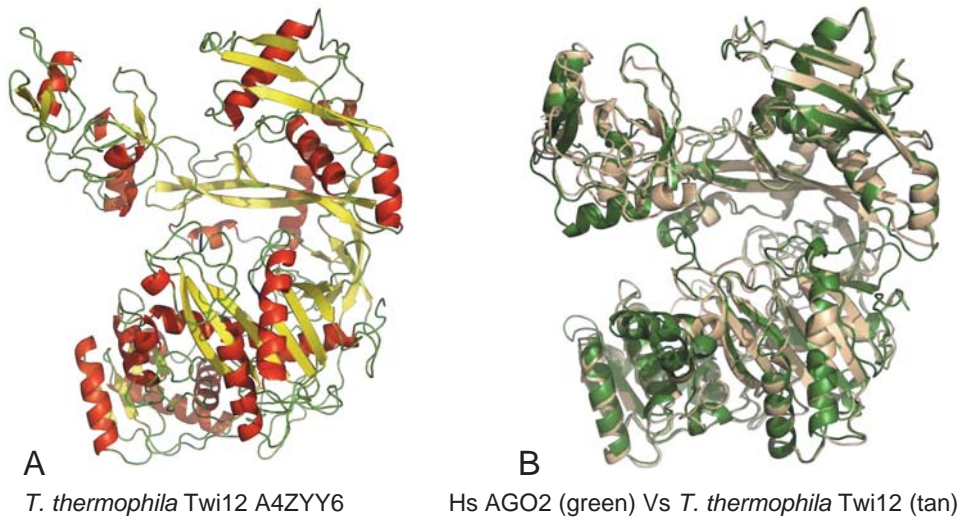
similar and too few to recreate an ancestor and so UniProtKB:A0A0G4FCF4 and A0A0G4HQX1 have been included in the final tree as orphans, the former retains a Gly rich N terminal insert which the latter lacks and both retain a PIWI-like C terminal signature. *V. brassicaformis* has just one argonaute sequence with ~500 residue insert at the N terminal. This ~500 residue insert is modelled by I-TASSER to be well away from the core of the argonaute protein, which otherwise is predicted to fold convincingly. We cannot be sure if the insert is an artefact of sequencing or a genuine insert and so the number of residue changes per site may appear larger than is warranted. *V. brassicaformis* (UniProtKB:A0A0G4ETG9) is also included as an orphan, this too retains a PIWI-like C terminal signature. All of the apicomplexan sequences grouped outside of the alveolates and so could not be allocated an ancestor with confidence. But they all appear PIWI-like as well.

The PIWI protein plays an essential part in the life cycle of ciliates. When the micronucleus undergoes genetic rearrangement prior to conjugation a huge amplification of genetic material occurs (which results in a new macronucleus) and the micronuclear DNA, (now interspersed amongst the new macronuclear DNA), needs to be deleted. To do this different species have utilised piRNA known as small-scan RNA (scnRNA) in opposite ways for the same purposes. scnRNA has the 2' O-methylation at the 3' end typical of piRNA and in *Tetrahymena* these are generated from the germ line genome and target germ line DNA in the somatic genome for elimination (Feng and Guang, 2013). In *Oxytrichia* the scnRNAs come from the somatic DNA and target somatic genes for retention. This demonstrates that piRNA/PIWI has the capability to mark genomic DNA either for destruction or for retention, as well as for switching on and off transcription (Feng and Guang, 2013).

*Tetrahymena thermophila* has at least 9 PIWI-like proteins but only one is known to be essential for growth (Twi12). On our tree of ciliates Twi12 had a branch label score indicating 1.6 residue changes per site. This automatically flagged the sequence for structural prediction because this implies that some amino acids have changed several times. We know that it is divergent and that Twi12 doesn't have a canonical catalytic site - (Ser-Glu-Asp-Tyr rather than Asp-Glu-Asp-His) and is not capable of slicing (Couvillion *et al.*, 2012). The majority of eukaryotic argonautes are not capable of target slicing even though most of them retain the catalytic tetrad, so functions other than slicing are common. Twi12 has 15% sequence identity with the



human AGO2 (which does have slicing capability) and 17% with human PIWI2. The predicted structural alignment with Hs AGO2 shows 87% structural homology (see fig 4c.3).



**Fig. 4c.3 The divergent argonaute Twi12 from *T. thermophila*.**

**A.** *T. thermophila* (UniProtKB:A4ZYY6) as modelled by I-TASSER. **B.** Although Twi12 shows 1.6 residue changes per site in the tree of ciliates and has 15% primary homology with the human argonaute AGO2 protein sequence it aligns structurally (FATCAT) with 87% of its residues in equivalent geometric position as the solved structure of the human argonaute (PDB:4OLA).

Twi12 loads only small RNAs processed from the 3' end of tRNA (tsRNA also known as tRFs), but binding of the tsRNA is essential for nuclear localisation of Twi12 bound to XRN2 (an exonuclease involved in processing 5.8S RNA). XRN2 does not function unless bound to Twi12, so tsRNA has an essential function in pre-rRNA processing (Couvillion *et al.*, 2012). It is not only ciliate PIWI-like argonautes that bind tsRNAs, in *Dicer*<sup>-/-</sup> yeast (*Schizosaccharomyces pombe*) the single argonaute (AGO-like), will load processed tRNAs (Kawaji *et al.*, 2008; Halic and Moazed, 2010), and humans also load tRNA fragments preferentially into AGO3 and AGO4 (slicing deficient) (Lee *et al.*, 2009). One suggestion for loading processed tRNAs is presumed to be simply as a block on AGO3 and 4 thus keeping the catalytic argonaute (AGO2) free for cleavage duties (Lee *et al.*, 2009). Given the importance of tsRNAs to *T. thermophila* this seems a rather simplistic explanation. Very recently the role of processed tRNAs is becoming clearer in animals (Kumar *et al.*, 2014; Telonis *et al.*, 2015; Venkatesh *et al.*, 2016), plants (Loss-Morais *et al.*, 2013) and in oomycetes (Åsman *et al.*, 2014).



We made four ancestral reconstructions from the protein ciliate sequences; ‘Ciliates group 1a and 1b’ and ‘Ciliates group 2a and 2b’ because they appear to have arisen from duplications from two original genes. These ancestors proved more fruitful than any other in terms of retrieving barely homologous sequences particularly those from archaea.

The ‘group 1a’ sequences ancestor C terminal signature is LYFL, group 1b shows LFFI, group 2a has LYYL, all typical PIWI-like signatures but group 2b has a C terminal signature of LHFL which is not something that we had seen before; the His residue is polar and relatively bulky though I-TASSER still models it tucked up inside and close to the 5' RNA binding pocket. All the predicted ancestral sequences retain a plausible catalytic tetrad in their predicted structures even though this is not necessary for the extant function of most argonautes. What is also interesting about group 2b is that the posterior probabilities for the creation of the ancestral sequence demonstrates very clearly that the number of residue changes per site doesn't reflect changes at all sites equally. A logo of posterior probability for the whole ancestor is shown in appendix III (supplementary material S1).

#### **4c.4.1.3. Rhizaria**

Due to the paucity of sequenced rhizaria we can only include the proteins available thus far as orphans where we are unable to assign them to an ancestor. What is interesting about the rhizaria is that using any of the sequences (two from *Plasmodiophora brassicae* UniProtKB:A0A0G4IM91 and A0A0G4IME3, two from *Spongospora subterranean* UniProtKB:A0A0H5QZP3 (fragment) and A0A0H5R3N1) in a BLAST search finds mostly vertebrate AGO-like sequences as closest hits (~40% sequence identity). Additionally the pairs look like recent duplications and the full-length sequences all have the AGO-like C terminal signature.

BLASTing only within ‘rhizaria’ in the NCBI database uncovered three sequences from *Reticulomyxa filosa* annotated as AUB, PIWI-like and ‘hypothetical’ (accession numbers have been changed to UniProt for consistency UniProtKB:X6P008 (fragment), X6M501 and X6M1Z6). This would cause some excitement if it were correct because, aside from excavates and metazoa, no other group has been found to harbour both AGO-like and PIWI-like argonautes. Although only one sequence is marked as a fragment, the two longer proteins are predicted to fold as argonautes but missing the C terminal. This comprises an essential part of the PIWI domain including

the last of the catalytic tetrad and the C terminal signature, so we cannot even be sure that these are functional. However the very ‘clean’ appearance of the structural predictions (*i.e.* lack of inserts) and conserved DED of the catalytic tetrad leads us to believe that the issues are sequencing rather than genetic. We eventually added these back in as orphans even though they are incomplete because of the placement of one of the plant ancestors more closely with rhizaria. The *R. filosa* fragments however grouped in the PIWI side of the tree, away from other rhizaria. It could be that this is because they are incomplete so this needs further investigation.

#### 4c.4.2. Amoebozoa

The Amoebozoa fall into two main groups; firstly the free-living slime molds (Mycetozoa) that have a multicellular capability when food is scarce. The second group includes the parasitic single-celled entamoeba and (in a tree comprising only of Amoebozoa), the single-cell soil amoeba *Acanthamoeba castellanii*. There are too few genes sequenced to be really helpful in terms of reconstruction of a putative Amoebozoan ancestor. *A. castellanii* PIWI (UniProtKB:L8HDC7) appears to have a very good tertiary structural prediction with a high C score in I-TASSER (0.82) but it could be compromised as it lacks an entire  $\beta$  sheet from the PIWI domain and consequently the first (sequentially) catalytic Asp is missing. It is described as a fragment and so this lack could be an artefact of sequencing.

These two groups (Mycetozoa and Entamoeba) are quite different from each other, except that they are all PIWI-like. The slime molds retain the PIWI-like C terminal signature (mostly LYFL, *A. castellanii* retains LYYL) but the Entamoeba sequences have a longer C terminal and the PIWI signature is not observed but then neither is the AGO-like signature, in fact the last residues are His-Pro-Pro-Tyr (HPPY). This extra sequence at the C terminal is predicted to be away from the argonaute core rather than tucked up inside close to the 5' RNA binding pocket as the canonical PIWI and AGO C terminal residues are.

Of the three available Entamoeba genomes; *Entamoeba nuttalli*, *E. histolytica* and *E. dispar* each have three PIWI-domain containing proteins that are virtually identical between the three species indicating a recent split. The differences between the sets of proteins suggest that the duplications arose in their common ancestor. For this reason we simply used one set as a representative of all of the sequenced Entamoeba, namely *E. dispar*. One of the sequences (UniProtKB:B0EI01) lacks a PAZ domain

more reminiscent of the short bacterial homologs (Makarova *et al.*, 2009) and so we have included UniProtKB:B0ET3Z and B0ENU6 as orphans. Both of these have divergent catalytic tetrads and lack any recognisable C terminal signature however PIWI-like argonautes demonstrating both chromatin modification and the RNA-induced transcriptional silencing have been described in *Entamoeba* (Zhang *et al.*, 2011).

The cellular slime moulds have a greater number of residue changes per sequence between the species than does the parasitic *Entamoeba* but still an insufficient number of species for ASR, which requires a minimum of four sequences. So we have chosen *Polysphondylium pallidum* as a representative cellular slime mold that harbours two distinct sequences, UniProtKB:D3B338 and D3BMP8. Both have extra residues at the N terminal (~500 and ~220 respectively), which I-TASSER predicts to lie away from the core of the argonaute. *E. dispar* and *P. pallidum* are included as orphans in the final trees. All are grouped on the PIWI side of the tree and so amoebozoans appear to have lost the AGO-like gene.

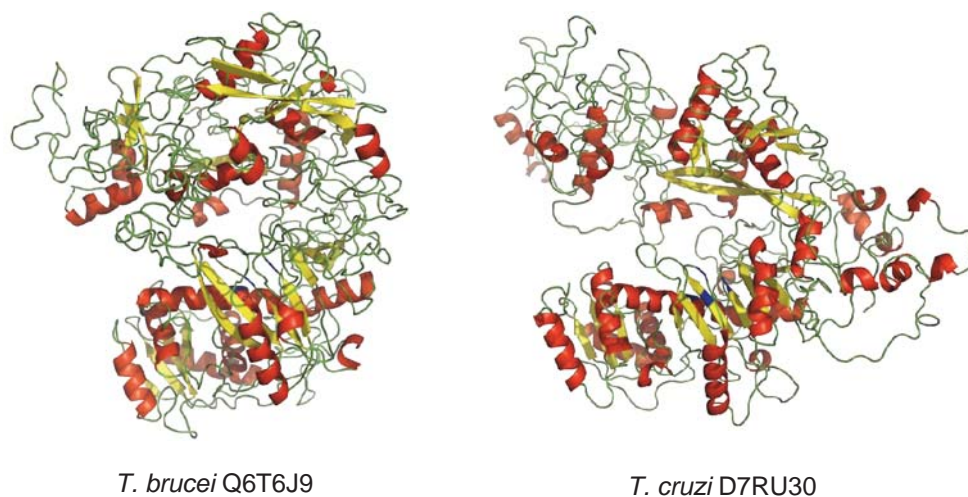
#### **4c.4.3. Excavates**

The excavates include some major parasitic agents but aside from the trypanosomids there are few sequenced organisms. Our ancestral calculations for this group are based mainly on *Lieshmania* and *Trypanosoma* although examples from *Heterolobosea* and *Giardia* are examined as orphans.

Some kinetoplasts, e.g., *T. brucei* and *T. congolense*, have clearly demonstrated RNAi capability (Ngô *et al.*, 1998) but some trypanosomes (e.g., *T. cruzi*) were said to lack AGOs (El-Sayed *et al.*, 2005) and are RNAi incompetent (DaRocha *et al.*, 2004). There remains a variety of *T. cruzi* that retains proteins that fall within the PIWI-like clade annotated ‘AGO-PIWI-Like’ (fig. 4c.4). These were named PIWI-tryp (Garcia Silva *et al.*, 2010b) so called because the PAZ domain is divergent and are said to lack the catalytic sites. RNA binding is not precluded by lack of the PAZ domain, the PAZ domain is protective of the 3' RNA but not essential (Hur *et al.*, 2013) and catalytic capability is not required for all of the functions of an argonaute.

We find that there are two clearly different types of argonautes within the kinetoplasts generally, but that *T. cruzi* retains only the PIWI-like argonaute that appears to come from a common ancestor of PIWI-like proteins within the kinetoplasts and they do have the catalytic residues in the anticipated 3-D structures. There are a

group of kinetoplast sequences that retain the C terminal signature of AGO-like proteins and these species have known RNAi capability though many have a divergent catalytic tetrad. This may be the reason why *T. cruzi* has been found to be RNAi incompetent (DaRocha *et al.*, 2004). Given that *T. cruzi* harbours tsRNAs even though it does not appear to generate canonical argonaute guide RNAs (Garcia Silva *et al.*, 2010a; Franzén *et al.*, 2011), and the fact that the protein is retained, implies a function as yet unknown.



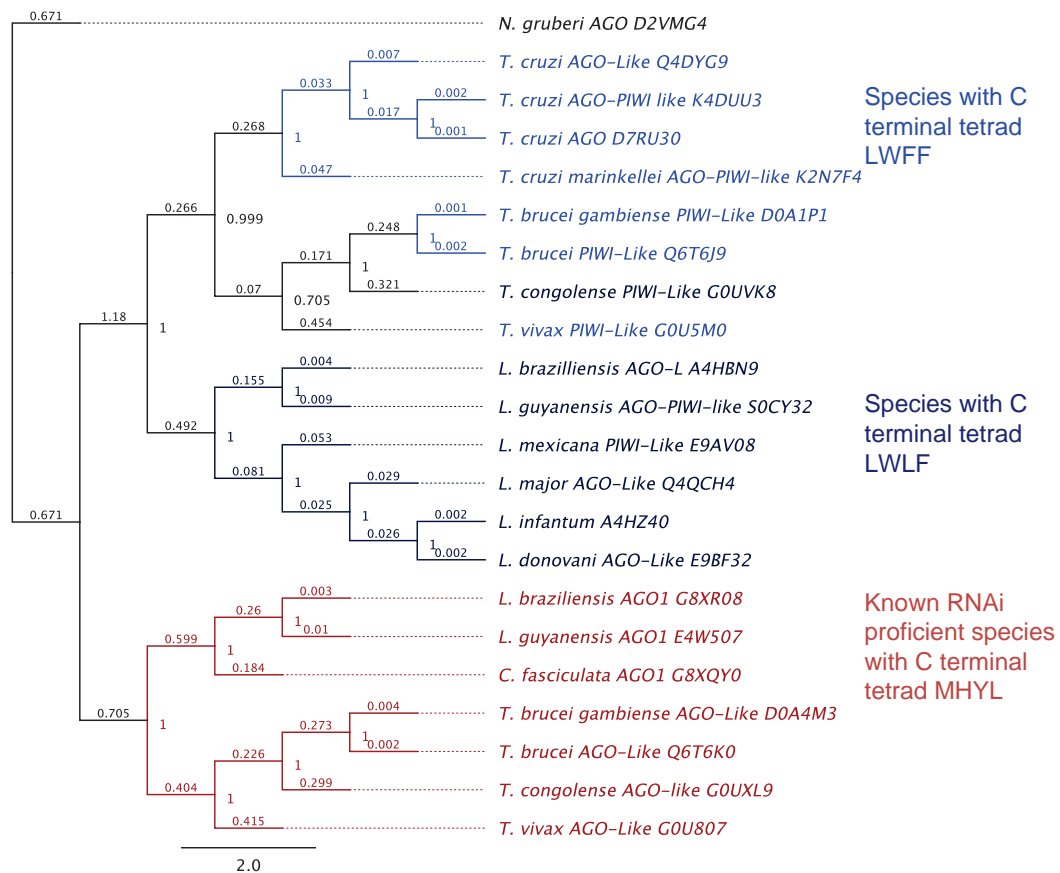
**Fig. 4c.4 A comparison between the canonical argonaute and PIWI-tryp within Trypanosomes.**

*T. brucei* is known to utilise the RNAi pathway. In contrast *T. cruzi* lacks RNA interference capability but retains an argonaute homolog that we predict to be from a PIWI-like ancestral sequence. This figure shows the PIWI-Like argonaute from *T. brucei* (UniProtKB:Q6T6J9). The AGO-like sequence from *T. brucei* (UniProtKB:Q6T6K0) is shown in fig. 4c.1.

Included in the ancestral reconstruction but not shown in the tree (fig. 4c.5), are two sequences from the obligate parasite of insect gastrointestinal tracts *Angomonas deanei* (UniProtKB:S9U4Q6 and S9VGD5). The sequences are both short, the former is modelled as comprising of all of the domains while the latter has an N-terminal and mid domain more reminiscent of prokaryote argonautes. They group within the AGO-like kinetoplasts and retain an AGO-like C terminal signature. We find that the kinetoplast sequences cannot be resolved in large eukaryote trees but suggest that this is because recurrent mutations have obscured the relationships (Mossel and Steel, 2004) rather than convincingly demonstrating a single ancestor in this group.

If we look at the last four residues of the C terminal we find that the methionine (conserved at the fourth from last residue in almost all of the AGO-like argonaute sequences that we have found) is present only in sequences from those species known to be RNAi competent (an hypothesis that should be tested). Those species have an additional argonaute that has the PIWI-like C terminal tetrad implicit of PIWI-like

proteins group separately. It is more parsimonious that kinetoplasts originally had two (at least) varieties of argonaute, one PIWI-like and one AGO-like and that the AGO-like argonaute has been lost in some species the same way that it has been lost in ciliates. We find it difficult to imagine convergent evolution in the kinetoplasts and metazoa (and in other species) that retain the fourth from last residue (either Leu or Met). A simpler explanation is that retention of both AGO and PIWI in kinetoplasts and metazoa and loss of one or the other (or both) in all other species. There are clearly two groups of argonaute proteins (fig. 4c.5) although we acknowledge that point remains unresolved in our final trees.



**Fig. 4c.5 Mr Bayes tree of annotated argonaute proteins in trypanosomid protozoans.**

Although the trypanosomid argonaute proteins cannot be resolved in large trees including all eukaryotes and appear monophyletic, they do fall into two distinct groups. It can be seen that with the exception of *C. fasciculata*, the species that are RNAi competent also retain an argonaute with a PIWI-like C terminal signature. This tree was rooted by the heterolobosea *Naegleria gruberi* AGO in order that posterior probability at the nodes could be shown.

The database annotations of some of the sequences are potentially misleading where they are marked AGO-like when the AGO-like protein sequences fall quite separately. *Leishmania* also have sequences annotated as AGO-like that fall within the PIWI-like clade, as well as sequences that fall within the AGO-like clade. Indeed some are annotated as AGO-PIWI-like safely covering either. As functionality can evolve and

has had plenty of time to do so, it seems likely that these variants may not fulfil the accepted roles that are currently known.

The heterolobosea *N. gruberi* is a free living excavate considered to be an ancient representative of early eukaryotes (Fritz-Laylin *et al.*, 2010) and harbours two different argonautes. *N. gruberi* UniProtKB:D2VMG4 is annotated as AGO but falls within the PIWI-like clade and retains a PIWI-like C terminal signature. The second *Naegleria* sequence (UniProtKB:D2V7J4) has a hitherto unseen C terminal signature of Pro-Phe-Phe-Lys (PFFL) and is unresolved and resides on a branch of its own in the trees. Two sequences are insufficient to make any ancestor and so they are included as orphans.

*Giardia intestinalis* is a diplomonad closely related to the kinetoplasts but the *Giardia* sequences could not be resolved in any of the large trees. *Giardia* is a major animal parasite harbouring possibly five duplications of the argonaute protein. The sequences found are very similar and so only one; UniProtKB:A8BCK6 has been included as an orphan. It has been suggested that antigenic variation in *Giardia* is regulated at the posttranscriptional level by RNAi (Prucca *et al.*, 2008), usually the province of the AGO-like argonautes, but UniProtKB:A8BCK6 grouped within PIWI-like proteins in our tree of ancestors and orphans. All of the *Giardia* sequences retain the C terminal Pro-Phe-Phe-Ile (PFFI) very similar to the unresolved *Naegleria* sequence.

#### **4c.4.4. Red and green algae**

It would be instructive to find PIWI-like argonautes in the Charophyta grouping closest to land plants but although mitochondrial and chloroplast genomes have been sequenced we could not find genomic sequences to search for them.

We had more success with the Chlorophytes; *Chlamydomonas reinhardtii* has three full-length copies (UniProtKB:A8J0N0, A8JAG8, A8J0M9) and *Volvox carteri* has one complete sequence (UniProtKB:D8U2I8). These all group with the AGO-like proteins even though they have a PIWI-like C terminal signature. The cold adapted green algae *Coccomyxa subellipsoidea* (UniProtKB:I0YVB2) sequence also groups with the *Chlamydomonas*, this has an AGO-like C terminal signature. *Micromonas pusilla*, (the only member of the genus *Micromonas*), a pico-eukaryote (tiny), (UniProtKB:C1DZY0) also has an AGO-like C terminal and groups with the AGO-like



sequences. *M. pusilla* was included as an orphan and was closest to one of our plant ancestors (chromatin modifiers of the 4, 6, 8, 9 clade).

*Chondrus crispus* (Red algae) (UniProtKB:R7QFY1, R7QNI5 and R7QGN7) are annotated as AGO1, 2 and 3, they also have a PIWI-like C terminal signature but group with the AGO-like side of the tree when added as orphans with the ancestors. So it does look as though our identified C terminal signature lacks consensus within algae.

#### **4c.4.5. Land Plants**

We included a selection of plant sequences (totalling 220 sequences) in our initial tree. The plant sequences fell between metazoan PIWI-like and metazoan AGO-like proteins. It appears most likely (given the red and green algal results) that land plants lost the PIWI type protein and yet still required the remaining argonautes to provide the same (or similar) functions, although to fall between the metazoa AGO-like and PIWI-like argonautes seemed rather unlikely. In our tree of ancestors calculated here (where we have removed all sequences that we do not confidently classify as argonautes), the plants remain firmly grouped with the metazoa and fungi AGO-like sequences and away from the metazoan and ciliate PIWI.

In plants 2'O-methylation of RNA is common e.g. HEN1 methylates the 3' end of plant miRNA and siRNA. In metazoa these RNAs have a 2nt overhang produced by Dicer cleavage. However rasiRNAs (in plants) and piRNAs (in animals) are 2'O methylated at their 3' end (Saito *et al.*, 2007). Consequently the RNA binding pocket structure needs to be wider in PIWI to accommodate this additional chemical group and this has been confirmed by two independent and alternate approaches (Simon *et al.*, 2011; Tian *et al.*, 2011a). This means that plant AGO structures must already have the wider PIWI-like RNA binding pocket. Plants and fungi both lack canonical PIWI-like argonautes but transcriptional control via chromatin and histone modification has been shown for both (Lippman and Martienssen, 2004; Vaucheret, 2008).

Most plant species have ten or more AGO-like argonautes. Barley (*Hordeum vulgare* var. *distichum*) has 70 putative argonautes, many of those are fragments (and some may be contamination as discussed later), goatgrass (*Aegilops tauschii*) has at least 21 verified argonautes and rice (*Oryza sativa* subsp. *Japonica*) appears to have at least 34 sequences that look to be both complete and genuine. The grasses however are known to have undergone a number of genome wide duplications and so we have used

as our basis for analysis the duplications that have occurred in *Arabidopsis* resulting in ten well studied AGOs (Vaucheret, 2008; Carbonell *et al.*, 2012).

The annotations assigned to the plant argonautes seem non-systematic, but an extensive review by (Kim *et al.*, 2011) showed that for *Arabidopsis thaliana* the plant argonautes can be grouped into three functional ‘clades’ (Vaucheret, 2008). Within each clade are members that are specific to seed development. AGOs 4, 6, 8\* and 9 are chromatin modifiers guided by 24nt RNAs and consistently group at the base of the phylogenetic trees. AGO8 may be a pseudogene (described in the supplementary material of (Takeda *et al.*, 2008)), but is expressed at low levels (Zhang *et al.*, 2016) and is described by Ensemble (Kersey *et al.*, 2015) as ‘known protein coding’. It is full length and modelled by I-TASSER as a complete argonaute and in evolutionary terms is still relevant.

AGOs 2, 3 and 7, are grouped as RNA binders guided by 21nt RNA. However AGO7 is involved in tasiRNA biogenesis (from the *Tas3* locus which target mRNAs encoding auxin response factors) by binding miR390 which guides AGO7 to cleave TAS3a (Adenot *et al.*, 2006), so AGO7 does have cleavage capability. AGO2 is involved in antiviral defence (Harvey *et al.*, 2011), and AGO3 has been shown to bind 24nt RNAs (mostly transposon derived) and interact directly in chromatin modification (Zhang *et al.*, 2016). So, although AGO 2 and 3 are originally duplicated from the same gene, and fall within the same clade of RNA binders, AGO3 has diverged in terms of functionality - with the caveat that it may also be involved in the post-transcriptional modification pathway as AGO3 is found in the cytosol as well as the nucleus. AGOs 1, 5 and 10 form a clade of argonautes whose principle function is (at least in *Arabidopsis*) cleavage of RNA. Our point is that although all of the plant sequences do fall neatly into three clades, we cannot properly ascribe function to them all because some of the argonautes in other species may also have diverged.

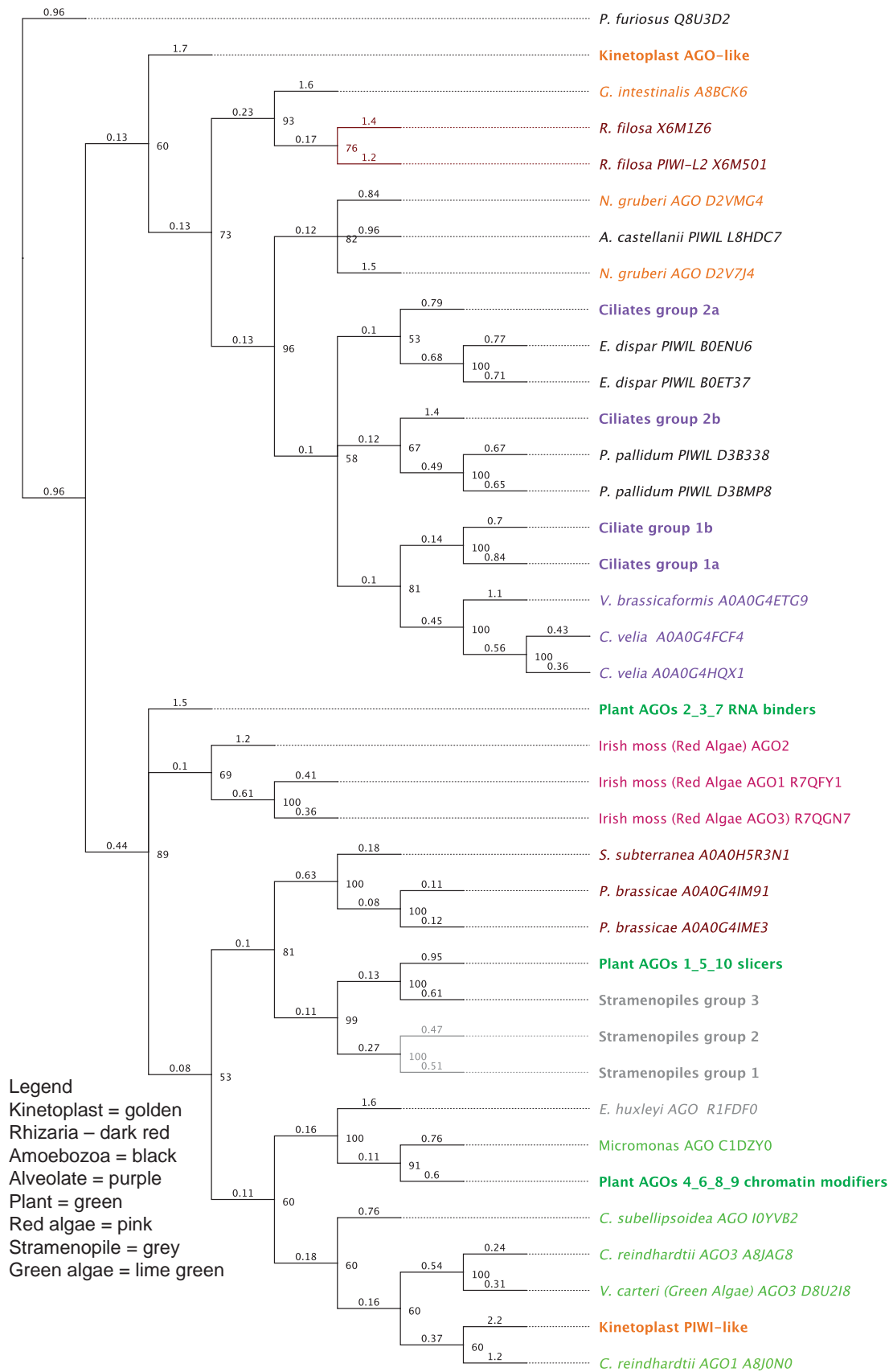
Most plant argonautes haven’t been allocated classification numbers. Of those that have, proteins from tomato, rice, and goatgrass (*A. tauschii*) loosely follow the *Arabidopsis* numbering in that AGO4 falls within the clade of chromatin modifiers described for *Arabidopsis* (AGO4, 6, 8 and 9), however rice appears to have at least 19 annotated argonautes and AGOs11-14, 17, 18 plus AGO1A, B, C and D fall within the *Arabidopsis* clade of slicing capable argonautes (AGO1, 5 and 10) yet no rice sequences are annotated as AGO5 or 10 - so the numbering appears to need functional relevance.

Within our calculated trees of just the plant sequences AGOs 4, 6, 8 and 9 appear more ancient and the AGOs with slicing capability (AGO1, 5 and 10), more recent. Almost all plant AGOs have a complete catalytic tetrad though we know from studies of human AGO that this may not be sufficient for nuclease activity (Schürmann *et al.*, 2013). What is clear from our trees is that early branching species such as *Physcomitrella patens* and the spike moss *Selaginella moellendorffii* already had AGO proteins in each of the three clades.

We have used the three AGO clades that naturally grouped in our trees (that match those described for *Arabidopsis*) to construct three ancestral protein sequences to see how they compare to the ancestors of the other major groups. Sequences from clade AGO4, 6, 8, and 9 (chromatin modifiers) totalled 47 sequences, clade AGO2, 3 and 7 (RNA binders) totalled 43 sequences and clade AGO1, 5, and 10 (slicers) was derived from 107 sequences. What is notable is that pine and spruce in general appear to lack AGOs. The only convincing argonaute sequences that we found were from the white spruce *Picea glauca*, (UniProtKB:Q4PLA9) and the Canary Island pine *Pinus canariensis* (UniProtKB:A0H0C4W3U3, previously AJA90779 and not found in Pfam). However this might simply be that insufficient sequencing has been completed because the authors of the *P. glauca* sequence note that there were five gene copies in the genome for the argonaute protein but this appears to be the only full-length protein (Tahir *et al.*, 2006). All three plant ancestors retain a catalytic tetrad and an AGO-like C terminal signature and group on the AGO-like side of the tree as anticipated. So plants harbour only AGO-like argonautes – the only PIWI-like sequence is barley (UniProtKB:F2DNY6), which we are convinced is contamination (most likely with rotifers) and was left out of any plant ancestor.

#### **4c.5. Ancestral trees**

Because trees lose information at deep time we have made ancestors that represent a larger number of sequences in order to simplify the trees. Here we have made ancestors for eukaryote groups other than metazoa and fungi (that we have reviewed elsewhere) to show how the AGO-like and PIWI-like proteins may have evolved. The ‘orphans’ that we cannot confidently place within a particular group are shown with the ancestors in a tree rooted by *P. furiosus* (UniProtKB:Q8U3-D2), a hyperthermophilic archaeon.



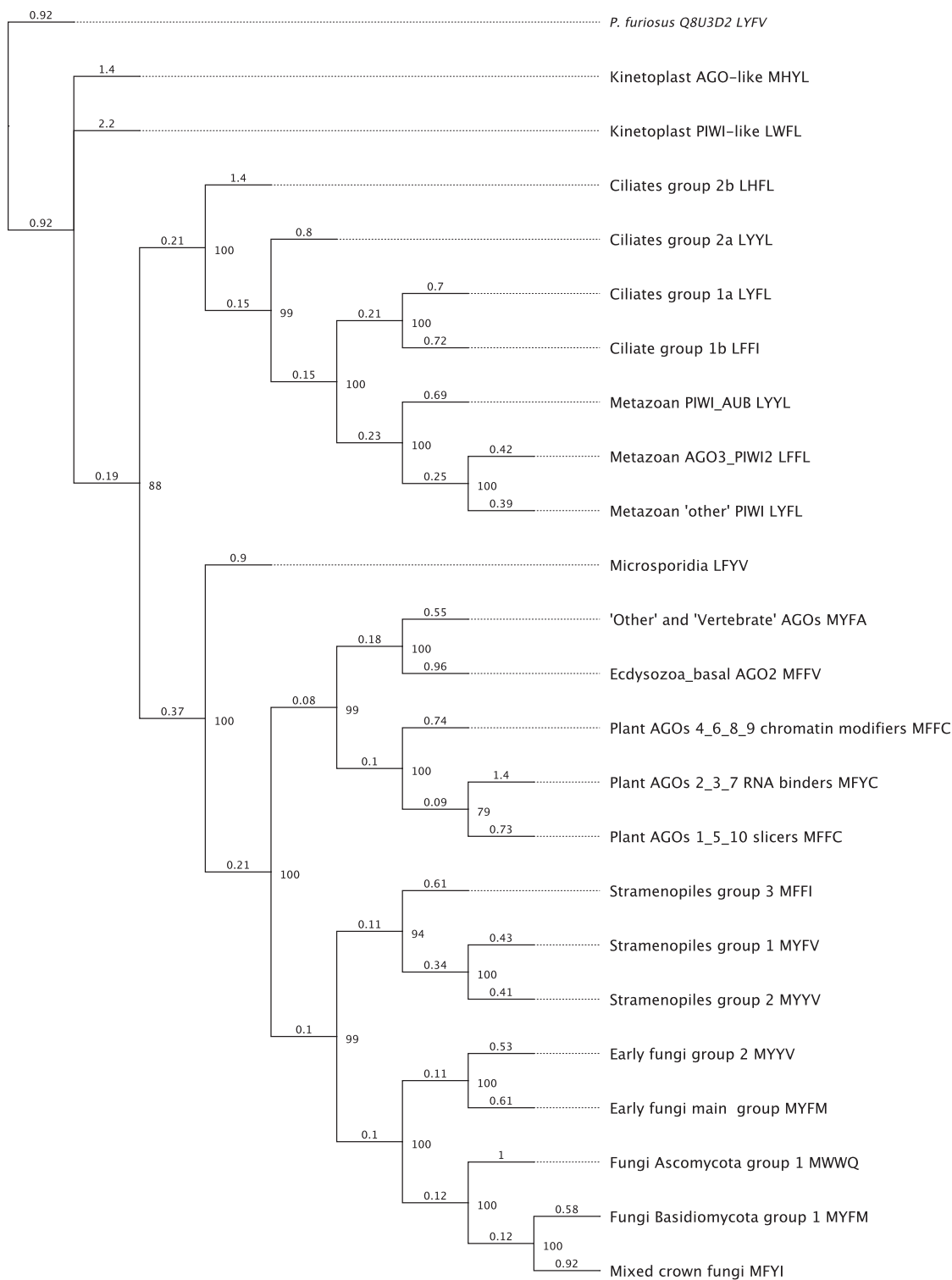
**Fig. 4c.6 A rooted tree from the calculated ancestral sequences.**

The ancestors (in bold) represent 375 sequences used in the recreations, the plant ancestors are made from 197 representative sequences where there are thousands in the data banks. Also included are the orphans that we were unable to place.

However the C terminal signature is not enough to separate AGO from PIWI within the kinetoplasts because contrary to expectations the ancestral kinetoplast PIWI-Like ancestor has sided with the AGO-Like sequences and vice versa! The PIWI-L kinetoplast ancestor has nestled with the algae sequences that also have a PIWI-L C terminal signature and AGO-like grouping.

We rooted the tree in order to demonstrate, the posterior probability though it is possible that *P. furiosus* is not sufficiently remote for an impartial root because it has the PIWI-like C terminal signature. The I-TASSER structural predictions and additional information (catalytic tetrad, C terminal signature and any additional notes) regarding the ancestors for each group is available in appendix III (supplementary material S2).

Each ancestral reconstruction was calculated from an unrooted tree but here we wanted to show the posterior probability at the nodes (and this is not shown on unrooted trees). In our original tree (that contained ~10% of sequences that we later discounted), the plant sequences grouped between metazoa AGO and metazoa PIWI which seemed most unlikely. Figure 4c.7 shows a phylogenetic tree calculated from the ancestors that we have reviewed here together with the metazoan and fungal ancestors that we have inferred previously. The tree shows that the plant ancestors remain close to the metazoan AGO ancestors but clearly on the AGO-like side of the tree. By most of our criteria microsporidia comprise of genuine AGO-like argonautes but they do retain a PIWI-like C terminal signature and are clearly different from fungi. Kinetoplasts remain unresolved but do not appear to have arisen from a duplication within just kinetoplasts. At deep times the C terminal signature breaks down as a means of distinguishing between AGO and PIWI. We had already noticed this within the algae but the microsporidia are also conflicted and this is apparent in the tree of all ancestors.



**Fig. 4c.7 Tree of all eukaryote ancestral reconstructions.**

This tree represents more than 700 sequences that have been grouped by similarities in structure, sequence, catalytic tetrad and C terminal signature prior to contributing to an ancestral reconstruction. We are not using a single gene to predict phylogeny but rather to simplify the relationship between the argonaute proteins using the ancestors as a proxy for many more sequences. The C terminal signature is shown on the labels.



## 4c.6. Discussion

We use I-TASSER to demonstrate that using tertiary information is relatively straightforward, and this is currently the best predictor of tertiary structure. It uses more information than just primary sequence structure, and allows much better alignments; based on the tertiary structure. We have also found several annotation issues that need to be followed up. Our search has been hampered by variable nomenclature resulting from the exponential growth in sequencing with a severe lagging in expert annotation. There are several synonyms for the proteins that we have examined and these are slowly being updated but we frequently found argonautes annotated as a variety of translation initiation factors (TIF, eIF2), ‘stem cell self renewal’, CnjA protein, ‘uncharacterised’ and argonaute-like or PIWI-like or both. In some cases sequences were just plain wrong. Barley (UniProtKB:F2DNY6) (Matsumoto *et al.*, 2011) annotated as ‘PIWI-like protein’ (which we confirm) but groups within basal metazoa, closest to rotifers. We also show that the PIWI-like proteins are found in excavates, alveolates, amoebzoa, as well as in metazoa – so they are apparently much older than just in metazoa.

Although we have forced the groups to coalesce at the base of each group resulting in more than one ancestor for each kingdom we also created a tree of all of the sequences that survived our selection process. This numbered 710 sequences that we can be confident are genuine argonautes and excluded around 50 sequences from the original sequences retrieved by BLAST searches. Not unexpectedly this tree contained many long branches with just one or two sequences on them essentially they are simply unresolved and represent sequences that we have already identified as having poor primary homology. In this tree however the sequences used have already undergone our selection process and structural prediction, catalytic tetrad and C terminal signature has been used to confirm that these are all structural homologs. Twi12 (UniProtKB:A4ZYY6) has already been identified as a divergent but genuine PIWI-L argonaute from *T. thermophila* (a ciliate from the group known as alveolates) but in our large tree of confirmed argonaute sequences it groups within a sub tree of fungi. Without structural analysis or prior knowledge it could reasonably be presumed to have been included in error. In the case of barley (UniProtKB:F2DNY6) there is no prior research but we were able to confirm that it is a *bona fide* PIWI-L argonaute – just not from barley. Our point is that it shows that the trees cannot be used to resolve such issues.

It is not ideal to place our own criteria for inclusion in one clade or another which is why we broke the trees into quite large groups but unless research has already been carried out (as in the case of Twi12) we cannot assign groups by tree placement from such large and deep trees.

We contend that there were at least two types of argonaute already in the common ancestor. This is supported by the two versions of the argonaute found in each of bacteria and archaea. And that stramenopiles, plants and fungi have lost PIWI-like proteins, and alveolates and amoebozoa have lost AGO-like proteins with Kinetoplasts and metazoa retaining both.

Duplication and rapid evolution of the duplicated sequence freed from its functional constraint is a common evolutionary mechanism. There is evidence from aphids (*A. pisum*) that there is an accelerated evolution of the duplicated gene and also differential expression implying an alternate functionality (Ortiz-Rivas *et al.*, 2012). So we propose that some of our sequences will have simply evolved novel functions over time as we know has happened with Twi12. We may also note the increasing use of small RNAs that may never have been involved with Dicer processing at all. It could be that the RNAs that have been there since the last universal common ancestor (LUCA) were the first argonaute guide RNAs. Many of the small RNAs arise from within DNA that codes for functions that are new to fine-tune gene expression today. So the argonaute system of control over gene expression as well as endogenous genes, endogenous and invasive parasitic nucleic acid is a dynamic system able to adapt.

Our main finding is that using the information that can be gained by looking at structural prediction and finding the extra evidence that can resolve such issues described here can be done at minimal cost without extensive expertise in programming or experimental bench biology using our methodology.

## Acknowledgements

We thank I-TASSER for the provision of free computer hours for structural prediction via their University of Michigan server <http://zhanglab.ccmb.med.umich.edu/I-TASSER/>.

## Supporting Information

**S1 Fig.** The FastML logo showing the posterior probability for the sequence at node 1 for the group ‘stramenopiles group 2’.

**S2 Table.** The 3-D structures for each eukaryote ancestor described and additional information.

Posterior probability logos of all of the ancestors are available in wikispaces please contact [tonidaly@mac.com](mailto:tonidaly@mac.com) for access.

**Supplementary material can be found in Appendix III.**

# Chapter Five: Conclusion

## 5. Conclusion

The growing number of uncharacterised sequences in public databases has turned the prediction of protein function into a challenging research field. Automated annotation methods are often error-prone due to the small subset of proteins with experimentally verified function but bench biology to determine structure and function requires a high level of skill and is time consuming and expensive. One goal of this thesis was to develop an *in silico* method of protein investigation using three-dimensional structural prediction to verify annotation. The field of annotation expertise is then thrown open to a wide range of interested persons from a greater variety of backgrounds.

Fortunately there are an increasing number of proteins whose 3-D-structure have been determined or inferred. This is necessary because the algorithms for structural prediction do rely on their being at least a partially solved structure. Finally the more we learn about protein evolution, the more we accept a dynamic view of evolution where proteins may eventually adopt new functions (and tertiary structures may evolve through time).

Once the identity of a protein is known it can be used for evolutionary studies and included with closely related sequences to infer an ancestral sequence. The ancestral sequence can then be used both as a BLAST query to find more remote homologs and also to represent groups of sequences in order to simplify trees that are known to lose information at deeper divergences. I have demonstrated the method with two very different proteins and the information found in each case was used to infer an evolutionary pathway for those proteins.

### 5.1. The chosen proteins

The Major Vault Protein is precisely structured. Although extra sequence at the N terminal may be tucked up inside the vault particle and C terminal excess on the exterior, inserts within the protein are rare and probably disruptive. The vault particle seems to be involved in many pathways but essential for none and there is good reason

to think that the monomer (or its constituent ‘modules’) had, or has, a purpose prior to, or other than, the formation of the particle. On the other hand the argonaute has a large variety of inserts but manages to function despite them. Some have evolved new (to science) functions utilising a variety of RNAs that I think we are only just beginning to appreciate. It would seem that the ‘RNA world’ is not history after all.

The original interest in these proteins was because of the RNA component. The RNA is essential to the argonaute but vault RNA seems to have an interesting biology of its own. Both proteins are found in bacteria. Argonautes are found in archaea but so far MVP monomers are absent. Depending on your point of view bacteria and / or archaea gave rise to eukaryotes or bacteria and archaea derive from eukaryotes (making them akaryotes) (Penny *et al.*, 2014; Forterre, 2015; Mariscal and Doolittle, 2015; Martin *et al.*, 2015). I will look at the links with the past and show that structural prediction and linking evidence to support our impressions can also help us here.

As the project progressed I could see that the results could be manipulated to suit the story that I was telling. Of course we want to be objective and impartial because it is beyond the remit of most biologists to benchmark every algorithm (which are subject to change and hopefully improvement) it could be easy to fall into the trap of doing what has been done before just because it was done before. There are plenty of benchmarking papers but a method that suits one protein is not necessarily suitable for another and poor scores are not necessarily a reason to exclude a sequence. Rather a sequence should be excluded when it fails on balance of the evidence.

## **5.2. Links with the past: Major Vault Protein**

The evolutionary history of the vault MVP may help to identify possible past functions and illuminate current thoughts on function. The capability to remove toxins from a cell would be advantageous and increase the likelihood of the vault RNP being conserved. The big picture questions are this; are vaults ancestral, retained in some species, but fallen into disrepair, or lost beyond all recognition in others? Alternatively have they been comprised of parts that had other functions such as TolA and the stomatin core and have come together in a fairly remote eukaryote and vaults formed thereafter? If we could be certain which species have functional vaults, and which don’t appear to have need for them, or possibly maintain the monomer for another purpose, we should be able to clarify their role.

Traditionally only primary sequences and BLAST results have been used to study homology of proteins. However, there are several reasons why tertiary (3-D) structures should also be used. We mentioned the loss of information mathematically from the Markov models and so we need more evidence to support our trees. As well as the predicted appearance of the very specifically shaped MVP monomer we have used the retention of the capability of monomers to dock laterally along their length as additional evidence to support the determination of homology.

### 5.2.1. Bacteria

Vault monomers have been found in some bacteria but not in any archaea thus far. Does this mean that they have been lost in many bacterial species and archaea, never reached archaea, or in neither but later acquired by some bacteria? BLAST results with known MVP sequences frequently identify TolA proteins from bacteria (fig. 5.1). These are part of Tol/Pal system of colicin detoxification. Generally these proteins have limited primary homology (~10%) with the C terminal (coil and cap) of the vault monomer. This demonstrates exactly the kind of barely-related sequence that the method is looking for.



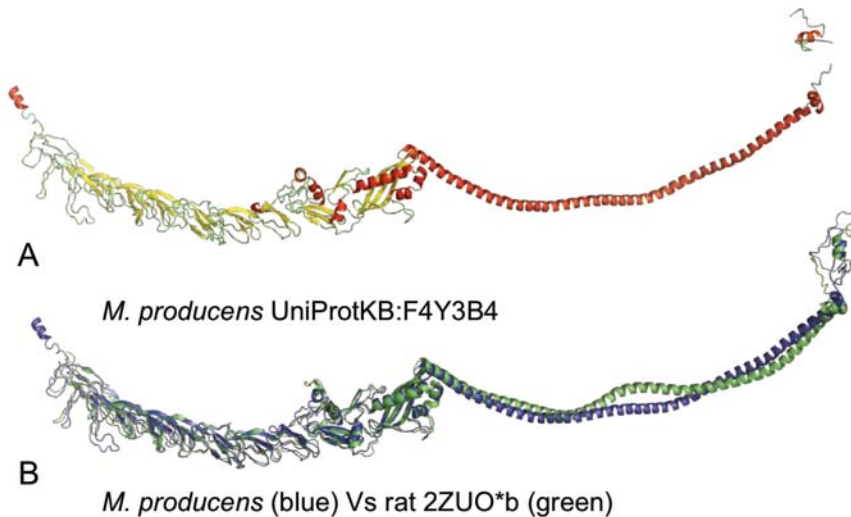
**Fig. 5.1 *Escherichia coli* TolA (UniProtKB:P19934)**

This protein is 421 amino acids and aligns with the cap-helix of the rat MVP (submitted to I-TASSER using the rat template 2ZUO\*b as a constraint). If we compare the sequence homology of the first 300 residues with the equivalent residues of the amoebozoia ancestral sequence that we calculated then the homology is 20% identical residues.

It is plausible that flexible proteins such as TolA, can be conveniently predicted to fold by I-TASSER (Roy *et al.*, 2010; Yang *et al.*, 2015) when submitted with the rat template as a constraint. Does the fact that it **can** fold to the vault shape mean that it **does**? It probably depends on the environment that it finds itself in. We know that the edges of each vault monomer are hydrophobic and so are attracted to one another and self-assemble to shield each from an aqueous environment. TolA is a membrane anchored protein, that is thought to span the periplasm of gram negative bacteria and might happily exist individually in that environment, but would congregate in a watery environment.



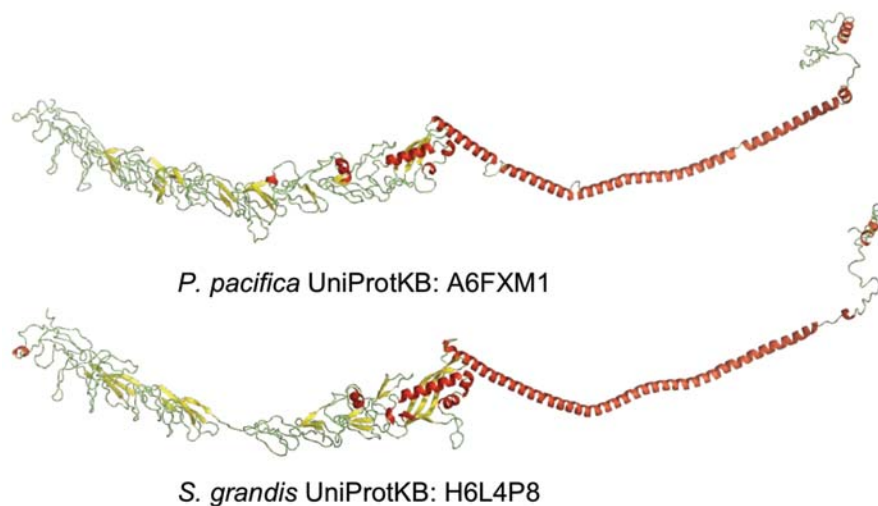
However the evidence for bacterial MVP is stronger than this; a number of cyanobacteria species harbour sequences with 54% homology with rat MVP annotated as ‘colicin uptake protein’, e.g., *Moorea producens* 3L formerly *Cyanobacteria lyngbya* UniProtKB:F4Y3B4. This seems extraordinarily high homology between rat and bacteria and with the barley contamination described in both MVP and argonautes in mind, the bacterial homology needs to be treated with caution. The I-TASSER 3-D structural prediction of this sequence is unmistakably MVP-like (fig. 5.2).



**Fig. 5.2 Bacterial MVP monomer.**

A. The I-TASSER predicted structure of the colicin uptake protein from cyanobacteria currently known as *Moorea producens* 3L (UniProtKB:F4Y3B4). B. A FATCAT (Ye and Godzik, 2003) comparison of F4Y3B4 (blue) with the known MVP structure from the rat (shown in green). Figures rendered in PyMol version 1.3.

As the proteins from the cyanobacteria are surprisingly close to the rat homologs, (if not from contamination), it could be argued that these have been acquired via horizontal transfer. In fact MVP homologs are found in a variety of bacteria (see fig. 5.3), and there are more homologs appearing as more bacteria are sequenced. Aside from the cyanobacteria, they tend to be mostly gram-negative predatory gliding bacteria and from a marine environment.



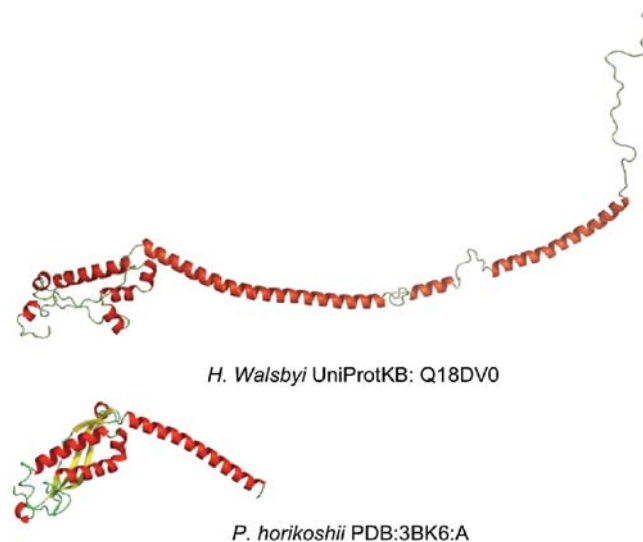
**Fig. 5.3 Full size bacterial MVP monomers.**

*Plesiocystis pacifica*, *Saprospira grandis* and other bacterial homologs have passed the test of docking laterally in ROSIE but these would make rather large organelles if they form in the manner of the rat vault compared with the size of the bacteria.

Given the size of the vault particle it seems strange that bacteria would harbour such a large particle. *Naegleria gruberi* have monomers that appear capable of forming a smaller particle (Daly *et al.*, 2013a) yet the MVP found so far in bacteria are full length. It could be argued that researchers would have observed these if they formed, but it has to be remembered that the discovery of the vault particle in the rat liver was serendipitous because they are pure protein and difficult to visualise (Kedersha and Rome, 1986). Why bacteria may have vault particles is probably simpler to solve. Not only are they involved in the detoxification process in some manner but they are also markers of radiation resistance in terms of cancer treatment (Herlevsen *et al.*, 2007) which may have been an advantage to cells when the sun posed a bigger problem in terms of cellular damage than now.

### 5.2.2. Archaea

Although structures homologous to the TolA protein and the stomatin core combined were found in archaea, whole MVP monomers have not, as yet, been observed. The proteins found to date are not expected to make vaults, but are reminiscent of the cap-helix that is fundamental to the vault construction and is similar to the TolA proteins involved with the bacterial colicin detoxification pathway. This would make sense and align with suggested vault functions. In fact the sequence from *Haloquadratum walsbyi* (UniProtKB: Q18DV0) has now been annotated ‘Mvp-type potassium channel superfamily protein’ (previously voltage-gated potassium channel) (Bolhuis *et al.*, 2006).



**Fig. 5.4 Putative archaea homolog sequences.**

There are a number of archaea with ‘bits’ of protein with known or predicted folds similar to MVP, e.g. *H. walsbyi* Mvp-type potassium channel superfamily protein (structure predicted by I-TASSER) and a monomer from the solved crystal structure of the stomatin core from *P. horikashii* (PDB: 3BK6:A). They can be found using some of the ancestral recreations as a BLAST query.

The group that produced the 3.5Å crystal structure (Tanaka *et al.*, 2009), identified the shoulder region as a homolog of the stomatin core structure from *Pyrococcus horikashii*. It is a member of the ‘band 7 flotillins’ that include prohibitin, flotillin, HlfK/C and podicin, these proteins are known to associate with lipid rafts and vault particles do as well (Kowalski *et al.*, 2007).

So our conundrum is thus; has MVP never formed in archaea, has it been lost, or are there simply insufficient species sequenced and it may yet be found? Archaea are generally classed as extremophiles, capable of living where other species find it too tough. It seems likely that they have developed pathways to withstand their environment and that the parts that we find are remnants of a system that they no longer need.

Without further studies we cannot say if the ancient particles, or even the extant particles outside of metazoa, have an RNA component. We cannot even be sure that whole vaults are formed. In terms of *in silico* investigation I found that the MVP sequence with the highest confidence score from I-TASSER and the lowest score for docking in ROSIE, both of which together, are highly predictive of vault formation, was from *Leishmania major* (UniProtKB:Q4QJJ7).

One of the features of the vault monomer is the highly conserved C terminal described in chapter 2b and reproduced here (fig. 5.5). In the ancestral reconstruction and in the individual sequences where reconstructions could not be made, the C

terminal ‘hook’ was not present in the more ancient species; green algae, microalgae, heterolobosea (*N. gruberi*). I believe that the hook is a new modification that allows vault particles to hook onto microtubules to facilitate rapid movement. They have been photographed attaching to microtubules via the end cap rather than the barrel side (Eichenmüller *et al.*, 2003) and they are large to move easily by diffusion.



**Fig. 5.5 The highly conserved C terminal from MVP.**

If the sequence from the section alone is used as a BLAST query it brings up nothing but vault particles.

Because the vault particle is involved with multi-drug resistance and cancer generally there will be many more research hours spent on this particle but what is especially exciting is the use that it could be put to in the future described in chapter 1.

### **5.3. Links with the past: Argonaute Family Proteins**

There is not any debate about the antiquity of the argonaute proteins. They are ubiquitous and must have been in the last ancestor of all known domains (Makarova *et al.*, 2009; Swarts *et al.*, 2015). What is different is that bacteria and archaea have argonautes that fall into two groups, short (that lack a PAZ domain and are frequently on an operon with another kind of nuclease), and long version which have a PAZ domain and are more similar to eukaryote argonautes (Makarova *et al.*, 2009). Additionally bacterial and archaeon argonautes are generally guided by DNA rather than RNA (Swarts *et al.*, 2014). This shouldn't be a surprise because if argonautes evolved to deal with parasitic nucleic acid then parasitic RNA, especially viral RNA is much more frequently found amongst eukaryotes and invasive DNA in prokaryotes. We must be careful here that we are not ascribing the defence of the Dark Arts as the only function of an argonaute. Although parasitic nucleic acid will always have been a burden, the function of creating a daughter cell is surely more fundamental (Alié *et al.*, 2011).

What is only recently becoming apparent is that the argonautes have even more varied functions and likely many more as yet unknown. We have seen in chapter 4c that Twi12 of *Tetrahymena thermophila* does not play any part in the usual *modus operandi*

of argonautes but yet is essential for the function of an exonuclease involved in 5.8S RNA processing (Couvillion *et al.*, 2012). The point to be made here is that Twi12 uses processed tRNA as its guide. tRNA cleavage appears to be a conserved pathway in eukaryotes (Thompson *et al.*, 2008), but has also been found in the bacterium *Streptomyces coelicolor* so this may be a very early source of small RNA.

There are some cases where a species has an acceptable argonaute but lacks small RNA. The placazoan *Hordeum vulgare* var. *distichum* is such an example, there seems a distinct lack of miRNA and few tRNA genes with no tRNAs processed into tRNA-derived RNA fragments either (Hertel *et al.*, 2009). The microRNA pathway is not the only pathway that utilises argonautes and so loss of miRNA processing in minimalist metazoa *T. adhaerens* is unremarkable but a lack of small RNA altogether seems unlikely. Why would *Trichoplax* maintain an argonaute that it cannot use?

Small RNAs processed from other well-described RNAs has not traditionally been considered part of the repertoire of guide RNAs. Frequently these are specifically excluded from study because their functions are known and would only add to the ‘noise’ blocking out the ‘signal’ for newly discovered small RNAs. It seems that researchers blocking the familiar RNAs may be missing the signal entirely! The complexity of the regulation of gene expression has entered a new age that is not new at all but has carried on right before us.

The primary sequence homology between argonautes (AGO or PIWI) is much lower than between corresponding MVP monomers. This is more to do with inserted sequence that doesn’t affect the binding of nucleic acid (bacteria and archaea bind DNA but can also bind RNA). As cleavage is only a small part of the repertoire of skills that an argonaute can have, it may function with a degenerate catalytic centre and so in the study I may have discarded sequences that do have a part to play simply because we have erred on the side of caution. The stated intent though, was to seek out those sequences that are on the edge of believable and see what can be learned from them.

What has been learnt from the argonaute study is that there appears to have been two different argonautes in the common ancestor not only of eukaryotes but also of bacteria and archaea. If we look at the C terminal signature of the long version of the prokaryote proteins that have been through the protocol, 11 of 20 retain the PIWI-like C terminal signature, just one has Met at the 4<sup>th</sup> from last residue resembling an AGO-like signature, three archaea have the Pro at the 4<sup>th</sup> from last residue that we saw in one of

the *Naegleria* sequences and in the *Giardia* sequences. One residue, in a sequence of ~850 cannot be the difference but it certainly has helped.

More work of this type needs to be carried out in the deepest of times. We have barely brushed the surface but it seems inescapable that a large and complex protein was in the last common ancestor of all life as we know it, driven by RNA. Whether it was a PIWI-like argonaute that then duplicated and became two flavours or whether there were always two, one dedicated to the Dark Arts and one dedicated to procreation we cannot yet say.

## 5.4. Challenges of the method

Many more servers were used than have been described and many are benchmarked elsewhere. This section describes some of the limitations and advantages of the servers and algorithms that are central to the method. The servers have also been improved upon over time which additionally makes comparing results in terms of numerical values less meaningful. New services are becoming available and the pipeline or workflow is designed to adapt to this as well as to different types of protein.

All of the databases used have been extensively upgraded over the duration of the research but probably none more so than Pfam. Where initially there were two levels of confidence depending on manual or automated pipelines the new Pfam (primarily aligned to UniProtKB which reduces the confusion of a plethora of accession numbers) now has one level and has extensive trees of structurally related domains. Extensive use was made of Pfam in chapters 4b and 4c.

### 5.4.1. BLASTp

BLASTp searches are designed to find sequence homology and so this research stretches that capability when looking for loss of sequence homology. For MVP described in chapter two I paid attention to the E scores and bit scores since the homology between vault monomers is high. Given the 'bit score'  $S'$  or the E score;  $E = mn2^{-S'}$  'n' has less comparative affect than the size of the database being searched. Especially for the argonaute sequences since most of the sequences (MVP, AGO and PIWI) are of similar length (~860 residues) and the argonaute search space is so much larger. If the size of the search space was known then the bit score could be used to calculate the significance but the search space is growing rapidly and so the significance is not repeatable. The theme 'lack of repeatability' is a recurring one, which is why



‘scores’ in general have been used cautiously.

Because my interest is in argonaute sequences with very little primary homology, they are either not picked up at all in a BLAST search regardless of the scoring matrix (particularly where the database is very large), or would traditionally be ignored due to a high E value. So the scores become irrelevant when the size of the database is either so great that sequences with merit are outside the thousand ‘hit’ limit, or the database has to be narrowed until it is small enough that such a loss does not occur and then the E value is artificially low. Comparing bit score or E values is then meaningless.

#### **5.4.2. MSA**

Many sequence alignment algorithms were trialled and sometimes sequences went through a number of them where the results were unexpected (e.g. *X. tropicalis* PIWI3 described in chapter 4a). Geneious align was the obvious choice for the bulk of the work since I had downloaded the Geneious suite, however Geneious align is painfully slow and a high number of sequences do not align at all. The scoring matrix can be adjusted but there is an unexpected utility in that sequences that Geneious refuses to align will automatically flag themselves as requiring submission to I-TASSER for structural prediction. This method of using Geneious to prioritise sequences for structural prediction was used extensively for the ciliate AGOs in chapter 4c.

MUSCLE will align practically anything. In this case the outcome may not be as accurate but particularly for the more remote sequences in chapter 5c this was essential for building trees from sequences that had already been rejected by Geneious but identified by I-TASSER as *bona fide* argonautes and have to be aligned for tree building and submission for ASR.

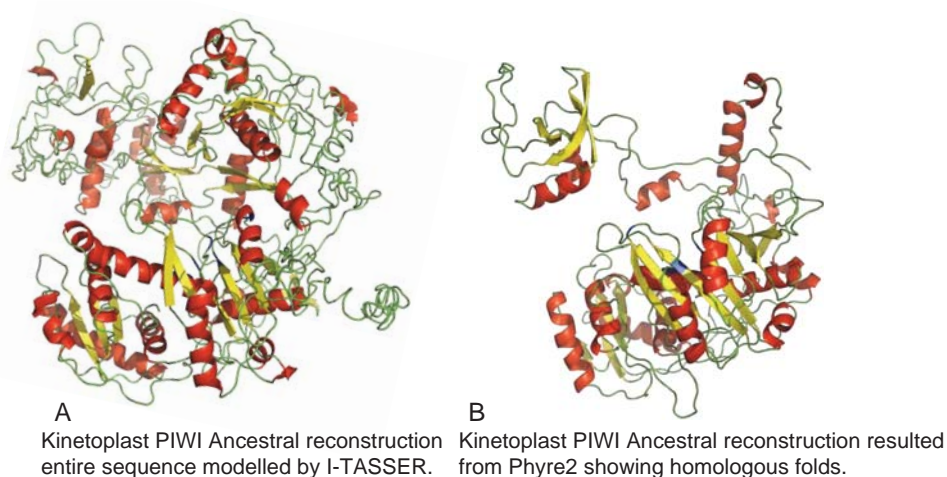
#### **5.4.3. Structural prediction**

Structural prediction from primary protein sequence has become a mainstream technique and I argue that it should be essential to augment bench work, evolutionary studies and annotation. Two well-respected servers are used extensively and have been described in chapters two and four. Here I outline the strengths and weaknesses of each for the purpose of my aim.

Phyre2 is very much faster than I-TASSER (~7 hours computing time as opposed to ~50 hours of computing time) but gives less information. This is entirely suitable for the bulk of the work but where there are significant deviations from the known structures in the Protein Data Bank I-TASSER will still result the top five

estimated structures, modelling the entire submitted sequence. Where the sequence lacks primary homology and other information is needed to support the structure prediction especially where the C score is poor (e.g. catalytic tetrad and C terminal signature). Phyre2 will not provide a result of the entire sequence where I-TASSER will.

An example is the Kinetoplast PIWI ancestor. This is a case where we know from the literature that trypanosomes at least do have argonaute proteins (Ngô *et al.*, 1998; Garcia Silva *et al.*, 2010b). The trypanosome AGO and PIWI are difficult to resolve and multiple mutations over time make it complicated to reconstruct ancestors. Two ancestors were reconstructed for the kinetoplasts as a whole because the sequences fell clearly into two groups (described in chapter 4c). The treatment of the ancestral kinetoplast PIWI-like sequence by I-TASSER and Phyre2 are shown below (fig 5.6).



**Fig. 5.6 I-TASSER and Phyre2 comparison.**

**A.** Shows the full sequence modelled by I-TASSER with a C score of  $-2.02$  (outside the score for a convincing fold). **B.** The result of the identical input into Phyre2 where only fold matches are included in the result indicates 40% coverage with 100% confidence.

Both servers provide valuable information and section 5.4.6 (ASR) demonstrates how this can be resolved in order to be confident that the kinetoplast PIWI ancestor is a good representation of the sequences that went into its calculation.

#### 5.4.4. RosettaDock (ROSIE)

The ROSIE server generates 1,000 possible structures and displays the ten with the lowest energy score. *In vivo* chaperones are often required for tertiary or quaternary protein structure but in the case of the vault particle it is known that the MVP proteins self assemble spontaneously, so the energy score should be low.

The drawback with the online server is that it can only cope with a maximum of 600 residues. However it is free to download and relatively easy to use and allows bigger files (in terms of number of residues) but each job takes ~50 hours and uses all of the computing power of a Macbook pro 2.7 GHz i7 rendering it unusable for everyday tasks. Other docking algorithm servers were compared but RosettaDock (ROSIE) gave the best results. The docking work is described extensively in chapter 2a (Daly *et al.*, 2013a).

#### **5.4.5. Tree calculation**

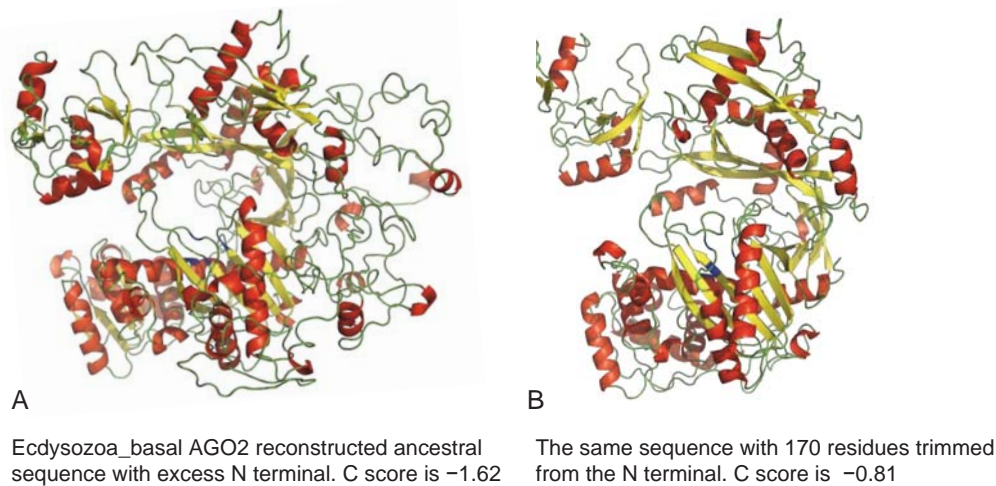
The tree calculations are carried out by MrBayes running via the Geneious platform. It is hungry in terms of computing power but the only algorithm (of MrBayes, Geneious, PHYML and FastTree) to reliably produce the same tree for the same input MSA. The average calculation time for each ancestral tree was ~120 hours. The final tree of 710 sequences which had passed as *bona fide* sequences too ~500 hours. The excess of time (that would have taken considerably less using any other algorithm) I consider time well invested.

#### **5.4.6. Ancestral Sequence Reconstruction (ASR)**

In chapter 2b I describe various limitations of three reconstruction algorithms; PAML4 (Phylogenetic Analysis by Maximum Likelihood) (Yang, 2007), FastML (Ashkenazy *et al.*, 2012) and Mega5 (Tamura *et al.*, 2011). These algorithms are easy to use but unrealistically short (PAML4) or long (FastML). To obtain the most reasonable sequence (neither long nor short) using (predominantly) FastML requires a workaround, i.e. the removal of inserts occurring in less than 10% of sequences, because FastML will fill all gaps. I was alerted to the existence of ‘PRANK’ (from the Löytynoja lab at the University of Helsinki (Löytynoja and Goldman, 2010)), now updated and included in the ‘Wasabi’ suit (Veidenberg *et al.*, 2015) which automatically removes inserts where 9% or less sequences have residues. Comparing Wasabi with manual removal of gaps resulted in very similar MSA suitable for submission to FastML but was much faster and less user intensive so is a very recent improvement to the original method.

Even so the majority of the resultant ancestor had long N terminals where the alignment algorithms use gaps to try and make the best job of a disparate group of residues. The N terminals of all the argonautes (AGO or PIWI) have very little sequence similarity. It is not possible to manually remove them and Wasabi does not do the job any better. This results in long N terminals in the reconstructed ancestor that is not

modelled at all well by I-TASSER and results in an artificially poor C score. Taking off the N terminal from the I-TASSER .pdb file manually and putting this sequence back through I-TASSER improves the C score without any loss of core structure bringing it within the cut off criteria for a correct fold (see fig. 5.7).

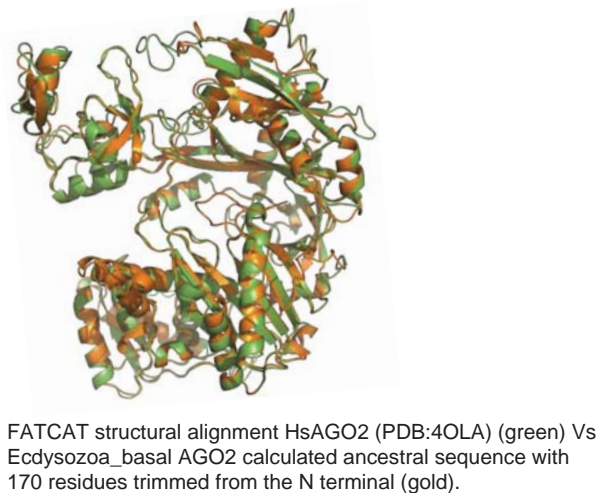


**Fig. 5.7 An I-TASSER comparison between raw and trimmed ASR node 1 sequences.**

The I-TASSER C score for many of the ancestral reconstructions is less than the score anticipated for a correct fold. **A.** Shows the predicted structure for the calculated raw ancestral sequence (1122 residues). **B.** Has had 170 residues trimmed from the N terminal (952 residues) without any loss of structure and improving the C score.

#### 5.4.7. FATCAT

Extraneous N terminal is also removed from the ancestral .pdf files done prior to structural alignment with FATCAT improving the FATCAT score as well (see fig. 5.8).



**Fig. 5.8 FATCAT structural alignment of the truncated ancestor with HsAGO2.**

The FATCAT scores for all of the ancestral reconstructions are quite high even with the raw sequence but truncating the residues that are obvious artefacts of the process improves that score as well. Here 94.5% of residues in the ancestral sequence (gold) are in an equivalent position to those of HsAGO2 (green). The raw score for the alignment is 2165.59 (given because the P score is zero) and our cut off criteria is  $>1500$ .

The P score is almost always zero unless the alignment is poor and so the raw score (higher the better) gives a greater sense of similarity when trying to compare models. FATCAT is used to check marginal sequences (described in the workflow in the preface to chapter four) before inclusion in a clade for ancestral reconstruction and is also used for every ancestor.

FATCAT also has an unexpected utility in that it can be used to position two MVP monomers adjacent to one another for submission to the RosettaDock. Aligning a copy of a structural prediction of any MVP monomer with each of 2ZUO\*b (\*describes the monomer in the solved structure) and 2ZUO\*d of the rat crystal structure (part of PDB:4V60) ensures that the monomers are side-by-side one MVP monomer space apart. FATCAT is free, simple to use and described in chapter 2a along with RosettaDock (as it was called) and the requirement for the spacing for docking is explained.

#### **5.4.8. Philanthropy**

In excess of 30,000 hours of computing time has been given freely by I-TASSER from Michigan. Not just to me but to all comers as long as everyone plays by the rules and submits one job at a time. This has been the mainstay of the project but free wall time from FastML, FATCAT and all of those mentioned in chapter one and in the glossary developed into a chain of events to help categorise a protein, not just MVP and argonautes, but is applicable to any protein that has a solved sequence (in part or whole).

The philanthropy demonstrated throughout the field of computational biology shows that people are keen to share their ideas, encourage people to use their product, give feedback and it is a bit of a thrill to get an email saying that yes, you are quite right, that didn't work or there is an error and we will fix it, take your advise, delete those errors sequences or amalgamate files and all of these things have happened. I am encouraged also that all of my papers have been cited (other than by my colleagues) which shows that there is creeping acceptance that 'BLAST, MSA, tree', is no longer enough, each process has flaws and each incremental flaw can add up to an error.

The point is that you do not need special training, you do not need a big institution (though that surely helps), but you can plug away and contribute to the bigger picture. In order to speed things up I also had the use of NeSI's high performance computing facility-running I-TASSER and many more hours of wall time were spent there. I am thankful although conscious that the armchair biologist doesn't have the luxury of that level of access.

## 5.5. The last word

The process developed here needs to be more widely tested as a general process. There are also some anomalies, for instance the early branching fungi do, indeed branch early in a tree of all fungal species, but when they are made into an ancestors, they appear much newer. Initially I thought that this might be because the sequences were too few but if that were the case then microsporidia should suffer from the same issue and it doesn't. We do need to know why these anomalies arise.

The evolutionary history of the argonautes needs a project of its own. Because the CRISPR system described in chapter 3 seems better adapted to deal with unwelcome nucleic acid it seems curious that prokaryotes have argonautes at all. Argonautes may have been retained after their spilt from their ancestor for cell regeneration and yet it has been shown that they do function in nicking even double stranded DNA (Makarova *et al.*, 2009). It does seem that RNA is far from being a relic of the 'RNA world' but is in the driving seat of so many more reactions.



## **Glossary**

### **Sources of Data**

#### **UniProt**

UniProt (Consortium, 2015) is supported by four data bases; UniProt Knowledgebase (UniProtKB), UniProt Reference Clusters (UniRef), UniProt Archive (UniParc), and Proteomes. Proteomes stores 43,505 species with complete sequenced genes as of January 2016. Searching the UniProtKB database reveals hundreds of similar sequences from a variety of species from viruses to prokaryotes and higher eukaryotes for the MVP monomer and thousands for the argonautes (AGO and PIWI).

#### **NCBI**

NCBI (National Centre for Biotechnology Information) makes available resources from many databases and in terms of protein sequences specifically from GenBank, RefSeq and the third party annotation database (TPA), as well as records from SwissProt, protein information resource (PIR), protein research foundation (PRF), and protein data bank (PDB). The sequences found in the UniProtKB database are crosschecked against the NCBI database, using accession numbers from the information given in UniProt. In some instances NCBI gives more information and also produces results from species that had not been identified in UniProt.

#### **Ensembl**

Ensembl presents annotated genetic information in great detail and is also used to undertake DNA as well as protein BLASTs and alignments. Ensembl is a joint project between European Bioinformatics Institute (EBI), and the Wellcome Trust Sanger Institute (WTSI) (Kersey *et al.*, 2015). The Ensembl default BLAST algorithm uses BLOSUM80.

#### **Pfam**

The above databases all link to Pfam (Finn *et al.*, 2010; Finn *et al.*, 2016). Pfam gives a visual annotation of sequences taken from the UniProt database and uses MSA and hidden Markov models (HMMs) (a statistical method that uses observable data to identify 'hidden' patterns and thus identify domain architecture from MSA).

## **Servers**

### **Structural prediction (I-TASSER)**

I-TASSER (Yang *et al.*, 2015) has been used throughout the research. I-TASSER is extensively described in chapter 3 and is very simple to use. The usual time taken for a sequence of ~800 residues is 50 hours (plus queue time). This is not slow for a task of this complexity; the rate-limiting factor is that only one sequence per IP address may be submitted at a time. Over a year of computer time has been used with I-TASSER runs.

### **Phyre2**

Phyre2 (Kelley *et al.*, 2015) is a very much faster server for structural prediction generally requiring less than seven hours to reach a conclusion (and equally easy to use), it lists the structures that are most similar to the submitted sequence. Phyre2 does not allow for *ab initio* prediction to the extent that I-TASSER does but is most useful for the bulk of the sequences where primary homology would already predict a genuine example of the protein under investigation. Slow or expensive methods of structural prediction can then be saved for sequences that come to attention due to other anomalies.

### **FATCAT**

To give a visual comparison of how similar a structural prediction of a sequence would look to a solved structure I made extensive use of FATCAT (Flexible structure AlignmentT by Chaining Aligned fragment pairs allowing Twists) (Ye and Godzik, 2003). As well as providing visual information and a ‘similarity score’ expressed as a percentage of residues similarly placed in three-dimensional space.

### **RosettaDock (ROSIE)**

For proteins that need to dock in order to complete their quaternary structure RosettaDock (Lyskov and Gray, 2008) is an ideal platform fulfilling the criteria of free and simple to use. The server has now moved to <http://rosie.graylab.jhu.edu/docking2> and changed its name to ROSIE (Rosetta Online Server that Includes Everyone), select ‘docking 2’ to submit a file to ROSIE (Lyskov *et al.*, 2013).

### **FastML**

Ancestral sequence reconstruction via FastML requires .fasta files for MSA and .newick files for trees. It is very easy to use, it will take the input MSA and build its own tree but the tree is not as consistent as a MrBayes tree. Other ASR algorithms were trialled (described in chapter 2b) but the argonaute study was completed solely using FastML.

## **‘In house’ data compilation**

Geneious Pro from Biomatters Ltd. is a bioinformatics platform and was originally downloaded as a trial, (<http://www.geneious.com/>) running on a miniMac 2.7 GHz Intel core i7. Geneious has been the workhorse for this project labouring for many hours at a time with large and unwieldy data sets gradually trimmed into workable units.

The platform was used principally for aligning sequences in order to determine the sequence similarity between various pairs and also to create MSAs for tree building and for ASR. Trees were built using MrBayes (Huelsenbeck and Ronquist, 2001) running within the Geneious platform.

## **Geneious alignments**

Geneious was used to generate pairwise alignments and multiple sequence alignments (MSA) and to create phylogenetic trees from those alignments. Over time iterations have changed and multiple alignments suffer from an increased possibility that the alignment is not the best that it could be, in order that it can be completed in a reasonable time. The Geneious algorithm is particularly precise and will perform an alignment with 500 iterations that can take many hours to perform.

## **MUSCLE**

**M**ultiple Sequence Comparison by **L**og- **E**xpectation (Edgar, 2004b; Edgar, 2004a) comprises of three parts; an initial alignment which is a rough draft, this is improved in the next stage, and then refined in the final stage. It is faster than the Geneious algorithm and has the advantage (for our purposes) that it will force an alignment where Geneious would not. Other algorithms both within the Geneious platform and online were trialled but especially in terms of the argonaute proteins, that in some cases had very little primary homology, MUSCLE was indispensable.

## **Geneious trees**

Geneious uses either neighbour-joining, or UPGMA (Unweighted Pair Group Method with Arithmetic mean), to build phylogenetic trees. Neither is entirely satisfactory. UPGMA assumes a constant rate of mutation (molecular clock), and this does not appear likely for any of the test proteins because they are under positive constraint, so neighbour-joining (NJ) was used. This does not assume a constant rate, but uses the least branch length at each stage of the algorithm. The Geneious tree builder was used for quick analysis but frequently does not produce an identical tree for an identical input.

Online servers can also provide a rapid service with the same drawback in terms of repeatability.

### **MrBayes trees**

MrBayes (Huelsenbeck and Ronquist, 2001) combines prior probability with maximum likelihood producing the most likely phylogenetic tree for the given data. MrBayes would produce the same tree for the same MSA when either unrooted or rooted by the same sequence and consistency is important. For ASR input MrBayes files were always used rather than the FastML tree builder even though MrBayes is computationally hungry from the point of view of a standard desktop computer.

## References

- Aboobaker, A. A. & Blaxter, M. L. 2003. Hox gene loss during dynamic evolution of the nematode cluster. *Current Biology*, 13, 37-40. Available: DOI 10.1016/s0960-9822(02)01399-4.
- Abramov, Y. A., Shatskikh, A. S., Maksimenko, O. G., Bonaccorsi, S., Gvozdev, V. A. & Lavrov, S. A. 2016. The differences between Cis-and Trans-Gene inactivation caused by heterochromatin in *Drosophila*. *Genetics*, 202, 93-106. Available: DOI 10.1534/genetics.115.181693.
- Adenot, X., Elmayan, T., Lauressergues, D., Boutet, S., Bouché, N., Gascioli, V. & Vaucheret, H. 2006. DRB4-Dependent TAS3 trans-Acting siRNAs Control Leaf Morphology through AGO7. *Current Biology*, 16, 927-932. Available: DOI 10.1016/j.cub.2006.03.035.
- Alié, A., Leclère, L., Jager, M., Dayraud, C., Chang, P., Le Guyader, H., Quéinnec, E. & Manuel, M. 2011. Somatic stem cells express Piwi and Vasa genes in an adult ctenophore: Ancient association of "germline genes" with stemness. *Developmental Biology*, 350, 183-197. Available: DOI 10.1016/j.ydbio.2010.10.019.
- Altschul, S. F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J. 1990. Basic local alignment search tool. *J Mol Biol* 215, 403-410.
- Alva, V., Söding, J. & Lupas, A. N. 2015. A vocabulary of ancient peptides at the origin of folded proteins. *eLife*.
- Amort, M., Nachbauer, B., Tuzlak, S., Kieser, A., Schepers, A., Villunger, A. & Polacek, N. 2015. Expression of the vault RNA protects cells from undergoing apoptosis. *Nat Commun*, 6. Available: DOI 10.1038/ncomms8030.
- Anderson, D. H., Kickhoefer, V. A., Sievers, S. A., Rome, L. H. & Eisenberg, D. 2007. Draft crystal structure of the vault shell at 9-A resolution. *PLoS biology*, 5.
- Aravin, A. & Tuschl, T. 2005. Identification and characterization of small RNAs involved in RNA silencing. *FEBS Letters*, 579, 5830-5840. Available: DOI 10.1016/j.febslet.2005.08.009.
- Ashkenazy, H., Penn, O., Doron-Faigenboim, A., Cohen, O., Cannarozzi, G., Zomer, O. & Pupko, T. 2012. FastML: A web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Research*, 40, W580-W584.
- Åsman, A. K. M., Vetukuri, R. R., Jahan, S. N., Fogelqvist, J., Corcoran, P., Avrova, A. O., Whisson, S. C. & Dixelius, C. 2014. Fragmentation of tRNA in *Phytophthora infestans* asexual life cycle stages and during host plant infection. *BMC Microbiology*, 14. Available: DOI 10.1186/s12866-014-0308-1.
- Ben, J., Zhang, Y., Zhou, R., Zhang, H., Zhu, X., Li, X., Zhang, H., Li, N., Zhou, X., Bai, H., Yang, Q., Li, D., Xu, Y. & Chen, Q. 2013. Major Vault Protein regulates class A scavenger receptor-mediated Tumor Necrosis Factor- $\alpha$  synthesis and apoptosis in macrophages. *The Journal of Biological Chemistry*, 288, 20076-20084. Available: DOI 10.1074/jbc.M112.449538.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. 2000. The Protein Data Bank. *Nucleic Acids Research*, 28, 235-242.
- Berriman, M., Haas, B. J., Loverde, P. T., Wilson, R. A., Dillon, G. P., Cerqueira, G. C., Mashiyama, S. T., Al-Lazikani, B., Andrade, L. F., Ashton, P. D., Aslett, M. A., Bartholomeu, D. C., Blandin, G., Caffrey, C. R., Coghlan, A., Coulson, R., Day, T. A., Delcher, A., Demarco, R., Djikeng, A., Eyre, T., Gamble, J. A., Ghedin, E., Gu, Y., Hertz-Fowler, C., Hirai, H., Hirai, Y., Houston, R., Ivens, A., Johnston, D. A., Lacerda, D., MacEdo, C. D., McVeigh, P., Ning, Z., Oliveira, G., Overington, J. P., Parkhill, J., Perte, M., Pierce, R. J., Protasio, A. V., Quail, M. A., Rajandream, M. A., Rogers, J., Sajid, M., Salzberg, S. L., Stanke, M., Tivey, A. R., White, O., Williams, D. L., Wortman, J., Wu, W., Zamanian, M., Zerlotini, A., Fraser-Liggett, C. M., Barrell, B. G. & El-Sayed, N. M. 2009. The genome of the blood fluke *Schistosoma mansoni*. *Nature*, 460, 352-358. Available: DOI 10.1038/nature08160.
- Bohmert, K., Camus, I., Bellini, C., Bouchez, D., Caboche, M. & Banning, C. 1998. AGO1 defines a novel locus of *Arabidopsis* controlling leaf development. *Embo Journal*, 17, 170-180. Available: DOI 10.1093/emboj/17.1.170.
- Bolhuis, H., Palm, P., Wende, A., Falb, M., Rampp, M., Rodriguez-Valera, F., Pfeiffer, F. & Oesterhelt, D. 2006. The genome of the square archaeon *Haloquadratum walsbyi*: Life at the limits of water activity. *BMC Genomics*, 7. Available: DOI 10.1186/1471-2164-7-169.
- Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R. & Hannon, G. J. 2007. Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell*, 128, 1089-1103. Available: DOI 10.1016/j.cell.2007.01.043.

- Buehler, D. C., Toso, D. B., Kickhoefer, V. A., Zhou, Z. H. & Rome, L. H. 2011. Vaults engineered for hydrophobic drug delivery. *Small*, 7, 1432-1439.
- Burki, F., Kudryavtsev, A., Matz, M. V., Aglyamova, G. V., Bulman, S., Fiers, M., Keeling, P. J. & Pawlowski, J. 2010. Evolution of Rhizaria: New insights from phylogenomic analysis of uncultivated protists. *BMC Evolutionary Biology*, 10.
- Capella-Gutiérrez, S., Marcet-Houben, M. & Gabaldón, T. 2012. Phylogenomics supports microsporidia as the earliest diverging clade of sequenced fungi. *BMC Biology*, 10, 1-14. Available: DOI 10.1186/1741-7007-10-47.
- Carbonell, A., Fahlgren, N., Garcia-Ruiz, H., Gilbert, K. B., Montgomery, T. A., Nguyen, T., Cuperus, J. T. & Carrington, J. C. 2012. Functional analysis of three Arabidopsis argonautes using slicer-defective mutants. *Plant Cell*, 24, 3613-3629. Available: DOI 10.1105/tpc.112.099945.
- Casañas, A., Querol-Audí, J., Guerra, P., Pous, J., Tanaka, H., Tsukihara, T., Verdaguer, N. & Fita, I. 2013. New features of vault architecture and dynamics revealed by novel refinement using the deformable elastic network approach. *Acta Crystallographica Section D: Biological Crystallography*, 69, 1054-1061. Available: DOI 10.1107/s0907444913004472.
- Cerutti, H. & Casas-Mollano, J. A. 2006. On the origin and functions of RNA-mediated silencing: From protists to man. *Current Genetics*, 50, 81-99. Available: DOI 10.1007/s00294-006-0078-x.
- Chak, L.-L. & Okamura, K. 2014. Argonaute-dependent small RNAs derived from single-stranded, non-structured precursors. *Frontiers in Genetics*, 5, 172. Available: DOI 10.3389/fgene.2014.00172.
- Champion, C. I., Kickhoefer, V. A., Liu, G., Moniz, R. J., Freed, A. S., Bergmann, L. L., Vaccari, D., Raval-Fernandes, S., Chan, A. M., Rome, L. H. & Kelly, K. A. 2009. A vault nanoparticle vaccine induces protective mucosal immunity. *Plos One*, 4.
- Chen, Y. P., Pettis, J. S., Zhao, Y., Liu, X., Tallon, L. J., Sadzewicz, L. D., Li, R., Zheng, H., Huang, S., Zhang, X., Hamilton, M. C., Pernal, S. F., Melathopoulos, A. P., Yan, X. & Evans, J. D. 2013. Genome sequencing and comparative genomics of honey bee microsporidia, *Nosema apis* reveal novel insights into host-parasite interactions. *BMC Genomics*, 14. Available: DOI 10.1186/1471-2164-14-451.
- Chendrimada, T. P., Finn, K. J., Ji, X., Baillat, D., Gregory, R. I., Liebhaber, S. A., Pasquinelli, A. E. & Shiekhattar, R. 2007. MicroRNA silencing through RISC recruitment of eIF6. *Nature*, 447, 823-8. Available: DOI 10.1038/nature05841.
- Chothia, C. 1992. One thousand families for the molecular biologist. *Nature*, 357, 543-544. Available: DOI 10.1038/357543a0.
- Chugani, D. C., Kedersha, N. L. & Rome, L. H. 1991. Vault immunofluorescence in the brain: New insights regarding the origin of microglia. *Journal of Neuroscience*, 11, 256-268.
- Chugani, D. C., Rome, L. H. & Kedersha, N. L. 1993. Evidence that vault ribonucleoprotein particles localize to the nuclear pore complex. *Journal of Cell Science*, 106, 23-29.
- Collins, L. & Penny, D. 2005. Complex spliceosomal organization ancestral to extant eukaryotes. *Molecular Biology and Evolution*, 22, 1053-1066. Available: DOI 10.1093/molbev/msi091.
- Collins, L. J. 2011. *THE RNA INFRASTRUCTURE: An Introduction to ncRNA Networks*. Landes Bioscience and Springer Science+Business Media.
- Collins, L. J., Poole, A. M. & Penny, D. 2003a. Using ancestral sequences to uncover potential gene homologues. *Appl Bioinformatics*, 2, S85-95.
- Collins, L. J., Poole, A. M. & Penny, D. 2003b. Using ancestral sequences to uncover potential gene homologues. *Appl Bioinformatics*, 2.
- Consortium, T. U. 2015. UniProt: a hub for protein information. *Nucleic Acids Research*, 43, D204-D212. Available: DOI 10.1093/nar/gku989.
- Copeland, C. S., Marz, M., Rose, D., Hertel, J., Brindley, P. J., Santana, C. B., Kehr, S., Attolini, C. S.-O. & Stadler, P. F. 2009. Homology-based annotation of non-coding RNAs in the genomes of *Schistosoma mansoni* and *Schistosoma japonicum*. *BMC Genomics*, 10, 1-13. Available: DOI 10.1186/1471-2164-10-464.
- Cora, E., Pandey, R. R., Xiol, J., Taylor, J., Sachidanandam, R., McCarthy, A. A. & Pillai, R. S. 2014. The MID-PIWI module of Piwi proteins specifies nucleotide- and strand-biases of piRNAs. *RNA*, 20, 773-781. Available: DOI 10.1261/rna.044701.114.
- Cornman, R. S., Chen, Y. P., Schatz, M. C., Street, C., Zhao, Y., Desany, B., Egholm, M., Hutchison, S., Pettis, J. S., Lipkin, W. I. & Evans, J. D. 2009. Genomic analyses of the microsporidian *Nosema ceranae*, an emergent pathogen of honey bees. *PLoS Pathogens*, 5. Available: DOI 10.1371/journal.ppat.1000466.
- Couvillion, M., Bounova, G., Purdom, E., Speed, T. & Collins, K. 2012. A *Tetrahymena* Piwi bound to mature tRNA 3' fragments activates the exonuclease Xrn2 for RNA processing in the nucleus. *Molecular Cell*, 48, 509-520.



- Cox, D. N., Chao, A., Baker, J., Chang, L., Qiao, D. & Lin, H. 1998. A novel class of evolutionarily conserved genes defined by piwi are essential for stem cell self-renewal. *Genes and Development*, 12, 3715-3727.
- Czech, B. & Hannon, G. J. 2016. One Loop to Rule Them All: The Ping-Pong Cycle and piRNA-Guided Silencing. *Trends in Biochemical Sciences*. Available: DOI 10.1016/j.tibs.2015.12.008.
- Daly, T., Chen, X. S. & Penny, D. 2011. How old are RNA networks. *Advances in Experimental Medicine and Biology*, 722, 255-273.
- Daly, T. K., Sutherland-Smith, A. J. & Penny, D. 2013a. Beyond BLASTing: Tertiary and Quaternary Structure Analysis Helps Identify Major Vault Proteins. *Genome Biology and Evolution*, 5, 217-232. Available: DOI 10.1093/gbe/evs135.
- Daly, T. K., Sutherland-Smith, A. J. & Penny, D. 2013b. *In silico* resurrection of the Major Vault Protein suggests It Is ancestral in modern eukaryotes. *Genome Biology and Evolution*, 5, 1567-1583. Available: DOI 10.1093/gbe/evt113.
- Dalzell, J. J., McVeigh, P., Warnock, N. D., Mitreva, M., Bird, D. M., Abad, P., Fleming, C. C., Day, T. A., Mousley, A., Marks, N. J. & Maule, A. G. 2011. RNAi effector diversity in nematodes. *PLoS Neglected Tropical Diseases*, 5. Available: DOI 10.1371/journal.pntd.0001176.
- DaRocha, W. D., Otsu, K., Teixeira, S. M. R. & Donelson, J. E. 2004. Tests of cytoplasmic RNA interference (RNAi) and construction of a tetracycline-inducible T7 promoter system in *Trypanosoma cruzi*. *Molecular and Biochemical Parasitology*, 133, 175-186. Available: DOI 10.1016/j.molbiopara.2003.10.005.
- Deng, X. W., Caspar, T. & Quail, P. H. 1991. cop1: a regulatory locus involved in light-controlled development and gene expression in *Arabidopsis*. *Genes and Development*, 5, 1172-1182.
- Dickenson, N. E., Moore, D., Suprenant, K. A. & Dunn, R. C. 2007. Vault ribonucleoprotein particles and the central mass of the nuclear pore complex. *Photochemistry and Photobiology*, 83, 686-691.
- Diestra, J. E., Condom, E., García del Muro, X., Scheffer, G. L., Pérez, J., Zurita, A. J., Muñoz-Segui, J., Vigués, F., Scheper, R. J., Capellá, G., Germà-Lluch, J. R. & Izquierdo, M. A. 2003. Expression of multidrug resistance proteins p-glycoprotein, multidrug resistance protein 1, breast cancer resistance protein and lung resistance related protein in locally advanced bladder cancer treated with neoadjuvant chemotherapy: Biological and clinical implications. *Journal of Urology*, 170, 1383-1387.
- Dornan, D., Bheddah, S., Newton, K., Inice, W., Frantz, G. D., Dowd, P., Koeppen, H., Dixit, V. M. & French, D. M. 2004. COP1, the negative regulator of p53, is overexpressed in breast and ovarian adenocarcinomas. *Cancer Research*, 64, 7226-7230. Available: DOI 10.1158/0008-5472.can-04-2601.
- Drinnenberg, I. A., Fink, G. R. & Bartel, D. P. 2011. Compatibility with killer explains the rise of RNAi-deficient fungi. *Science*, 333, 1592.
- Edgar, R. C. 2004a. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *Bmc Bioinformatics*, 5, 113-113. Available: DOI 10.1186/1471-2105-5-113.
- Edgar, R. C. 2004b. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32, 1792-1797.
- Eichenmüller, B., Kedersha, N., Solovyeva, E., Everley, P., Lang, J., Himes, R. H. & Suprenant, K. A. 2003. Vaults bind directly to microtubules via their caps and not their barrels. *Cell Motility and the Cytoskeleton*, 56, 225-236.
- El-Sayed, N. M., Myler, P. J., Bartholomeu, D. C., Nilsson, D., Aggarwal, G., Tran, A. N., Ghedin, E., Worthey, E. A., Delcher, A. L., Blandin, G., Westenberger, S. J., Caler, E., Cerqueira, G. C., Branche, C., Haas, B., Anupama, A., Arner, E., Åslund, L., Attipoe, P., Bontempi, E., Bringaud, F., Burton, P., Cadag, E., Campbell, D. A., Carrington, M., Crabtree, J., Darban, H., Da Silveira, J. F., De Jong, P., Edwards, K., Englund, P. T., Fazelina, G., Feldblyum, T., Ferella, M., Frasch, A. C., Gull, K., Horn, D., Hou, L., Huang, Y., Kindlund, E., Klingbeil, M., Kluge, S., Koo, H., Lacerda, D., Levin, M. J., Lorenzi, H., Louie, T., Machado, C. R., McCulloch, R., McKenna, A., Mizuno, Y., Mottram, J. C., Nelson, S., Ochaya, S., Osoegawa, K., Pai, G., Parsons, M., Pentony, M., Pettersson, U., Pop, M., Ramirez, J. L., Rinta, J., Robertson, L., Salzberg, S. L., Sanchez, D. O., Seyler, A., Sharma, R., Shetty, J., Simpson, A. J., Sisk, E., Tammi, M. T., Tarleton, R., Teixeira, S., Van Aken, S., Vogt, C., Ward, P. N., Wickstead, B., Wortman, J., White, O., Fraser, C. M., Stuart, K. D. & Andersson, B. 2005. The genome sequence of *Trypanosoma cruzi*, etiologic agent of chagas disease. *Science*, 309, 409-415+435. Available: DOI 10.1126/science.1112631.
- Fabian, M. R., Sonenberg, N. & Filipowicz, W. 2010. Regulation of mRNA translation and stability by microRNAs. *Annual Review of Biochemistry*.

- Faehnle, C. R., Elkayam, E., Haase, A. D., Hannon, G. J. & Joshua-Tor, L. 2013. The Making of a Slicer: Activation of Human Argonaute-1. *Cell Reports*, 3, 1901-1909. Available: DOI 10.1016/j.celrep.2013.05.033.
- Fahlgren, N., Bollmann, S. R., Kasschau, K. D., Cuperus, J. T., Press, C. M., Sullivan, C. M., Chapman, E. J., Hoyer, J. S., Gilbert, K. B., Grönwald, N. J. & Carrington, J. C. 2013. *Phytophthora* have distinct endogenous small RNA populations that include short interfering and microRNAs. *Plos One*, 8, e77181.
- Farazi, T. A., Juranek, S. A. & Tuschl, T. 2008. The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members. *Development*, 135, 1201-1214. Available: DOI 10.1242/dev.005629.
- Farrow, A. L., Rana, T., Mittal, M. K., Misra, S. & Chaudhuri, G. 2011. *Leishmania*-induced repression of selected non-coding RNA genes containing B-box element at their promoters in alternatively polarized M2 macrophages. *Molecular and Cellular Biochemistry*, 350, 47-57.
- Feng, X. Z. & Guang, S. H. 2013. Non-coding RNAs mediate the rearrangements of genomic DNA in ciliates. *Science China Life Sciences*, 56, 937-943. Available: DOI 10.1007/s11427-013-4539-4.
- Fernandez-Fuentes, N., Dybas, J. M. & Fiser, A. 2010. Structural Characteristics of Novel Protein Folds. *PLoS Comput Biol*, 6, e1000750.
- Finn, R. D., Coghill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J. & Bateman, A. 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, 44, D279-D285. Available: DOI 10.1093/nar/gkv1344.
- Finn, R. D., Mistry, J., Tate, J., Coghill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L. L., Eddy, S. R. & Bateman, A. 2010. The Pfam protein families database. *Nucleic Acids Research*, 38, D211-D222. Available: DOI 10.1093/nar/gkp985.
- Forterre, P. 2015. The universal tree of life: An update. *Frontiers in Microbiology*, 6. Available: DOI 10.3389/fmicb.2015.00717.
- Franzellitti, S., Capuzzo, A., Viarengo, A. & Fabbri, E. 2011. Interactive effects of nickel and chlorpyrifos on Mediterranean mussel cAMP-mediated cell signaling and MXR-related gene expressions. *Comparative Biochemistry and Physiology - C Toxicology and Pharmacology*, 154, 377-382.
- Franzén, C. 2005. How do microsporidia invade cells? *Folia Parasitologica*, 52, 36-40.
- Franzén, O., Arner, E., Ferella, M., Nilsson, D., Respuela, P., Carninci, P., Hayashizaki, Y., Åslund, L., Andersson, B. & Daub, C. O. 2011. The short non-coding transcriptome of the protozoan parasite *Trypanosoma cruzi*. *PLoS Neglected Tropical Diseases*, 5.
- Fritz-Laylin, L. K., Prochnik, S. E., Ginger, M. L., Dacks, J. B., Carpenter, M. L., Field, M. C., Kuo, A., Paredez, A., Chapman, J., Pham, J., Shu, S., Neupane, R., Cipriano, M., Mancuso, J., Tu, H., Salamov, A., Lindquist, E., Shapiro, H., Lucas, S., Grigoriev, I. V., Cande, W. Z., Fulton, C., Rokhsar, D. S. & Dawson, S. C. 2010. The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. *Cell*, 140, 631-642.
- Gao, F., Roy, S. W. & Katz, L. A. 2015. Analyses of alternatively processed genes in ciliates provide insights into the origins of scrambled genomes and may provide a mechanism for speciation. *mBio*, 6. Available: DOI 10.1128/mBio.01998-14.
- Garcia Silva, M. R., Frugier, M., Tosar, J. P., Correa-Dominguez, A., Ronalte-Alves, L., Parodi-Talice, A., Rovira, C., Robello, C., Goldenberg, S. & Cayota, A. 2010a. A population of tRNA-derived small RNAs is actively produced in *Trypanosoma cruzi* and recruited to specific cytoplasmic granules. *Molecular and Biochemical Parasitology*, 171, 64-73.
- Garcia Silva, M. R., Tosar, J. P., Frugier, M., Pantano, S., Bonilla, B., Esteban, L., Serra, E., Rovira, C., Robello, C. & Cayota, A. 2010b. Cloning, characterization and subcellular localization of a *Trypanosoma cruzi* argonaute protein defining a new subfamily distinctive of trypanosomatids. *Gene*, 466, 26-35.
- Garcia-Silva, M. R., Frugier, M., Tosar, J. P., Correa-Dominguez, A., Ronalte-Alves, L., Parodi-Talice, A., Rovira, C., Robello, C., Goldenberg, S. & Cayota, A. 2010. A population of tRNA-derived small RNAs is actively produced in *Trypanosoma cruzi* and recruited to specific cytoplasmic granules. *Molecular and Biochemical Parasitology*, 171, 64-73.
- Gilbert, W. 1986. Origin of life: The RNA world. *Nature*, 319, 618. Available: DOI 10.1038/319618a0.
- Goldsmith, L. E., Pupols, M., Kickhoefer, V. A., Rome, L. H. & Monbouquette, H. G. 2009. Utilization of a protein "shuttle" to load Vault nanocapsules with gold probes and proteins. *ACS Nano*, 3, 3175-3183. Available: DOI 10.1021/nn900555d.

- Grimson, A., Srivastava, M., Fahey, B., Woodcroft, B. J., Chiang, H. R., King, N., Degan, B. M., Rokhsar, D. S. & Bartel, D. P. 2008. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature*, 455, 1193-1197. Available: DOI 10.1038/nature07415.
- Grützmann, K., Szafranski, K., Pohl, M., Voigt, K., Petzold, A. & Schuster, S. 2014. Fungal alternative splicing is associated with multicellular complexity and virulence: A genome-wide multi-species study. *DNA Research*, 21, 27-39. Available: DOI 10.1093/dnares/dst038.
- Gullberg, M., Tolf, C., Jonsson, N., Mulders, M. N., Savolainen-Kopra, C., Hovi, T., Van Ranst, M., Lemey, P., Hafenstein, S. & Lindberg, A. M. 2010. Characterization of a putative ancestor of coxsackievirus B5. *Journal of Virology*, 84, 9695-9708.
- Haag, K. L., James, T. Y., Pombert, J.-F., Larsson, R., Schaer, T. M. M., Refardt, D. & Ebert, D. 2014. Evolution of a morphological novelty occurred before genome compaction in a lineage of extreme parasites. *Proceedings of the National Academy of Sciences of the United States of America*, 111, 15480-15485. Available: DOI 10.1073/pnas.1410442111.
- Halic, M. & Moazed, D. 2010. Dicer-Independent Primal RNAs Trigger RNAi and Heterochromatin Formation. *Cell*, 140, 504-516. Available: DOI 10.1016/j.cell.2010.01.019.
- Hall, B. G. 2005. Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. *Molecular Biology and Evolution*, 22, 792-802. Available: DOI 10.1093/molbev/msi066.
- Hamill, D. R. & Suprenant, K. A. 1997. Characterization of the sea urchin major vault protein: A possible role for vault ribonucleoprotein particles in nucleocytoplasmic transport. *Developmental Biology*, 190, 117-128. Available: DOI 10.1006/dbio.1997.8676.
- Hansen, T. B., Venø, M. T., Jensen, T. I., Schaefer, A., Damgaard, C. K. & Kjems, J. 2016. Argonaute-associated short introns are a novel class of gene regulators. *Nat Commun*, 7. Available: DOI 10.1038/ncomms11538.
- Hanson-Smith, V., Kolaczowski, B. & Thornton, J. W. 2010. Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. *Molecular Biology and Evolution*, 27, 1988-1999. Available: DOI 10.1093/molbev/msq081.
- Harvey, J. J. W., Lewsey, M. G., Patel, K., Westwood, J., Heimstädt, S., Carr, J. P. & Baulcombe, D. C. 2011. An antiviral defense role of AGO2 in plants. *Plos One*, 6. Available: DOI 10.1371/journal.pone.0014639.
- Helbo, A., Søgaaard, Treppendahl, M., Aslan, D., Dimopoulos, K., Nandrup-Bus, C., Holm, M. S., Andersen, M. K., Liang, G., Kristensen, L. S. & Kirsten, G. 2015. Hypermethylation of the VTRNA1-3 promoter is associated with poor outcome in lower risk myelodysplastic syndrome patients. *Genes*, 6, 977-990. Available: DOI 10.3390/genes6040977.
- Henikoff, S. & Henikoff, J. G. 1992. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89, 10915-10919.
- Herlevsen, M., Oxford, G., Owens, C. R., Conaway, M. & Theodorescu, D. 2007. Depletion of major vault protein increases doxorubicin sensitivity and nuclear accumulation and disrupts its sequestration in lysosomes. *Molecular Cancer Therapeutics*, 6, 1804-1813.
- Herrmann, C., Zimmermann, H. & Volknandt, W. 1997. Analysis of a cDNA encoding the major vault protein from the electric ray *Discopyge ommata*. *Gene*, 188, 85-90. Available: DOI 10.1016/S0378-1119(96)00781-0.
- Hertel, J., de Jong, D., Marz, M., Rose, D., Tafer, H., Tanzer, A., Schierwater, B. & Stadler, P. F. 2009. Non-coding RNA annotation of the genome of *Trichoplax adhaerens*. *Nucleic Acids Research*, 37, 1602-1615. Available: DOI 10.1093/nar/gkn1084.
- Hobbs, J. K., Shepherd, C., Saul, D. J., Demetras, N. J., Haaning, S., Monk, C. R., Daniel, R. M. & Arcus, V. L. 2012. On the origin and evolution of thermophily: Reconstruction of functional precambrian enzymes from ancestors of *Bacillus*. *Molecular Biology and Evolution*, 29, 825-835. Available: DOI 10.1093/molbev/msr253.
- Holm, L. & Sander, C. 1997. An evolutionary treasure: Unification of a broad set of amidohydrolases related to urease. *Proteins-Structure Function and Genetics*, 28, 72-82. Available: DOI 10.1002/(sici)1097-0134(199705)28:1.
- Hu, H. D., Ye, F., Zhang, D. Z., Hu, P., Ren, H. & Li, S. L. 2010. ITRAQ quantitative analysis of multidrug resistance mechanisms in human gastric cancer cells. *Journal of Biomedicine and Biotechnology*, 2010. Available: DOI 10.1155/2010/571343.
- Hu, Y., Stenlid, J., Elfstrand, M. & Olson, A. 2013. Evolution of RNA interference proteins dicer and argonaute in Basidiomycota. *Mycologia*, 105, 1489-1498.
- Huang, T. & Zhang, X. 2012. Contribution of the Argonaute-1 Isoforms to Invertebrate Antiviral Defense. *Plos One*, 7. Available: DOI 10.1371/journal.pone.0050581.

- Huelsenbeck, J. P. & Ronquist, F. 2001. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics*, 17, 754-755.
- Huffman, K. E. & Corey, D. R. 2005. Major vault protein does not play a role in chemoresistance or drug localization in a non-small cell lung cancer cell line. *Biochemistry*, 44, 2253-2261.
- Hur, J. K., Zinchenko, M. K., Djuranovic, S. & Green, R. 2013. Regulation of Argonaute slicer activity by guide RNA 3'-end interactions with the N-terminal lobe. *Journal of Biological Chemistry*, 288, 7829-7840. Available: DOI 10.1074/jbc.M112.441030.
- Illergård, K., Ardell, D. H. & Elofsson, A. 2009. Structure is three to ten times more conserved than sequence—A study of structural response in protein cores. *Proteins: Structure, Function, and Bioinformatics*, 77, 499-508. Available: DOI 10.1002/prot.22458.
- Irimia, M., Rukov, J. L., Penny, D. & Roy, S. W. 2007. Functional and evolutionary analysis of alternatively spliced genes is consistent with an early eukaryotic origin of alternative splicing. *Bmc Evolutionary Biology*, 7. Available: DOI 10.1186/1471-2148-7-188.
- Itou, D., Shiromoto, Y., Shin-ya, Y., Ishii, C., Nishimura, T., Ogonuki, N., Ogura, A., Hasuwa, H., Fujihara, Y., Kuramochi-Miyagawa, S. & Nakano, T. 2015. Induction of DNA methylation by artificial piRNA production in male germ cells. *Current Biology*, 25, 901-906. Available: DOI 10.1016/j.cub.2015.01.060.
- Izquierdo, M. A., Scheffer, G. L., Flens, M. J., Giaccone, G., Broxterman, H. J., Meijer, C. J. M., van der Valk, P. & Scheper, R. J. 1996. Broad distribution of the multi-drug resistance-related vault lung resistance in normal human tissues and tumors. *American Journal of Pathology* 148, 877-887.
- Jekely, G., Paps, J. & Nielsen, C. 2015. The phylogenetic position of ctenophores and the origin(s) of nervous systems. *Evodevo*, 6. Available: DOI 110.1186/2041-9139-6-1.
- Juliano, C., Wang, J. & Lin, H. 2011. Uniting germline and stem cells: the function of Piwi proteins and the piRNA pathway in diverse organisms. *Annual review of genetics*, 45, 10.1146/annurev-genet-110410-132541. Available: DOI 10.1146/annurev-genet-110410-132541.
- Kar, U. K., Srivastava, M. K., Andersson, Ö., Baratelli, F., Huang, M., Kickhoefer, V. A., Dubinett, S. M., Rome, L. H. & Sharma, S. 2011. Novel CCL21-Vault nanocapsule intratumoral delivery inhibits lung cancer growth. *PLoS ONE*, 6, e18758.
- Kawaji, H., Nakamura, M., Takahashi, Y., Sandelin, A., Katayama, S., Fukuda, S., Daub, C. O., Kai, C., Kawai, J., Yasuda, J., Carninci, P. & Hayashizaki, Y. 2008. Hidden layers of human small RNAs. *BMC Genomics*, 9. Available: DOI 10.1186/1471-2164-9-157.
- Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P. & Drummond, A. 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28, 1647-1649. Available: DOI 10.1093/bioinformatics/bts199.
- Kedersha, N. L. & Rome, L. H. 1986. Isolation and characterization of a novel ribonucleoprotein particle - large structures contain a single species of small RNA. *Journal of Cell Biology*, 103, 699-709.
- Keeling, P. J. 2004. Diversity and evolutionary history of plastids and their hosts. *American Journal of Botany*, 91, 1481-1493.
- Keeling, P. J. 2009. Chromalveolates and the evolution of plastids by secondary endosymbiosis1. *Journal of Eukaryotic Microbiology*, 56, 1-8. Available: DOI 10.1111/j.1550-7408.2008.00371.x.
- Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protocols*, 10, 845-858. Available: DOI 10.1038/nprot.2015.053.
- Kersey, P. J., Allen, J. E., Armean, I., Boddu, S., Bolt, B. J., Carvalho-Silva, D., Christensen, M., Davis, P., Falin, L. J., Grabmueller, C., Humphrey, J., Kerhornou, A., Khobova, J., Aranganathan, N. K., Langridge, N., Lowy, E., McDowall, M. D., Maheswari, U., Nuhn, M., Ong, C. K., Overduin, B., Paulini, M., Pedro, H., Perry, E., Spudich, G., Tapanari, E., Walts, B., Williams, G., Tello-Ruiz, M., Stein, J., Wei, S., Ware, D., Bolser, D. M., Howe, K. L., Kulesha, E., Lawson, D., Maslen, G. & Staines, D. M. 2015. Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Research*.
- Kickhoefer, V. A., Garcia, Y., Mikyas, Y., Johansson, E., Zhou, J. C., Raval-Fernandes, S., Minoofar, P., Zink, J. I., Dunn, B., Stewart, P. L. & Rome, L. H. 2005. Engineering of vault nanocapsules with enzymatic and fluorescent properties. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 4348-4352.
- Kickhoefer, V. A., Han, M., Raval-Fernandes, S., Poderycki, M. J., Moniz, R. J., Vaccari, D., Silvestry, M., Stewart, P. L., Kelly, K. A. & Rome, L. H. 2009. Targeting Vault nanoparticles to specific cell surface receptors. *ACS Nano*, 3, 27-36. Available: DOI 10.1021/nn800638x.



- Kickhoefer, V. A., Liu, Y., Kong, L. B., Snow, B. E., Stewart, P. L., Harrington, L. & Rome, L. H. 2001. The telomerase/vault-associated protein TEP1 is required for vault RNA stability and its association with the vault particle. *Journal of Cell Biology*, 152, 157-164. Available: DOI 10.1083/jcb.152.1.157.
- Kickhoefer, V. A., Rajavel, K. S., Scheffer, G. L., Dalton, W. S., Scheper, R. J. & Rome, L. H. 1998. Vaults are up-regulated in multidrug-resistant cancer cell lines. *Journal of Biological Chemistry*, 273, 8971-8974. Available: DOI 10.1074/jbc.273.15.8971.
- Kickhoefer, V. A., Searles, R. P., Kedersha, N. L., Garber, M. E., Johnson, D. L. & Rome, L. H. 1993. Vault ribonucleoprotein particles from rat and bullfrog contain a related small RNA that is transcribed by RNA polymerase III. *Journal of Biological Chemistry*, 268, 7868-7873.
- Kickhoefer, V. A., Siva, A. C., Kedersha, N. L., Inman, E. M., Ruland, C., Streuli, M. & Rome, L. H. 1999a. The 193-kD vault protein, VPARP, is a novel poly(ADP-ribose) polymerase. *Journal of Cell Biology*, 146, 917-928. Available: DOI 10.1083/jcb.146.5.917.
- Kickhoefer, V. A., Stephen, A. G., Harrington, L., Robinson, M. O. & Rome, L. H. 1999b. Vaults and telomerase share a common subunit, TEP1. *Journal of Biological Chemistry*, 274, 32712-32717. Available: DOI 10.1074/jbc.274.46.32712.
- Kim, W. K., Eamens, A. L. & Waterhouse, P. M. 2011. RNA processing activities of the *Arabidopsis* Argonaute protein family. In: Grabowski, P. P. (ed.) *RNA Processing*. Pittsburg: In Tech.
- Kitazono, M., Sumizawa, T., Takebayashi, Y., Chen, Z. S., Furukawa, T., Nagayama, S., Tani, A., Takao, S., Aikou, T. & Akiyama, S. I. 1999. Multidrug resistance and the lung resistance-related protein in human colon carcinoma SW-620 cells. *Journal of the National Cancer Institute*, 91, 1647-1653.
- Klattenhoff, C. & Theurkauf, W. 2008. Biogenesis and germline functions of piRNAs. *Development*, 135, 3-9. Available: DOI 10.1242/dev.006486.
- Kolli, S., Zito, C. I., Mossink, M. H., Wiemer, E. A. C. & Bennett, A. M. 2004. The major vault protein is a novel substrate for the tyrosine phosphatase SHP-2 and scaffold protein in epidermal growth factor signaling. *Journal of Biological Chemistry*, 279, 29374-29385.
- Kong, L., Hao, Q., Wang, Y., Zhou, P., Zou, B. & Zhang, Y. X. 2015. Regulation of p53 expression and apoptosis by vault RNA2-1-5p in cervical cancer cells. *Oncotarget*, 6, 28371-28388. Available: DOI 10.18632/oncotarget.4948.
- Kong, L. B., Siva, A. C., Rome, L. H. & Stewart, P. L. 1999. Structure of the vault, a ubiquitous cellular component. *Structure*, 7, 371-379.
- Kosloff, M. & Kolodny, R. 2008. Sequence-similar, structure-dissimilar protein pairs in the PDB. *Proteins*, 71, 891-902. Available: DOI 10.1002/prot.21770.
- Kowalski, M. P., Dubouix-Bourandy, A., Bajmoczy, M., Golan, D. E., Zaidi, T., Coutinho-Sledge, Y. S., Gygi, M. P., Gygi, S. P., Wiemer, E. A. C. & Pier, G. B. 2007. Host resistance to lung infection mediated by major vault protein in epithelial cells. *Science*, 317, 130-132.
- Kruger, K., Grabowski, P. J., Zaug, A. J., Sands, J., Gottschling, D. E. & Cech, T. R. 1982. Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. *Cell*, 31, 147-157. Available: DOI 10.1016/0092-8674(82)90414-7 [Accessed 2016/02/08].
- Kumar, P., Anaya, J., Mudunuri, S. B. & Dutta, A. 2014. Meta-analysis of tRNA derived RNA fragments reveals that they are evolutionarily conserved and associate with AGO proteins to recognize specific RNA targets. *BMC Biology*, 12. Available: DOI 10.1186/s12915-014-0078-0.
- Kunkeaw, N., Jeon, S. H., Lee, K., Johnson, B. H., Tanasanvimon, S., Javle, M., Pairojkul, C., Chamgramol, Y., Wongfieng, W., Gong, B., Leelayuwat, C. & Lee, Y. S. 2013. Cell death/proliferation roles for nc886, a non-coding RNA, in the protein kinase R pathway in cholangiocarcinoma. *Oncogene*, 32, 3722-3731. Available: DOI 10.1038/nc.2012.382.
- Ladewig, E., Okamura, K., Flynt, A. S., Westholm, J. O. & Lai, E. C. 2012. Discovery of hundreds of mirtrons in mouse and human small RNA data. *Genome Research*, 22, 1634-1645. Available: DOI 10.1101/gr.133553.111.
- Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. 2001. Identification of novel genes coding for small expressed RNAs. *Science*, 294, 853-858. Available: DOI 10.1126/science.1064921.
- Lai, C. Y., Wiethoff, C. M., Kickhoefer, V. A., Rome, L. H. & Nemerow, G. R. 2009. Vault nanoparticles containing an adenovirus-derived membrane lytic protein facilitate toxin and gene transfer. *ACS Nano*, 3, 691-699. Available: DOI 10.1021/nn8008504.
- Landgraf, P., Rusu, M., Sheridan, R., Sewer, A., Iovino, N., Aravin, A., Pfeffer, S., Rice, A., Kamphorst, A. O., Landthaler, M., Lin, C., Succi, N. D., Hermida, L., Fulci, V., Chiaretti, S., Foa, R., Schliwka, J., Fuchs, U., Novosel, A., Mueller, R.-U., Schermer, B., Bissels, U., Inman, J., Phan, Q., Chien, M., Weir, D. B., Choksi, R., De Vita, G., Frezzetti, D., Trompeter, H.-I., Hornung, V.,

- Teng, G., Hartmann, G., Palkovits, M., Di Lauro, R., Wernet, P., Macino, G., Rogler, C. E., Nagle, J. W., Ju, J., Papavasiliou, F. N., Benzing, T., Lichter, P., Tam, W., Brownstein, M. J., Bosio, A., Borkhardt, A., Russo, J. J., Sander, C., Zavolan, M. & Tuschl, T. 2007. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, 129, 1401-1414. Available: DOI 10.1016/j.cell.2007.04.040.
- Lange, C., Walther, W., Schwabe, H. & Stein, U. 2000. Cloning and initial analysis of the human multidrug resistance-related MVP/LRP gene promoter. *Biochemical and Biophysical Research Communications*, 278, 125-133. Available: DOI.doi.org/10.1006/bbrc.2000.3782.
- Lara, P. C., Lloret, M., Clavo, B., Apolinario, R. M., Henriquez-Hernandez, L. A., Bordon, E., Fontes, F. & Rey, A. 2009. Severe hypoxia induces chemo-resistance in clinical cervical tumors through MVP over-expression. *Radiation Oncology*, 4. Available: DOI 29 10.1186/1748-717x-4-29.
- Lara, P. C., Pruschy, M., Zimmermann, M. & Henriquez-Hernandez, L. A. 2011. MVP and Vaults: A role in the radiation response. *Radiation Oncology*, 148.
- Lazarowski, A. & Czornyj, L. 2011. Potential role of multidrug resistant proteins in refractory epilepsy and antiepileptic drugs interactions. *Drug Metabolism and Drug Interactions*, 26, 21-26.
- Lee, H. S., Lee, K., Jang, H. J., Lee, G. K., Park, J. L., Kim, S. Y., Kim, S. B., Johnson, B. H., Zo, J. I., Lee, J. S. & Lee, Y. S. 2014. Epigenetic silencing of the non-coding RNA nc886 provokes oncogenes during human esophageal tumorigenesis. *Oncotarget*, 5, 3472-3481.
- Lee, K., Kunkeaw, N., Jeon, S. H., Lee, I., Johnson, B. H., Kang, G.-Y., Bang, J. Y., Park, H. S., Leelayuwat, C. & Lee, Y. S. 2011. Precursor miR-886, a novel noncoding RNA repressed in cancer, associates with PKR and modulates its activity. *RNA*, 17, 1076-1089. Available: DOI 10.1261/rna.2701111.
- Lee, Y. S., Shibata, Y., Malhotra, A. & Dutta, A. 2009. A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes & Development*, 23, 2639-2649. Available: DOI 10.1101/gad.1837609.
- Li, J. Y., Volkandt, W., Dahlstrom, A., Herrmann, C., Blasi, J., Das, B. & Zimmermann, H. 1999. Axonal transport of ribonucleoprotein particles (vaults). *Neuroscience*, 91, 1055-1065.
- Light, S., Sagit, R., Sachenkova, O., Ekman, D. & Elovsson, A. 2013. Protein Expansion Is Primarily due to Indels in Intrinsically Disordered Regions. *Molecular Biology and Evolution*, 30, 2645-2653.
- Lim, R. S. M., Anand, A., Nishimiya-Fujisawa, C., Kobayashi, S. & Kai, T. 2014. Analysis of *Hydra* PIWI proteins and piRNAs uncover early evolutionary origins of the piRNA pathway. *Developmental Biology*, 386, 237-251. Available: DOI 10.1016/j.ydbio.2013.12.007.
- Lippman, Z. & Martienssen, R. 2004. The role of RNA interference in heterochromatic silencing. *Nature*, 431, 364-370. Available: DOI 10.1038/nature02875.
- Liu, K., Chen, C., Guo, Y., Lam, R., Bian, C., Xu, C., Zhao, D. Y., Jin, J., MacKenzie, F., Pawson, T. & Min, J. 2010. Structural basis for recognition of arginine methylated Piwi proteins by the extended Tudor domain. *Proceedings of the National Academy of Sciences of the United States of America*, 107, 18398-18403. Available: DOI 10.1073/pnas.1013106107.
- Liu, Y. J., Hodson, M. C. & Hall, B. D. 2006. Loss of the flagellum happened only once in the fungal lineage: phylogenetic structure of Kingdom Fungi inferred from RNA polymerase II subunit genes. *Bmc Evolutionary Biology*, 6, 74-74. Available: DOI 10.1186/1471-2148-6-74.
- Loss-Morais, G., Waterhouse, P. M. & Margis, R. 2013. Description of plant tRNA-derived RNA fragments (tRFs) associated with argonaute and identification of their putative targets. *Biology Direct*, 8. Available: DOI 10.1186/1745-6150-8-6.
- Löytynoja, A. & Goldman, N. 2010. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *Bmc Bioinformatics*, 11, 1-7. Available: DOI 10.1186/1471-2105-11-579.
- Luby-Phelps, K. 2000. Cytoarchitecture and physical properties of cytoplasm: Volume, viscosity, diffusion, intracellular surface area. *International Review of Cytology*, 192, 189-221.
- Lyskov, S., Chou, F. C., Conchúr, S. Ó., Der, B. S., Drew, K., Kuroda, D., Xu, J., Weitzner, B. D., Renfrew, P. D., Sripakdeevong, P., Borgo, B., Havranek, J. J., Kuhlman, B., Kortemme, T., Bonneau, R., Gray, J. J. & Das, R. 2013. Serverification of molecular modeling applications: The Rosetta Online Server that Includes Everyone (ROSIE). *Plos One*, 8. Available: DOI 10.1371/journal.pone.0063906.
- Lyskov, S. & Gray, J. J. 2008. The RosettaDock server for local proteinprotein docking. *Nucleic Acids Research*, 36, 233-238. Available: DOI 10.1093/nar/gkn216.
- Majorek, K. A., Dunin-Horkawicz, S., Steczkiewicz, K., Muszewska, A., Nowotny, M., Ginalski, K. & Bujnicki, J. M. 2014. The RNase H-like superfamily: new members, comparative structural analysis and evolutionary classification. *Nucleic Acids Research*, 42, 4160-4179. Available: DOI 10.1093/nar/gkt1414.



- Makarova, K. S., Wolf, Y. I., van der Oost, J. & Koonin, E. V. 2009. Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements. *Biology Direct*, 4, 29. Available: DOI 10.1186/1745-6150-4-29.
- Mándoky, L., Géczi, L., Doleschall, Z., Bodrogi, I., Csuka, O., Kásler, M. & Bak, M. 2004. Expression and prognostic value of the lung resistance-related protein (LRP) in germ cell Testicular Tumors. *Anticancer Research*, 24, 1097-1104.
- Mariscal, C. & Doolittle, W. F. 2015. Eukaryotes first: How could that be? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370. Available: DOI 10.1098/rstb.2014.0322.
- Martin, F., Kohler, A., Murat, C., Balestrini, R., Coutinho, P. M., Jaillon, O., Montanini, B., Morin, E., Noel, B., Percudani, R., Porcel, B., Rubini, A., Amicucci, A., Amselem, J., Anthouard, V., Arcioni, S., Artiguenave, F., Aury, J. M., Ballario, P., Bolchi, A., Brenna, A., Brun, A., Buée, M., Cantarel, B., Chevalier, G., Couloux, A., Da Silva, C., Denoeud, F., Duplessis, S., Ghignone, S., Hilselberger, B., Iotti, M., Marçais, B., Mello, A., Miranda, M., Pacioni, G., Quesneville, H., Riccioni, C., Ruotolo, R., Splivallo, R., Stocchi, V., Tisserant, E., Viscomi, A. R., Zambonelli, A., Zampieri, E., Henrissat, B., Lebrun, M. H., Paolocci, F., Bonfante, P., Ottonello, S. & Wincker, P. 2010. Périgord black truffle genome uncovers evolutionary origins and mechanisms of symbiosis. *Nature*, 464, 1033-1038. Available: DOI 10.1038/nature08867.
- Martin, W. F., Garg, S. & Zimorski, V. 2015. Endosymbiotic theories for eukaryote origin. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370. Available: DOI 10.1098/rstb.2014.0330.
- Matsumoto, N. M., Prabhakaran, P., Rome, L. H. & Maynard, H. D. 2013. Smart vaults: Thermally-responsive protein nanocapsules. *Acs Nano*, 7, 867-874. Available: DOI 10.1021/nn3053457.
- Matsumoto, T., Tanaka, T., Sakai, H., Amano, N., Kanamori, H., Kurita, K., Kikuta, A., Kamiya, K., Yamamoto, M., Ikawa, H., Fujii, N., Hori, K., Itoh, T. & Sato, K. 2011. Comprehensive sequence analysis of 24,783 Barley full-length cDNAs derived from 12 clone libraries. *Plant Physiology*, 156, 20-28. Available: DOI 10.1104/pp.110.171579.
- Matunis, M. J., Coutavas, E. & Blobel, G. 1996. A novel ubiquitin-like modification modulates the partitioning of the Ran-GTPase-activating protein RanGAP1 between the cytosol and the nuclear pore complex. *Journal of Cell Biology*, 135, 1457-1470.
- Mochizuki, K. & Gorovsky, M. A. 2004. Conjugation-specific small RNAs in *Tetrahymena* have predicted properties of scan (scn) RNAs involved in genome rearrangement. *Genes and Development*, 18, 2068-2073. Available: DOI 10.1101/gad.1219904.
- Mossel, E. & Steel, M. 2004. A phase transition for a random cluster model on phylogenetic trees. *Mathematical Biosciences*, 187, 189-203.
- Mossink, M. H., Van Zon, A., Fränzel-Luiten, E., Schoester, M., Kickhoefer, V. A., Scheffer, G. L., Scheper, R. J., Sonneveld, P. & Wiemer, E. A. C. 2002a. Disruption of the murine major vault protein (MVP/LRP) gene does not induce hypersensitivity to cytostatics. *Cancer Research*, 62, 7298-7304.
- Mossink, M. H., Van Zon, A., Fränzel-Luiten, E., Schoester, M., Scheffer, G. L., Scheper, R. J., Sonneveld, P. & Wiemer, E. A. C. 2002b. The genomic sequence of the murine major vault protein and its promoter. *Gene*, 294, 225-232.
- Moult, J., Pedersen, J. T., Judson, R. & Fidelis, K. 1995. A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function and Genetics*, 23, ii-iv.
- Mrazek, J., Toso, D., Ryazantsev, S., Zhang, X., Zhou, Z. H., Fernandez, B. C., Kickhoefer, V. A. & Rome, L. H. 2014. Polyribosomes are molecular 3-D nanoprinters that orchestrate the assembly of vault particles. *Acs Nano*, 8, 11552-11559. Available: DOI 10.1021/nn504778h.
- Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. 1995. SCOP - A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247, 536-540. Available: DOI 10.1006/jmbi.1995.0159.
- Musilova, K. & Mraz, M. 2015. MicroRNAs in B-cell lymphomas: how a complex biology gets more complex. *Leukemia*, 29, 1004-1017. Available: DOI 10.1038/leu.2014.351.
- Nakanishi, K., Ascano, M., Gogakos, T., Ishibe-Murakami, S., Serganov, A. A., Briskin, D., Morozov, P., Tuschl, T. & Patel, D. J. 2013. Eukaryote-specific insertion elements control human Argonaute slicer activity. *Cell Reports*, 3, 1893-1900.
- Nakanishi, K., Weinberg, D. E., Bartel, D. P. & Patel, D. J. 2012. Structure of yeast Argonaute with guide RNA. *Nature*, 486, 368. Available: DOI 10.1038/nature11211.
- Nandy, C., Mrazek, J., Stoiber, H., Grasser, F. A., Huttenhofer, A. & Polacek, N. 2009. Epstein-Barr Virus-Induced Expression of a Novel Human Vault RNA. *Journal of Molecular Biology*, 388, 776-784. Available: DOI 10.1016/j.jmb.2009.03.031.

- Ng, B. C., Yu, M., Gopal, A., Rome, L. H., Monbouquette, H. G. & Tolbert, S. H. 2008. Encapsulation of semiconducting polymers in vault protein cages. *Nano Letters*, 8, 3503-3509.
- Ngô, H., Tschudi, C., Gull, K. & Ullu, E. 1998. Double-stranded RNA induces mRNA degradation in *trypanosoma brucei*. *Proceedings of the National Academy of Sciences of the United States of America*, 95, 14687-14692.
- Nowotny, M., Gaidamakov, S. A., Crouch, R. J. & Yang, W. 2005. Crystal structures of RNase H bound to an RNA/DNA hybrid: Substrate specificity and metal-dependent catalysis. *Cell*, 121, 1005-1016. Available: DOI 10.1016/j.cell.2005.04.024.
- Obara, S., Iwataki, Y. & Mikami, K. 2000. Identification of a possible stem-cell-maintenance gene homologue in the unicellular eukaryote *Paramecium caudatum*. *Proceedings of the Japan Academy, Series B*, 76, 57-62. Available: DOI 10.2183/pjab.76.57.
- Olsen, P. H. & Ambros, V. 1999. The lin-4 regulatory RNA controls developmental timing in *Caenorhabditis elegans* by blocking LIN-14 protein synthesis after the initiation of translation. *Developmental Biology*, 216, 671-680. Available: DOI 10.1006/dbio.1999.9523.
- Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. 1997. CATH - a hierarchic classification of protein domain structures. *Structure*, 5, 1093-1108. Available: DOI 10.1016/S0969-2126(97)00260-8.
- Ortiz-Rivas, B., Jaubert-Possamai, S., Tanguy, S., Gauthier, J. P., Tagu, D. & Claude, R. 2012. Evolutionary study of duplications of the miRNA machinery in aphids associated with striking rate acceleration and changes in expression profiles. *Bmc Evolutionary Biology*, 12.
- Palakodeti, D., Smielewska, M., Lu, Y. C., Yeo, G. W. & Graveley, B. R. 2008. The PIWI proteins SMEDWI-2 and SMEDWI-3 are required for stem cell function and piRNA expression in planarians. *RNA*, 14, 1174-1186. Available: DOI 10.1261/rna.1085008.
- Parfrey, L. W., Grant, J., Tekle, Y. I., Lasek-Nesselquist, E., Morrison, H. G., Sogin, M. L., Patterson, D. J. & Katz, L. A. 2010. Broadly sampled multigene analyses yield a well-resolved eukaryotic tree of life. *Systematic Biology*, 59, 518-533.
- Paspalas, C. D., Perley, C. C., Venkitaramani, D. V., Goebel-Goody, S. M., Zhang, Y. F., Kurup, P., Mattis, J. H. & Lombroso, P. J. 2009. Major Vault Protein is expressed along the nucleus-neurite axis and associates with mRNAs in cortical neurons. *Cerebral Cortex*, 19, 1666-1677. Available: DOI 10.1093/cercor/bhn203.
- Pelin, A., Selman, M., Aris-Brosou, S., Farinelli, L. & Corradi, N. 2015. Genome analyses suggest the presence of polyploidy and recent human-driven expansions in eight global populations of the honeybee pathogen *Nosema ceranae*. *Environmental Microbiology*, 17, 4443-4458. Available: DOI 10.1111/1462-2920.12883.
- Penny, D., Collins, L. J., Daly, T. K. & Cox, S. J. 2014. The Relative Ages of Eukaryotes and Akaryotes. *Journal of Molecular Evolution*, 79, 228-239. Available: DOI 10.1007/s00239-014-9643-y.
- Persson, H., Kvist, A., Vallon-Christersson, J., Medstrand, P., Borg, A. & Rovira, C. 2009. The non-coding RNA of the multidrug resistance-linked vault particle encodes multiple regulatory small RNAs. *Nature Cell Biology*, 11, 1268-U265. Available: DOI 10.1038/ncb1972.
- Poderycki, M. J., Kickhoefer, V. A., Kaddis, C. S., Raval-Fernandes, S., Johansson, E., Zink, J. I., Loo, J. A. & Rome, L. H. 2006. The vault exterior shell is a dynamic structure that allows incorporation of vault-associated proteins into its interior. *Biochemistry*, 45, 12184-12193.
- Prucca, C. G., Slavin, I., Quiroga, R., Elías, E. V., Rivero, F. D., Saura, A., Carranza, P. G. & Luján, H. D. 2008. Antigenic variation in *Giardia lamblia* is regulated by RNA interference. *Nature*, 456, 750-754. Available: DOI 10.1038/nature07585.
- Prusiner, S. B., Scott, M. R., DeArmond, S. J. & Cohen, F. E. 1998. Prion protein biology. *Cell*, 93, 337-348. Available: DOI 10.1016/S0092-8674(00)81163-0 [Accessed 2016/03/14].
- Qian, W., He, X., Chan, E., Xu, H. & Zhang, J. 2011. Measuring the evolutionary rate of protein-protein interaction. *Proceedings of the National Academy of Sciences*. Available: DOI 10.1073/pnas.1104695108.
- Quinn, J. J. & Chang, H. Y. 2016. Unique features of long non-coding RNA biogenesis and function. *Nature Reviews Genetics*, 17, 47-62. Available: DOI 10.1038/nrg.2015.10.
- Rajasethupathy, P., Antonov, I., Sheridan, R., Frey, S., Sander, C., Tuschl, T. & Kandel, E. R. 2012. A role for neuronal piRNAs in the epigenetic control of memory-related synaptic plasticity. *Cell*, 149, 693-707. Available: DOI 10.1016/j.cell.2012.02.057.
- Reddy, T. B. K., Thomas, A. D., Stamatis, D., Bertsch, J., Isbandi, M., Jansson, J., Mallajosyula, J., Pagani, I., Lobos, E. A. & Kyrpides, N. C. 2014. The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. *Nucleic Acids Research*, 43, D1099-1106. Available: DOI 10.1093/nar/gku950.

- Reis, E. V., Pereira, R. V., Gomes, M., Jannotti-Passos, L. K., Baba, E. H., Coelho, P. M. Z., Mattos, A. C. A., Couto, F. F. B., Castro-Borges, W. & Guerra-Sá, R. 2014. Characterisation of major vault protein during the life cycle of the human parasite *Schistosoma mansoni*. *Parasitology International*, 63, 120-126. Available: DOI 10.1016/j.parint.2013.10.005.
- Rome, L. H. & Kickhoefer, V. A. 2013. Development of the Vault particle as a platform technology. *Acs Nano*, 7, 889-902. Available: DOI 10.1021/nn3052082.
- Ross, R. J., Weiner, M. M. & Lin, H. 2014. PIWI proteins and PIWI-interacting RNAs in the soma. *Nature*, 505, 353-359.
- Roy, A., Kucukural, A. & Zhang, Y. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. *Nature protocols*, 5, 725-738.
- Saito, K., Sakaguchi, Y., Suzuki, T., Siomi, H. & Siomi, M. C. 2007. Pimet, the Drosophila homolog of HEN1, mediates 2'-O-methylation of Piwi-interacting RNAs at their 3'- ends. *Genes and Development*, 21, 1603-1608. Available: DOI 10.1101/gad.1563607.
- Schadendorf, D., Makki, A., Stahr, C., Van Dyck, A., Wanner, R., Scheffer, G. L., Fiens, M. J., Scheper, R. & Henz, B. M. 1995. Membrane transport proteins associated with drug resistance expressed in human melanoma. *American Journal of Pathology*, 147, 1545-1552.
- Scheffer, G. L., Schroeijers, A. B., Izquierdo, M. A., Wiemer, E. A. C. & Scheper, R. J. 2000. Lung resistance-related protein/major vault protein and vaults in multidrug-resistant cancer. *Current Opinion in Oncology*, 12, 550-556.
- Schirle, N. T. & MacRae, I. J. 2012. The crystal structure of human argonaute2. *Science*, 336, 1037-1040. Available: DOI 10.1126/science.1221551.
- Schirle, N. T., Sheu-Gruttadauria, J. & MacRae, I. J. 2014. Structural basis for microRNA targeting. *Science*, 346, 608-613. Available: DOI 10.1126/science.1258040.
- Schürmann, N., Trabuco, L. G., Bender, C., Russell, R. B. & Grimm, D. 2013. Molecular dissection of human Argonaute proteins by DNA shuffling. *Nature Structural and Molecular Biology*, 20, 818-826.
- Sędziewska, K. A., Fuchs, J., Temsch, E. M., Baronian, K., Watzke, R. & Kunze, G. 2011. Estimation of the *Glomus intraradices* nuclear DNA content. *New Phytologist*, 192, 794-797.
- Shimamoto, Y., Sumizawa, T., Haraguchi, M., Gotanda, T., Jueng, H. C., Furukawa, T., Sakata, R. & Akiyama, S. 2006. Direct activation of the human major vault protein gene by DNA-damaging agents. *Oncology reports*, 15, 645-652.
- Sierra, R., Matz, M. V., Aglyamova, G., Pillet, L., Decelle, J., Not, F., de Vargas, C. & Pawlowski, J. 2013. Deep relationships of Rhizaria revealed by phylogenomics: A farewell to Haeckel's Radiolaria. *Molecular Phylogenetics and Evolution*, 67, 53-59. Available: DOI 10.1016/j.ympev.2012.12.011.
- Signorovitch, A. Y., Dellaporta, S. L. & Buss, L. W. 2005. Molecular signatures for sex in the Placozoa. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 15518-15522. Available: DOI 10.1073/pnas.0504031102.
- Sigova, A., Rhind, N. & Zamore, P. D. 2004. A single Argonaute protein mediates both transcriptional and posttranscriptional silencing in *Schizosaccharomyces pombe*. *Genes & Development*, 18, 2359-2367. Available: DOI 10.1101/gad.1218004.
- Simon, B., Kirkpatrick, J. P., Eckhardt, S., Reuter, M., Rocha, E. A., Andrade-Navarro, M. A., Sehr, P., Pillai, R. S. & Carlomagno, T. 2011. Recognition of 2'-O-Methylated 3'-end of piRNA by the PAZ domain of a Piwi protein. *Structure*, 19, 172-180. Available: DOI 10.1016/j.str.2010.11.015 [Accessed 2016/05/13].
- Skinner, D. E., Rinaldi, G., Koziol, U., Brehm, K. & Brindley, P. J. 2014. How might flukes and tapeworms maintain genome integrity without a canonical piRNA pathway? *Trends in Parasitology*, 30, 123-129. Available: DOI 10.1016/j.pt.2014.01.001.
- Spanu, P. D., Abbott, J. C., Amselem, J., Burgis, T. A., Soanes, D. M., Stüber, K., Van Themaat, E. V. L., Brown, J. K. M., Butcher, S. A., Gurr, S. J., Lebrun, M. H., Ridout, C. J., Schulze-Lefert, P., Talbot, N. J., Ahmadinejad, N., Ametz, C., Barton, G. R., Benjdia, M., Bidzinski, P., Bindschedler, L. V., Both, M., Brewer, M. T., Cadle-Davidson, L., Cadle-Davidson, M. M., Collemare, J., Cramer, R., Frenkel, O., Godfrey, D., Harriman, J., Hoede, C., King, B. C., Klages, S., Kleemann, J., Knoll, D., Koti, P. S., Kreplak, J., López-Ruiz, F. J., Lu, X., Maekawa, T., Mahanil, S., Micali, C., Milgroom, M. G., Montana, G., Noir, S., O'Connell, R. J., Oberhaensli, S., Parlange, F., Pedersen, C., Quesneville, H., Reinhardt, R., Rott, M., Sacristán, S., Schmidt, S. M., Schön, M., Skamnioti, P., Sommer, H., Stephens, A., Takahara, H., Thordal-Christensen, H., Vigouroux, M., Wessling, R., Wicker, T. & Panstruga, R. 2010. Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. *Science*, 330, 1543-1546.

- Srivastava, M., Begovic, E., Chapman, J., Putnam, N. H., Hellsten, U., Kawashima, T., Kuo, A., Mitros, T., Salamov, A., Carpenter, M. L., Signorovitch, A. Y., Moreno, M. A., Kamm, K., Grimwood, J., Schmutz, J., Shapiro, H., Grigoriev, I. V., Buss, L. W., Schierwater, B., Dellaporta, S. L. & Rokhsar, D. S. 2008. The *Trichoplax* genome and the nature of placozoans. *Nature*, 454, 955-960.
- Stadler, P. F., Chen, J. J. L., Hackermüller, J., Hoffmann, S., Horn, F., Khaitovich, P., Kretzschmar, A. K., Mosig, A., Prohaska, S. J., Qi, X., Schutt, K. & Ullmann, K. 2009a. Evolution of Vault RNAs. *Molecular Biology and Evolution*, 26.
- Stadler, P. F., Chen, J. J. L., Hackermüller, J., Hoffmann, S., Horn, F., Khaitovich, P., Kretzschmar, A. K., Mosig, A., Prohaska, S. J., Qi, X. D., Schutt, K. & Ullmann, K. 2009b. Evolution of Vault RNAs. *Molecular Biology and Evolution*, 26, 1975-1991. Available: DOI 10.1093/molbev/msp112.
- Steiner, E., Holzmann, K., Pirker, C., Elbling, L., Micksche, M., Sutterlüty, H. & Berger, W. 2006. The major vault protein is responsive to and interferes with interferon-gamma-mediated STAT1 signals. *Journal of Cell Science*, 119, 459-469.
- Stephen, A. G., Raval-Fernandes, S., Huynh, T., Torres, M., Kickhoefer, V. A. & Rome, L. H. 2001. Assembly of Vault-like particles in insect cells expressing only the Major Vault Protein. *Journal of Biological Chemistry*, 276, 23217-23220.
- Stewart, P. L., Makabi, M., Lang, J., Dickey-Sims, C., Robertson, A. J., Coffman, J. A. & Suprenant, K. A. 2005. Sea urchin vault structure, composition, and differential localization during development. *BMC Developmental Biology*, 5.
- Suk, K., Choi, J., Suzuki, Y., Ozturk, S. B., Mellor, J. C., Wong, K. H., MacKay, J. L., Gregory, R. I. & Roth, F. P. 2011. Reconstitution of human RNA interference in budding yeast. *Nucleic Acids Research*, 39. Available: DOI 10.1093/nar/gkq1321.
- Suprenant, K. A. 2002. Vault ribonucleoprotein particles: Sarcophagi, gondolas, or safety deposit boxes? *Biochemistry*, 41, 14447-14454.
- Suprenant, K. A., Bloom, N., Fang, J. W. & Lushington, G. 2007. The major vault protein is related to the toxic anion resistance protein (TelA) family. *Journal of Experimental Biology*, 210, 946-955. Available: DOI 10.1242/jeb.001800.
- Swarts, D. C., Hegge, J. W., Hinojo, I., Shiimori, M., Ellis, M. A., Dumrongkulraksa, J., Terns, R. M., Terns, M. P. & Van Der Oost, J. 2015. Argonaute of the archaeon *Pyrococcus furiosus* is a DNA-guided nuclease that targets cognate DNA. *Nucleic Acids Research*, 43, 5120-5129. Available: DOI 10.1093/nar/gkv415.
- Swarts, D. C., Jore, M. M., Westra, E. R., Zhu, Y., Janssen, J. H., Snijders, A. P., Wang, Y., Patel, D. J., Berenguer, J., Brouns, S. J. J. & Van Der Oost, J. 2014. DNA-guided DNA interference by a prokaryotic Argonaute. *Nature*, 507, 258-261. Available: DOI 10.1038/nature12971.
- Tahir, M., Law, D. A. & Stasolla, C. 2006. Molecular characterization of PgAGO, a novel conifer gene of the ARGONAUTE family expressed in apical cells and required for somatic embryo development in spruce. *Tree Physiology*, 26, 1257-1270.
- Takeda, A., Iwasaki, S., Watanabe, T., Utsumi, M. & Watanabe, Y. 2008. The mechanism selecting the guide strand from small RNA duplexes is different among Argonaute proteins. *Plant and Cell Physiology*, 49, 493-500. Available: DOI 10.1093/pcp/pcn043.
- Tam, O. H., Aravin, A. A., Stein, P., Girard, A., Murchison, E. P., Cheloufi, S., Hodges, E., Anger, M., Sachidanandam, R., Schultz, R. M. & Hannon, G. J. 2008. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature*, 453, 534-538.
- Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. & Kumar, S. 2011. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, 28, 2731-2739.
- Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. 2013. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Molecular Biology and Evolution*, 30, 2725-2729. Available: DOI 10.1093/molbev/mst197.
- Tanaka, H., Kato, K., Yamashita, E., Sumizawa, T., Zhou, Y., Yao, M., Iwasaki, K., Yoshimura, M. & Tsukihara, T. 2009. The Structure of Rat Liver Vault at 3.5 Angstrom Resolution. *Science*, 323, 384-388. Available: DOI 10.1126/science.1164975.
- Telonis, A. G., Loher, P., Honda, S., Jing, Y., Palazzo, J., Kirino, Y. & Rigoutsos, I. 2015. Dissecting tRNA-derived fragment complexities using personalized transcriptomes reveals novel fragment classes and unexpected dependencies. *Oncotarget*, 6, 24797-24822. Available: DOI 10.18632/oncotarget.4695.
- Tews, D. S., Nissen, A., Külgen, C. & Gaumann, A. K. A. 2000. Drug resistance-associated factors in primary and secondary glioblastomas and their precursor tumors. *Journal of Neuro-Oncology*, 50, 227-237.



- Thompson, D. M., Lu, C., Green, P. J. & Parker, R. 2008. tRNA cleavage is a conserved response to oxidative stress in eukaryotes. *RNA*, 14, 2095-2103.
- Thompson, J. D., Linard, B., Lecompte, O. & Poch, O. 2011. A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods: Current Challenges and Future Perspectives. *Plos One*, 6, e18093.
- Tian, Y., Simanshu, D. K., Ma, J.-B. & Patel, D. J. 2011a. Structural basis for piRNA 2'-O-methylated 3'-end recognition by Piwi PAZ (Piwi/Argonaute/Zwille) domains. *Proceedings of the National Academy of Sciences of the United States of America*, 108, 903-910. Available: DOI 10.1073/pnas.1017762108.
- Tian, Y., Simanshu, D. K., Ma, J. B. & Patel, D. J. 2011b. Structural basis for piRNA 2'-O-methylated 3'-end recognition by Piwi PAZ (Piwi/Argonaute/Zwille) domains. *Proceedings of the National Academy of Sciences of the United States of America*, 108, 903-910. Available: DOI 10.1073/pnas.1017762108.
- Tisserant, E., Malbreil, M., Kuo, A., Kohler, A., Symeonidi, A., Balestrini, R., Charron, P., Duensing, N., Frei Dit Frey, N., Gianinazzi-Pearson, V., Gilbert, L. B., Handa, Y., Herr, J. R., Hijri, M., Koul, R., Kawaguchi, M., Krajinski, F., Lammers, P. J., Masclaux, F. G., Murat, C., Morin, E., Ndikumana, S., Pagni, M., Petitpierre, D., Requena, N., Rosikiewicz, P., Riley, R., Saito, K., San Clemente, H., Shapiro, H., Van Tuinen, D., Bécard, G., Bonfante, P., Paszkowski, U., Shachar-Hill, Y. Y., Tuskan, G. A., Young, P. W., Sanders, I. R., Henrissat, B., Rensing, S. A., Grigoriev, I. V., Corradi, N., Roux, C. & Martin, F. 2013. Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis. *Proceedings of the National Academy of Sciences of the United States of America*, 110, 20117-20122.
- Tolia, N. H. & Joshua-Tor, L. 2007. Slicer and the Argonautes. *Nature Chemical Biology*, 3, 36-43. Available: DOI 10.1038/nchembio848.
- Treppendahl, M. B., Qiu, X., Søggaard, A., Yang, X., Nandrup-Bus, C., Hother, C., Andersen, M. K., Kjeldsen, L., Möllgaard, L., Hellström-Lindberg, E., Jendholm, J., Porse, B. T., Jones, P. A., Liang, G. & Grønbaek, K. 2012. Allelic methylation levels of the noncoding VTRNA2-1 located on chromosome 5q31.1 predict outcome in AML. *Blood*, 119, 206-216. Available: DOI 10.1182/blood-2011-06-362541.
- Tsaousis, A. D., Kunji, E. R. S., Goldberg, A. V., Lucocq, J. M., Hirt, R. P. & Embley, T. M. 2008. A novel route for ATP acquisition by the remnant mitochondria of *Encephalitozoon cuniculi*. *Nature*, 453, 553-U11. Available: DOI 10.1038/nature06903.
- Vagin, V. V., Wohlschlegel, J., Qu, J., Jonsson, Z., Huang, X., Chuma, S., Girard, A., Sachidanandam, R., Hannon, G. J. & Aravin, A. A. 2009. Proteomic analysis of murine Piwi proteins reveals a role for arginine methylation in specifying interaction with Tudor family members. *Genes & Development*, 23, 1749-1762. Available: DOI 10.1101/gad.1814809.
- van Zon, A., Mossink, M. H., Schoester, M., Houtsmuller, A. B., Scheffer, G. L., Scheper, R. J., Sonneveld, P. & Wiemer, E. A. C. 2003. The formation of vault-tubes: A dynamic interaction between vaults and vault PARP. *Journal of Cell Science*, 116, 4391-4400.
- Van Zon, A., Mossink, M. H., Schoester, M., Scheffer, G. L., Scheper, R. J., Sonneveld, P. & Wiemer, E. A. C. 2001. Multiple human vault RNAs: Expression and association with the vault complex. *Journal of Biological Chemistry*, 276, 37715-37721. Available: DOI 10.1074/jbc.M106055200.
- van Zon, A., Mossink, M. H., Schoester, M., Scheper, R. J., Sonneveld, P. & Wiemer, E. A. C. 2004. Efflux kinetics and intracellular distribution of daunorubicin are not affected by major vault protein/lung resistance-related protein (vault) expression. *Cancer Research*, 64, 4887-4892. Available: DOI 10.1158/0008-5472.can-03-3891.
- Vastenhouw, N. L., Fischer, S. E. J., Robert, V. J. P., Thijssen, K. L., Fraser, A. G., Kamath, R. S., Ahringer, J. & Plasterk, R. H. A. 2003. A genome-wide screen identifies 27 genes involved in transposon silencing in *C. elegans*. *Current Biology*, 13, 1311-1316. Available: DOI 10.1016/s0960-9822(03)00539-6.
- Vasu, S. K., Kedersha, N. L. & Rome, L. H. 1993. cDNA cloning and disruption of the major vault protein alpha gene (mvpA) in *Dictyostelium discoideum*. *Journal of Biological Chemistry*, 268, 15356-15360.
- Vasu, S. K. & Rome, L. H. 1995. *Dictyostelium* vaults: Disruption of the major proteins reveals growth and morphological defects and uncovers a new associated protein. *Journal of Biological Chemistry*, 270, 16588-16594.
- Vaucheret, H. 2008. Plant ARGONAUTES. *Trends in Plant Science*, 13, 350-358. Available: DOI 10.1016/j.tplants.2008.04.007.
- Veidenberg, A., Medlar, A. & Löytynoja, A. 2015. Wasabi: An integrated platform for evolutionary sequence analysis and data visualization. *Molecular Biology and Evolution*.

- Venkatesh, T., Suresh, P. S. & Tsutsumi, R. 2016. TRFs: miRNAs in disguise. *Gene*, 579, 133-138. Available: DOI 10.1016/j.gene.2015.12.058.
- Vollmar, F., Hacker, C., Zahedi, R. P., Sickmann, A., Ewald, A., Scheer, U. & Dabauvalle, M. C. 2009. Assembly of nuclear pore complexes mediated by major vault protein. *Journal of Cell Science*, 122, 780-786. Available: DOI 10.1242/jcs.039529.
- Wang, B., Collins Iii, J. J. & Newmark, P. A. 2013. Functional genomic characterization of neoblast-like stem cells in larval *Schistosoma mansoni*. *eLife*, 2013. Available: DOI 10.7554/eLife.00768.
- Wang, M., Abad, D., Kickhoefer, V. A., Rome, L. H. & Mahendra, S. 2015. Vault Nanoparticles Packaged with Enzymes as an Efficient Pollutant Biodegradation Technology. *Acs Nano*, 9, 10931-10940. Available: DOI 10.1021/acsnano.5b04073.
- Watson, J. D., Laskowski, R. A. & Thornton, J. M. 2005. Predicting protein function from sequence and structural data. *Current Opinion in Structural Biology*, 15, 275-284. Available: DOI 10.1016/j.sbi.2005.04.003.
- Wei, H., Zhou, B., Zhang, F., Tu, Y., Hu, Y., Zhang, B. & Zhai, Q. 2013. Profiling and identification of small rDNA-derived RNAs and their potential biological functions. *Plos One*, 8, e56842. Available: DOI 10.1371/journal.pone.0056842.
- Wei, K. F., Wu, L. J., Chen, J., Chen, Y. F. & Xie, D. X. 2012. Structural evolution and functional diversification analyses of argonaute protein. *Journal of Cellular Biochemistry*, 113, 2576-2585.
- Weiberg, A., Wang, M., Lin, F.-M., Zhao, H., Zhang, Z., Kaloshian, I., Huang, H.-D. & Jin, H. 2013. Fungal small RNAs suppress plant immunity by hijacking host RNA interference pathways. *Science*, 342, 118-123. Available: DOI 10.1126/science.1239705.
- Weiss, L. M., Edlind, T. D., Vossbrinck, C. R. & Hashimoto, T. 1999. Microsporidian molecular phylogeny: The fungal connection. *Journal of Eukaryotic Microbiology*, 46, 17S-18S.
- Wicker, T., Oberhaensli, S., Parlange, F., Buchmann, J. P., Shatalina, M., Roffler, S., Ben-David, R., Dolezel, J., Simkova, H., Schulze-Lefert, P., Spanu, P. D., Bruggmann, R., Amselem, J., Quesneville, H., van Themaat, E. V. L., Paape, T., Shimizu, K. K. & Keller, B. 2013. The wheat powdery mildew genome shows the unique evolution of an obligate biotroph. *Nature Genetics*, 45, 1092-+. Available: DOI 10.1038/ng.2704.
- Wiemer, E. A., Van Zon, A., Mossink, M. H., Scheffer, G. L., Scheper, R. J. & Sonneveld, P. 2004. Vaults and drug resistance: What we learn from an MVP/LRP knockout mouse model. *AACR Meeting Abstracts*, 2004, 568-a-.
- Williams, P. D., Pollock, D. D., Blackburne, B. P. & Goldstein, R. A. 2006. Assessing the accuracy of ancestral protein reconstruction methods. *PLoS Computational Biology*, 2, 0598-0605.
- Wu, S. & Zhang, Y. 2007. LOMETS: A local meta-threading-server for protein structure prediction. *Nucleic Acids Research*, 35, 3375-3382.
- Xu, D., Zhang, J., Roy, A. & Zhang, Y. 2011. Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. *Proteins: Structure, Function, and Bioinformatics*, 79, 147-160. Available: DOI 10.1002/prot.23111.
- Yang, G. D., Huang, T. J., Peng, L. X., Yang, C. F., Liu, R. Y., Huang, H. B., Chu, Q. Q., Yang, H. J., Huang, J. L., Zhu, Z. Y., Qian, C. N. & Huang, B. J. 2013. Epstein-Barr Virus Encoded LMP1 Upregulates MicroRNA-21 to Promote the Resistance of Nasopharyngeal Carcinoma Cells to Cisplatin-Induced Apoptosis by Suppressing PDCD4 and Fas-L. *Plos One*, 8. Available: DOI 10.1371/journal.pone.0078355.
- Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J. & Zhang, Y. 2015. The I-TASSER suite: Protein structure and function prediction. *Nature Methods*, 12, 7-8. Available: DOI 10.1038/nmeth.3213.
- Yang, Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, 24, 1586-1591.
- Ye, Y. & Godzik, A. 2003. Flexible structure alignment by chaining aligned fragment pairs allowing twists. *Bioinformatics*, 19, ii246-ii255.
- Yi, C., Li, S., Chen, X., Wiemer, E. A. C., Wang, J., Wei, N. & Deng, X. W. 2005. Major vault protein, in concert with constitutively photomorphogenic 1, negatively regulates c-Jun-mediated activator protein 1 transcription in mammalian cells. *Cancer Research*, 65, 5835-5840.
- Yigit, E., Batista, P. J., Bei, Y., Pang, K. M., Chen, C.-C. G., Tolia, N. H., Joshua-Tor, L., Mitani, S., Simard, M. J. & Mello, C. C. 2006. Analysis of the *C. elegans* argonaute family reveals that distinct argonautes act sequentially during RNAi. *Cell*, 127, 747-757. Available: DOI 10.1016/j.cell.2006.09.033.
- Yu, Z., Chen, D., Su, Z., Li, Y., Yu, W., Zhang, Q., Yang, L., Li, C., Yang, S., Ni, L., Gui, Y., Mao, Z. & Lai, Y. 2014. MiR-886-3p upregulation in clear cell renal cell carcinoma regulates cell migration,

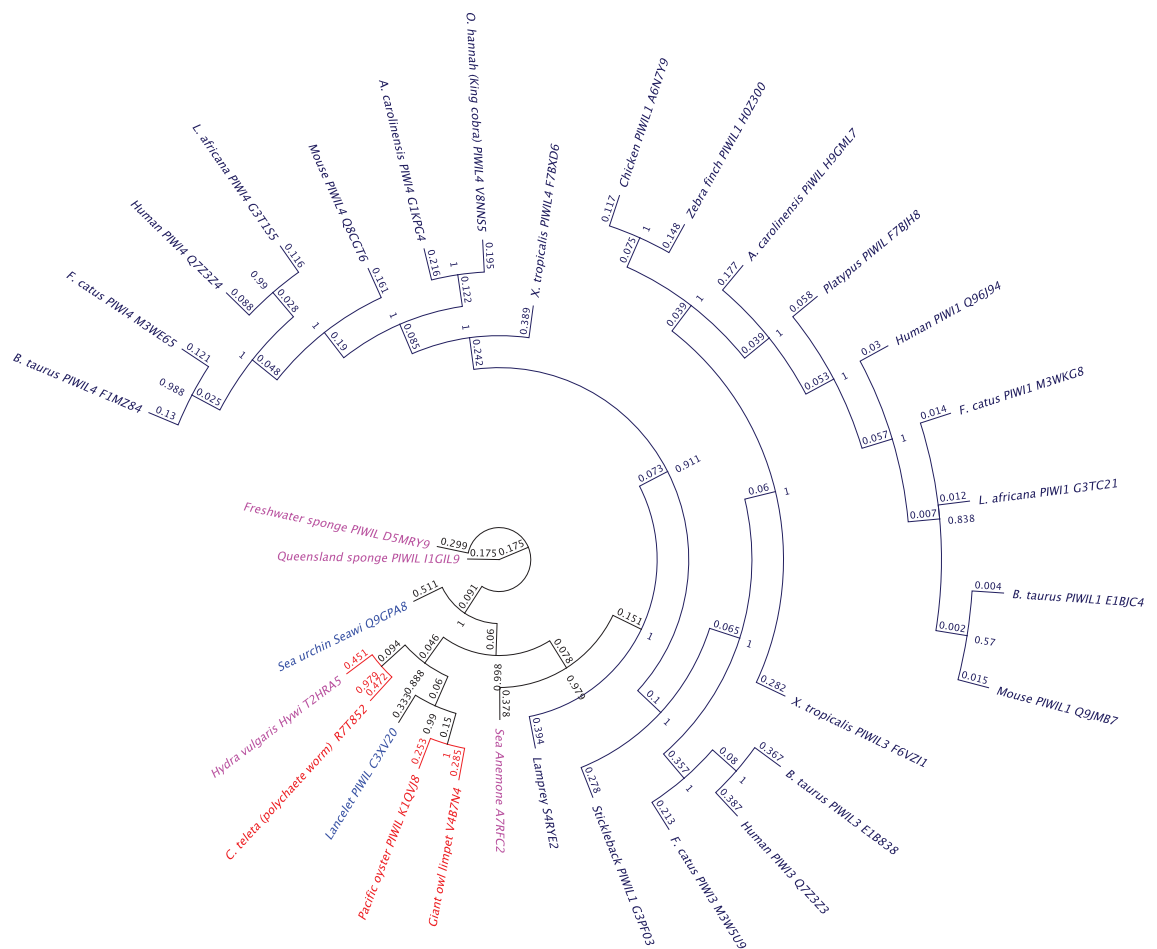


- proliferation and apoptosis by targeting PITX1. *International Journal of Molecular Medicine*, 34, 1409-1416. Available: DOI 10.3892/ijmm.2014.1923.
- Zamore, P. D. & Haley, B. 2005. Ribo-gnome: The big world of small RNAs. *Science*, 309, 1519-1524. Available: DOI 10.1126/science.1111444.
- Zhang, H., Alramini, H., Tran, V. & Singh, U. 2011. Nucleus-localized antisense small RNAs with 5'-polyphosphate termini regulate long term transcriptional gene silencing in *Entamoeba histolytica* G3 strain. *Journal of Biological Chemistry*, 286, 44467-44479. Available: DOI 10.1074/jbc.M111.278184.
- Zhang, Y. 2008. I-TASSER server for protein 3-D structure prediction. *BMC Bioinformatics*, 9.
- Zhang, Z., Liu, X., Guo, X., Wang, X.-J. & Zhang, X. 2016. *Arabidopsis* AGO3 predominantly recruits 24-nt small RNAs to regulate epigenetic silencing. *Nature Plants*, 2, 16049. Available: DOI 10.1038/nplants.2016.49 nature.com/articles/nplants201649#supplementary-information.
- Zhenbao, Fotouhi-Ardakani, N., Wu, L., Maoui, M., Wang, S., Banville, D. & Shen, S. H. 2002. PTEN associates with the vault particles in HeLa cells. *Journal of Biological Chemistry*, 277, 40247-40252.
- Zheng, Y. 2013. Phylogenetic analysis of the Argonaute protein family in platyhelminths. *Molecular Phylogenetics and Evolution*, 66, 1050-1054.
- Zurita, A. J., Diestra, J. E., Condom, E., García del Muro, X., Scheffer, G. L., Scheper, R. J., Pérez, J., Germà-Lluch, J. R. & Izquierdo, M. A. 2003. Lung resistance-related protein as a predictor of clinical outcome in advanced testicular germ-cell tumours. *British Journal of Cancer*, 88, 879-886.



## Appendix I

**S1** A comparison between a MrBayes trees before (this page), and after (following page) removal from the Multiple Sequence Alignment (MSA) of stretches of sequence insert found in less than 10% of species. This is done to prevent FastML from filling the gaps and producing unreasonably long ancestors. The tree was submitted unrooted for the purpose of ASR but have both been rooted here by *Amphimedon queenslandica* (Queensland sponge) which has the least sequence in common with the others in order to show the posterior probability at the nodes and the number of residue changes per site on the branch labels. The multiple sequence alignment (MSA) that produced this full-length sequence tree was 1,032 residues long.

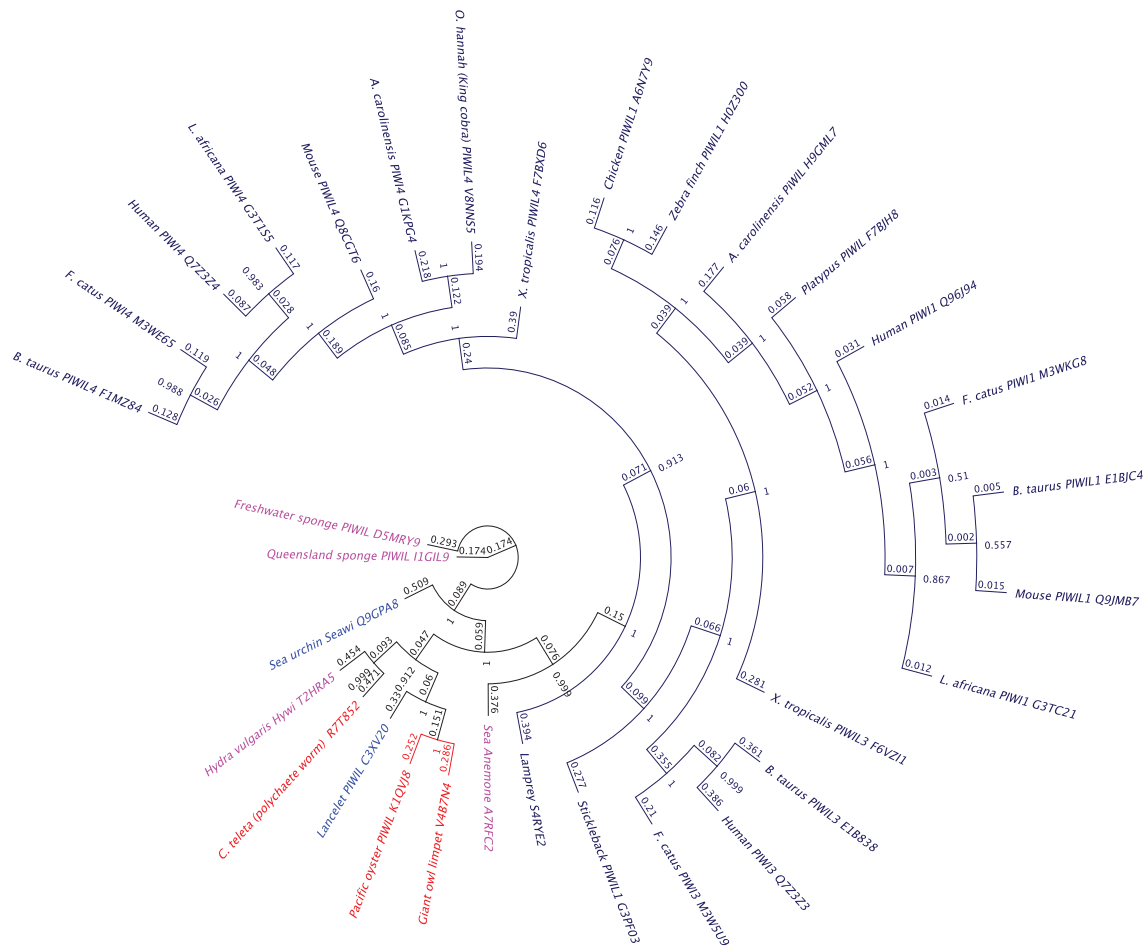


Colour scheme  
 Basal metazoa = pink  
 Lophotrochozoa = red  
 Deuterostomes = bright blue  
 Vertebrates = dark blue

## Appendix I

This is the same tree - but with some sequence removed to reduce the extent of the gaps. This is where it is more parsimonious that the residues represent an insert in less than 10% of species rather than a loss from the remainder of the species.

The MSA that produced this tree was 888 residues, i.e. 144 residues have been removed from the total length of the MSA. We anticipate that the changes per residue will be altered by this treatment but not the relationships between sequences.



Colour scheme  
 Basal metazoa = pink  
 Lophotrochozoa = red  
 Deuterostomes = bright blue  
 Vertebrates = dark blue

**S2 Table 4a.1. Tree summary.** Each MSA comprised of the same sequences plus a different outgroup for each one which was used as a root

Tree Root	Fate of the 'flippers'			AGO	PIWI	Sequences grouping with root (Accession numbers given where there is another sequence with the same name)	Isolated sequences (Accession numbers given where there is another sequence with the same name)
Archaea root <i>P. furiosus</i>	<i>S. mediterranea</i> PIWIL1-2				✓		<i>C. elegans</i> PRG1 and 2 <i>C. elegans</i> ERGO1
	<i>S. mansoni</i> AGO2C (2 of 3 are shown)		✓				
	<i>C. elegans</i> ERGO1		✓				
Trypanosome <i>T. brucei</i> AGO-L	<i>S. mediterranea</i> PIWIL1-2				✓		<i>S. mediterranea</i> PIWIL1 and 2 <i>C. elegans</i> ERGO1
	<i>S. mansoni</i> AGO2C		✓				
	<i>C. elegans</i> ERGO1		✓				
Trypanosome <i>T. brucei</i> PIWI-L	<i>S. mediterranea</i> PIWIL1-2				✓		<i>S. mediterranea</i> PIWIL1 and 2
	<i>S. mansoni</i> AGO2C		✓				
	<i>C. elegans</i> ERGO1		✓				
Slime mold <i>D. discoideum</i>	<i>S. mediterranea</i> PIWIL1-2		✓				<i>C. elegans</i> PRG1 and 2
	<i>S. mansoni</i> AGO2C		✓				
	<i>C. elegans</i> ERGO1		✓				
Atypical microsporidium <i>M. daphniae</i>	<i>S. mediterranea</i> PIWIL1-2				✓	<i>C. elegans</i> ALG3 and TAG76	Complete insect AGO2 'clade' <i>M. leidyl</i> AGO K9MVN9, <i>P. bachei</i> AGO3 <i>S. mansoni</i> both AGO2C sequences;
	<i>S. mansoni</i> AGO2C				✓		
	<i>C. elegans</i> ERGO1				✓		
Early fungi (AGO-L) <i>B. dendrobatidis</i>	<i>S. mediterranea</i> PIWIL1-2				✓	<i>S. mansoni</i> both AGO2C sequences; C4QPD0, C4QPD1 and C4QPD2 (virtually identical to C4QPD1 and left out of trees for clarity)	<i>M. leidyl</i> AGO K9MVN9, <i>P. bachei</i> AGO3 Both comb jellyfish
	<i>S. mansoni</i> AGO2C		With root				
	<i>C. elegans</i> ERGO1				✓		
Comb jellyfish <i>P. bachei</i> AGO3	<i>S. mediterranea</i> PIWIL1-2				✓	<i>M. leidyl</i> (comb jelly) AGO K9MVN9	
	<i>S. mansoni</i> AGO2C		Isolated / neither side				
	<i>C. elegans</i> ERGO1		Isolated / neither side				
Comb jellyfish <i>P. bachei</i> PIWI1	<i>S. mediterranea</i> PIWIL1-2		✓			<i>P. bachei</i> PIWI2 <i>H. robusta</i> (leech) TIG277	
	<i>S. mansoni</i> AGO2C		✓				
	<i>C. elegans</i> ERGO1		✓				
Queensland sponge AGO-L	<i>S. mediterranea</i> PIWIL1-2				✓	Freshwater sponge AGO D5MRZ0	<i>C. savignyi</i> (sea squirt) H2YU99
	<i>S. mansoni</i> AGO2C				✓		
	<i>C. elegans</i> ERGO1				✓		
Queensland sponge PIWI-L	<i>S. mediterranea</i> PIWIL1-2		All sequences are on one side only, i.e. this tree shows AGO sequences as arising from a duplication of a PIWI protein			Freshwater sponge PIWI D5MRY9	Sea urchin seaweed (PIWI-L)
	<i>S. mansoni</i> AGO2C						
	<i>C. elegans</i> ERGO1						

S3 Unrooted tree of all metazoan PIWI used to create the three ancestors depicted in fig. 4a.5

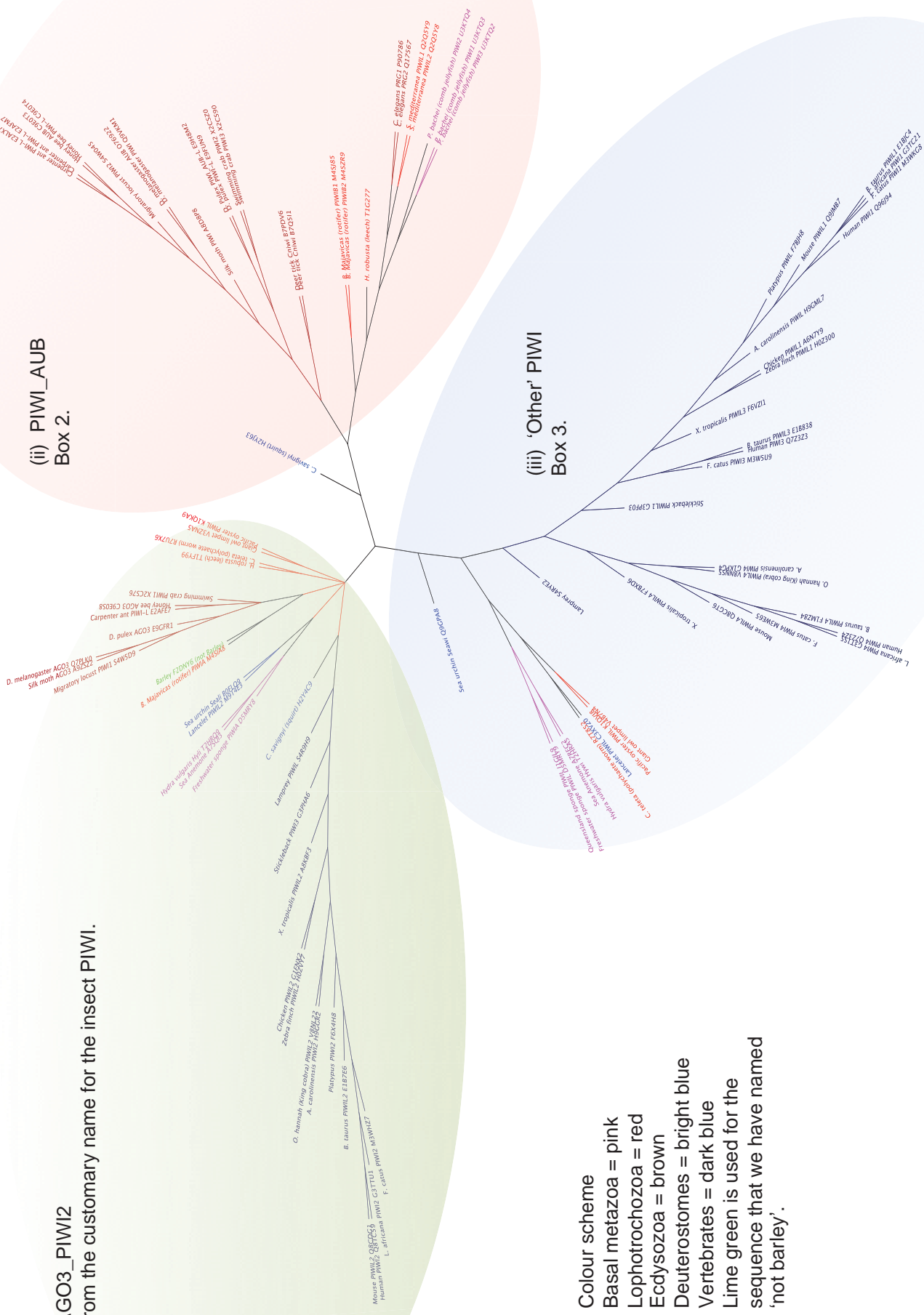
(i) Box 1. AGO3\_PIWI2

So called from the customary name for the insect PIWI.  
Details in

(ii) PIWI\_AUB  
Box 2.

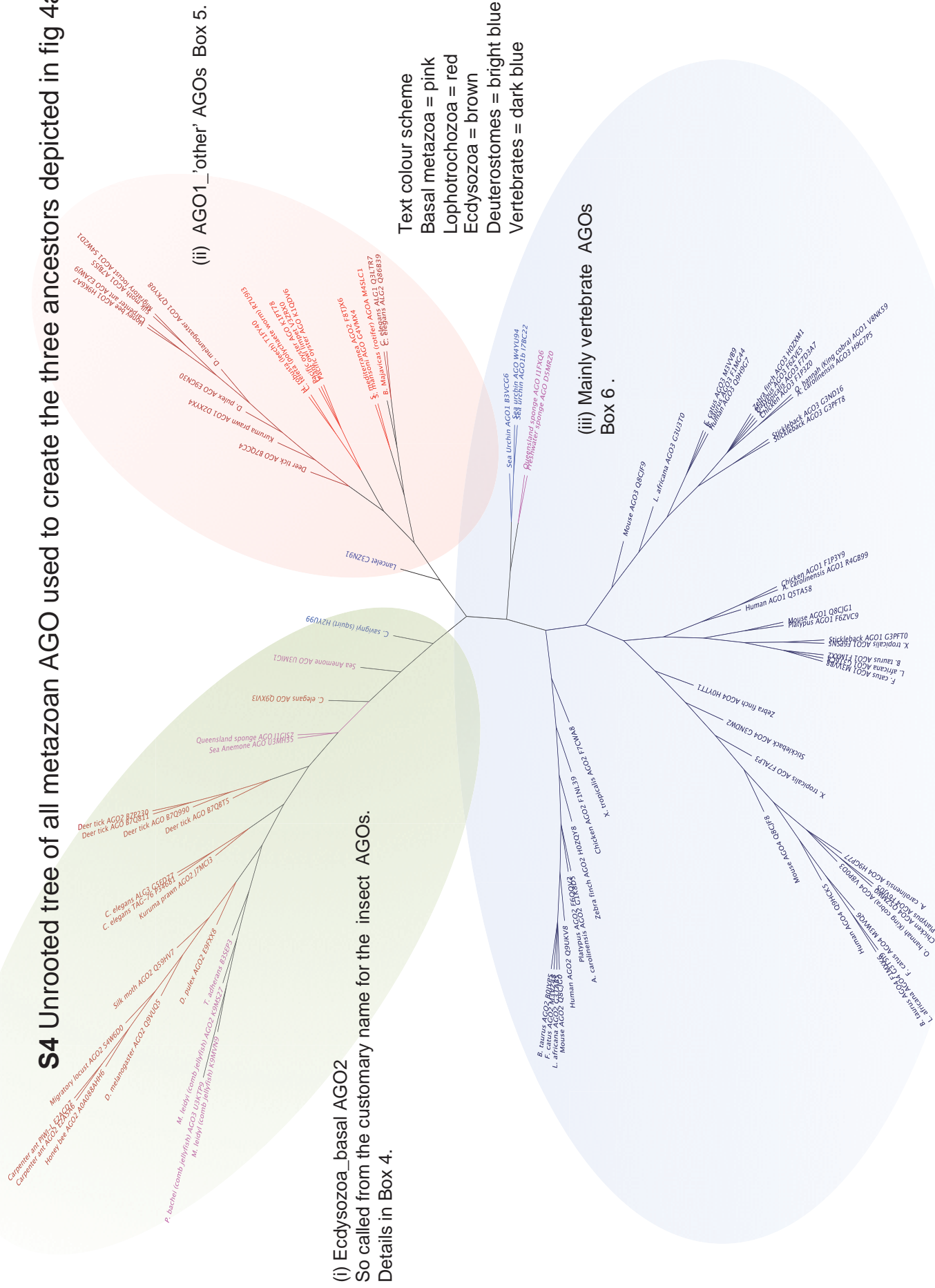
(iii) 'Other' PIWI  
Box 3.

- Colour scheme
- Basal metazoa = pink
  - Lophotrochozoa = red
  - Ecdysozoa = brown
  - Deuterostomes = bright blue
  - Vertebrates = dark blue
  - Lime green is used for the sequence that we have named 'not barley'.





**S4** Unrooted tree of all metazoan AGO used to create the three ancestors depicted in fig 4a.7



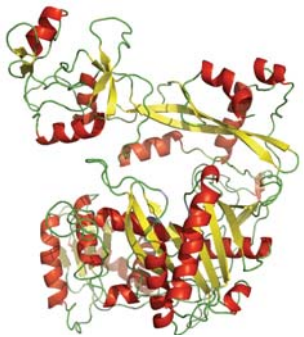

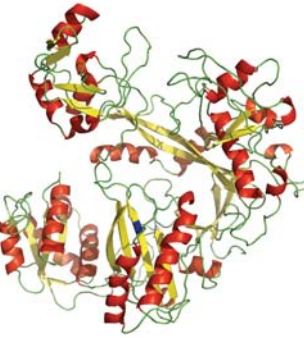
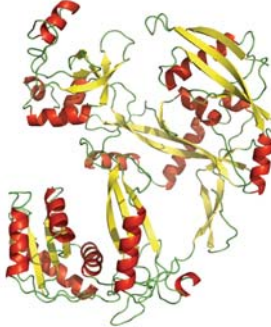

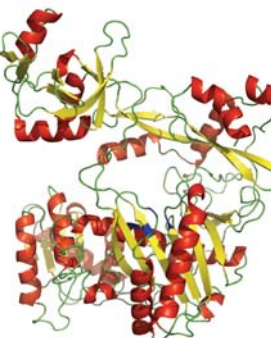


## Appendix II

**S2** *R. irregularis* I-TASSER results. Red type indicates structures unlikely to retain function, blue indicates lack of confidence. Catalytic tetrad is marked in dark blue where residues are conserved and light blue where similar residues are in their place. These structural predictions are in the order that they appear in Fig. 4b.5 Fates of the *R. irregularis* expansion.

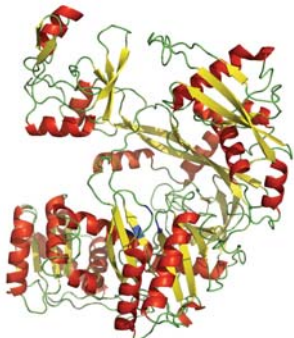

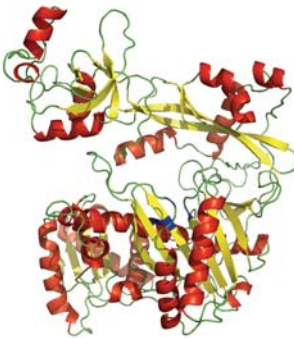
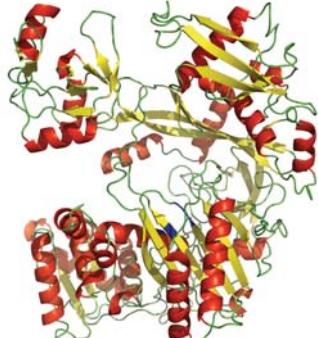

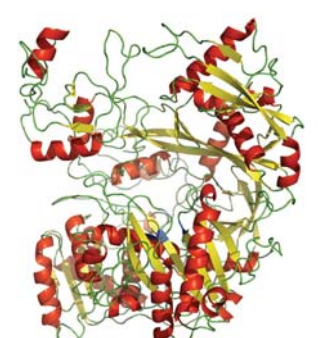
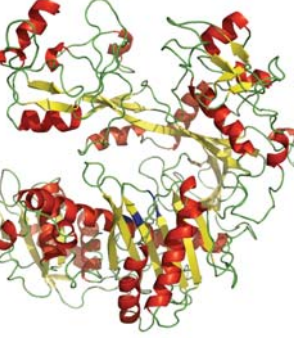

Accession number, I-TASSER C score, catalytic residues, annotation, notes, Pfam domain annotation	I-TASSER Predicted Structure	Accession number, I-TASSER C score, catalytic residues, annotation, notes, Pfam domain annotation	I-TASSER Predicted Structure
<p>U9U3P8 Confidence score (C score) 1.54 DEDH PIWI-Like A large portion of the N terminal domain is absent (circled) but the PAZ domain is intact. Pfam – lacks N domain. 663 residues</p>		<p>U9UAG7 C score 0.30 DEDH 'Uncharacterised' PAZ and N domains are poor. Pfam – lacks part of PAZ and N domain is missing. 544 residues – this is not annotated as a fragment and is more similar to short bacterial proteins.</p>	
<p>U9SQW1 C score 0.77 DEDH AGO Pfam finds a lack of MID domain however FATCAT aligns this with HsAGO2 crystal structure with 811 of 837 residues.</p>		<p>U9SKF0<sup>a</sup> C score 0.24 DPA 'Uncharacterised' DPA at the catalytic site may not impede RNA binding however there is some loss of structure in the PAZ domain (circled)<sup>a</sup>. Pfam – finds a lack of MID domain but a FATCAT alignment with HsAGO2 finds it almost entirely complete. 897 residues</p>	
<p>U9USE1 C score -1.53 DEDH AGO The excess C terminal accounts for the poor C score. It is unlikely to affect function. This sequence has the signature of a PIWI protein (LFYL) if it were not for the excess sequence. Pfam are unable to define the excess sequence. 1094 residues</p>		<p>U9UIX7 C score 1.30 DEDH 'Uncharacterised' PAZ domain is poor Pfam – PAZ and N domain is missing and additionally has some extra PIWI domain. 489 residues – this is not annotated as a fragment and is more similar to short bacterial proteins.</p>	
<p>U9UV71 C score -1.72 DE-- Catalytic residues not present 'Uncharacterised' Superficially OK but structurally not similar. Pfam finds a MID domain and an incomplete PIWI domain. 923 residues</p>		<p>U9TQS7 C score -0.27 DEDH 'Uncharacterised' Pfam finds some of the MID domain missing and in this case when aligned with HsAGO2 via FATCAT we find that the RNA binding pocket may be compromised. 862 residues</p>	

## Appendix II



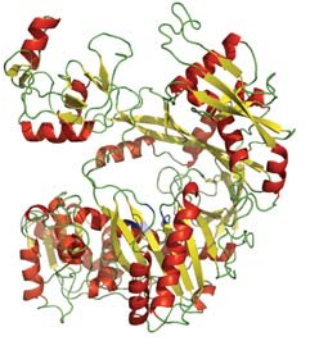
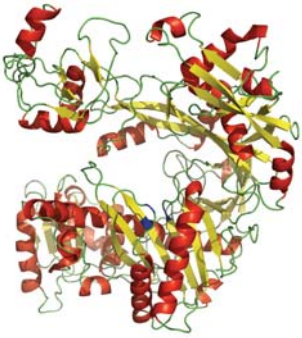
Accession number, I-TASSER C score, catalytic residues annotation, notes, Pfam domain annotation	I-TASSER Predicted Structure	Accession number, I-TASSER C score, catalytic residues, annotation, notes, Pfam domain annotation	I-TASSER Predicted Structure
<p>U9SXZ8</p> <p>C score 1.53</p> <p>DEDH</p> <p>‘Uncharacterised’</p> <p>PAZ domain is completely lacking</p> <p>403 residues - this is not annotated as a fragment and is more similar to short bacterial proteins</p>		<p>U9SSJ1</p> <p>C score 0.69</p> <p>DEDH</p> <p>AGO</p> <p>Pfam finds all domains complete.</p> <p>803 residues</p>	
<p>U9SVO2</p> <p>C score 1.13</p> <p>SDNR</p> <p>AGO</p> <p>This is currently in our ‘not sure’ basket. Structurally the RNA binding pockets are complete as is the PIWI domain.</p> <p>Pfam finds that the N domain is missing but this may not impede function.</p> <p>683 residues</p>		<p>U9SXX2</p> <p>C score 1.56</p> <p>DEDH</p> <p>‘Uncharacterised’</p> <p>Pfam finds all domains complete, however there are two <math>\beta</math> sheets missing (circled) although binding within the PAZ pocket may not be essential.</p> <p>801 residues</p>	
<p>U9STT8</p> <p>C score -0.07</p> <p>DGN-</p> <p>AGO</p> <p>The coil is missing that would have the histidine so we find the PIWI domain incomplete.</p> <p>Pfam finds some N domain missing, other domains are present but the PIWI domain has an extra PIWI segment.</p> <p>728 residues</p>		<p>U9U974</p> <p>C score 0.24</p> <p>RED-</p> <p>‘Uncharacterised’</p> <p>Too much structure is lacking including catalytic sites.</p> <p>Pfam finds incomplete PAZ and MID domains, though they state that PIWI is complete which is not what we see here.</p> <p>628 residues – this is annotated as a fragment.</p>	
<p>U9T923</p> <p>C score 2.00 (this is the highest possible score)</p> <p>DEDH</p> <p>AGO</p> <p>Pfam finds all domains complete.</p> <p>816 residues</p>		<p>U9SW52</p> <p>C score 1.30</p> <p>DEDH</p> <p>AGO</p> <p>Parts of the N domain are missing.</p> <p>Pfam finds that the N domain is missing. Again this may not impede function and this protein is annotated as an argonaute.</p> <p>678 residues</p>	



## Appendix II

Accession number, I-TASSER C score, catalytic residues, annotation, notes, Pfam domain annotation	I-TASSER Predicted Structure	Accession number, I-TASSER C score, catalytic residues, annotation, notes, Pfam domain annotation	I-TASSER Predicted Structure
<p>U9TIW8 C score 1.90 DEDH AGO Pfam finds all domains complete. 824 residues</p>		<p>U9T5B0 C score 0.66 DEDH AGO Pfam finds the N domain missing, this may not impede function. 692 residues</p>	
<p>U9T5E9 C score 1.13 DEDH AGO There is some N domain lacking. Pfam finds the N domain is lacking. 690 residues</p>		<p>U9URB1 C score -1.79 DEDH AGO Pfam finds all domains complete, yet the I-TASSER C score is particularly low. This sequence has greater primary homology with HsAGO2 than does U9T923, which has a very high score and we can see no reason for the discrepancy. 843 residues</p>	
<p>U9TXA2 C score 0.64 DEDH AGO Pfam finds the N domain missing. 757 residues – this is annotated as a fragment although has been upgraded from uncharacterised to AGO within the time frame of the study.</p>		<p>U9UWV0 C score -1.73 DDDH AGO There is some <math>\beta</math> sheet missing from the PAZ domain, which may not preclude RNA binding. Pfam finds all domains complete but once again the I-TASSER score is poor. 949 residues</p>	
<p>U9SLX6 C score 0.22 DEDH AGO Pfam finds all domains complete. 842 residues</p>		<p>U9SN94 C score -0.37 DEDH AGO Pfam finds all domains complete. 870 residues</p>	

## Appendix II

Accession number, I-TASSER C score, catalytic residues, annotation, notes, Pfam domain annotation	I-TASSER Predicted Structure	Accession number, I-TASSER C score, catalytic residues, annotation, notes, Pfam domain annotation	I-TASSER Predicted Structure
<p>U9SI69</p> <p>C score 0.64</p> <p>SEDH</p> <p>AGO</p> <p>Pfam finds all domains complete.</p> <p>827 residues</p>		<p>U9UQ28</p> <p>C score 0.24</p> <p>SERR</p> <p>‘Uncharacterised’</p> <p>The structure isn’t all that convincing and the catalytic residues aren’t either.</p> <p>Pfam finds the MID domain is missing and the PIWI domain lacks integrity</p> <p>580 residues</p>	
<p>U9SUD1</p> <p>C score 0.34</p> <p>SEDH</p> <p>AGO</p> <p>This is still likely to have some function.</p> <p>Pfam finds all domains complete.</p> <p>834 residues</p>		<p>U9SL93</p> <p>C score 0.90</p> <p>DEDS</p> <p>AGO</p> <p>This is still likely to have some function, there is a histidine very close to the serine which may play a part but this isn’t essential to RNA binding.</p> <p>Pfam finds all domains complete.</p> <p>831 residues</p>	

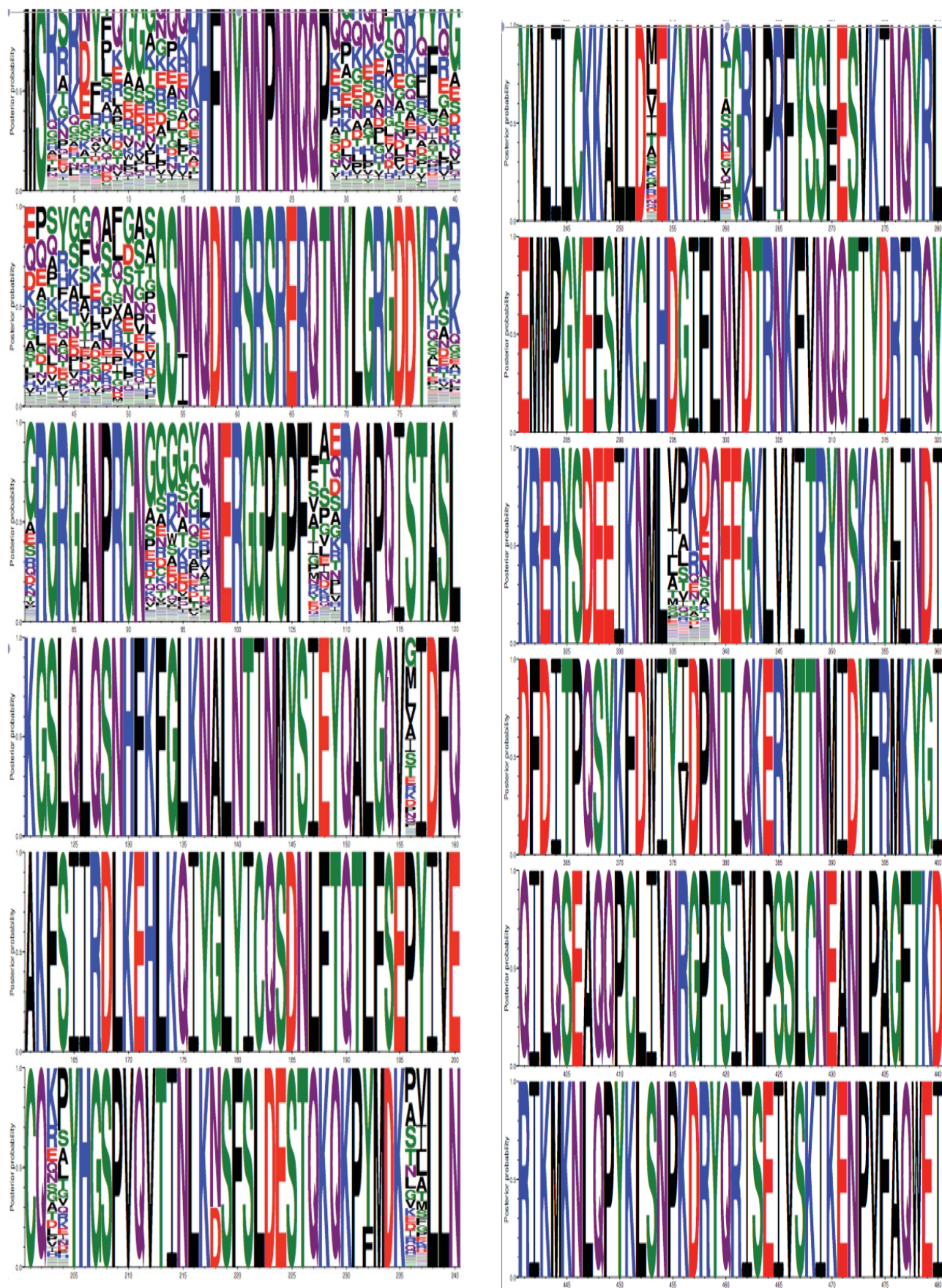
<sup>a</sup> A caveat must be applied here because the lack of apparent structure could be an artefact of the structural prediction or of the PyMol interpretation of the pdb file. This applies to all of the predicted structures but in this case the residues are there but we can’t be certain whether they do or do not form a helix. The FATCAT alignment shows the *R. irregularis* residues aligning with the helix entirely.

Some sequences are annotated as fragments and are marked as such in the detail. Some are simply MID and PIWI domains which is a situation common in bacterial argonautes



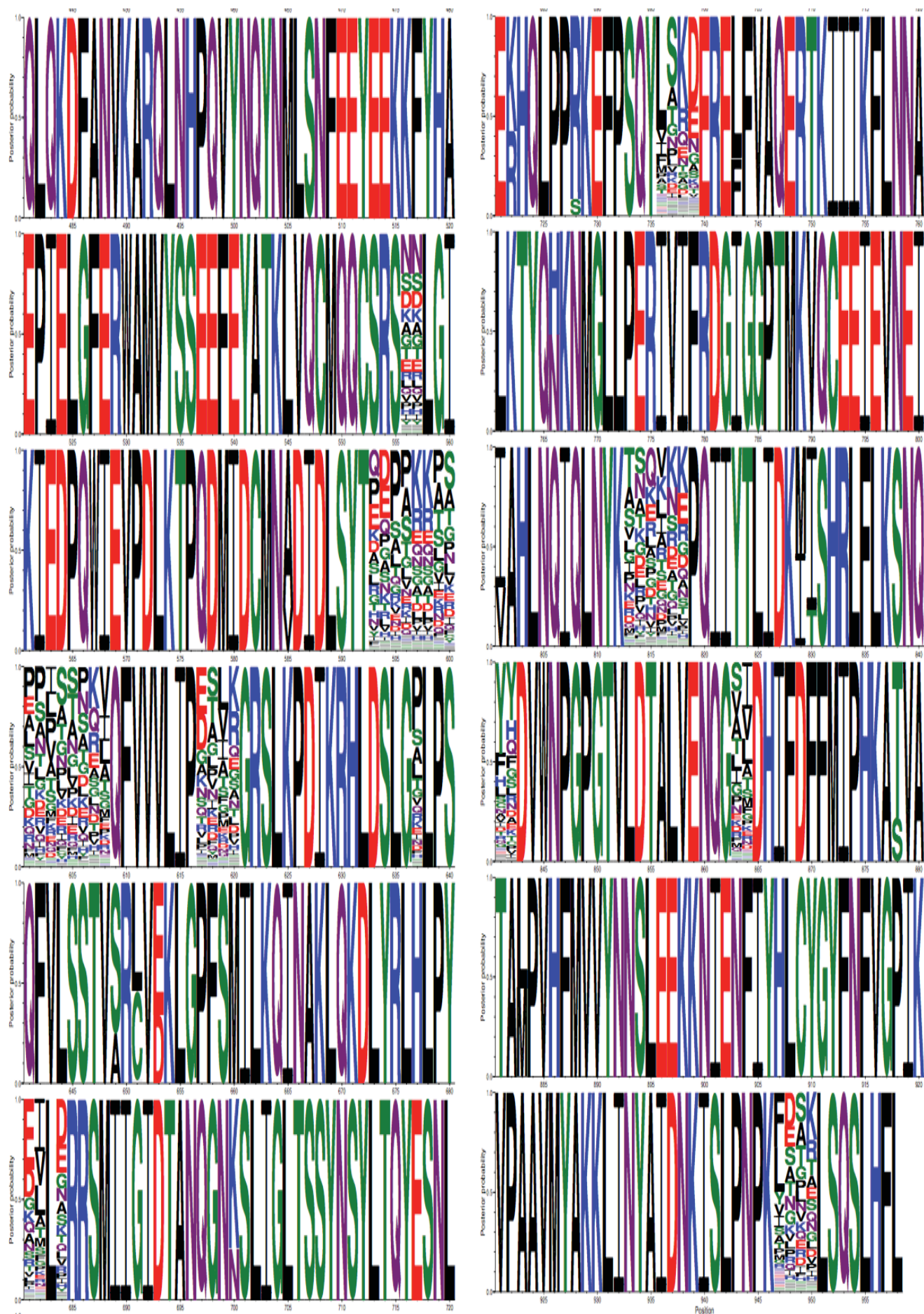
## APPENDIX III

**S1** Logo of the posterior probability for the reconstruction of node 1 for the sequences that went to make up the ancestor ‘Ciliate 2b’. Panels read from top to bottom and continues from the top of the adjacent panel. 100% probability is indicated by whole letters.



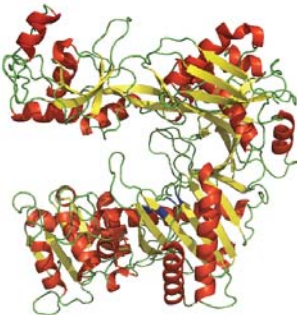
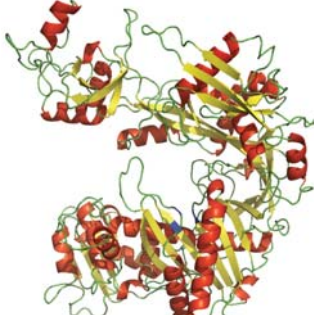
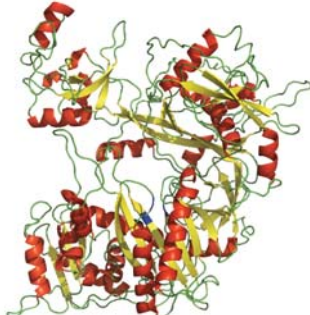
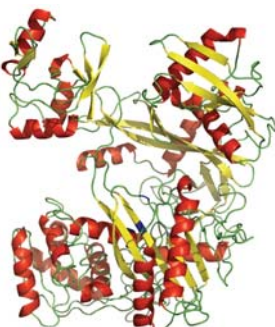
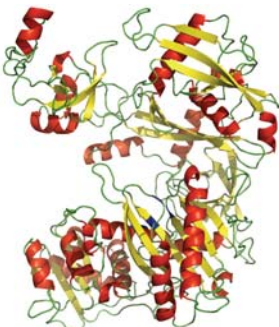
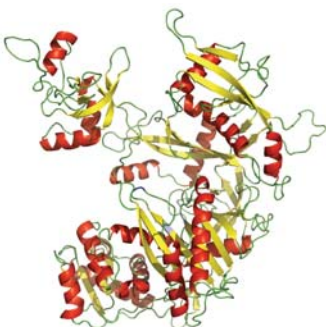
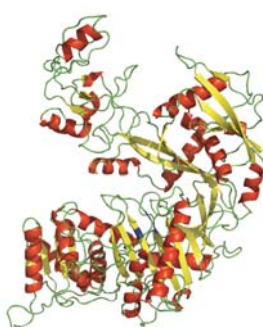


# APPENDIX III

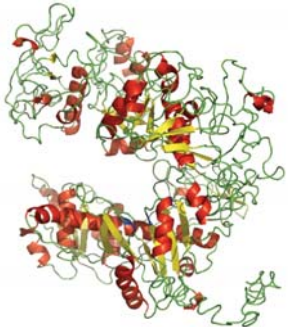
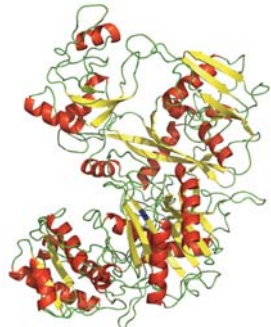
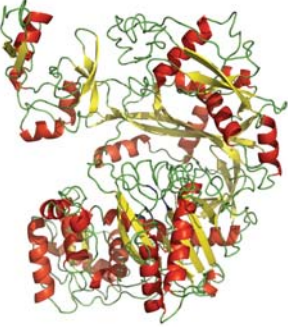

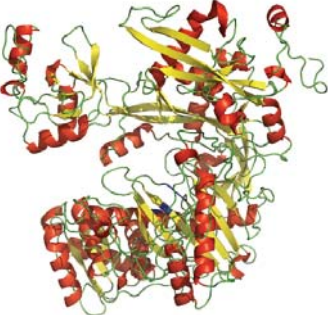


## APPENDIX III

### S2 Ancestral sequences calculated at node 1 for each ancestor.

Group, catalytic sequence, C terminal signature and notes	I-TASSER Predicted Structure	Group, catalytic sequence, C terminal signature and notes	I-TASSER Predicted Structure
Stramenopile ancestral sequences			
Stramenopiles group 1 DEDH MYFV Has the RGGG repeats at the N terminal I-TASSER C score -0.5		Stramenopiles group 2 DEDH MYYV Does not have RGGG repeats at the N terminal I-TASSER C score -0.86	
Stramenopiles group 3 DEDH MFFI Has the RGGG repeats at the N terminal I-TASSER C score -1.69			
Ciliate ancestral sequences			
Ciliates group 1a DEDH LYFL I-TASSER C score 1.49		Ciliates group 1b DEDH LFFI I-TASSER C score 1.55	
Ciliates group 2a EEQE LYFL I-TASSER C score -0.89		Ciliates group 2b DEDK LHFL I-TASSER C score -1.04	

## APPENDIX III

Group, catalytic sequence, C terminal signature and notes	I-TASSER Predicted Structure	Group, catalytic sequence, C terminal signature and notes	I-TASSER Predicted Structure
Kinetoplast ancestral sequences			
Kinetoplast PIWI- like ancestor  DEDH  LWFL  I-TASSER C score -2.02		Kinetoplast AGO- like ancestor  DERS  MHYL  I-TASSER C score -1.71	
Plant ancestral sequences			
Plant 4_6_8_9 chromatin modifiers  DEDH  MFFC  I-TASSER C score 0.09		Plant 2_3_7 RNA binders  DERH  MFYC  I-TASSER C score -1.66	
Plant 1_5_10 Slicers  DEDH  MFFC  I-TASSER C score -1.88			





**MASSEY UNIVERSITY**  
GRADUATE RESEARCH SCHOOL

**STATEMENT OF CONTRIBUTION  
TO DOCTORAL THESIS CONTAINING PUBLICATIONS**

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

**Name of Candidate:** Toni K. Daly

**Name/Title of Principal Supervisor:** Professor David Penny

**Name of Published Research Output and full reference:**

Toni K Daly et al. (2013) Beyond BLASTing: Tertiary and quaternary structure analysis helps identify Major Vault Proteins. *Genome Biology and Evolution* 5: 217-232

**In which Chapter is the Published Work:** Chapter 2a

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate: **90%**  
and / or
- Describe the contribution that the candidate has made to the Published Work:

**Toni Daly**

Digitally signed by Toni Daly  
DN: cn=Toni Daly, o=Massey University,  
ou=Institute of Fundamental Sciences,  
email=t.daly1@massey.ac.nz, c=NZ  
Date: 2016.04.25 11:10:17 +1200

Candidate's Signature

**25/4/2016**

Date

Principal Supervisor's signature

**19/9/2016**

Date



**MASSEY UNIVERSITY**  
**GRADUATE RESEARCH SCHOOL**

**STATEMENT OF CONTRIBUTION  
TO DOCTORAL THESIS CONTAINING PUBLICATIONS**

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

**Name of Candidate:** Toni K. Daly

**Name/Title of Principal Supervisor:** Professor David Penny

**Name of Published Research Output and full reference:**

Toni K Daly et al. (2013) In silico resurrection of the Major Vault Protein suggests it is ancestral in modern eukaryotes. *Genome Biology and Evolution* 5 (8): 1567-1583.

**In which Chapter is the Published Work:** Chapter 2b

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate: **90%**  
and / or
- Describe the contribution that the candidate has made to the Published Work:

**Toni Daly**

Digitally signed by Toni Daly  
DN: cn=Toni Daly, o=Massey University,  
ou=Institute of Fundamental Sciences,  
email=t.daly1@massey.ac.nz, c=NZ  
Date: 2016.04.25 11:47:59 +12'00'

Candidate's Signature

**25/4/2016**

Date

Principal Supervisor's signature

**19/9/2016**

Date



**MASSEY UNIVERSITY**  
GRADUATE RESEARCH SCHOOL

**STATEMENT OF CONTRIBUTION  
TO DOCTORAL THESIS CONTAINING PUBLICATIONS**

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

**Name of Candidate:** Toni K. Daly

**Name/Title of Principal Supervisor:** Professor David Penny

**Name of Published Research Output and full reference:**

How old are RNA networks?

Daly, T., Chen, X. S., & Penny, D. (2011). How old are RNA networks? , Advances in Experimental Medicine and Biology 722, 255-273.

**In which Chapter is the Published Work:** Chapter 3

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate: 80%  
and / or
- Describe the contribution that the candidate has made to the Published Work:

**Toni Daly**

Digitally signed by Toni Daly  
DN: cn=Toni Daly, o=Massey University,  
ou=Institute of Fundamental Sciences,  
email=t.daly1@massey.ac.nz, c=NZ  
Date: 2016.04.23 09:53:36 +1200

Candidate's Signature

**23/4/16**

Date

*E. J. Penny*

Principal Supervisor's signature

*19/9/2016*

Date





**MASSEY UNIVERSITY**  
**GRADUATE RESEARCH SCHOOL**

**STATEMENT OF CONTRIBUTION  
TO DOCTORAL THESIS CONTAINING PUBLICATIONS**

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

**Name of Candidate:** Toni K. Daly

**Name/Title of Principal Supervisor:** Professor David Penny

**Name of Published Research Output and full reference:**

Toni Daly, et al. (2016) Long Long AGO: The evolutionary history of Argonaute and PIWI in metazoa by ancestral protein inference and structure prediction. (submitted).

**In which Chapter is the Published Work:** Chapter 4a

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate: 90%  
and / or
- Describe the contribution that the candidate has made to the Published Work:

**Toni Daly**  
Digitally signed by Toni Daly  
DN: cn=Toni Daly, o=Massey University,  
ou=Institute of Fundamental Sciences,  
email=t.daly1@massey.ac.nz, c=NZ  
Date: 2016.04.25 11:50:05 +12'00'

Candidate's Signature

**25/04/2016**

Date

*E. D. Penny*

Principal Supervisor's signature

*19/9/2016*

Date



**MASSEY UNIVERSITY**  
**GRADUATE RESEARCH SCHOOL**

**STATEMENT OF CONTRIBUTION  
TO DOCTORAL THESIS CONTAINING PUBLICATIONS**

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

**Name of Candidate:** Toni K. Daly

**Name/Title of Principal Supervisor:** Professor David Penny

**Name of Published Research Output and full reference:**

Chapter 5b. Toni K Daly et al. (2016) Argonaute gain and loss during fungal evolution. (submitted).

**In which Chapter is the Published Work:** Chapter 4b

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate: 90%  
and / or
- Describe the contribution that the candidate has made to the Published Work:

**Toni Daly**

Digitally signed by Toni Daly  
DN: cn=Toni Daly, o=Massey University,  
ou=Institute of Fundamental Sciences,  
email=t.daly1@massey.ac.nz, c=NZ  
Date: 2016.04.25 11:51:32 +1200

Candidate's Signature

**25/04/2016**

Date

*E. D. Penny*

Principal Supervisor's signature

**19/9/2016**

Date



**MASSEY UNIVERSITY**  
**GRADUATE RESEARCH SCHOOL**

**STATEMENT OF CONTRIBUTION  
TO DOCTORAL THESIS CONTAINING PUBLICATIONS**

(To appear at the end of each thesis chapter/section/appendix submitted as an article/paper or collected as an appendix at the end of the thesis)

We, the candidate and the candidate's Principal Supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the candidate's contribution as indicated below in the *Statement of Originality*.

**Name of Candidate:** Toni K. Daly

**Name/Title of Principal Supervisor:** Professor David Penny

**Name of Published Research Output and full reference:**

Toni Daly et al. Argonautes origins in eukaryotes. (in preparation).

**In which Chapter is the Published Work:** Chapter 4c

Please indicate either:

- The percentage of the Published Work that was contributed by the candidate: 90%  
and / or
- Describe the contribution that the candidate has made to the Published Work:

**Toni Daly**  
Digitally signed by Toni Daly  
DN: cn=Toni Daly, o=Massey University,  
ou=Institute of Fundamental Sciences,  
email=t.daly1@massey.ac.nz, c=NZ  
Date: 2016.04.25 11:53:03 +12'00'

Candidate's Signature

**25/04/2016**

Date

*E. D. Penny*

Principal Supervisor's signature

*19/9/2016*

Date