Eukaryotic Signature Proteins:
Guides to pathogenic eukaryotic parasites

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

PhD

in

Genetics

At Massey University, Palmerston North,

New Zealand

Jian Han

2012

Abstract

Eukaryotic Signature Proteins (ESPs) are proteins that delineate the eukaryotes from the archaea and bacteria. They have no homologues in any prokaryotic genome, but their homologues are present in all main branches of eukaryotes. ESPs are thus likely to have descended from ancient proteins that have existed since the first eukaryotic cell. This project looks at ESPs of some eukaryotic parasites and human (*Homo sapiens*) as their host organism and focuses on *Giardia lamblia*, a fresh water pathogenic basal eukaryote. The ESP datasets from *Giardia* and two other parasites, *Trichomonas vaginalis* and *Plasmodium falciparum*, as well as the host human were calculated in light of available genomic data and the datasets contained a range of proteins associated with membrane, cytoskeleton, nucleus and protein synthesis.

ESPs have great potential in phylogenetic studies since these proteins are present in all eukaryotes and are expected to have a slow and constant rate of evolution. Phylogenetic analyses were performed on the 18 eukaryotic organisms including some basal eukaryotes, and also for mammals, using orthologues of the all ESPs from these organisms. Strategies such as concatenating sequences and constructing consensus networks were tested to evaluate their potential with large numbers of ESP alignments. The results were promising, and ESPs hold great potential for their use in future phylogenetic analyses of eukaryotes.

RNA interference is hypothesised to be an ancient mechanism for gene regulation and like the ESPs, it is typically found in all main branches of eukaryotes. High throughput sequencing data from *Giardia* and *Trichomonas* small RNAs (15-29mers) were re-analysed showing two length peaks for *Giardia* RNAs: a "larger peak" and an "ultra small peak", the former of which is likely to be the product of the enzyme Dicer, which processes miRNA. The "ultra small peak" but not the "larger peak" was also found in *Trichomonas*. The two peaks possibly represent two different mechanisms of RNA interference (RNAi) in these parasites, but analysis of potential target sites from the Dicer-processed RNAs has not yet shown any indication that ESPs are regulated any differently from other parasite proteins.

Sugar metabolic pathways including glycolysis and citric acid cycle were searched for ESPs, this was done to determine the relationship between the conservation of

eukaryotic metabolic pathways and conservation of individual proteins. However no ESPs were identified from these pathways because *Giardia* has enzymes that show more similarity to those from prokaryotes than eukaryotes. These enzymes are significantly different from that of the host's, and these alternative enzymes offer potential as novel drug targets. In addition, ESPs that are present from host but lost in some parasites were analysed, and these ESPs are involved in many understudied pathways. It is these differences which can provide a guide in determining which pathways we should examine when designing drug targets.

Overall, numerous proteomic similarities and differences in ESPs were identified between host and parasite. These proteins show potential for future evolutionary studies, and will guide future directions in ancestral eukaryotic regulation and metabolism.

Acknowledgements

It has been a very challenging yet rewarding journey towards the completion of this thesis. I am thankful to everyone who helped me throughout my work, and kept my life interesting during my study.

Foremost I would express my sincere gratitude towards my supervisor, Dr. Lesley Collins. Bioinformatics was never my forte during my undergraduate studies. But thanks to Dr. Collins, I have learnt many bioinformatics skills during the four years of my doctorial study. These skills will be very useful in my future career.

I would thank my co-supervisor Dr. Patrick Biggs, who has been tirelessly commenting on my "broken" English, it was very fortunate to have someone who can speak English like the Queen. Dr. Biggs has also been outstanding in helping me setting up databases.

I also thank my other co-supervisor David Penny, who has been helping me with the writing and cracking jokes from time to time.

I express my warm thanks to my colleagues who helped me during different parts of the project. Dr. Tim White, who is an absolute computing genius, has provided generous computing assistance; Dr. Simon Hills and Bojian Zhong have given a helping hand on my phylogenetics studies.

I thank my friends/flatmates (Nick, Sophie, Ping, Bryn, Sam, Ryan, Suz, Nat and Justin) and others, my weekends would be very dull without you guys. Also thanks to my snowboarding buddy Max for keeping me alive from giant snowballs.

Special thanks to my parents, both doctors, for their financial and emotional support. They provided me with plenty of encouragement.

Finally, thanks to everyone in the boffin lounge, the environment and work ethic has been wonderful here. Thanks to Massey University, Palmerston North for providing the working space.

This work was funded by Health Research Council (HRC) - Emerging Researcher Grant (Dr. L. Collins) 07/168. Eukaryotic Signature Proteins - Guides to Modern Eukaryotic Parasites.

Table of Contents

Abstract iii

Acknowledgements v

Table of Contents vii

List of Figures xii

List of Tables xiv

Terminology xv

Chapter 1: Introduction 1

1.1 Eukaryotic signature proteins 1

1.2 Parasites involved in the project 4

1.2.1 *Giardia lamblia*, a unique organism 4

1.2.2 *Trichomonas* and *Plasmodium* 8

1.2.3 Current RNA work on *Giardia* and *Trichomonas* 10

1.3 Thesis structure 11

1.3.1 Generating a new ESP dataset – Chapter 2 12

1.3.2 Phylogenetic analysis using ESPs – Chapter 3 12

1.3.3 Metabolic analysis of *Giardia* – Chapter 4 14

1.3.4 Small RNAs in *Giardia* and *Trichomonas* – Chapter 5 15

1.3.5 Summary 16

Chapter 2: Collecting Eukaryotic Signature Proteins 17

2.1 Introduction 17

2.1.1 BLAST statistics 18

2.2 Material and methods 19

2.2.1 Selection of species for analysis 19

2.2.2 ESP calculations 27

2.2.3 Assigning Gene Ontology terms 29

2.2.4 Database construction and management 30

2.3 Results and Discussion 32
2.3.1 The *Giardia* ESP dataset 32
2.3.2 Comparison with Hartman's dataset 37
2.3.3 Using E-value as an alternative to bit-score as cut-off 38
2.3.4 The *Plasmodium* and *Trichomonas* ESP datasets 41
2.3.5 Human (*Homo sapiens*) ESP dataset 41
2.3.6 Human ESPs in parasites 43
2.3.7 Differences and similarities between parasite ESP datasets 47
2.3.8 Other groups of proteins 48
2.4 Conclusions 49
2.4.1 ESP calculation conclusions 49
2.4.2 Database updates 50
2.4.3 Implications for current models of evolution 50
Supplementary material for Chapter 2 53
S2.1 ESP calculation protocol and Perl scripts 53
S2.2 List of 274 *Giardia* ESPs 58
S2.3 List of 37 *Giardia* proteins which are conserved in all organisms 68
S2.4 List of 44 *Escherichia* proteins which are conserved in all bacteria and not found in archaea 70
S2.5 Poster 72
Chapter 3: Phylogenetic analysis using ESPs 75
3.1 Introduction 75
3.1.1 Overview 75
3.1.2 The current phylogenetic system 75
3.1.3 How deep phylogenetic analysis was done in the past 77
3.1.4 The ESP approach 78
3.2 Method 79

3.2.1 Phylogenetic software 79
3.2.2 Phylogenetic methods 80
3.2.3 Analysis procedure 81
3.3 Results 83
3.3.1 ML trees of ESP 83
3.3.2 Bayesian analysis 85
3.3.3 Unexpected tree shapes 86
3.3.4 Consensus tree 88
3.3.5 Divide trees based on topology comparisons with expected tree 91
3.3.6 Consensus tree with split tree, software results can be deceptive 93
3.3.7 Tree building by concatenating sequences 97
3.3.8 Tree built with different model 100
3.3.9 Relationship between protein function and its phylogenetic usefulness 101
3.3.10 Phylogenetic analysis of mammal species using ESP 103
3.4 Discussion 104
3.4.1 ESPs as candidates for evolutionary studies 104
3.4.2 Limitations 105
3.4.3 Conclusion and Future work 106
Supplementary material for Chapter 3 109
S3.1 SplitsTree consensus network explanation 109
S3.2 Perl script used in this chapter 111
Chapter 4: Reconstruction of metabolic pathways in *Giardia* 115
4.1 Introduction 115
4.2 Materials and Methods 118
4.3 Results 120
4.3.1 Glycolysis and Gluconeogenesis 120
4.3.2 Tricarboxylic acid cycle 126

4.3.3 Oxidative phosphorylation 128
4.3.4 Other metabolic pathways 131
4.4 Discussion 132
Supplementary material for Chapter 4 137
S1. Enzymes of glycolysis pathway in *Giardia* 137
S2. Enzymes of citric acid cycle in *Giardia* 141
S3. Enzymes of oxidative phosphorylation in *Giardia* 143
S4 Perl script used 144
Chapter 5: Non-coding RNAs of *Giardia* and *Trichomonas* and their relationship to ESPs 145
5.1 Introduction to small ncRNAs 145
5.2 Methods 148
5.2.1 Sample preparation and sequencing 148
5.2.2 Adaptor trimming and mapping 149
5.2.3 Finding mapped RNA targeting sites 151
5.3 Results and Discussion 152
5.3.1 Summary of number of RNAs yielded after each step 152
5.3.2 Small RNAs of *Giardia* 152
5.3.3 Small RNAs of Trichomonas 154
5.3.4 *Giardia* mapping results 155
5.3.5 *Trichomonas* mapping results 156
5.3.6 GC content 157
5.3.7 Determination of whether the “ultra small peak” of *Giardia* is a result of secondary processing of longer RNAs 158
5.3.8 Possible target sites of *Giardia* 26 and 27mers 158
5.4 Conclusion 160
Supplementary material for Chapter 5 162

S5.1 Abstract for 3rd Next Generation Sequencing Conference 162

S5.2 Abstract for IV International Giardia and Cryptosporidium Conference 163

Final words 165

References 171

List of Figures

Chapter 1

Figure 1. *Giardia* trophozoites as viewed by an electron microscope 4
Figure 2. *Giardia* life cycle 5
Figure 3. Electron microscopy of *Trichomonas* 8
Figure 4. *Plasmodium* (trophozoite ring form) inside erythrocytes 9

Chapter 2

Figure 1. Phylogenetic relationship of selected archaeal species 21
Figure 2. Phylogenetic relationship of selected bacterial species 23
Figure 3. Phylogenetic position of eukaryotic organisms chosen for this project 24
Figure 4. Procedure used for calculating ESPs 28
Figure 5. Illustration of *Giardia* database layout 31
Figure 6. Human ESP and GO term 46

Chapter 3

Figure 1. Phylogenetic position of eukaryotic organisms chosen for this project 76
Figure 2. Unrooted ML tree of protein GL50803_93275 (Translational activator GCN1) from different species 84
Figure 3. DensiTree output of Bayesian analysis of protein GL50803_93275 85
Figure 4. Unrooted ML tree of orthologues for GL50803_7896 from different species showing effect of including an incorrect gene paralogue 87
Figure 5. Unrooted ML tree of orthologues of GL50803_15339 from different species showing effect of including different *Ciona* paralogues 88
Figure 6. Unrooted consensus tree built using 267 ML trees 89
Figure 7. Unrooted average consensus tree built using 267 ML trees 91
Figure 8. Box plot of gene length distribution. 92
Figure 9. Consensus network Type 1 93
Figure 10. Consensus network Type 2 94
Figure 11. Consensus network Type 3 95
Figure 12. Average consensus 96
Figure 13. Unrooted tree generated using the WAG+Γ4+I model 98
Figure 14. Unrooted tree generated with *Giardia* removed 99
Figure 15. Unrooted tree generated using the Dayhoff model 100
Figure 16. Phylogenetic tree of mammalian species 104

Chapter 4

Figure 1. Glycolysis in *Giardia* 121
Figure 2. A possible ethanol fermenting pathway in *Giardia* 123
Figure 3. TCA cycle enzymes in *Giardia*. 127
Figure 4. The Oxidative phosphorylation pathway in *Giardia* 130
Figure 5. Pentose phosphate pathway in *Giardia* 131
Figure 6. Alanine and aspartate in *Giardia* 132
Figure S1. KEGG diagram of glycolytic enzymes in *Giardia* 140
Figure S2. TCA cycle enzymes in *Giardia* 142

Chapter 5

Figure 1. Micro RNA and siRNA mechanism of action 146
Figure 2. Why adaptor trimming was performed 150
Figure 3. Summary of analysis procedure 151
Figure 4. Length and 5’ nucleotide distribution for *Giardia* ncRNA 153
Figure 5. Length and 5’ nucleotide distribution for *Trichomonas* ncRNA 154
Figure 6. Length and 5’ nucleotide distribution for mapped *Giardia* ncRNA 155
Figure 7. Length and 5’ nucleotide distribution for mapped *Trichomonas* ncRNA 156

List of Tables

Chapter 1

Table 1. Antigiardial drugs and their targets 6
Table 2. Some types of ncRNAs 10

Chapter 2

Table 1. List of archaeal species used in study 20
Table 2. List of eubacterial species used in study 22
Table 3. List of Eukaryotic species used in study 26
Table 4. Categories of *Giardia* ESPs 33
Table 5. Proteins with multiple copies in ESP dataset 34
Table 6. *Giardia* ESPs with homologues from Hartman dataset 37
Table 7. Comparison between using E-value and bit-score as cut-offs 39
Table 8. Summary of the number of ESPs obtained for each organism 41
Table 9. The 15 most abundant GO terms for human ESPs (updated version) 42
Table 10. The 15 most abundant GO terms for 8000 random human proteins 43
Table 11. The 40 most abundant GO terms for human ESPs 45
Table 12. ESP between parasites 47
Table 13. ESP vs other parasites' proteome 48

Chapter 3

Table 1. Function and phylogenetic utility 102

Chapter 4

Table 1. Bacteria-like *Giardia* enzymes in glycolysis pathway 125
Table 2. *Giardia* glycolytic enzyme candidates maintained in all eukaryotes 135

Chapter 5

Table. 1 list of categories for mapped RNAs 152
Table 2. Number of *Giardia* and *Trichomonas* RNAs remained after each step 152
Table 3. GC content of *Giardia* and *Trichomonas* small RNAs 157
Table 4. Number of overlapping 16 and 17mers 158
Table 5. Loci of 26 and 27mers in relation to genes 159

Terminology

3'UTR (three prime untranslated region): Region of mRNA downstream of the termination codon. In metazoans this region is where miRNA binds to regulate gene expression.

5'UTR (five prime untranslated region): Region of mRNA upstream of the starting codon that often contain regulatory elements such as ribosome binding sites.

Akaike Information Criterion (AIC): Measure used in model testing based on the goodness of fit to a statistical model

Basal Eukaryote: A unicellular eukaryotic which is believed to have diverged early during the evolution of eukaryotes, e.g. *Giardia lamblia*.

Bayesian inference: A tree searching method which is statistically similar to maximum likelihood, the aim is to find the tree with maximum posterior probability; can allow complex models of evolution to be implemented.

BLAST (Basic Local Alignment Search Tool): Software which enables comparison of amino acid or nucleotide sequences.

Blastp: A BLAST program which compares protein queries with protein databases.

Cellular signature structure (CSS): Cell organelles or complex found in eukaryotes but not prokaryotes, e.g. mitochondria, Golgi apparatus, spliceosome.

Excavata: A eukaryotic supergroup that contains the morphological feature of a ventral feeding groove. This supergroup includes Diplomonads (*Giardia lamblia*) and Parabasalia (*Trichomonas vaginalis*).

Eukaryotic Signature Protein (ESP): A protein with no homologues in prokaryotic (archaea and bacteria) genomes, but it has homologues which are present in all the major branches of eukaryotes.

Gene Ontology (GO): A project aimed to unify the representation of gene attributes across all species by using a controlled vocabulary to assign their functions. Website: http://www.geneontology.org.

Long branch attraction (LBA): A phenomenon observed when highly divergent lineages are grouped together, regardless of their true evolutionary relationships. The long branches of a tree will group together regardless of the true tree topology.

Maximum likelihood (ML) inference: A tree searching method which aims to find the tree with highest probability to produce the observed data.

Messenger RNA (mRNA): RNA transcribed from DNA, after mRNA processing (e.g. Capping, intron splicing) the mature mRNA is translated into protein by the ribosome.

Micro RNA (miRNA): ~21-22 base pair (bp) single stranded RNA processed by the Dicer or Drosha proteins, which regulates gene expression by means of complimentary binding to the target mRNA.

Non-coding RNA (ncRNA): RNA that does not code for proteins, but may have a function such as regulating, modifying or processing other RNAs.

Perl: A dynamic programming language. Able to perform various bioinformatic tasks especially data mining and can connect with MySQL databases to enable fast and automated database management and queries.

Small interfering RNA (siRNA): ~21-26 bp double stranded RNA processed by Dicer which regulates gene expression by means of complimentary binding to the target mRNAs.