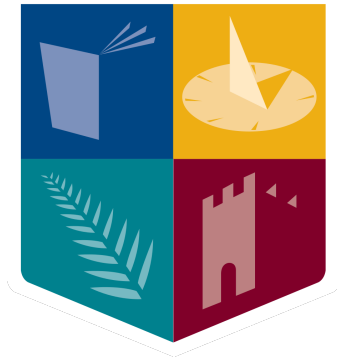


# **Next generation sequencing technology and phylogenomics enhance the resolution of deep node phylogenies: A study of the Protostomia**

A thesis submitted to the National University of Ireland Maynooth  
for the degree of **Doctor of Philosophy**



**Maynooth University**  
National University of Ireland Maynooth

---

Presented by:  
**Robert Carton B.Sc. (Hons) Genetics and Cell Biology**

Department of Biology  
NUIM  
Maynooth  
Co. Kildare, Ireland

**March 2017**

**Supervisor:** Prof. Davide Pisani B.Sc., PhD (Bristol)  
**Co-supervisor:** Dr. David Fitzpatrick B.Sc., PhD (NUI)  
**Head of Department:** Prof. Paul Moynagh, B.A., PhD (Dublin)

## Table of Contents

<b>Index of Figures</b> .....	<b>xv</b>
<b>Index of Tables</b> .....	<b>xviii</b>
<b>Abbreviations</b> .....	<b>xx</b>
<b>Acknowledgements</b> .....	<b>xxiii</b>
<b>Declaration</b> .....	<b>xxv</b>
<b>Abstract</b> .....	<b>xxvi</b>

# Chapter 1: The Phylogenomic Era

<b>1.1 Introduction to Phylogenetics</b> .....	<b>1</b>
1.1.1 Darwin’s Theory.....	1
1.1.2 Homology.....	4
<b>1.2 The Evolution of Phylogenetics</b> .....	<b>7</b>
1.2.1 Morphology.....	7
1.2.2 The Advent of Molecular Data.....	8
1.2.3 Parsimony.....	9
1.2.4 Maximum Likelihood.....	9
1.2.5 Bayesian Inference.....	12
1.2.6 Stochastic and Systematic Error.....	14
<b>1.3 Bioinformatics Methodology</b> .....	<b>17</b>
1.3.1 Phylogenetic Reconstruction.....	17
1.3.2 Signal Dissection.....	19
1.3.2.1 The Slow / Fast Technique.....	19
1.3.2.2 Dayhoff Recoding.....	20
1.3.2.3 Taxon Pruning.....	21
1.3.3 The Basic Local Alignment Search Tool (BLAST).....	22
1.3.4 The Markov Clustering Algorithm (MCL).....	25
1.3.5 Gene Ontology.....	27
1.3.6 The Molecular Clock.....	30
1.3.6.1 Progressing from Strict to Relaxed Clocks.....	30

1.3.6.2 Bayesian Clock Models.....	32
1.3.6.3 The Fossil Record.....	34
1.3.7 Total Evidence Dating.....	37
<b>1.4 The Emergence of Phylogenomics.....</b>	<b>37</b>
1.4.1 Sequencing Technologies.....	37
1.4.2 The Polymerase Chain Reaction.....	39
1.4.3 Next Generation Sequencing.....	40
1.4.4 The Decline of Sequencing Costs.....	41
1.4.5 The Sequence Read Archive.....	43
1.4.6 Applying de Bruijn Graphs to De-Novo Short Read Assemblies.....	45
1.4.7 From Phylogenetics to Phylogenomics.....	46
1.4.8 Genomics versus Transcriptomics.....	48
1.4.8.1 Comparison of Sequencing Costs.....	49
1.4.8.2 Ortholog Coverage in Datasets.....	49
1.4.8.3 Scale and Complexity of Alternative Libraries.....	50
1.4.8.4 Repeats, Redundancy, and Isoforms.....	51
<b>1.5 The Protostomia.....</b>	<b>53</b>
1.5.1 Superphylum Ecdysozoa.....	55
1.5.2 The Platyzoa.....	58
1.5.3 Superphylum Lophotrochozoa.....	59

<b>1.6 Thesis Aims</b> .....	<b>64</b>
1.6.1 Chapter 2.....	64
1.6.2 Chapter 3.....	64
1.6.3 Chapter 4.....	65
1.6.4 Chapter 5.....	65
1.6.5 Chapter 6.....	66

## Chapter 2: Phylogenomics and the Case of the Rapidly Evolving Tardigrada

<b>2.1 Introduction</b> .....	<b>67</b>
2.1.1 Phylum Tardigrada.....	67
2.1.2 Significance of the Tardigrades.....	69
2.1.3 Conflicting Phylogenetic Theories.....	72
2.1.4 Long Branch Attraction within the Ecdysozoa.....	77
2.1.5 Dating the Tardigrada Origins with Newly Sequenced Taxa.....	77
2.1.6 Aims of this Study.....	78
<b>2.2 Materials and Methods</b> .....	<b>79</b>
2.2.1 Specimen Collection.....	81
2.2.1.1 Pycnogonid <i>Pycnogonum littorale</i> .....	81
2.2.1.2 Opilione.....	83
2.2.1.3 Limulus.....	83
2.2.1.4 Oniscidea.....	83
2.2.1.5 Onychophoran <i>Epiperipatus sp</i> .....	83
2.2.1.6 Halicyptus.....	83
2.2.1.7 Meiopriapulid.....	84
2.2.1.8 Kinorhynch.....	84
2.2.1.9 Tardigrade <i>Hypsibius dujardini</i> .....	84
2.2.1.10 Taxa Downloaded from the Sequence Read Archive.....	86
2.2.2 DNA & RNA Extractions.....	87
2.2.2.1 DNA Concentration and Purity Analysis.....	87

2.2.2.2 DNA Integrity Analysis.....	88
2.2.2.3 RNA Concentration and Purity Analysis.....	90
2.2.2.4 RNA Integrity Analysis.....	90
2.2.3 Genome and Transcriptome Sequencing.....	94
2.2.4 Data Quality Control.....	95
2.2.5 Transcriptome Assembly and Translation.....	98
2.2.6 Ortholog Mapping.....	101
2.2.7 Dataset Summary.....	106
2.2.8 Phylogenetic Reconstruction.....	106
2.2.9 Model Testing: Bayesian Cross Validation.....	107
2.2.10 Tardigrade Dataset: Signal Dissection.....	109
2.2.10.1 Slow / Fast Analysis.....	109
2.2.10.2 Dayhoff Recoding.....	111
2.2.10.3 Taxon Pruning.....	112
2.2.11 Divergence Time Estimation.....	112
<b>2.3 Results.....</b>	<b>114</b>
2.3.1 Tardigrade Phylogeny.....	114
2.3.1.1 CAT Model.....	114
2.3.1.2 GTR Model.....	115
2.3.1.3 CAT-GTR Model.....	115
2.3.1.4 Best Fitting Model for the Data.....	119
2.3.1.5 Tardigrade Slow / Fast Analyses.....	120
2.3.1.6 Dayhoff Recoding.....	121
2.3.1.7 Taxon Pruning.....	123

2.3.2 Tardigrade Divergence Time Estimation.....	125
2.3.2.1 Dating the Alternative Phylogenies.....	125
<b>2.4 Discussion.....</b>	<b>129</b>
2.4.1 Assessing De-Novo Assembly Methods.....	129
2.4.2 Ortholog Mapping.....	131
2.4.3 Phylogenomic Datasets.....	132
2.4.4 Phylogenomics and Systematic Error.....	133
2.4.5 The Importance of Model Testing.....	135
2.4.6 The Uncertainty of the Molecular Clock.....	136
2.4.6.1 Gaps in the Fossil Record.....	136
2.4.6.2 Fossil Identification.....	137
2.4.6.3 Crown and Stem Groups.....	138
<b>2.5 Conclusions.....</b>	<b>139</b>



## Chapter 3: Chaetognatha: The Mosaic Metazoans

<b>3.1 Introduction</b> .....	<b>143</b>
3.1.1 Chaetognatha: Ancient Predators.....	143
3.1.2 Deuterostomes or Protostomes?.....	145
3.1.3 Competing Phylogenetic Hypotheses.....	145
3.1.4 The Chaetognath Fossil Record.....	150
3.1.5 Lineage Characteristics and Ascribing the Amiskwia Fossil.....	152
3.1.6 Aims of this Study.....	154
<b>3.2 Materials and Methods</b> .....	<b>155</b>
3.2.1 Specimen Collection, gDNA Extraction, Sequencing, and Assembly..	157
3.2.2 Data Quality Control - Contaminant Screening.....	157
3.2.3 Ortholog Mapping.....	158
3.2.4 Dataset Summary.....	158
3.2.5 Phylogenetic Reconstruction.....	159
3.2.6 Model Testing: Bayesian Cross Validation.....	160
3.2.7 Chaetognath Dataset: Signal Dissection.....	161
3.2.7.1 Slow / Fast Analysis.....	161
3.2.7.2 Dayhoff Recoding.....	162
3.2.7.3 Taxon Pruning.....	162
3.2.8 Morphological Dataset.....	162
3.2.9 Divergence Time Estimation.....	163
3.2.10 Total Evidence Dating.....	164

<b>3.3 Results</b> .....	<b>166</b>
3.3.1 Chaetognath Phylogeny.....	166
3.3.1.1 CAT Model.....	166
3.3.1.2 GTR Model.....	168
3.3.1.3 CAT-GTR Model.....	170
3.3.1.4 Best Fitting Model for the Data.....	172
3.3.1.5 Protostome Phylogeny.....	173
3.3.1.6 Slow / Fast Analyses.....	175
3.3.1.7 Dayhoff Recoding.....	176
3.3.1.8 Taxon Pruning.....	177
3.3.1.9 Morphological Phylogeny.....	177
3.3.2 Chaetognath Divergence Time Estimation.....	179
3.3.2.1 Chaetognatha Origins under Clock Models.....	179
3.3.2.2 Total Evidence Dating.....	181
<b>3.4 Discussion</b> .....	<b>182</b>
3.4.1 Disparity Between Rocks and Clocks.....	182
3.4.2 Total Evidence Dating.....	183
3.4.3 Disagreement Amongst Signal Dissection Experiments.....	184
<b>3.5 Conclusions</b> .....	<b>186</b>

## Chapter 4: Arthropod Terrestrialization

<b>4.1 Introduction</b> .....	<b>190</b>
4.1.1 Arthropod Terrestrialization.....	190
4.1.2 Terrestrialization: A Complex Timeline.....	193
4.1.3 Aims of this Study.....	194
<b>4.2 Materials and Methods</b> .....	<b>196</b>
4.2.1 Generation of Molecular Libraries.....	196
4.2.2 Transcriptome Assembly and Translation.....	197
4.2.3 Ortholog Mapping.....	197
4.2.4 Divergence Time Estimation.....	198
<b>4.3 Results</b> .....	<b>199</b>
4.3.1 Divergence Time Estimation.....	199
<b>4.4 Discussion</b> .....	<b>202</b>
<b>4.5 Conclusions</b> .....	<b>203</b>

# Chapter 5: A Phylostratigraphic Study of Protein Family Evolution Across the Metazoa with Focus on the Protostomia

<b>5.1 Introduction</b> .....	<b>205</b>
5.1.1 Preliminary Study.....	205
5.1.2 Protein Families .....	206
5.1.3 Aims of the Study.....	207
<b>5.2 Materials and Methods</b> .....	<b>208</b>
5.2.1 Specimen Collection.....	210
5.2.1.1 Taxa Downloaded from the Sequence Read Archive.....	212
5.2.2 DNA & RNA Extractions.....	213
5.2.3 DNA & RNA Concentration and Integrity Analyses.....	213
5.2.4 Genome and Transcriptome Sequencing.....	214
5.2.5 Data Quality Control.....	215
5.2.6 Transcriptome Assembly and Translation.....	215
5.2.7 MCL: Protein Family Generation.....	217
5.2.8 Distribution of Protein Families Across a Metazoan Supertree.....	220
5.2.9 BLAST2GO: Annotating Protein Families.....	223
<b>5.3 Results</b> .....	<b>227</b>
5.3.1 Data Quality and Assembly Statistics.....	227
5.3.2 MCL Protein Clustering.....	227

5.3.3 Protein Families.....	227
5.3.4 Rate of New Protein Family Acquisition.....	230
5.3.5 Protein Family Annotation.....	232
<b>5.4 Discussion.....</b>	<b>241</b>
5.4.1 Adjusting the Balance of Sequenced Arthropods.....	241
5.4.2 Data Quality in the SRA.....	242
5.4.3 Future Improvements to Experimental Design.....	244
<b>5.5 Conclusions.....</b>	<b>245</b>

## Chapter 6: Thesis Discussion

<b>6.1 Phylogenomics, an Important Step Forward</b> .....	<b>251</b>
<b>6.2 Experimental Chapter Summaries</b> .....	<b>252</b>
6.2.1 Chapter 2.....	252
6.2.2 Chapter 3.....	252
6.2.3 Chapter 4.....	253
6.2.4 Chapter 5.....	254
<b>6.3 Suggested Alterations to Protostome Phylogeny</b> .....	<b>255</b>
<b>6.4 Reflections on the Cambrian Explosion</b> .....	<b>257</b>
<b>6.5 Discussion on Thesis Findings</b> .....	<b>258</b>
<b>6.6 Future Work</b> .....	<b>261</b>
6.6.1 SRA Data Quality Standards.....	261
6.6.2 Supplementing the Foundations of Phylogenomic Datasets.....	261
6.6.3 Phylostratigraphic Investigations of Protein Families.....	262
6.6.4 Total Evidence Dating of the Chaetognatha.....	262
6.6.5 Chaetognaths and Lophotrochozoan Phylogenomic Dataset.....	263

<b>Bibliography</b> .....	<b>264</b>
<b>Appendices</b> .....	<b>290</b>
<b>Supplementary Material</b> .....	<b>CD</b>
<b>Publications</b> .....	<b>296</b>

# Index of Figures

## Chapter 1

<b>Figure 1.1</b> The First Phylogenetic Trees.....	2
<b>Figure 1.2</b> Global versus Local Similarity.....	24
<b>Figure 1.3</b> The MCL Process.....	26
<b>Figure 1.4</b> Gene Ontology Example.....	29
<b>Figure 1.5</b> International Chronostratigraphic Chart.....	36
<b>Figure 1.6</b> The Decline of Sequencing Costs.....	42
<b>Figure 1.7</b> The Growth of the Sequence Read Archive.....	44
<b>Figure 1.8</b> The Protostomia.....	54
<b>Figure 1.9</b> The Arthropod Subphyla: Mandibulata versus Myriochelata.....	56
<b>Figure 1.10</b> Alternative Tardigrade Hypotheses.....	57
<b>Figure 1.11</b> Superphylum Lophotrochozoa.....	62

## Chapter 2

<b>Figure 2.1</b> Anatomy of the Tardigrades.....	68
<b>Figure 2.2</b> Alternative Tardigrade Hypotheses.....	73
<b>Figure 2.3</b> Flowchart Detailing Materials and Methods of Chapter 2.....	80
<b>Figure 2.4</b> Sequenced Taxa [A - H] and Focus of this Study [I].....	85
<b>Figure 2.5</b> Gel Electrophoresis of the <i>Opilione sp.</i> gDNA.....	89
<b>Figure 2.6</b> Bioanalyzer Readings.....	92



<b>Figure 2.7</b> Chromatogram Visualization and Corresponding Phred Scores.....	96
<b>Figure 2.8</b> Paralog Removal from Putative Orthologs.....	105
<b>Figure 2.9</b> Tardigrade CAT Phylogeny.....	116
<b>Figure 2.10</b> Tardigrade GTR Phylogeny.....	117
<b>Figure 2.11</b> Tardigrade CAT-GTR Phylogeny.....	118
<b>Figure 2.12</b> Tardigrade Dayhoff Recoding CAT-GTR.....	122
<b>Figure 2.13</b> Tardigrade Taxon Pruning CAT-GTR.....	124
<b>Figure 2.14</b> Tardigrade Molecular Clocks.....	126
<b>Figure 2.15</b> Comparing De-Novo Assembly Methods.....	130

## Chapter 3

<b>Figure 3.1</b> Chaetognath Phylogenies from Molecular Studies.....	149
<b>Figure 3.2</b> The Fossil Record of the Chaetognatha and Amiskwia.....	151
<b>Figure 3.3</b> Lineage Characteristics of the Chaetognatha.....	153
<b>Figure 3.4</b> Flowchart Detailing Materials and Methods of Chapter 3.....	156
<b>Figure 3.5</b> Chaetognath CAT Phylogeny.....	167
<b>Figure 3.6</b> Chaetognath GTR Phylogeny.....	169
<b>Figure 3.7</b> Chaetognath CAT-GTR Phylogeny.....	171
<b>Figure 3.8</b> Protostome CAT-GTR Phylogeny.....	174
<b>Figure 3.9</b> Chaetognath Morphological Phylogeny.....	178
<b>Figure 3.10</b> Chaetognath Molecular Clock.....	180
<b>Figure 3.11</b> Crown Group Lineages and Stem Group Fossils.....	182

## Chapter 4

**Figure 4.1** The Fossil Record of Terrestrial Arthropods.....192

**Figure 4.2** Molecular Clock Results: Re-viewing the Terrestrial Timeline.....201

## Chapter 5

**Figure 5.1** Flowchart Detailing Materials and Methods of Chapter 5.....209

**Figure 5.2** Specimens Collected and Sequenced for Chapter 5.....211

**Figure 5.3** Supertree Metazoa.....221

**Figure 5.4** Flowchart for Protein Family Annotation.....226

**Figure 5.5** Distribution of Metazoan Protein Families.....229

**Figure 5.6** Rate of Protein Family Acquisition.....231

## Chapter 6

**Figure 6.1** Revised Protostome Relationships.....256

# Index of Tables

## Chapter 1

<b>Table 1.1</b> Genomics versus Transcriptomics.....	52
---	----

## Chapter 2

<b>Table 2.1</b> Tardigrade Study: Transcriptomes Downloaded from the SRA.....	86
<b>Table 2.2</b> Tardigrada Study: Assembled and Translated Transcripts.....	100
<b>Table 2.3</b> Tardigrada Slow / Fast Dataset.....	110
<b>Table 2.4</b> Tardigrade Dataset: Molecular Clock Calibrations.....	113
<b>Table 2.5</b> Tardigrade Dataset: BCV.....	119
<b>Table 2.6</b> Tardigrade Slow / Fast Results.....	121
<b>Table 2.7</b> Summary of Ecdysozoan Divergence Dates under the CIR and U-GAMMA Models.....	128

## Chapter 3

<b>Table 3.1</b> Chaetognath Dataset.....	159
<b>Table 3.2</b> Chaetognath Slow / Fast Dataset.....	161
<b>Table 3.3</b> Chaetognath Dataset: Molecular Clock Calibrations.....	164
<b>Table 3.4</b> Chaetognath Dataset: BCV.....	172
<b>Table 3.5</b> Chaetognath Slow / Fast Results.....	175

## Chapter 4

<b>Table 4.1</b> Terrestrialization Study: Assembled and Translated Transcripts.....	197
<b>Table 4.2</b> Terrestrialization Study Molecular Clock Calibrations.....	198
<b>Table 4.3</b> Terrestrialization Study Divergence Time Estimation.....	200

## Chapter 5

<b>Table 5.1</b> Specimens Collected for Protein Family Study.....	210
<b>Table 5.2</b> Transcriptomes Downloaded from the SRA.....	212
<b>Table 5.3</b> Protein Families Study Assembled and Translated Transcripts.....	216
<b>Table 5.4</b> Influence of Inflation Rate on a 847,637 x 847,637 Protein Clustering Matrix.....	219
<b>Table 5.5</b> Prominent Functions of Metazoa Protein Families.....	233
<b>Table 5.6</b> Prominent Functions of Eumetazoa Protein Families.....	234
<b>Table 5.7</b> Prominent Functions of Bilateria Protein Families.....	235
<b>Table 5.8</b> Prominent Functions of Protostomia Protein Families.....	236
<b>Table 5.9</b> Prominent Functions of Ecdysozoa Protein Families.....	237
<b>Table 5.10</b> Prominent Functions of the Secernentea & Chromadorea Ancestor Protein Families.....	238
<b>Table 5.11</b> Prominent Functions of Arthropoda Protein Families.....	239
<b>Table 5.12</b> Prominent Functions of the Anopheles & Aedes Ancestor Protein Families.....	240

## Abbreviations

3-(N-morpholino)propanesulfonic acid (MOPS)

Absorbance (*A*)

Adenine (*A*)

Base pair (bp)

Basic local alignment search tool (BLAST)

Bayesian cross validation (BCV)

Biological process (BP)

Cellular location (CL)

Complimentary DNA (cDNA)

Credibility Intervals (CI)

Cytosine (*C*)

Deoxyribonucleic acid (DNA)

Diethylpyrocarbonate (DEPC)

Directed acyclic graph (DAG)

DNA Database of Japan (DDBJ)

Expectation value (E. value)

Expressed sequence tag (EST)

European Bioinformatics Institute (EBI)

General time reversible (GTR)

Gene ontology (GO)

Genomic DNA (gDNA)

Guanine (*G*)

High scoring pair (HSP)

International Nucleotide Sequence Data Collaboration (INSD)

Kilo base pairs (kbp)

Kishino and Hasegawa test (KH test)

Last common ancestor (LCA)

Long branch attraction (LBA)

Markov chain Monte Carlo (MCMC)

Markov cluster algorithm (MCL)

Maximum likelihood (ML)

Maximum parsimony (MP)

Maximum segment pair (MSP)

Messenger RNA (mRNA)

MicroRNA (miRNA)

Million years ago (MYA)

Mitochondrial DNA (mtDNA)

Molecular function (MF)

Multiple sequence alignment (MSA)

National Centre for Bioinformatics (NCBI)

Nearest neighbour joining (NNJ)

Next generation sequencing (NGS)

Non-redundant protein database (nr)

Nucleotide (nt)

Peptide (pep)

Polymerase chain reaction (PCR)

Posterior probability (PP)

Quality control (QC)

Ribonucleic acid (RNA)

Relative centrifugal force (RCF)  
Revolutions per minute (RPM)  
Sequence read archive (SRA)  
Shimodaira and Hasegawa test (SH test)  
Standard deviation (SD)  
Subtree pruning re-grafting (SPR)  
Terabases (Tb)  
Thymine (T)  
Total evidence dating (TED)  
Un-correlated gamma (U-GAMMA)  
Uracil (U)  
White Noise (WN)

---

## Symbols

Micro ( $\mu$ ) ( $10^{-6}$ )  
Millimolar (mM)  
Molar (M)  
Nanometer (nm)  
Nanogram (ng)  
Tera (T) ( $10^{12}$ )

## Acknowledgements

First and foremost, to my parents Tony and Cindy. Without such incredible support from them throughout my life I would never have reached this juncture. I don't think it is possible to overstate how fortunate I am to have them. Having great parents is probably the most important lottery one can win.

Secondly to Jenny, without you I would have given up, probably several times over. At times throughout this PhD of deep frustration and isolation you were there to bet on me and convince me to bet on myself. Thank you.

A big thank you to my PI and supervisor Davide Pisani who gave me such an awesome opportunity to work on his project and contribute to the scientific community. And for his knowledge and advice that steered me down this very long road. Thanks to James McNerney and David Fitzpatrick for taking over supervisor duties. The help is very much appreciated because I know nobody likes extra paperwork.

To my friends in the Maynooth lab: Lahcen, AJ, Bob, Karen, Aoife, Sinead, Leanne, and David. Unfortunately I didn't get to spend as much time working with you as I would have liked, but the time spent was immensely enjoyable. To the University of Bristol and the palaeobiology group who welcomed me in and treated me as one of their own. It has been an invaluable opportunity to learn from the likes of Jakob Vinther and Phil Donoghue, even with the consideration that Phil's favourite "Disney" film is Despicable Me 2. Also, a special thank you to Al and Suki!



Thanks to Eoin Mulville for his hard work with numerous DNA and RNA extractions. Eoin went way out of his way to help with my project.

Thanks to Queens University of Belfast for their accommodation and assistance when collecting specimens. However, standing in the middle of a freezing cold lough, at 6am, in November, was a pleasure I hope I will not encounter again.

Further acknowledgments to the Science Foundation of Ireland for funding this project and paying me for coffee breaks and discussing Noel's House Party for inordinate amounts of time with James Fleming.

This project would simply not have been possible without server access. A thank you to the Irish Centre for High End Computing, the University of Bristol servers, and of course the Maynooth bioinformatics server Darwin, our little engine that could.

Almost finally, don't forget where you came from. A shout out to Mary O'Connell's bioinformatics lab that has since moved from Dublin City University to University of Leeds. It was an honour to be part of the lab with Mary, Claire, Tom, Andrew, and Mark. The enthusiasm cultivated in that group kindled my interest in research.

Finally, for Chester, the defining example of how to play the cards you are dealt.

## Declaration

I certify that the Thesis is my own work and I have not obtained a Degree in this University or elsewhere on the basis of this Doctoral Thesis.

Signed \_\_\_\_\_

Robert Carton

## Abstract

This *magnum opus* concerns the generation of genomic level data, through next generation sequencing technologies, and the application of these new molecular libraries to various aspects of protostome evolution.

In Chapter 1 I introduce the most important contribution to the field of evolution: Darwin's Theory of natural selection, the keystone to our current understanding and methodologies. Following this I discuss the first applications of such knowledge to morphological and molecular data, the theory behind the bioinformatics techniques used to analyse such information, and how recent advancements in sequencing technologies have opened the door to large-scale studies of evolution. After these principles are established a summary of the clade of animals that are the focus of this thesis: the Protostomia is provided.

Chapter 2 is a study of a remarkably adaptable group of ecdysozoans called the tardigrades or "water bears". Their rapidly evolving nature has made the phylogenetic affinity of the Tardigrada ambiguous with three alternative hypotheses contesting their placement. A phylogenomic approach was implemented in order to clarify their position in conjunction with signal dissection experiments to minimize systematic error. The origin of the Tardigrada was also investigated using a series of molecular clocks. Important findings from this chapter include evidence that the tardigrade-nematode grouping is a systematic artifact known as long branch attraction, their true affinity lying with the onychophorans, and that the tardigrade lineage diverged some 480 million years ago in the Lower Ordovician period, slightly older than previously thought (Rota-Stabelli *et al.* 2013).

Chapter 3 comprises a detailed study of one of the first animal predators to originate: the Chaetognatha. Molecules and morphology have clashed on its bilaterian affinity and molecular studies remain in wide disagreement as to their exact position with the Protostomia. The newly sequenced *Parasagitta sp.* genome was incorporated in to a pre-existing ecdysozoan dataset and phylogenomic reconstruction methods and divergence time estimation experiments were implemented to uncover the eventful 520 million year evolutionary history of these ancient carnivores.

Results from these experiments show that the chaetognaths are unequivocally protostomes with a deuterostome-like development and that their true placement within this group is as basal lophotrochozoans. Moreover a large discrepancy discovered between the age of the fossil and extant chaetognaths points to an extinction event within the lineage that had previously not been reported. This signifies a remarkable example of an animal that has undergone a complete role reversal in the food chain during its 500 million year reign: from ancient predators to contemporary prey.

Chapter 4 describes my involvement in a collaborative work on a palaeobiological exploration of arthropod terrestrialization. New molecular libraries from the Crustacea and Myriapoda were used to investigate the independent colonization events of the arthropod subphyla. Our results from this study support Erwin *et al.* (2011) and dos Reis *et al.* (2015) findings of a Cryogenian origination for the Metazoa and a Lower Cambrian radiation of animal lineages. The arachnids were the last of the terrestrial arthropods to colonize land in the Upper Ordovician (450 MYA), with the hexapods colonizing land in the Lower Ordovician (483 MYA) broadly in agreement with the fossil record. However the Myriapods colonized land twice, initially in the Cambrian (543 MYA) and then in the Lower Ordovician (473

MYA). The implications of which suggest that terrestrial ecosystems capable of supporting life existed as far back as the Cambrian time period over 500 MYA.

Finally Chapter 5 details the application of phylostratigraphy to a large-scale study of novel protein families spanning the Metazoa with a focus on the Protostomia and Ecdysozoa. Twenty-eight taxa from next generation sequencing experiments contributed to this work and were the subject of homology searches using BLAST and protein family clustering using MCL, which were then distributed across a forty-eight node cladogram ranging from the roots of the Animal Kingdom to the tips of the arthropod subphyla. The rate of protein family acquisition has increased in ancient high-level taxonomic nodes compared to that of the younger nodes and lineages in the tree, suggesting that protein families functioning within extant animals have existed for a considerable amount of time before these animals diverged. The tweaking of these families, more so than the gradual increase in family numbers, has influenced the evolution and diversification of the animal lineages existing today.

### 1.1 Introduction to Phylogenetics

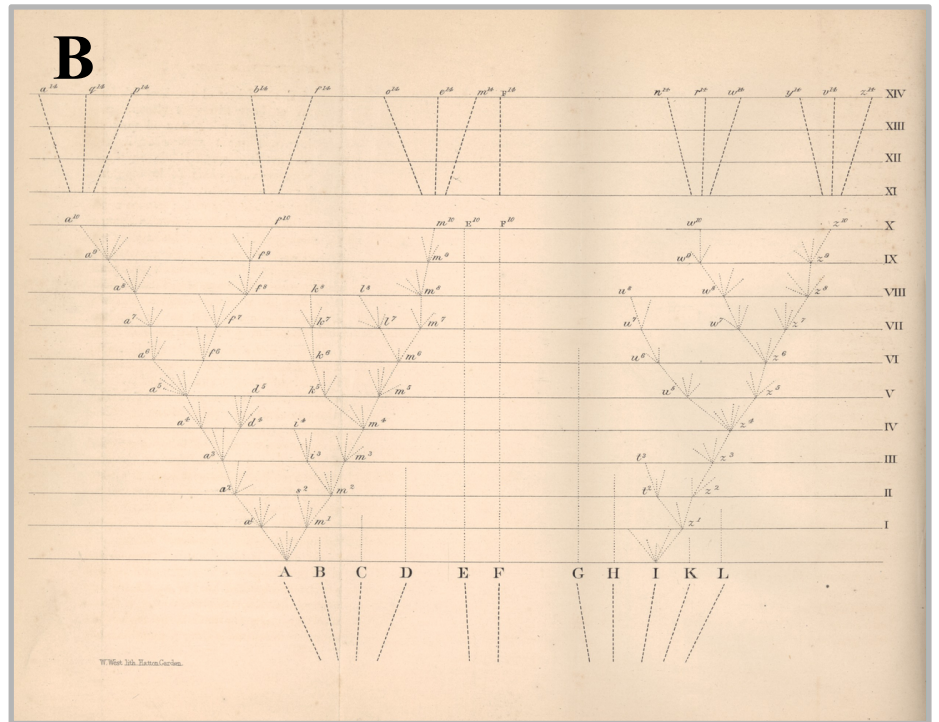
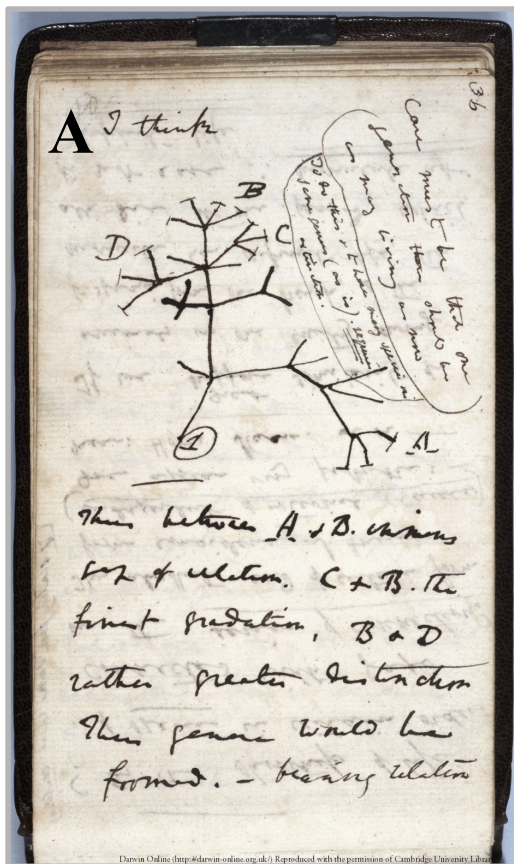
#### 1.1.1 Darwin's Theory of Evolution

The works of Charles Darwin (1809 – 1882) are considered to be the most important contributions to the field of evolutionary biology and indeed, some of the most influential findings to the scientific community as a whole. Darwin is credited for offering the first scientifically sound explanation for the mechanism of evolutionary inheritance. Writing in his most famous manuscript:

*On the Origins of the Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*

(Darwin, 1859), Darwin describes a system of inheritance whereby extant lineages are descendants of a shared extinct ancestor from which they originated. This system takes the form of a branching pattern, whereby living lineages radiate from ancestors that have themselves previously radiated from an even older ancestor. This branching pattern of evolutionary inheritance by common ancestry would be later coined *phylogeny* by Ernst Haeckel (1834 – 1919). The word phylogeny derives from the Greek words “*phylê*” and “*geneia*”, translating to “race” “origins”. Musings on the branching pattern of evolutionary inheritance by way of common ancestry took the form of tree drawings, the first of which can be found in Darwin's Notebook B (Darwin, 1837), drawn shortly after his voyage around the world on the H. M. S.

Beagle [Figure 1.1 A]. Darwin would later go on to include a more descriptive phylogenetic tree in *On The Origins of the Species* (Darwin, 1859) [Figure 1.1 B]. To this day, over 150 years later, studies of evolutionary inheritance are still described using phylogenetic trees.



**Figure 1.1: The First Phylogenetic Trees**

**Figure 1.1 A:** The first piece of physical evidence for the use of a tree structure to explain inheritance. Darwin writes “Thus between A & B immense gap of relation. C & B the finest graduation, B & D rather greater distinction. Thus genera would be formed - bearing relation”. Describing thoughts on how to structure radiations from speciation and map their relation in 1839. Source: [Darwin-online.org.uk](http://Darwin-online.org.uk)

**Figure 1.1 B:** A more sophisticated illustration of evolutionary inheritance by way of common ancestry, from *On the Origins of the Species* (Darwin, 1859).

This idea of a phylogenetic tree explained the common features shared between similar organisms. They had inherited these features from a common ancestor. The further one travels towards the roots of the tree, the older the age. The tips of the tree exist in the present time. The most groundbreaking revelation from *On The Origins of*

*the Species* was Darwin's explanation for the mechanism that drives this system of inheritance from a last common ancestor (LCA). He called it *natural selection*. The term "natural selection" describes a process whereby the survival of a species is directly influenced by adaptations to their environment and ability to reproduce. These adaptations are inherited from one generation to the next eventually becoming fixed at a population wide level. Those that did not possess such an adaptation or trait struggled to survive in challenging environments and thus experienced lesser life spans, allotting them a smaller period of time to breed. Thus over time the section of the population missing such a necessary trait to survive dwindles as fewer and fewer live long enough to produce offspring that would inherit their "flawed" characteristics. Natural selection is the driving force behind evolution but what was not known until long after the discovery of DNA and its role in inheritance, is that the process begins with a change at the molecular level of an organism. The newly acquired traits that we speak of are the product of spontaneous DNA mutations that change the genetic code, and consequently, its protein product (Eyre-Walker & Keightley, 2007). These non-synonymous mutations will often be deleterious and cause severe biological disorders at the molecular level resulting in death. However, seldom these mutations, or a series of these mutations over time, will amount to a beneficial advantage at the phenotypic level giving the organism a selective advantage in their environment and increase their chances of living long enough to pass on these adaptations. Hundreds of millions of years of evolution by way of natural selection has woven a tangled web of phylogenetic relationships which evolutionary biologists are keen to tease apart.



### 1.1.2 Homology

Richard Owen (1804 - 1892) conceptualized homology in 1843 defining it as “The same organ in different animals under every variety of form and function” (Owen, 1843). Essentially this was the first description of the same anatomical features in different animals. This term was applied to Darwin’s findings and re-defined in the context of evolution: homologous traits amongst lineages are deemed so because of their inheritance from a shared common ancestor, i.e. evidence that these lineages are related to some degree (Munkres, 1984). Ergo such traits will be similar, or even identical, depending on how much time has passed since the speciation event and the usefulness of the trait. Homologous traits allow us to trace the speciation events and radiating lineages of a common ancestor. It is important to distinguish homology from sequence similarity as homologous characters can be, and usually are, very similar, but similar sequences are not necessarily homologous (Fitch, 2000). Inferring the latter can lead to incorrect phylogenetic assumptions (Jensen, 2001). Homology exists in three main forms: orthology, paralogy, and xenology (Fitch, 2000). Orthology is of most relevance to this thesis; it traces homology through speciation events. For example, a gene inherited from a last common ancestor by two or more lineages would be deemed an ortholog. Orthology is a critical concept when supplementing phylogenetic datasets with newly sequenced molecular libraries, as they are markers of speciation events. Therefore the phylogeny of a set of orthologous sequences is exactly the same as the phylogeny of the taxa from which they are sourced (Fitch, 2000). Ortholog identification and mapping strategies for newly sequenced taxa are outlined in **Materials and Methods 2.2.6**. Paralogy comprises the divergence of two or more sequences following a duplication event. Paralogs present a problem for homolog studies as genes can duplicate within the same organism meaning paralogs

are not necessarily indicative of speciation events and are therefore considered as homoplastic traits when identifying homology through orthology (Koonin, 2005). Gene duplications are extremely common in molecular biology as they are one of the primary mechanisms of adaptation through novel and sub functionalization (Innan & Kondrashov, 2010). Indeed there is evidence that the entire genome of the ancestral vertebrate duplicated at least twice (Dehal & Boore, 2005).

As such, one must be vigilant of erroneously including paralogs in ones analysis. With this in mind, it is important to discuss the nuances of paralogy which are classified in three forms: out-paralogs, in-paralogs, and pseudo-paralogs (Koonin, 2005) and moreover how they are identified and avoided in this thesis. Out-paralogs are described as gene duplication products occurring *before* a speciation event (Koonin, 2005). Therefore the threat of out-paralogs in deep node datasets can be negated through strict sequence similarity search criteria using BLAST (Altschul, 1990). This is because the paralogous gene product from a gene duplication event will invariably be under less selective constraints than its relative ortholog (Hurles, 2004) as changes in the genetic code of the paralog are needed to undergo adaptations of novel or sub functionalization or even the more dramatic changes required for function loss via pseudogenesis facilitated by genetic drift (Lynch *et al.* 2000). This, in addition to the old age of deep node datasets used in this thesis, means that dozens if not hundreds of millions of years have passed since the formation of out-paralogs in the molecular libraries used and over time the sequence composition of these duplicates should have changed enough to be weeded out by strict BLAST (Altschul, 1990) searches which have been incorporated into the ortholog mapping strategy [Materials and Methods 2.2.6].

In-paralogs are defined as gene duplications occurring lineage specifically *after* a speciation event (Koonin, 2005). This makes in-paralogs more problematic than out-paralogs as their lineage specific and post-speciation nature means they could have duplicated recently in extant lineages and not enough time may have passed for these sequences to change substantially to the original gene. This can make discerning them from the correct ortholog difficult, even with strict sequence similarity searches. A tree building strategy was implemented in this thesis to avoid in-paralogs, a stage in the ortholog mapping process [**Materials and Methods 2.2.6**] where paralogs very similar to their respective ortholog are pruned based on their branch length and position in a gene tree.

Pseudo-paralogs are a product of horizontal gene transfer (Koonin, 2005), a non-conventional form of gene inheritance amongst the metazoans (similar to that of xenologs discussed below) and so are not accounted for in the methods. In the unlikely event of a pseudo-paralog being included in the large gene number datasets used in the following studies its artifactual discrepancy would be so small that it would be drowned out by the phylogenetic signal of hundreds of concatenated orthologs.

Xenology is thought to be a rare occurrence in the Animal Kingdom but there are some possible cases (Boto, 2014). It is the mechanism whereby genes are shared through horizontal gene transfer, the transfer of genes via non-inheritance e.g. virus gene transfer or sharing of genes between bacteria to promote antibiotic resistance (Dzidic & Bedekovic, 2003). Although not a problem concerning metazoans for the most part, xenologs create a major issue for phylogenetic reconstruction studies of the prokaryotes (Philippe & Douady, 2003).

## 1.2 The Evolution of Phylogenetics

### 1.2.1 Morphology

Before the discovery of the genetic structure and code (Watson & Crick, 1953; Crick *et al.* 1961; Nirenberg & Matthaei, 1961) studies of homology were based entirely on morphological characteristics. The premise was simple, similar traits were likely homologous and thus such similar traits found in different species must have originated in a common ancestor (Stevens, 1984). While a logical assumption on the surface, this was not an accurate reflection of natural inheritance and led to the invasion of homoplasy into studies, obfuscating the true evolutionary paths that biologist were trying to uncover. Such homoplastic characters were often caused by loss of traits, perhaps the most famous example pertaining to the primates. The ancestral primate possessed a tail, which has been conserved in most extant primates but lost in others, most noticeably *Homo sapiens* (Fleagle, 2013). Another example of homoplasy mistaken for homology is in the case of convergent, or parallel, evolution. In such a case similar traits are found amongst lineages but are not the product of ancestral inheritance. Instead these traits evolved independently, often by chance in response to a common environmental hurdle, such as the Diptera and Aves. Both have flight capabilities but are from opposing bilaterian groups: the protostomes (Yeates & Wiegmann, 1999) and deuterostomes (Sibley & Ahlquist, 1990) respectively.

Contemporary morphological studies are far more sophisticated however in identifying such homoplasy (Martin & Luo, 2005 and Luo *et al.* 2007) and the concept of morphological homology is still of essential importance in the field of molecular evolution when studying divergence time estimation as molecular clocks must be grounded in the fossil record (Yang & Donoghue, 2016).

### 1.2.2 The Advent of Molecular Data

It was not until the late 1960s that molecular data was applied to phylogenetics (Fitch & Margoliash, 1967). This development allowed scientists to investigate evolution from the underlying genotypic level as opposed to the morphologically based phenotypic level, revealing a new source of phylogenetic signal previously hidden. Up until the last decade molecular evolution projects were hampered by data restrictions because the generation of molecular data was painstaking and expensive (Sanger *et al.* 1977). Consequently datasets involving metazoan representatives were restricted to small libraries such as ribosomal subunits (Aguinaldo *et al.* 1997), mitogenomic studies (Miya & Nishida, 2000) and libraries consisting of limited numbers of genes and taxa (Pisani *et al.* 2004).

The application of homology changes slightly when moving from morphological to molecular studies. Morphological homology concerns physical traits and whether they are homologous or not. Molecular homology tends to refer to the relationship of many individual characters (nucleotide or amino acid residues) in a multiple sequence alignment (MSA) and whether they are homologous to one another (Thompson *et al.* 1994 and Edgar, 2004). A phylogenetic tree is reconstructed based on the level of homology observed across the sum of the residues for each taxon in the MSA. For the purpose of this thesis the MSA software used was MUSCLE (Edgar, 2004), a sequence aligner used in many large-scale phylogenetic investigations (Savard *et al.* 2006; Dunn *et al.* 2008; Wheeler *et al.* 2009; Pyron & Wiens, 2011; Floudas *et al.* 2012; Zhang *et al.* 2014; Hug *et al.* 2016) and considered an adequate tool for phylogenomic applications (Yang & Rannala, 2012 and Philippe *et al.* 2017).

### **1.2.3 Parsimony**

The first widely used approach to phylogenetic inference using molecular data was with the utilization of parsimony (Edwards & Cavalli-Sforza, 1963). Parsimony takes a path of least resistance approach to phylogeny, assuming the simplest scenario is the correct one. The objective of maximum parsimony is to identify the most realistic tree, with the least number of steps (character changes) that accurately represents the characters given. Essentially the phylogenetic tree with the least number of mutations is considered the most valid. Parsimony fell out of favour as a tool for molecular studies as increasingly sophisticated models, less susceptible to homoplasy and more efficient in describing complex evolutionary scenarios, came to the fore (Yang & Rannala, 1997; Guindon & Gascuel, 2003; Kolaczkowski & Thornton, 2004; Gadagkar & Kumar, 2005). Consequently, most parsimonious tree reconstruction methods are not pursued in this thesis.

### **1.2.4 Maximum Likelihood**

Maximum likelihood (Edwards & Cavalli-Sforza, 1964) introduced a more sophisticated probabilistic search for choosing the best phylogenetic representative of a multiple sequence alignment. ML calculates the likelihood (L) of observing the given data (D) based on two hypotheses: a model of evolution (M), and a tree that is condition to change (T). This can be presented as the following formula:

$$L = p ( D | M, T )$$

The data (D) is typically a nucleotide or amino acid multiple sequence alignment. ML applies these parameters to randomly generate phylogenetic trees of conditional topology and branch length. It then searches through some of the trees generated (“tree space”) in an attempt to find the best tree, i.e. the tree with the highest likelihood. Tree space can be thought of as a three dimensional graph of peaks and troughs. The peaks represent trees with high likelihoods whereas the troughs are areas of improbable trees that do not model the MSA particularly well. The tree with the best likelihood is known as the global maximum, the highest peak on the graph of tree space (Guindon & Gascuel, 2003). ML runs its search through tree space by taking a hill climbing approach, a mathematical method whereby not all the possible trees are considered (Guindon & Gascuel, 2003). Hill climbing involves arbitrarily beginning at a random point on the graph of tree space, although the operator can be directed with a non-randomized approach under nearest neighbour joining (NNJ) and subtree pruning re-grafting (SPR) methods (Felsenstein, 2004) but are not applied in this thesis. The ML operator, in the case of this study PhyML (Guindon *et al.* 2010), searches in a single direction until it travels up a slope and reaches a peak. The likelihood value of this peak is logged. Random searches in tree space continue with the aim of locating a peak that is taller than the highest likelihood logged. Given enough time there is considerable chance of finding the global maximum, if not then local maximum trees, which are not the most optimal tree but still represent a very high likelihood of the given data (Fan *et al.* 1998).

Hill climbing approaches are not just advantageous to ML but a requirement as permuting every tree possible is computationally exhaustive and even numerically impossible for datasets of twenty or more taxa because of the combinatorial tree

explosion: where the number of possible trees for a character matrix grows exponentially for each taxon added (Boudali & Duga, 2005).

Since there is no guarantee that the global maximum will be found, it is important to test the credibility of the tree that the ML operator estimates to be the most probable phylogenetic reconstruction for the data. The credibility of phylogenetic estimations from ML can be evaluated using a technique called bootstrapping (Efron, 1979), first applied to phylogenetics by Felsenstein (1985).

Bootstrapping is a statistical resampling technique with replacement, not to be confused with the older statistical technique of jackknifing which resamples without replacement (Felsenstein, 1985) - therefore replicating a diminished matrix. While not useful in this circumstance, a diminishing resampling of a matrix does have its own benefits, see **Materials and Methods 2.2.9** where jackknifing aids in alleviating the computational burden of Bayesian cross validations.

Bootstrapping resamples the matrix (multiple sequence alignment) with replacement a specified number of times to create a specified number of dataset replicates. These datasets are then run under ML, each producing a tree that the operator estimates is representative of the highest likelihood. A consensus tree is formed from all the “best” trees from the replicated datasets. The tree generated by ML from the original matrix is compared to the consensus tree and each node is assigned a bootstrap support value (BP) based on the parity between the two trees. In cases where a definitive topology must be chosen between a conflicting “best supported” tree and consensus tree a KH test (Kishino & Hasegawa, 1989) can be useful in selecting the “true” tree. The KH test compares two topologies in respect to an aligned set of molecular sequences and estimates the variance of their different log-likelihoods (Kishino & Hasegawa, 1989). When comparing more than two topologies a modified



version of the KH test named the SH test (Shimodaira & Hasegawa, 1999) is more suitable as it takes the account the additional number of tests required when considering a multitude of topologies.

A ML strategy was implemented in this thesis as part of the gene tree building step of the ortholog identification and mapping process using PhyML (Guindon *et al.* 2010), see **Materials and Methods 2.2.6**.

### 1.2.5 Bayesian Inference

Bayesian mathematics was devised over 250 years ago by Thomas Bayes in 1763. Phylogenetic inference with the application of Bayesian probability is the most popular contemporary method of tree reconstruction (Yang & Rannala, 2012 and Chen *et al.* 2014). In this thesis, all major phylogenetic trees and divergence time estimations have been reconstructed with the application of a Bayesian framework of probability.

Bayesian probabilistic methods are similar to likelihood methods previously described, both have likelihood functions incorporated within, but with a significant difference: Bayesian inference estimates phylogenetic trees using a prior probability distribution, an initial observation a priori to any information about the data being taken into account (Felsenstein, 2004). Bayesian models employ MCMC chains that alter their prior probability estimations through observations made during the process; they then proceed to apply this knowledge to further observations forming posterior probabilities (Yang & Rannala, 1997). Whereas ML attempts to improve on its estimations in a less sophisticated way when applying hill climbing models to tree space graphs (Edwards & Cavalli-Sforza, 1964). However ML is still highly relevant

to Bayesian methods as its function is altered with a prior probability in order to calculate posterior probabilities (Felsenstein, 2004). Bayes' theorem is described by the following formula:

$$p[ H | D ] = \frac{p[ H ] \times p[ D | H ]}{p[ D ]}$$

Where H and D are the hypothesis and data respectively,  $p[ H | D ]$  is the posterior probability distribution,  $p[ H ]$  is the prior probability (PP) and  $p[ D ]$  is the normalizing constant that ensures the PP distribution integrates to 1 (Vapnik, 1998). The hypothesis is represented by a topology, branch lengths and a substitution model, while the data consists of the MSA representing the taxa.

If Bayesian methods have existed for over 250 years then why have they only rose to prominence in evolutionary studies relatively recently? The answer was mentioned earlier: the Markov chain Monte Carlo (MCMC) (Metropolis *et al.* 1953 and Hastings, 1970). The application of MCMC algorithms to Bayesian methods made a previously incalculable parameter, the random sampling of the PP, calculable (Yang & Rannala, 1997). The PP occupies a very small portion of tree space, thus finding it with previously applied methods such as hill climbing models is statistically unlikely (Douady *et al.* 2003). MCMC bypasses this issue because regardless of the starting points of two independent chains, they will converge on the PP (Lemey *et al.* 2009). In order for this to be possible, MCMC chains require an interval of time to “warm up” as two chains can theoretically start anywhere on the graph and thus need time to assess observations, recalculate, and learn in order to reach proximity. This is often referred to as the “burn in” (Lartillot *et al.* 2009).

With the knowledge of how both Bayesian inference and ML operate, one can conclude that the Bayesian approach has a greater probability of identifying the true phylogeny of a particular dataset as it is not reliant on stumbling across the global maximum, instead taking a more deliberate approach of constantly observing, testing, and reassessing parameters in order to incrementally raise the likelihood of finding the best tree. A further advantage of Bayesian probability over ML is that the former does not require additional resampling tests such as bootstrapping (Efron, 1979 and Felsenstein, 1985), to assess the confidence of the tree, which is computationally costly.

### **1.2.6 Stochastic and Systematic Error**

Errors leading to false topologies in phylogenetic reconstruction studies are categorized into two forms: stochastic and systematic.

Stochastic, or sampling, errors tend to occur in gene restricted datasets (Doerr *et al.* 2012) where lack of data, and thus low phylogenetic signal has an obscuring influence on the topology. This was a persistent problem in small phylogenetic datasets pre-NGS technology, see work from Giribet *et al.* (2001) and Pisani *et al.* (2004) for examples of this error in an ecdysozoan datasets, as acquiring enough genes to adequately represent the taxa of interest was an expensive and long process (Sanger *et al.* 1977).

Stochastic error has been virtually eradicated in the phylogenomic era as the addition of genes to appropriate taxa erases the bias completely (Jeffroy *et al.* 2006). Therefore the threat of stochastic error is negligible in phylogenomic studies such as those presented in this thesis.

Systematic error, occurring when the assumptions of sequence evolution models inaccurately describe the data (Philippe *et al.* 2017), is a more persistent problem. This is because unlike stochastic biases, the addition of more information does not remove the error but conversely increases it (Phillips *et al.* 2004; Philippe *et al.* 2005; Jeffroy *et al.* 2006; Rodriguez-Ezpeleta *et al.* 2007). Consequently underlying systematic errors are augmented in phylogenomic datasets. Systematic error can be reviewed under three categories: compositional bias, heterotachy and long branch attraction (Philippe *et al.* 2017).

Compositional bias groups sequences based on similarities between their nucleotide or amino acid composition instead of their respective homology. The most prominent form of compositional bias influencing phylogenomic studies is strand asymmetry associated with mitochondrial DNA (mtDNA) datasets (Rota-Stabelli & Telford, 2008). Since this type of data is not used in the following experiments, compositional bias is not a pertinent source of systematic error in this thesis.

Heterotachy describes the nature of rate heterogeneity (variance in site substitution across a molecular sequence over time) (Kolaczkowski & Thornton, 2004). This process is difficult to account for when applying evolutionary models to multiple sequence alignments, particularly concatenated superalignments, as the inconsistency of non-uniform site substitution is computationally expensive to account for (Kolaczkowski & Thornton, 2004). Using models that do not account for heterotachy can result in phylogenetic inconsistency (Philippe *et al.* 2005 and Rodriguez-Ezpeleta *et al.* 2007). Approaches to account for such pitfalls in evolutionary modeling (Ronquist & Huelsenbeck, 2003), do so at the cost of speed and computational power (Kolaczkowski & Thornton, 2004). Site-heterogeneous models of evolution,

incorporated into a Bayesian probability framework, can reduce the risk of this form of systematic error (Lartillot & Philippe, 2004 and Lartillot *et al.* 2009), see **Introduction 1.3.1** for a description of these models.

Long Branch Attraction (LBA) is a systematic error that biases phylogenetic studies involving rapidly evolving taxa. First described by Felsenstein in 1978, LBA influences a dataset by aggregating long branched taxa in a phylogeny by mistaking their commonality of saturated sites for similarity based on a shared ancestor. The very nature of LBA disrupts phylogenetic reconstruction and has generated false reconstructions in numerous studies across the tree of life (Brinkmann & Philippe 1999; Stiller & Hall, 1999; Reyes *et al.* 2000; Omilian & Taylor, 2001; Inagaki *et al.* 2004; Stefanovic *et al.* 2004; Brinkmann *et al.* 2005; Dabert *et al.* 2010; Boussau *et al.* 2014), particularly for its propensity to artificially inflate confidence values of nodes in the tree. Because of its systematic positively misleading nature, the use of large phylogenomic datasets does not nullify, but contrarily tends to amplify the artifact, although adding new taxa to “break up” long branches can be useful in reducing the error (Bergsten, 2005).

LBA can be aggravated by additional factors such as poor taxon sampling or distant outgroups (Lartillot *et al.* 2007), so careful choice of these parameters for the evolution study in question is important. Certain unavoidable aspects of phylogenetic reconstruction are prone to LBA such as designating an outgroup. The branch leading to the outgroup will be long by default and as a consequence can drag long branched in-groups towards the base of the tree (Philippe *et al.* 2005)

Reconstruction methods such as parsimony are particularly prone to LBA but maximum likelihood and the Bayesian approach are not immune either (Bergsten,

2005). Appropriate measures to reduce the effects of LBA include using a taxon rich dataset, slowly evolving genes, and choice of suitable probabilistic tree reconstruction models (Lartillot *et al.* 2007). However when studying rapidly evolving groups such as the tardigrades or nematodes (see Chapter 2), LBA cannot be avoided, even using large amounts of molecular data (Philippe *et al.* 2011a), and so the best approach is to minimize its deleterious effects through a number of signal dissection measures (Brinkmann & Philippe, 1999; Dayhoff *et al.* 1968; Aguinaldo *et al.* 1997).

## **1.3 Bioinformatics Methodology**

Bioinformatics describes the overlap between computer and biological sciences where software is developed for the express purpose of generating, processing, parsing or deciphering biological data (Higgs & Attwood, 2013). Described herein is the theory of the methodologies central to the experimental chapters in this thesis.

### **1.3.1 Phylogenetic Reconstruction**

The Bayesian framework for estimating phylogenies has been described but it is also important to discuss the models of evolution used in this thesis that are incorporated into the framework. Evolutionary models consist of three main elements: the tree, the rate of character exchange (the frequency or pattern of nucleotide or amino acid substitution), and the composition vector (Lio & Goldman, 1998). The latter two components are referred to as the underlying process of the model. The models used in this thesis are the general time reversible (GTR) (Tavaré, 1986) CAT (Lartillot & Philippe, 2004), and combined version of the two: CAT-GTR (Lartillot *et al.* 2009),

which takes advantage of the strengths of both models but at the cost of computational resources.

The GTR model of evolution was designed with nucleotides in mind as opposed to amino acids, it is a more sophisticated version of the Jukes-Cantor (JC) model which assigns equal weights to exchange rate frequencies between the nucleotide bases (Jukes, 1969). This early model was primitive because assuming equal probabilities of nucleotide substitution is biologically unrealistic for the basic reason that transitions are more likely to occur than transversions; purines (adenine and guanine) are more likely to be substituted for other purines and as opposed to pyrimidines (cytosine, thymine, and uracil) and vice-versa due to the difference in molecular structure of the two types of nucleotide bases (Kimura, 1980). The GTR model takes this in to account and additionally is flexible in estimating any type of nucleotide substitution (purine to pyrimidine for example) but at an appropriately weighted exchange rate frequency that takes into account the likelihood of it happening (Tavaré, 1986).

The CAT model is a mixture model that incorporates both discrete and indiscrete site heterogeneity (Lartillot & Philippe, 2004). Site heterogeneous aware models are considered to be more biologically accurate than site homogenous models as they account for differing rates of evolution within genes (Lartillot & Philippe, 2004 and Lartillot *et al.* 2007). CAT was devised for amino acid substitutions as the scaffold of the model is founded on a probability vector that assigns different likelihoods to the substitution of a character based on the biochemical profile of the twenty amino acids (Lartillot & Philippe, 2004). Incorporating these two models, forming CAT-GTR, entails availing of the site rate substitution flexibility of the GTR model with the site heterogeneity and amino acid probability profile of the CAT model, the combination of which has been demonstrated to be more resistant to systematic errors caused by

saturated sites in large amino acid character phylogenomic-scale datasets (Lartillot *et al.* 2007)

### **1.3.2 Signal Dissection**

Signal dissection is becoming a necessary step in the phylogenetic reconstruction of lineages prone to LBA, and can increase the robustness of phylogenetic reconstructions of otherwise ambiguously placed clades. There are three main approaches to tackle the LBA artifact: the slow / fast technique, Dayhoff recoding, and taxon pruning (Brinkmann & Philippe, 1999; Dayhoff *et al.* 1968; Aguinaldo *et al.* 1997).

#### **1.3.2.1 The Slow / Fast Technique**

For taxa influenced by LBA it can be useful to parse the dataset into categories based on their rate of evolution. This can be achieved through the slow/fast method outlined by Brinkmann & Philippe (1999). This is achieved by dividing the MSA into a series of monophyletic groups and calculating the number of changes per site (Swofford, 2002). The characters are then ordered in terms of how often they have changed and ranked in to categories of slow or fast rates of substitution (Brinkmann & Philippe, 1999). Once separated into these rate evolution categories, the fastest and slowest characters of the datasets can then be reconstructed as phylogenies and compared to the results of the original full size dataset. Major differences between the datasets can highlight the presence of LBA. It is possible to demonstrate the presence of LBA in a dataset by comparing phylogenies representing the fastest and slowest evolving sites to the original dataset. Under a scenario where LBA is in effect, one expects to see



the grouping of rapidly evolving clades in the phylogenies reconstructed from the fastest evolving sites and a gradual separation of these clades as the fastest characters are progressively stripped away.

Stripping away characters in a dataset alters the underlying phylogenetic signal as one could improve or reduce the signal depending on the characters removed. Generally speaking, the fastest evolving characters contain less signal than the comparatively slower evolving characters as these sites often become saturated from high levels of residue change causing the initial composition of the sequence to be masked or even lost (Wenzel & Siddall, 1999 and Xia *et al.* 2003). However although useful as a general rule, this is not strictly always the case and because the slow / fast method is somewhat blind - we cannot identify where the signal is, just rank the characters by their rate of evolution. The approach taken in this thesis was to compare a series of datasets of increasing rate of evolution and decreasing rate of evolution, in intervals of ten percent, in order to identify a pattern of change as opposed to arbitrarily testing a single subset of the data (see Borner *et al.* 2014).

### **1.3.2.2 Dayhoff Recoding**

The Dayhoff recoding method (Dayhoff *et al.* 1968) recodes similar amino acids into single characters based on their size, charge, polarity, and commonality in nature, in an attempt to account for signal saturation caused by the large number of substitutions in the proteins of rapidly evolving lineages. The drawback to this approach is the inevitable loss of phylogenetic signal from simplifying the characters in the dataset, but it can be useful in investigating potential artifacts in saturated data. There are three standard Dayhoff recoding recipes, Dayhoff 6; which homogenizes the twenty

letter amino acid IUPAC code into six characters, Dayhoff 4; recodes amino acids into four characters, and HP; which sorts the amino acids into two characters based on their polarity (Lartillot *et al.* 2009).

While the three substitution models increase in their simplicity and therefore decrease in signal, there is no objective procedure for testing which of them would fit the dataset of interest better according to the Phylobayes manual (Lartillot *et al.* 2009). Therefore a comprehensive reconstruction of the data under all three recipes is recommended.

### **1.3.2.3 Taxon Pruning**

The most heavy-handed approach of the signal dissection methods, taxon pruning consists of removing rapidly evolving taxa from the dataset outright in order to lower the risk of LBA (Aguinaldo *et al.* 1997). The obvious drawback of this method is the loss of taxon coverage for the clades of interest. One must create a balance between reducing the influence of LBA while not harming the overall taxon coverage of the dataset too much, as poor taxon sampling also reduces phylogenetic signal (Zwickl & Hillis, 2002). Therefore the method by which taxa are selected for pruning varies based on the dataset used. The approach in this thesis is to remove the fastest evolving taxa for the clades of interest while leaving at least one member to represent the clade in question. The pruned phylogeny can then be compared to the original phylogeny from the full dataset to examine if LBA is causing the grouping of rapidly evolving clades.

### 1.3.3 The Basic Local Alignment Search Tool (BLAST)

The Basic Local Alignment Search Tool (BLAST) (Altschul, 1990) is perhaps the most extensively used and significant algorithm in bioinformatics and molecular evolution. Inspired by a series of dynamic programming algorithms that ranked sequence similarity based on mutation steps (Needleman & Wunsch, 1970) and heuristic search models such as FASTP (Lipman & Pearson, 1985) its *modus operandi* is based on a premise known as maximal segment pairing (MSP): an approximate gauge of local similarity between alignments as opposed to extremely precise global similarity. Global similarity, implemented by the dynamic programming algorithms, is an approach where every residue is compared to each other [Figure 1.2 A] (Lobo, 2008). This method's greatest strength, being meticulously robust, is contrastingly its greatest drawback as such exactitude causes the search to be computationally exhaustive and impractical. As a consequence, dynamic programming algorithms, although theoretically useful, quickly became an unrealistic avenue for identifying biological sequence similarity (Altschul, 1990).

Local similarity [Figure 1.2 B] on the other hand observes sequence commonality based on segments of the sequence (words) instead of basing congruence on every residue, the effect of this approach markedly increases the speed of the search (Altschul, 1990; Altschul *et al.* 1997; Tatusova & Madden, 1999). This is the main advantage to the BLAST package (Altschul, 1990), query sequences that are broken into words of a chosen length (usually three characters long) and searched against a database, trawling through the database in orders of magnitude faster than whole query sequences being compared one residue at a time to every database sequence as the local search method can quickly discard database sequences that do not reach a similarity significance threshold match (Altschul, 1990). Another standout advantage

of the BLAST algorithm over its predecessors is its ability to weigh the results of its local similarity searches with statistics known as expectation values or E. values (Altschul *et al.* 1997). The E. value of a potential match or high scoring pair (HSP) can be summarized as the probability of finding a query sequence of length X in a database of size Y by random chance (Tatusova & Madden, 1999). Accordingly, the lower the E. value for a particular HSP the more statistically significant it can be considered. Setting a threshold for the E. value is a useful and an often-employed provision to prevent insignificant matches.

## A Global Similarity

Graph Calculation

0	A <sup>-1</sup>	T <sup>-2</sup>	G <sup>-3</sup>	G <sup>-4</sup>	C <sup>-5</sup>	A <sup>-6</sup>	T <sup>-7</sup>
T <sup>-1</sup>	-1	0	-1	-2	-3	-4	-3
C <sup>-2</sup>	-2	-1	-1	-2	-1	-2	-3
G <sup>-3</sup>	-3	-2	0	+1	0	-1	-2
C <sup>-4</sup>	-4	-3	-1	0	+2	+1	0
A <sup>-5</sup>	-3	-4	-2	-1	+1	+3	+2
A <sup>-6</sup>	-2	-3	-3	-2	0	+4	+3
T <sup>-7</sup>	-3	-1	-2	-3	-1	+3	+5

Traceback

0	A <sup>-1</sup>	T <sup>-2</sup>	G <sup>-3</sup>	G <sup>-4</sup>	C <sup>-5</sup>	A <sup>-6</sup>	T <sup>-7</sup>
T <sup>-1</sup>	-1	0	-1	-2	-3	-4	-3
C <sup>-2</sup>	-2	-1	-1	-2	-1	-2	-3
G <sup>-3</sup>	-3	-2	0	+1	0	-1	-2
C <sup>-4</sup>	-4	-3	-1	0	+2	+1	0
A <sup>-5</sup>	-3	-4	-2	-1	+1	+3	+2
A <sup>-6</sup>	-2	-3	-3	-2	0	+4	+3
T <sup>-7</sup>	-3	-1	-2	-3	-1	+3	+5

## B Local Similarity

Query Sequence

ATGGCAT  
 ATG  
 TGG  
 GGC  
 GCA  
 CAT

3  
Letter  
Word  
List

BLAST Search  
Resulting in HSP

- 1) TTATGGCATAAC  
                  GGC
- 2) TTATGGCATAAC  
           ATGGC
- 3) TTATGGCATAAC  
       ATGGCAT

BLAST Search  
No significant hit

- 4) TTGCCAAGCACAG  
                   GCA
- 5) Match falls below threshold and is discarded

Figure 1.2: Global versus Local Similarity

### Figure 1.2 A: Global Similarity

The residues of the query and database sequences are displayed on the X and Y axes. Global similarity graphs are read from top left corner to bottom right. The values in red are score deductions, gaining in size the further the residue is from the start of the graph. Identical sequences will follow a perfectly diagonal path in the graph, imperfect similarity will cause the path to move horizontal or vertical. These transitions are penalized, representing gaps in the sequence alignment, hence the horizontal and vertical point deductions. For simplicities sake, identical residues are awarded a score of +1 while differing residues are penalized -1. Each box on the graph is denoted a value based on adding the comparison score of the residues on the X and Y axes (+1 or -1) to the values in the boxes of its left corner, extreme left and above, with the most positive score being kept. Traceback: Once the value for each box has been computed, a path in the graph is traced starting from the bottom right hand corner. The path follows the most positive direction denoted by the individual boxes. Horizontal or vertical movements represent gaps in the sequence alignment.

### Figure 1.2 B: Local Similarity

The query sequence is segmented into a list of words of predefined length. 1) The database sequence is seeded with a word from the query. 2) The alignment is extended as the residues on either side match words from the list. 3) The full query sequence is found in the database, significant homology is defined as a HSP. 4) A different database sequence which shares similarity to one word from the list. 5) Since its neighboring residues do not match any other words from the list this match falls below the significance threshold. The BLAST algorithm does not proceed from the seeding step and moves to the next database sequence. A global similarity search could not identify the suitability of the second database sequence without first computing the graph and then finding the most suitable path for the alignment whereas a local search can quickly identify a poor match and move on thus saving time. In this case a global search would need to compute a 7 x 13 graph (query vs database sequences) as opposed to using a segment of the query sequence.

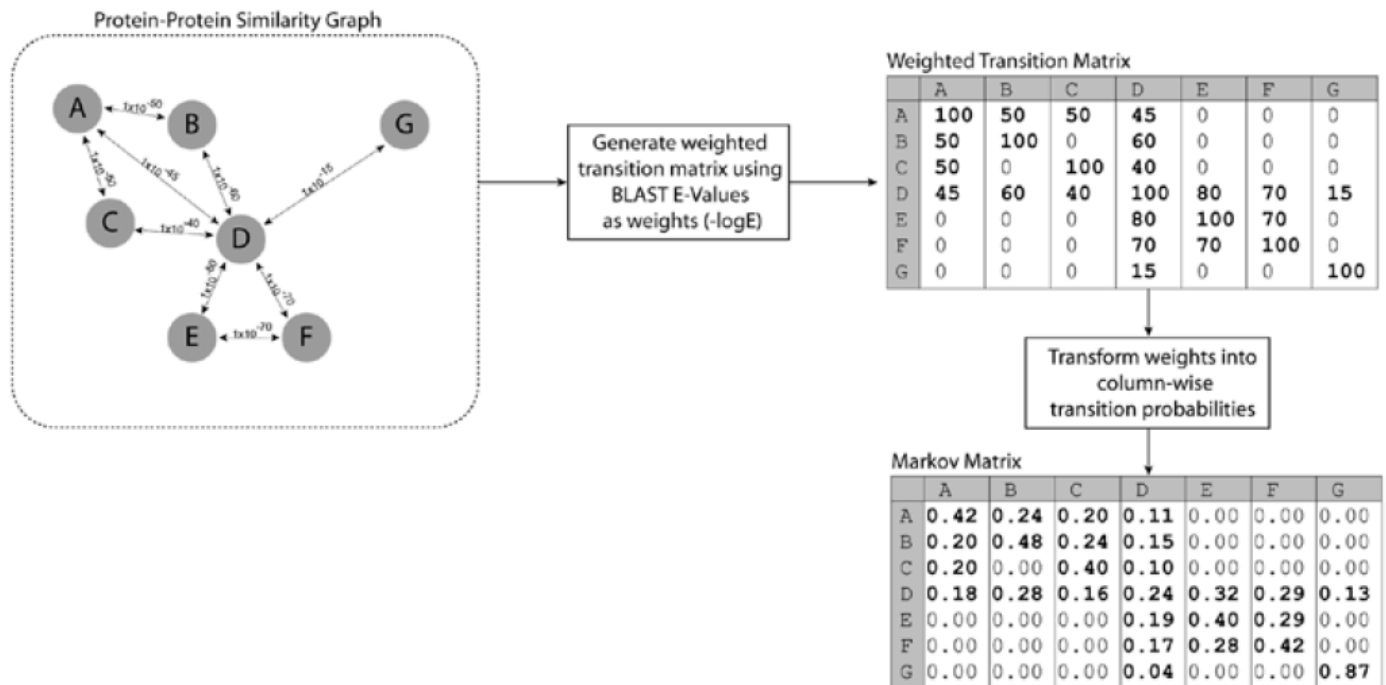
### 1.3.4 The Markov Clustering Algorithm (MCL)

The Markov Clustering Algorithm (Enright *et al.* 2002) was developed in order to cluster sequences based on their similarity, or in the case of this body of work to group proteins into families based on similar functionality and sequence structure. Theoretically this seems like a straightforward task, however practically speaking the multi-domain nature of most eukaryotic proteins (Liu & Rost, 2003) makes protein family grouping difficult. The issue lies with what are known as “promiscuous domains” found in many complex (eukaryotic) proteins (Basu *et al.* 2008). These are domains that are commonly shared amongst proteins that do not share similar function or sequence structure (Haggerty *et al.* 2014).

These promiscuous domains confuse the most commonly used and effective method of identifying similarity in molecular biology: BLAST (Altschul, 1990). BLAST cannot differentiate between promiscuous non-functionally similar protein domains, and relevant functionally similar protein domains, this often results in the grouping of dissimilar proteins into families based on their promiscuous protein domains (Marcotte *et al.* 1999). A solution to this problem was identified by Enright *et al.* (2002) using hidden Markov models and graph theory, they named it MCL. MCL (the Markov Clustering Algorithm) is one of the more robust and reliable engines for protein family detection (Brohée & van Helden, 2006) and has the capabilities of avoiding the main pitfalls of protein clustering explained above.

MCL revolves around a graph containing nodes and edges. Each node is a protein sequence and the edges that connect them are representative of the sequence similarity between one or more nodes (Enright *et al.* 2002) In MCL, edges are weighted by the E. value of a prior BLASTp step, the lower the E. value the shorter

the edge between protein nodes. These weighted edges are then translated into a Markov matrix of transition probabilities [Figure 1.3].



**Figure 1.3: The MCL Process**

Every protein sequence in a fasta file is designated to a node (circles labelled A-F). These nodes are compared to one another via BLAST. The respective generated E. values correspond to the length of the edges in the graph between the nodes (lines connecting circles). These weighted values are recorded in a table comparing similarities between proteins and translated to a Markov matrix. From Enright *et al* (2002).

Once the weighted edges have been converted to Markov probabilities (a stochastic matrix) MCL simulates “random walks” through the graph in the form of expansions and inflations. Expansions are essentially matrix squaring using the usual matrix product. Inflation is a method whereby each entry in the matrix (transition probability edge value) is raised to a power and the matrix itself is rescaled until it is stochastic (Markov) again (Enright *et al.* 2002). These operators make it possible to measure flow in the graph, areas of high flow correspond to high traffic of random walks (Enright *et al.* 2002). MCL considers areas of high traffic flow to be indicative of true

similarity between nodes (based on transition probabilities) and thus concentrates on these edges and diminishes edges in areas of weak flow.

Expansions and inflations cause random walk operators to run through the graph until there is little or no net change and a state of flow equilibrium is reached, this final form of the graph is considered the protein clusters (or families) (Enright *et al.* 2002).

### **1.3.5 Gene Ontology**

Gene ontology (GO) is a method of gene and gene product annotation based on a very specific vocabulary. GO was initially developed by a number of model organism sequencing consortiums (The FlyBase Consortium, 1999; Ball *et al.* 2000; Blake *et al.* 2000; Ringwald *et al.* 2000) that wanted to create a standard vocabulary for gene annotation that could be translated between projects, deciding on three major criteria; molecular function, biological process and cellular component (Gene Ontology Consortium, 2004). Molecular function specifies the function of the gene product but not where this product acts within the cell nor when it acts. Biological process describes a biological purpose that the gene product directly influences. According to the GO developers (Ashburner *et al.* 2001), a spider web of molecular functions can influence a single or even multiple biological processes but this is too complex and overarching for GO to directly link these two separate ontologies. Instead GO aims to identify the molecular function and biological process of genes separately instead of inferring one directly from the other.

Cellular component ontology relates to the specific whereabouts in the cell the gene product acts. Differing molecular functions can share the same cellular component and similarly separate biological processes can occur in the same location of the cell,



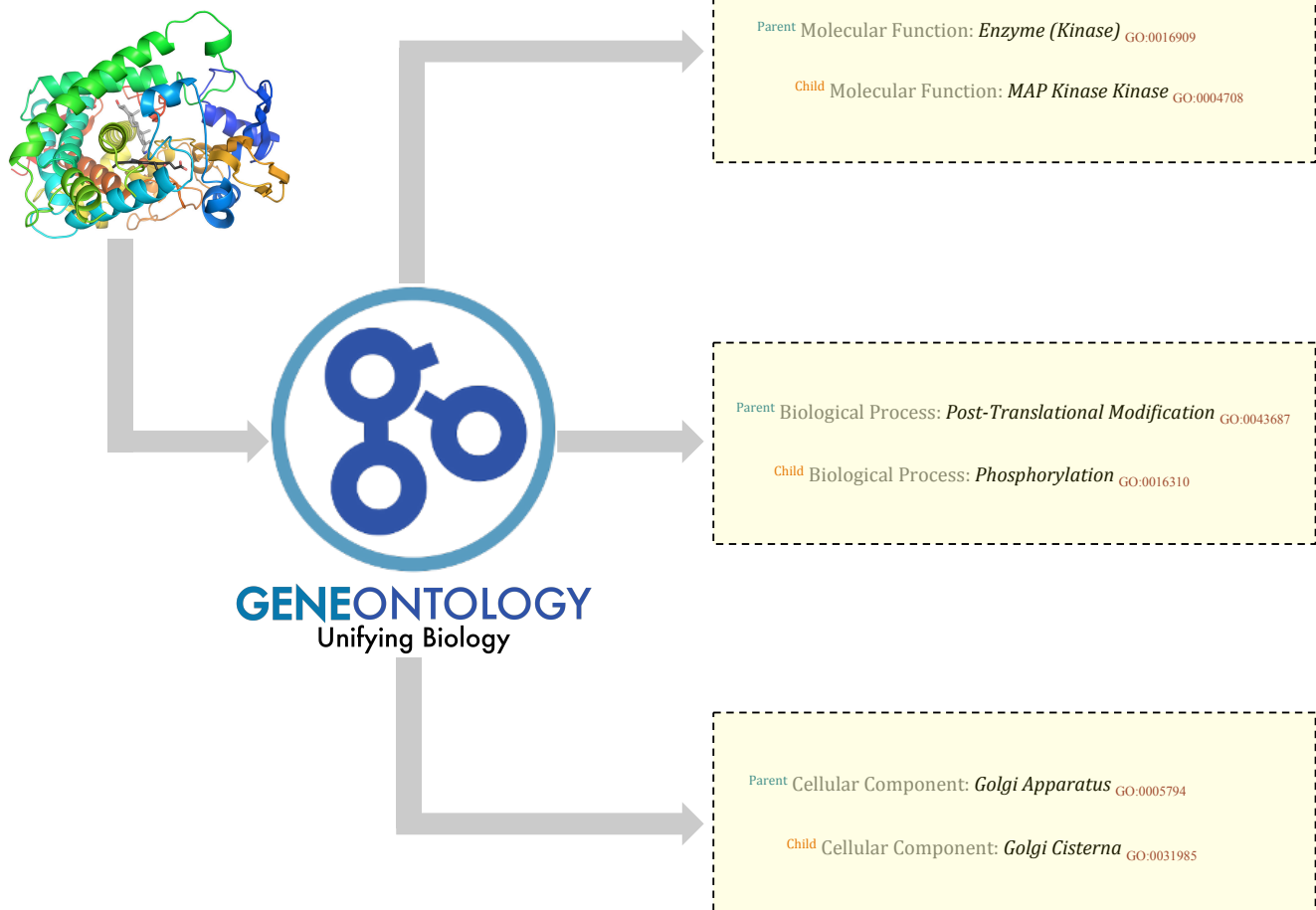
making it important to define cellular component as its own ontology. An example of how a single gene products function can be broken down into the three separate ontologies can be found in [Figure 1.4](#).

Ashburner *et al.* (2001) use a system of directed acyclic graphs (DAGs) to structure their gene ontology annotations. These consist of “parent” and related “child” categories. The parent is often a general or multifaceted description of one of the GO categories. The child categories identified within are more detailed and specific versions of these parent annotations. Child categories from the same parent will differ from one another but will always have a commonality with the shared parent annotation. Child annotations can also have multiple parents. These parent - child categories are also represented as different levels of GO, low levels (parents) describe broad terms while high levels (child) get more and more specific.

The benefits of the parent - child system of annotation is that it allows for multiple levels of specificity, while still being somewhat informative for gene products that are relatively unknown and as such may not have much information in the GO database. These genes can often still be annotated to a basic functional degree.

The GO database is available at <http://geneontology.org/page/go-database>.

## Gene Product - Protein



### Figure 1.4: Gene Ontology Example

An example of the three categories of gene ontology applied to the annotation of a single gene product. Each of the three ontology categories are broken down into parent (general annotations) and child (specific annotations). The nested child ontologies can extend to multiple specificities, only a single order of specificity is illustrated for simplicity. In this example the *molecular function* of the gene product is a MAP Kinase Kinase that influences the *biological process* of phosphorylation in the *cellular component* Golgi cisterna. The GO terms in this example are genuine, searched for and chosen by me to use as the example, and they along with the GO description provided can be found at <http://amigo.geneontology.org/amigo>

The GO annotation method is a useful tool for studying the function of protein families generated through MCL, outlined in the previous section. In chapter 5 the Markov Clustering Algorithm and gene ontology were applied to study macroevolutionary trends of protein family acquisition across the Metazoa with an emphasis on the Protostomia. MCL was used to cluster protein families which were then distributed amongst a supertree representing forty-nine taxa from the Animal Kingdom and GO was used to annotate nodes on the supertree which displayed an above average rate of protein family acquisition.

### **1.3.6 The Molecular Clock**

The molecular clock is a means of divergence time estimation using molecular data. It works on the premise that the genetic differences between lineages increase over time since the point of divergence from a last common ancestor (LCA). However, the description and preconceptions of the initial molecular clock are no longer accurate for contemporary divergence time estimations studies.

#### **1.3.6.1 Progressing from Strict to Relaxed Clocks**

The genetic changes between lineages occurring over time were initially considered to be undergoing at a constant rate, allowing the differences between lineages to be translated into a timescale. The first clocks took a biologically unrealistic homogenous approach to species dating studies where all lineages were assumed to evolve at an equal rate (Zuckerkandl & Pauling, 1962 & 1965). Using molecules such as DNA and proteins to study species origins was innovative, but this “globally constant” method created a litany of inaccuracies. The complex nature of evolution manifests itself in a multitude of forms that vary from species to species such as mutation rate, generation rate, and population size. These are now summarized as the lineage effects of molecular clocks, artifacts of strict clock methods (Ho, 2014). Imposing a singular rate of change on a dataset is an oversimplification of such effects and so resulting divergence time estimations from this method could not be trusted.

Contemporary clocks estimate divergence dates by use of rate heterogeneity models, facilitated by a Bayesian framework of probability, in conjunction with time calibration points from the fossil record (dos Reis *et al.* 2016 and Yang & Donoghue,

2016). This concept of taking account for rate variation amongst separate lineages, while crucially grounding estimations by constraining the divergence dates to fossil-based calibrations and accounting for the uncertainty of the fossil record, was first implemented by Sanderson (1996) and shortly afterwards by work from Thorne *et al.* (1998). The process is an approximation of evolutionary rates over time whereby rates among nodes are cross-correlated (compared against themselves) in intervals (auto-correlated clock models). The rate trajectory of every branch in a tree differs based on the data, but the rate of each branch is defined by a singular rate: the average rate between the beginning node and terminal node of the particular branch. Taking into account for rate heterogeneity in dating studies is considered a relaxed clock approach, in contrast to the strict clock method: assignment of a single rate of evolution across all lineages in the dataset.

Although a vast improvement over the previous homogenous rate models, the proceeding method is not without its drawbacks. Assuming a singular rate per branch is also not biologically realistic. Genes change over time and assuming an average rate of change along a branch instead of accounting for the actual rate distribution across the branch can skew the data, particularly if the genes in the dataset incur many changes (rapidly evolving genes) or if one studies deep node divergence dates (the older the genes the more potential changes that can occur). Additionally, one must consider that for contemporary phylogenomic datasets, the branch for each lineage can compose of hundreds of concatenated genes. This problem has been coined as gene effects (Ho, 2014) and have been somewhat nullified by other methods of divergence time estimation (Drummond *et al.* 2012 and Ronquist *et al.* 2012) which do not restrict branch lengths to a single rate of evolution but at the price of heavy

computational cost and a more restricted catalog of evolutionary models compared to more expansive methods (Lartillot *et al.* 2009).

### 1.3.6.2 Bayesian Clock Models

An in depth review of Bayesian clock methods (dos Reis *et al.* 2016) considers the Bayesian framework for divergence time estimates to be the superior method of species dating but it is not without its problems, mainly concerning its inability to separate the time and rate parameters which comprise the molecular sequence data analysed. Our divergence time estimations avail of such Bayesian clock models, specifically the un-correlated gamma distribution model (Drummond *et al.* 2006) often referred to as U-GAMMA, the auto-correlated CIR model (Lepage *et al.* 2007), and the lesser-used un-correlated White Noise model (WN) (Lepage *et al.* 2007), which was run in the chaetognath study of chapter 3.

The rate distribution models described herein were developed for phylogenetic and molecular clock applications in response to the unpredictability of evolutionary rates and calibration bounds. Uncertainty in rates is mainly down to expansive variability of molecular datasets, dependent on the taxa included and genes used. These parameters virtually ensure a diverse rate of evolution across the tree of any large-scale dataset which, as discussed previously, nullifies the usefulness of models that assume a constant rate of evolution across all lineages, i.e. strict clocks (Zuckerkandl & Pauling, 1962 & 1965). As for calibrations, ambivalence derives from disagreement in the fossil record, usually pertaining to dispute in ascribing morphological aspects of fossils, causing uncertainty in identification and classification, which in turn results in conflicting assignment of them as minimum bounds to certain internal nodes of the

tree. To address these issues, which have been inflated by the complexity of large-scale phylogenomic datasets, relaxed clock models were developed. The most commonly used of these models (Drummond *et al.* 2006 and Lepage *et al.* 2007) take a contrasting un-correlated versus auto-correlated approach.

The un-correlated gamma model, Drummond *et al.* (2006), assumes no correlation of rates for adjacent branches on the tree, instead the rate variance for every branch is determined independently; they are not influenced by their predecessor nodes. Gamma distributions are modeled to account for rate variation composing of an exponential and chi-squared distribution (Drummond *et al.* 2006). As such the un-correlated gamma model is less susceptible to topological artifacts because the rate of each branch is not dictated by previous nodes or previous branches in the tree, unlike auto-correlated models.

The WN model is also un-correlated, bases rate variation on a gamma distribution, but differs to the U-GAMMA model in how it applies variance, particularly to long branches in the analysis. The variance of the U-GAMMA model is squared with time whereas the variance of WN is linear which, according to Lepage *et al.* (2007), accounts for the propensity of un-correlated models to estimate smaller variance over long branches.

Auto-correlated models were briefly mentioned earlier, their process is an approximation of evolutionary rates over time whereby rates among nodes are cross-correlated (compared against themselves) in intervals (Drummond *et al.* 2006 and Lepage *et al.* 2007). The rate of each branch is determined by an assumption, a parametric distribution - the mean of which is a function of the parent branches rate of evolution. These parametric distributions can take the form of lognormal distribution where the rate variance is dependent on branch length or exponential distribution

where the rate variance is dependent on the ancestral node (Drummond *et al.* 2006). Because the rate of the considered branch is dependent on its predecessor, the topology of the tree should be unequivocal and the root of the tree must be designated an assigned rate known as the root prior.

The auto-correlated model used in divergence time estimation experiments for this thesis was the CIR model (Lepage *et al.* 2007) which itself is similar to the lognormal models of rate variation but with a stationary distribution. The CIR model for rate variation is based on its namesake (Cox *et al.* 1985) and applies the square of the Ornstein-Uhlenbeck model - a stationary Gaussian process (Aris-Brosou & Yang, 2003), in order to prevent augmented rates near the root of the tree. This is an important precaution when estimating divergence times of nodes close to the roots of the tree, in the case of this thesis the origins of major animal clades.

### **1.3.6.3 The Fossil Record**

The key to accurate divergence time estimations using molecular clocks is to constrain their calculations with calibrations from the fossil record (Donoghue & Benton, 2007). The central dogma of the fossil record is the law of strata superposition that applies to both fossils and geology. Devised by Nicholas Steno in 1669, it states that the oldest layers of rock are found deepest in the earth, with newer layers towards the top. Consequently fossils found in the deepest layers can be considered older than fossils found in shallower strata. Therefore based on the law of strata superposition, the age of a fossil is defined by the specific rock formation whence it was discovered. However this law alone is not enough to define the relative age of fossils by a quantifiable metric. The definitive dating procedure of fossils is

based on a universal geological timescale, a chronostratigraphic chart representing the relationship between time (chrono) and the layering arrangement of rocks in the earth (geological stratigraphy). The chronostratigraphic chart is divided into tiered stratigraphical units of time consisting of Eons, which are made up of Eras, which consist of Periods, which are in turn divided up into Epochs, which are summarized by the smallest units: Ages (Cohen *et al.* 2013). Each unit is designated a specific time period that correlates to a particular geological layer summarized in [Figure 1.5](#). These designations are the culmination of years of careful geological research that have resulted in a universal metric of temporal geology. As with any field of active research, the exact designations for these stratigraphical units of time evolve as new information comes to light, with the chronostratigraphic chart being continually updated. Therefore we are left with a highly accurate yet not strictly precise resource for understanding the age of rock formations. Without this standard, fossil dating would be an equivocal affair and divergence time estimation may not be possible due to the restricted capabilities of molecular data previously discussed.





# INTERNATIONAL CHRONOSTRATIGRAPHIC CHART

v 2016/04

www.stratigraphy.org

International Commission on Stratigraphy

System / Era

Eonhem / Era

System / Epoch

numerical age (Ma)

numerical age (Ma)

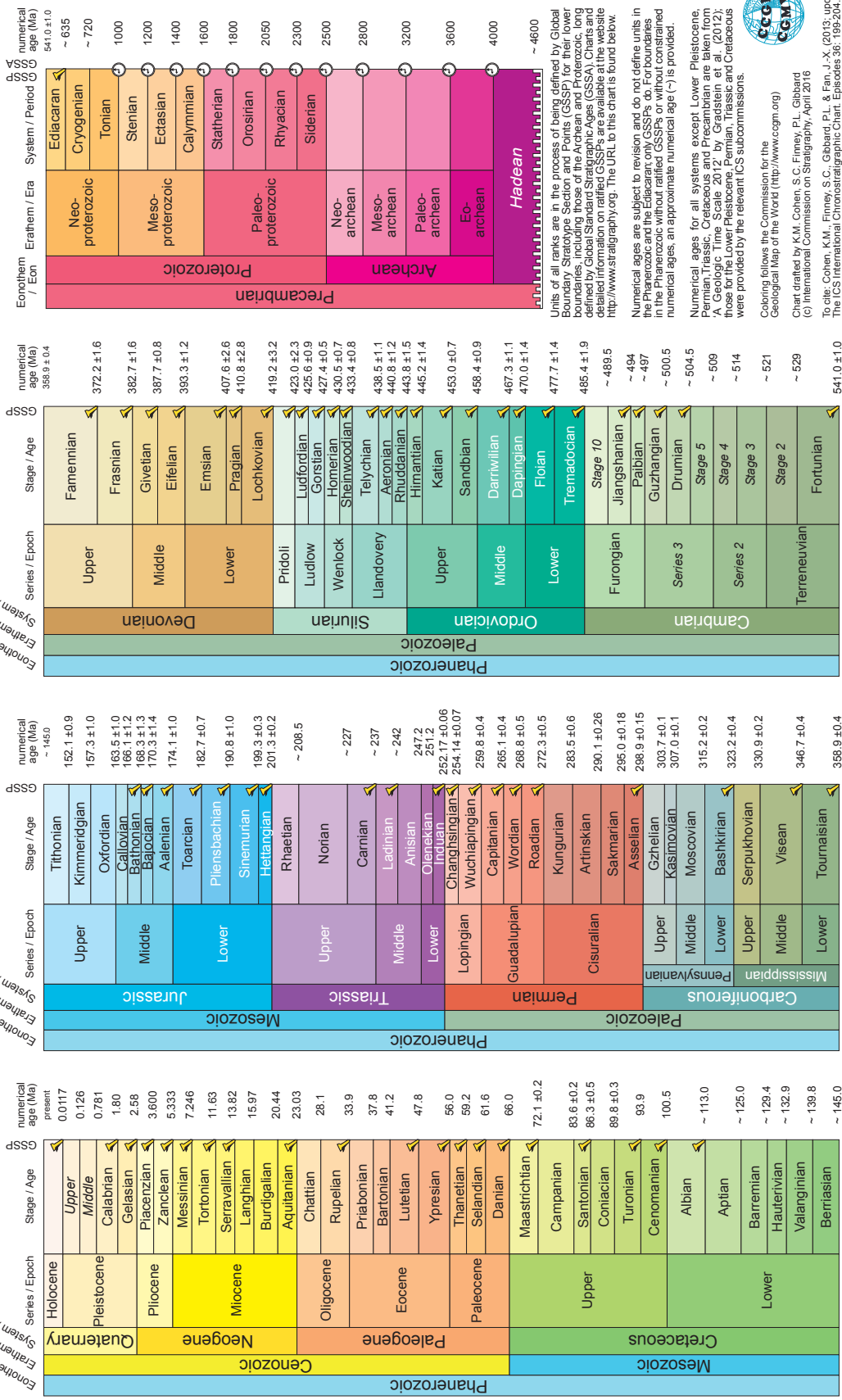


Figure 1.5: International Chronostratigraphic Chart

The Chronostratigraphic Chart 2014/04: The chart divides geological stratigraphies into units of time from 0 to 4,600 million years ago. The stratigraphies are assigned to tiered classifications of Eons, Eras, Periods, Epochs and Ages. The chronostratigraphic chart is considered the universal standard for the precise classification of temporal geological strata.

The chart is the product of the combined knowledge of the International Union of Geological Sciences (IUGS). The chart is provided by [www.stratigraphy.org](http://www.stratigraphy.org)

Units of all ranks are in the process of being defined by Global Stratigraphic Tables (GSSs) for the Phanerozoic and Proterozoic, including those of the Archean and Proterozoic, long defined by Global Standard Stratigraphic Ages (GSSA). Charts and detailed information on ratified GSSs are available at the website <http://www.stratigraphy.org>. The URL to this chart is found below.

Numerical ages are subject to revision and do not define units in the Phanerozoic and the Eoarchean, only GSSs do. For boundaries in the Phanerozoic without ratified GSSs or without constrained numerical ages, an approximate numerical age (\*) is provided.

Numerical ages for all systems except Lower Pleistocene, Permian, Triassic, Cretaceous and Precambrian are taken from 'A Geological Time Scale 2012' by Gradstein et al. (2012); those for the Lower Pleistocene, Permian, Triassic and Cretaceous were provided by the relevant ICS subcommissions.

Coloring follows the Commission for the Geological Map of the World (<http://www.cgmw.org>)

Chart drafted by K.M. Cohen, S.C. Finney, P.L. Gibbard

(c) International Commission on Stratigraphy, April 2016

To cite: Cohen, K.M., Finney, S.C., Gibbard, P.L. & Fan, J.-X. (2013, updated) The International Chronostratigraphic Chart. Episodes 36, 199-204.

URL: <http://www.stratigraphy.org/ICSchart/ChronostratChart2016-04.pdf>



### **1.3.7 Total Evidence Dating**

Total evidence dating (TED) is a method of reconstructing the evolutionary history of a dataset using morphological, molecular, and fossil data (Ronquist *et al.* 2012a). In the past, morphological / fossil and molecular experiments were stand-alone studies, often conflicting and rarely collaborating (Yang & Donoghue, 2016). In molecular dating studies the fossil record has been used almost singularly as a minimum bound for origin dating (Benton *et al.* 2015) despite the evolutionary information their ascribed characteristics contain.

TED takes advantage of these differing sources of evidence in order to maximize the evolutionary signal of a dataset and gauge a convergent estimate on an answer. Therefore TED is a robust method of dating a clade by applying multiple sources of evidence to complicated evolutionary timelines and is used to further examine the origins of the Chaetognatha in chapter 3.

## **1.4 The Emergence of Phylogenomics**

### **1.4.1 Sequencing Technologies**

The birth of modern day sequencing technologies started with Frederick Sanger and his two-dimensional homochromatography method of detecting base residues (Brownlee & Sanger, 1969). This in conjunction with the use of DNA polymerase to sequence small chains of nucleic acids or oligonucleotides, (Sanger *et al.* 1973) led to the famous Sanger method (Sanger *et al.* 1977).

The Sanger method sequenced DNA by manipulating DNA polymerase I in to adding nucleotides to a growing oligonucleotide chain using thymidylic acid (dT).

The main issues with the Sanger method of sequencing were the amount of time taken, scalability of time versus genome size, resolution accuracy, and expense of acquiring ddTPs making it a non-viable option for most organisms (Schuster, 2008). As a direct consequence of these drawbacks, only a narrow group of organisms were sequenced during the Sanger era, near exclusively model organisms. Model organisms, those that have been studied extensively under laboratory conditions, were considered suitable for sequencing studies in addition to their popularity because they met certain criteria such as small genome size (The *C. elegans* Sequencing Consortium, 1998 & Aparicio *et al.* 2002) rapid generation rate (Gibbs *et al.* 2004), suitability under laboratory conditions (Adams *et al.* 2000), usefulness in disease research (Blattner *et al.* 1997 & Holt *et al.* 2002) and some because of their genetic similarity to humans (Chinwalla *et al.* 2002) making them ideal candidates for medical studies.

This created an imbalanced resource of molecular information, where model organisms had been extensively covered through dedicated genome sequencing projects and most others had sparse molecular information from specific EST studies, if any at all. The knock on effects of these limitations for molecular evolution studies has been long lasting with most having to design experimentation around a “take what one can get” approach to data acquisition. This led to numerous phylogenetic studies based on very small datasets that often weren’t indicative of the true relationships between the organisms studied due to lack of data and sparse taxon sampling (Boore *et al.* 1995; Giribet *et al.* 2001; Mallaat *et al.* 2004; Pisani *et al.* 2004).

### 1.4.2 The Polymerase Chain Reaction

The aforementioned sequencing studies were not only made possible by the Sanger method of sequencing, but with the assistance of another major technological invention which has defined molecular biology: the polymerase chain reaction (PCR) (Saiki *et al.* 1985 and Mullis *et al.* 1986).

PCR is a highly sensitive method by which targeted oligonucleotides can be replicated and amplified to very large numbers.

The obvious advantage of PCR is that small concentrations of DNA can be amplified to much larger concentrations, however it can also accurately generate double stranded DNA from single stranded DNA or even RNA molecules (Saiki *et al.* 1985 and Mullis *et al.* 1986). This made sequencing studies not only easier in terms of collection of raw data but presented the opportunity to specifically target parts of a genome that had previously suffered from low coverage. The specificity of PCR relies on primers, small chemically synthesized DNA oligonucleotides, 10-20 bp long, that hybridize to the target DNA at the 5 prime ends of the forward and reverse strands in an ultra-specific manner (Wallace *et al.* 1979). This specificity is based on the exact compilation of the nucleic acids of the primer reflecting the reverse compliment of the target oligonucleotide. PCR operates in three main steps: Denaturing, in which DNA is de-hybridized into single strands using high temperature for a short period of time. Annealing in which primers are hybridized to the target DNA using DNA polymerase, the temperature at which this occurs depends on the primer composition. Finally extension, where free nucleotides are added to the reaction and DNA polymerase extends the opposing strand for forty PCR cycles that usually takes about ten minutes (Wallace *et al.* 1979).

### 1.4.3 Next Generation Sequencing

Next Generation Sequencing (NGS) techniques were invented in 2005 (Margulies *et al.* 2005). The key to NGS is its massively parallel method of sequencing millions of DNA fragments simultaneously. This sped up sequencing by orders of magnitude compared to the Sanger method that had dominated the previous thirty years. With the initial NGS systems genome sequencing took weeks instead of years, with modern day advancements a genome can be sequenced within twenty-four hours (Goodwin *et al.* 2016).

One of the more widely used methods of NGS is sequencing by Illumina Solexa (Bennett, 2004), a method made popular due to its high level of accuracy, massively-parallel capabilities that greatly accelerate the process, its novel method of detection by pyrosequencing, and crucially its cost effectiveness (Metzker, 2010). The molecular libraries generated for this thesis are products of Illumina sequencing.

Sequencing by Illumina involves shifting focus from large read sequencing to short read sequencing (Li *et al.* 2010). The key to Illumina sequencing is the unique type of slide used called the flowcell (Holt & Jones, 2008) that can facilitate millions of different chemical reactions simultaneously and the use of modified nucleotides to include fluorescently labeled reversible terminators (Canard & Sarfati, 1994 and Shendure *et al.* 2005).

The purified DNA is broken up into millions of fragments, usually up to 200bp long, the ends of which are attached to adapters. These adapters are attached to unique spots on the flowcell. Each fragment is then amplified into a cluster. The original fragments are then washed away leaving only the amplified clones. The sequencing by synthesis process begins and individual fluorescently labeled nucleotides compete against each other to hybridize to the clusters one step at a time in cycles. This process works on

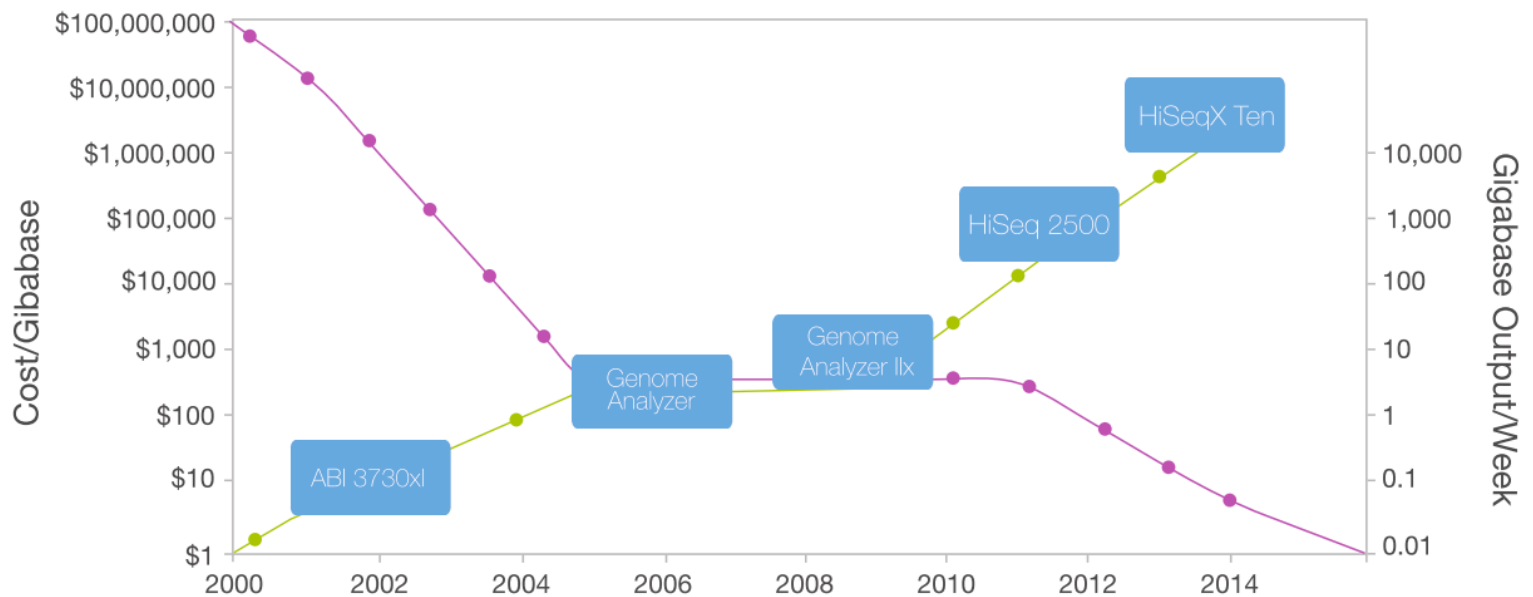
the chemical properties of nucleotides: adenines will only hybridize to thymines, cytosines to guanines and vice versa (Felsenfeld & Miles, 1967). The four types of bases possess a different type of fluorescent label that emit a frequency of light (colour) unique to the type of base they are. As mentioned, the altered free nucleotides have a secondary function: a reversible terminator. This terminator prevents more than one free nucleotide hybridizing to the bound DNA per cycle (Ju *et al.* 2006 and Turcatti *et al.* 2008). When the cycle is complete a picture of the slide is taken and the type of base added for each read cluster is identified by the particular colour emitted from the fluorescent label of the newly hybridized nucleotide. Between each cycle the terminators are removed from the hybridized modified nucleotides and the process repeats until the number of cycles covers the length of the reads. The sequencer analyses each photograph of every cycle and calls the bases of each read based on the frequency emitted and encodes a confidence score of that call depending on the intensity of the signal. A more detailed step by step guide to sequencing by synthesis and an accompanying diagram can be found in [Supplementary Material 1.1](#). NGS technology has made it possible for rapid de-novo genome and transcriptome sequencing, projects without the need of a reference genome (Zerbino *et al.* 2008; Simpson *et al.* 2009; Grabherr *et al.* 2001). This has greatly increased the production of molecular data.

#### **1.4.4 The Decline of Sequencing Costs**

NGS not only greatly increased the speed and accuracy of sequencing but also brought a dramatic cost reduction. For over thirty years sequencing projects consumed time and resources in the making and were very costly. The first published

human genome, using the Sanger method, was spread across an unprecedentedly large collaboration (Lander *et al.* 2001) taking thirteen years, three billion dollars and published on the same day as a private consortium (Venter *et al.* 2001).

Present day NGS projects cost as little as \$1,000 (<https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet-hiseq-x-ten.pdf>). This has drastically changed the scope of modern day genomics as sequencing projects are now within reach of most modestly funded research groups [Figure 1.6].



**Figure 1.6: The Decline of Sequencing Costs**

The dramatic rise of data output and concurrent falling cost of sequencing from 2000 - 2015. The Y-axes on both sides of the graph are logarithmic. (Source: [www.illumina.com/technology/next-generation-sequencing.html](http://www.illumina.com/technology/next-generation-sequencing.html)).

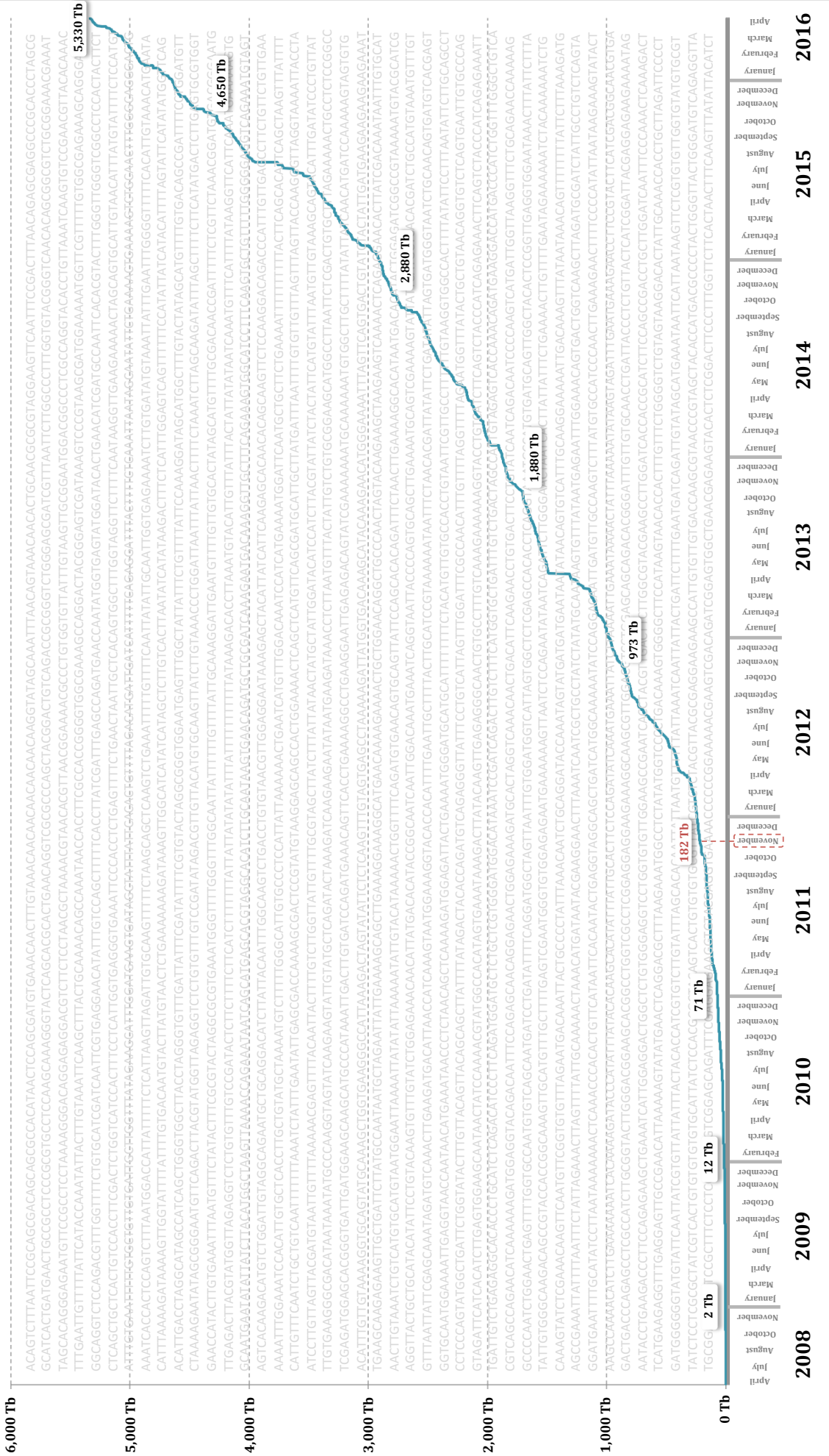
### 1.4.5 The Sequence Read Archive

The Sequence Read Archive (SRA) (Leinonen *et al.* 2011) is the online designated database for raw sequence data from published and unpublished NGS projects (<http://www.ncbi.nlm.nih.gov/sra>). The SRA is a branch of the International Nucleotide Sequence Data Collaboration (INSD) group that also includes the European Bioinformatics Institute (EBI) and the DNA Database of Japan (DDBJ). In addition to a raw sequence data deposit, the SRA contains information relating to the biological samples and experimentation of the raw data in question. Thus the SRA has been important in ensuring the availability of newly published NGS data for researchers around the world that wish to design new experiments with more data and reproduce previous ones.

The drastic growth of the SRA since its inception in 2007 is irrefutable evidence that we are now in the phylogenomic era. After nine years the SRA now boasts a data depository eclipsing 5,000 Tb and since the commencement of this research project in November 2011 it has seen a thirty fold increase in NGS data deposits [Figure 1.7]. With the ever-increasing amount of raw sequence data being published from NGS projects, the SRA has become an essential resource for collecting these data in an organized and easily accessible manner.



# The Growth of the SRA



**Figure 1.7: The Growth of the Sequence Read Archive (SRA)**  
 Monthly statistics spanning an eight year period representing the number of nucleotides deposited in the archive from sequencing projects. The X-axis denotes the timeline while the Y-axis denotes the number of nucleotides deposited in the SRA in terabases (Tb). 1Tb = 1.0E15 nucleotides. Data source: [http://www.ncbi.nlm.nih.gov/Traces/sra/sra\\_stat.cgi](http://www.ncbi.nlm.nih.gov/Traces/sra/sra_stat.cgi)  
 The nucleotide numbers at the end of each year are in bold on the graph while the figure during the commencement of this project is in red.

### 1.4.6 Applying de Bruijn Graphs to De-Novo Short Read Assemblies

NGS projects produce much shorter and far more numerous reads than the previous method (Sanger *et al.* 1977). Although these types of reads greatly increase the coverage of the sequenced molecular library, their size and volume makes them difficult to assemble using traditional overlap consensus methods (Bonfield *et al.* 1995). In order to organize this influx of sequence read data a new approach was needed. Instead of trying to match pairs of reads by overlapping them, reads are broken into even smaller pieces of constant length known as k-mers as part of the de Bruijn graph assembly method (Pevzner *et al.* 2001). Each of these k-mers are represented as a single node in the de Bruijn graph. If the k-mers represented by the two nodes overlap by  $k - 1$  nucleotides they are connected by an edge (Zerbino & Birney, 2008 and Compeau *et al.* 2011). This algorithm is applied to all nodes in the graph until edges are formed between many nodes forming paths, dictated by overlapping k-mers. Once the algorithm is finished these paths are considered to be individually assembled sequences. The largest variable that has to be considered under the de Bruijn graph approach is the choice of k-mer: the size of the sequence that all the reads will be broken into. The one rule that has to be followed is that the value for k must be at least 1 - the read length. After that it depends what software one uses. For example ABySS (Simpson *et al.* 2009) recommends the value of k be the interval between half the read length and the read length minus ten. It can then amalgamate multiple k-mer assemblies from this range into a chimeric assembly.

Arguably the biggest weakness of the de Bruijn graph algorithm is the construction of false paths in the graph from k-mers not representative of the true DNA sequence, caused by sequencing errors. The commonly used assembly software that employ de Bruijn graphs (Zerbion & Birney 2008; Simpson *et al.* 2009; Grabherr *et al.* 2011)

alleviate this issue by removing singleton nodes (k-mers that only appear once, with no overlap), trimming spurious edges that link only a few nodes and are poorly supported, discarding anomalous edges in the graph referred to as “bubbles”, and by discarding sequences that don’t match a certain length (Simpson *et al.* 2009 and Compeau *et al.* 2011).

#### **1.4.7 From Phylogenetics to Phylogenomics**

The development of NGS technology, its diminishing cost, adaptive short-read assembly methods, and the SRA has paved the way for large-scale, taxon rich studies (Borner *et al.* 2014; Fernandez *et al.* 2014; Zapata *et al.* 2015; Kocot *et al.* 2017; Simon *et al.* 2017). This has led to a natural movement of molecular evolution studies from small restrictive phylogenetic datasets to large, although often cumbersome (Jeffroy *et al.* 2006), phylogenomic datasets. Initially this was a slow movement as experimental design and data assembly methods caught up with the new short read technology but since 2012 the number of NGS experiments has increased at a near exponential rate [Figure 1.6 & Figure 1.7].

Small datasets have historically been problematic for most molecular evolution studies, particularly when studying deep node phylogenetics as they routinely suffered from stochastic, or sampling, errors (Nei, 1986) that obfuscate the small amount of underlying phylogenetic signal with random noise leading to faux and or poorly supported phylogenetic trees (Telford & Holland, 1993; Papillon *et al.* 2003; Pisani *et al.* 2004; Matus *et al.* 2006; Paps *et al.* 2009b). Such incorrect phylogenetic assumptions are often referred to as artifacts (Simmons & Freudenstein, 2002 and Altaba, 2009). These deleterious stochastic errors thrive on datasets comprised of few

or single gene studies with sparse taxon sampling, which made up the bulk of phylogenetic studies before the NGS era as molecular data was neither affordable to generate nor easy to come by via published research concerning non-model organisms (Ekblom & Galindo, 2010). Furthermore, data restrictions of the past had encouraged an overreliance on singular and limiting data sources, most notably mtDNA, that are prone to a specific type of systematic error known as compositional heterogeneity (Rota-Stabelli *et al.* 2013). Other systematic biases such as long branch attraction (LBA) thrive in under sampled datasets as breaking up the branches with the addition of more taxa can often repel it (Bergsten, 2005).

With taxon and data-rich studies coming to fore, the threat of stochastic errors from a lack of data sampling has been virtually eradicated. It would be naïve however to assume that phylogenomics solves the variety of hurdles encountered when working with molecular datasets (Jeffroy *et al.* 2006). Somewhat ironically, missing data is becoming a problem in phylogenomic super alignments (very large MSAs consisting of hundreds of concatenated genes for any number of taxa) as often there is a lack of consistency of ortholog coverage and perhaps also due to the stretching of sequence aligner capabilities such as MUSCLE (Edgar, 2004). The threat of systematic biases remain, particularly the influence of LBA on saturated sites, a very unwelcome side effect of rapidly evolving taxa at the molecular level that is augmented with data addition. In chapter 2 methods designed to ease their influence are discussed.

### 1.4.8 Genomics versus Transcriptomics

Two forms of NGS data were used in this thesis: genomes and transcriptomes. These data were generated de novo “of new”, i.e. sequenced from scratch without the framework of a reference genome.

Genomic data is used in the case of *H. duajrdini*, and *Parasagitta sp.* in chapters 2 and 3, while transcriptomes make up the bulk of the newly sequenced data throughout the rest of this thesis. Both data sources are discussed herein.

A genome consists of the full lexicon of an organisms genes regardless of whether they are expressed or not (Koboldt *et al.* 2013), an organized genomic library will contain a single representative of every gene. Transcriptomes consist of the transcript messenger RNAs (mRNAs) that only represent the genes being expressed at detectable levels at the time of RNA extraction (Burgess, 2016). Transcriptomes are an incomplete exhibit of a genome and inconsistently represent the genes based on expression levels. The higher the level of gene expression the more transcripts produced for that gene and vice-versa (Ma, 2006). Therefore transcriptomes will never represent the full catalog of an organism’s genes and will always contain an element of redundancy (in the context of datasets used for molecular evolution studies) as they will have multiple representatives of the same genes. Fortunately the sophisticated open-source transcriptome assembly software can take into account such redundancy and remove duplicate sequences making further data processing more efficient (Simpson *et al.* 2009 and Grabherr *et al.* 2011). Below is a discussion of the pros and cons of using de-novo genomic and transcriptomic libraries, followed by a summary table of this information and a rationalization for choice of data type generated for the experiments in this thesis.

#### 1.4.8.1 Comparison of Sequencing Costs

The first point of discussion when comparing genomic to transcriptomic data is the difference in sequencing costs. The cost of sequencing has declined greatly since the advent of NGS technology [Figure 1.6], however there is still significant difference in the expenses between sequencing genomes and transcriptomes. A comparison of seventy-six sequencing centers across North America offering NGS genome sequencing by Illumina Solexa (Bennett, 2004) reveals retail charges between \$1,000 and \$3,000, whereas a similar comparison concerning transcriptome sequencing over fifty-three centers reports retail charges between \$200 and \$400 [Table 1.1]. Based on these figures it is five to seven times more expensive to sequence a genome than it is a transcriptome. Most in-house molecular libraries were sequenced by Edinburgh Genomics who offered considerably more competitive prices than the above rates. However the cost difference between sequencing genomes and transcriptomes remained constant so it was more economically feasible to sequence transcriptomes than genomes.

#### 1.4.8.2 Ortholog Coverage in Datasets

Since we were identifying and mapping orthologs to an established dataset of genes we only need coverage for those genes as opposed to every gene in the genome. Theoretically a genomic approach ensures a complete coverage (although see Table 3.1 where the *Parasagitta sp.* genome only had a 55% coverage of the Philippe *et al.* (2011b) dataset) while a transcriptome approach ensures a decent coverage, but with complete coverage unlikely. Incomplete ortholog coverage of a particular taxon is deleterious to the phylogenetic signal of the dataset but not as much of a hindrance as

once thought (Philippe *et al.* 2017) particularly when working with NGS data and large superalignments such as those used in this thesis [**Materials and Methods 2.2.6 & 3.2.3**]. Moreover the datasets in question were initially built from expressed sequence tags (ESTs) a somewhat random availability of genes used for a variety of published molecular experiments. As such a transcriptomic approach, although not perfect, is a considerable improvement of ortholog coverage than the ESTs.

#### **1.4.8.3 Scale and Complexity of Alternative Libraries**

De novo assembly refers to the assembly of a genome or transcriptome without using a reference genome in order to accurately predict genes (Zerbino & Birney, 2008; Simpson *et al.* 2009; Grabherr *et al.* 2011). A de novo approach is of essential importance when treading new molecular ground of an organism never before sequenced. This approach is the cornerstone of our experiments since these works aim to improve phylogenetic reconstruction methods with new molecular catalogs of previously non-sequenced species.

Genomes contain an enormous amount of non-coding DNA that needs to be excluded from the assembly procedure. For example, the coding genes of the human genome represent just 1.5% of the total (Lander *et al.* 2001), the remaining being made up of non-functional DNA such as introns, endogenous retroviruses, transposable elements, repetitive sequences, pseudogenes, transcription and translator regulatory features, centromeres (structural components at the center of chromosomes), telomeres (capping and protecting the chromosome terminus), non-coding RNA molecules (most notably miRNAs), with the remaining DNA serving as a structural component for chromosomal integrity (Lander *et al.* 2001; Venter *et al.* 2001; Dunham *et al.*

2012; Plazzo & Gregory, 2014). Therefore the open source genome assembly software (Zerbino & Birney, 2008) must parse through the vast majority of the DNA in order to identify open reading frames (ORFs) indicative of functional genes (Penn *et al.* 2000 and Andrews & Rothnagel, 2014), which is a long and computationally intensive process. The validity of these ORFs needs to be certain as in these cases there is no reference genome to compare the predicted genes to. Instead this can be achieved by cross-referencing the predicted genes with genetic databases such as RefSeq (Pruitt *et al.* 2005). Transcriptomic libraries are much simpler to parse as by definition each sequence should be representative of an expressed gene, therefore far less computational resources are required to assemble transcriptomes.

#### **1.4.8.4 Repeats, Redundancy, and Isoforms**

When not studying gene expression profiles, transcript repetition is a hindrance to transcriptomic data, however redundant sequences can be easily removed by assembly software (Simpson *et al.* 2009 and Grabherr *et al.* 2011). Another issue of transcriptomic data are isoforms: different mRNA iterations of the same gene (Flintoft, 2013). The intron and exon makeup of genes is what makes isoforms possible, allowing genes to be spliced into different arrangements and variations, the primary reason as to why a single gene can code for multiple different proteins (Lee & Rio, 2015). Isoforms can be a problem when mapping newly sequences transcriptomes to large datasets of concatenated orthologs as they can be similar enough to one another to make it difficult to choose between them when matching the corresponding ortholog in the dataset. In such cases the isoform that is most similar to the dataset ortholog in terms of sequence similarity and length is chosen. Sequence



repeats and isoforms are less of a problem for high quality genomic libraries as they can identify the longest ORF for a particular gene resulting in a single sequence representative for each gene in the genome (Zerbino & Birney, 2008).

**Table 1.1 Genomics versus Transcriptomics**

A table describing some important benefits and drawbacks of the different forms of molecular data used in this thesis. \* Costs sourced from a list of North American sequencing centers courtesy of [scienceexchange.com](http://scienceexchange.com): 76 sequencing centers offering NGS of genomics, 53 sequencing centers offering NGS of transcriptomes. \*\* Examples taken from in-house sequencing projects [Supplementary Material 2.2].

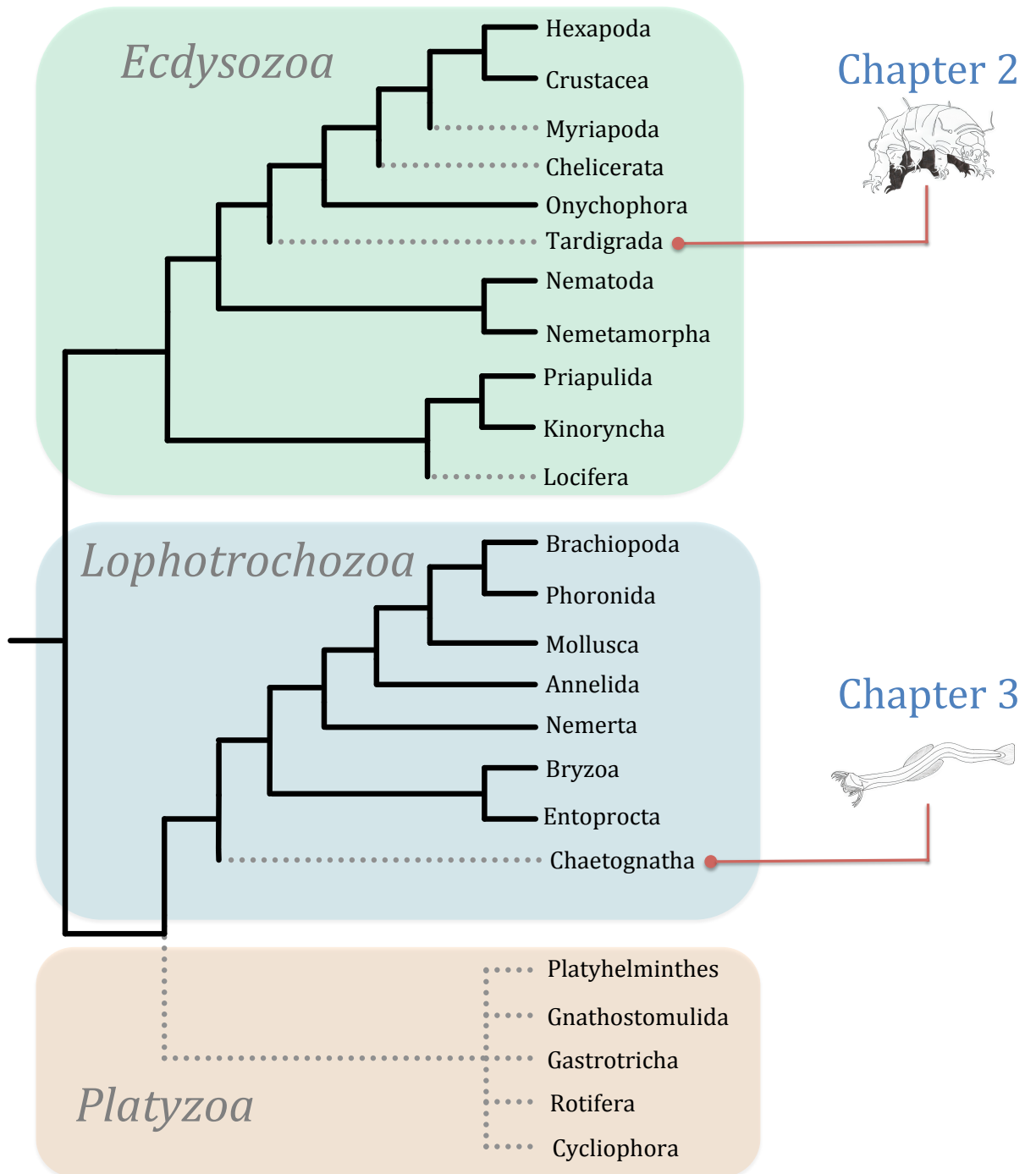
**Table 1.1: Genomics versus Transcriptomics**

	<b>Genomics</b>	<b>Transcriptomics</b>
<b>Sequencing Cost</b>	\$1,000 - \$3,000* per sample	\$200 - \$400* per sample
<b>Dataset Coverage</b>	Potential for 100% Coverage	Potential for High Coverage
<b>Library Size &amp; Complexity</b>	288 - 545 million reads [28.8 - 54.5 Gb pairs]** Full genomic library Requires more processing: ORF & Gene prediction One representative per gene	23 - 111 million reads [2.3 - 11.1 Gb pairs]** Incomplete genomic library No gene prediction required, instead gene verification Multiple representatives for each transcript and isoforms
<b>Open Source De novo Assemblers</b>	Laborious, inefficient, and requires software modding for animal genomes	Highly automated and efficient
<b>Computational Resources</b>	Extremely high demand	Manageable cost

## 1.5 The Protostomia

This body of work takes advantage of the extensive encyclopedia of evolutionary knowledge generated over the last 150 years in conjunction with incredible technological advancement in recent years to study one of the oldest groups of animals on the planet: the Protostomia. All animal life belongs in the Kingdom Metazoa (Lake, 1990) which itself is broadly divided into two subkingdoms the Parazoa, consisting of sponges (Riesgo *et al.* 2014), and the Eumetazoa which contains most animal life known today; anything that has an organized tissue system, a three germ layer embryonic development, and neurons (Nielson, 2001). The largest most prominent member of the Eumetazoa are the Bilateria, animals with a body plan defined by bilateral symmetry (Peterson & Ernisse, 2001 and Halanych *et al.* 2004) (although the consistency of this classification is under question with recent work from Cannon *et al.* (2016)), these animals are divided into two very large groups based on their embryonic development: the Deuterostomia (which gave rise to mammals, reptiles, and birds) (Ruggiero *et al.* 2015) and the Protostomia (Philippe *et al.* 2005) which later radiated into the most biodiverse collection of animals on the planet, the invertebrates (Erwin & Valentine, 2012). There also exists a small number of bilaterians that belong to neither of these two large clades (Davidson *et al.* 1995).

This body of work concerns the study of protostome evolution, bilateral animals with a pattern of embryonic development involving the blastopore morphing into the anus (Martín-durán *et al.* 2016), this is true for most cases but there is an exception to this classical distinction (see chapter 3). The Protostomia are made up of three superphyla: the Ecdysozoa, Lophotrochozoa, and Platyzoa (Philippe *et al.* 2005) [Figure 1.8].



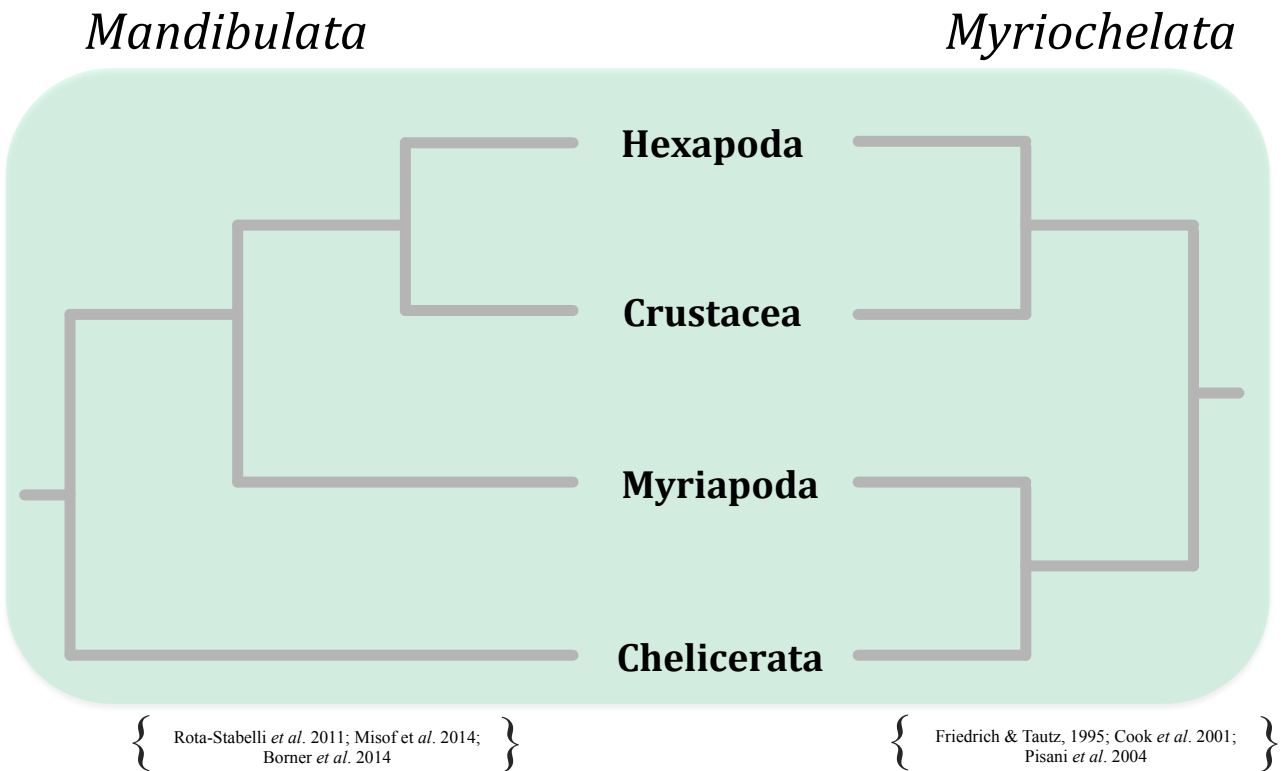
**Figure 1.8: The Protostomia**

A cladogram illustrating the superphyla of the protostomes.

Disputed topologies are in dotted lines. Only one of the hypotheses for each of the disputed groups is outlined for simplicities sake. The phylogeny for the Ecdysozoa is provided by Rota-Stabelli *et al.* 2011. The phylogeny of the Lophotrochozoa is provided by Paps *et al.* 2009a and Kocot *et al.* 2016. The Platyzoa is controversial clade of un-segmented protostomes proposed by Cavalier-Smith in 1998. The taxa studied in Chapters 2 and 3 are highlighted.

### 1.5.1 Superphylum Ecdysozoa

The Ecdysozoa constitute one of the two main subdivisions within the Protostomia (Aguinaldo *et al.* 1997). Their most prominent member is the Arthropoda, made up of the subphyla Hexapoda (insects), Crustacea (lobsters, crabs), Myriapoda (centipedes, millipedes), and the Chelicerata (spiders, scorpions) (Telford *et al.* 2008). Other members of the Ecdysozoa include the Onychophora (velvetworms), Tardigrada (water bears) studied in chapter 3, the Nematoida (roundworms), and the Scalidorpha (penis worms, mud dragons) (Telford *et al.* 2008). All animals belonging to the Ecdysozoa have the ability to molt, hence “ecdysis”. The phylogenetic relationships between many of the ecdysozoans is not agreed upon. Beginning with the arthropod subphyla, there is disagreement on two competing hypothesis: the Mandibulata, the grouping of the Pancrustacea (hexapods and crustaceans) with the myriapods, characterised by their similar jaws and made evident by molecular phylogenetic studies (Rota-Stabelli *et al.* 2011; Misof *et al.* 2014; Borner *et al.* 2014). Contrarily, the Myriochelata have been proposed by earlier studies (Friedrich & Tautz, 1995; Cook *et al.* 2001; Pisani *et al.* 2004) which group the myriapods with the chelicerates to a sister pancrustcean group [Figure 1.9]. The debate between these two hypotheses will be discussed later in the thesis when many of their members become part of phylogenetic reconstruction studies.

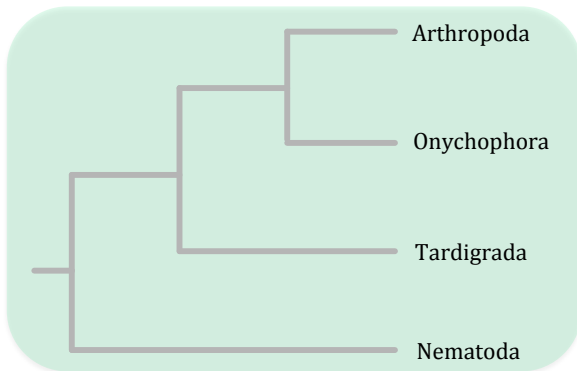


**Figure 1.9: The Arthropod Subphyla: Mandibulata versus Myriochelata**

Disagreement within the Arthropoda, centering on the relationships of the centipedes & millipedes (myriapods) and the spiders, ticks, scorpions, and marine based horseshoe crabs & sea spiders (chelicerates). The relationship of the other subphyla is well defined however; the hexapods (represented by flies, beetles and other insects) form a clade with the crustaceans (crabs, lobsters, shrimps) named the Pancrustacea.

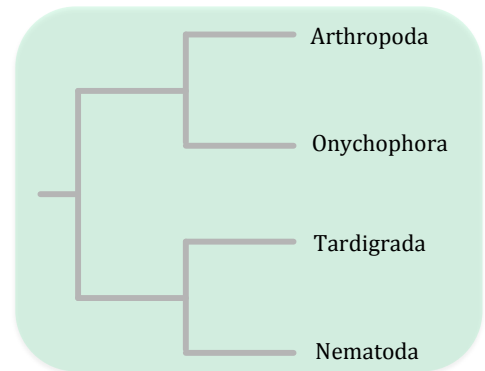
Also under debate within super phylum Ecdysozoa is the positioning of the Tardigrada, a conundrum discussed in-depth in chapter 2. There is evidence for a group of tardigrades, onychophorans, and arthropods coined the Panarthropoda (Campbell *et al.* 2010 and Rota-Stabelli *et al.* 2011) [Figure 1.10 A], a tardigrade - nematode grouping (Roeding *et al.* 2007; Lartillot & Philippe 2008; Meusemann *et al.* 2010; Borner *et al.* 2014) [Figure 1.10 B], and a tardigrade - arthropod grouping with the exclusion of the onychophorans (Smith & Ortega-Hernandez 2014 and Gross *et al.* 2015) known as the Tactopoda [Figure 1.10 C].

## A Panarthropoda



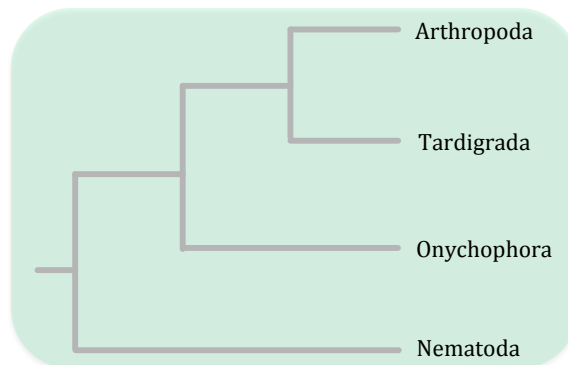
{ Rota-Stabeli *et al.* 2010 and Campbell *et al.* 2011 }

## B Tardigrada & Nematoda



{ Roeding *et al.* 2007; Lartillot & Philippe 2008; Meusemann *et al.* 2010; Borner *et al.* 2014 }

## C Tactopoda



{ Smith & Ortega-Hernandez 2014 and Gross *et al.* 2015 }

### Figure 1.10: Alternative Tardigrade Hypotheses

The three tardigrade topologies based on differing experimental support. The grouping of the tardigrades with the nematodes and the Panarthropoda are based on molecular datasets; mostly ESTs. The Tactopoda is supported entirely by morphological evidence.

The ecological and evolutionary relevance of Ecdysozoa can hardly be understated as this animal group includes most of the world's biodiversity and biomass (with the Arthropoda and the Nematoda respectively) (Minelli *et al.* 2013 and Hugot *et al.* 2001). Furthermore, the oldest, unambiguous, globally distributed evidence of bilaterian worm activity on Earth is represented by the 521 million year old

*Treptichnus pedum* trace fossils, that have been shown to represent feeding traces of macroscopic, priapulid-grade worms (Jensen *et al.* 1998). Finally, arthropods constitute the most abundant fossils in Cambrian strata and have had a key role in defining Earth biodiversity since then, representing some of the first examples of terrestrialization 426 MYA (Wilson & Anderson, 2004). Consequently these animals are important to our understanding of the origins and evolution of ancient animals and as a case study for animals who have undergone drastic ecological changes.

### **1.5.2 The Platyzoa**

The Platyzoa is a group of un-segmented, mostly microscopic, flat, acoelomate and pseudocoelomate (meaning no or unconventional body cavity) protostomes first proposed by Cavalier-Smith in 1998. The group consists mainly of the platyhelminthes (“flat worms” - worms without a body cavity), gnathostomulids (“jaw worms” - marine based microscopic worms), gastrotrichs (“hairyback worms”), rotifers (“wheel animals” - plankton), and cycliophorans (“symbions” - protostomes with sack-like bodies). The experimental evidence for a monophyletic group including all members of the proposed phyla is sparse and the confidence for the grouping is poor (Dunn *et al.* 2008 and Hejnol *et al.* 2009) with other studies suggesting a paraphyletic Platyzoa (Struck *et al.* 2014 & Laumer *et al.* 2015). These findings have led to researchers suggesting that the Platyzoa is an artificial grouping caused by systematic bias (Kocot *et al.* 2016). If such claims are factual it would mean redistributing the taxa discussed into the closely related Lophotrochozoa. The validity of the Platyzoa is investigated later in this thesis.

### 1.5.3 The Lophotrochozoa

The Lophotrochozoa (Halanych *et al.* 1995) make up the remaining protostomes, with their topology described from molecular phylogenetic reconstruction experiments from Paps *et al.* (2009a) and Kocot *et al.* (2016) [Figure 1.11]. The lophotrochozoans are animals characterized as non-molting protostomes, mostly consisting of the Brachiopoda “lamp shells” - marine based animals with a shell like appearance, Phoronida “horseshoe worms”, Mollusca (including the marine based squids and terrestrial snails), Annelida “ringed worms” such as earthworms, Nemertea “ribbon worms”, Bryozoa “moss animals”, Entoprocta, and Chaetognatha (“bristle jaws” - plankton (Giribet, 2008). It is important to note that the previously defined platyzoans: platyhelminthes, gnathostomulids, gastrotrichs, rotifers, and cycliophorans may not share a common ancestor and could be members of the Lophotrochozoa but the affinity of these groups remains uncertain as the validity of the Platyzoa is still under debate (Dunn *et al.* 2008; Hejnol *et al.* 2009; Struck *et al.* 2014; Laumer *et al.* 2015; Kocot *et al.* 2016).

Lophotrochozoans display an unusually extensive morphological diversification for a superphylum with considerable structural disparity between the Mollusca and Annelida despite sharing a common ancestor as part of a monophyletic group, making them particularly interesting in the field of evolutionary cladistics (Giribet, 2008). This suggests a pattern of considerable morphological adaptations have occurred in certain lophotrochozoan lineages since their origins.

The arrangement of the internal lophotrochozoan phyla is uncertain, possibly as a consequence of the aforementioned unusual range of morphological adaptations across the superphylum making it difficult to identify the true topologies within. There have been many proposed groupings such as the Eutrochozoa - the organization

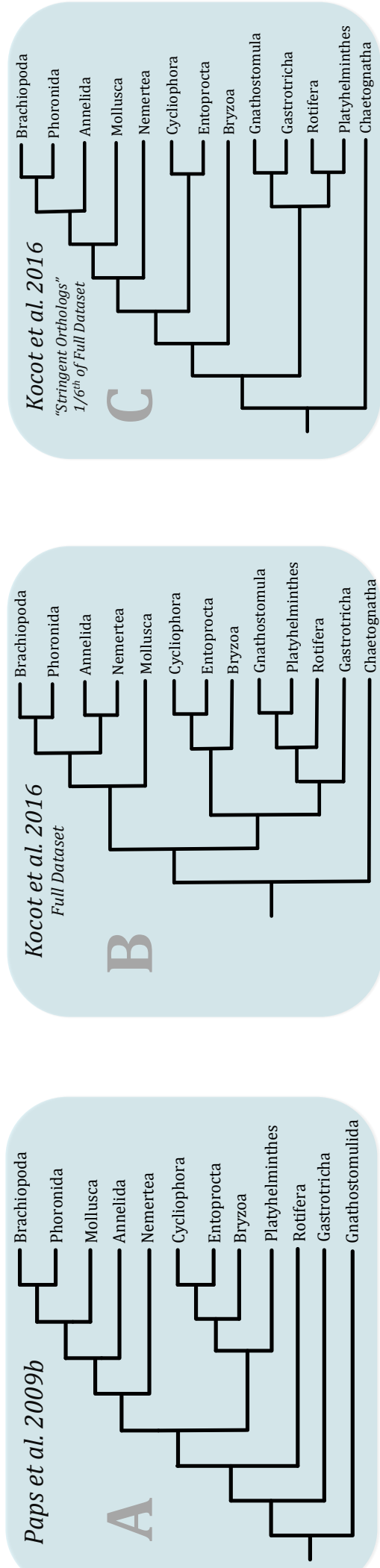


of the Mollusca, Annelida, and Nemertea based on the similarities of their coelomic sacs and type of mesoderm formation (Peterson & Eernisse, 2001), the Neotrochozoa - the grouping of the Mollusca and Annelida with the exclusion of the Nemertea based on a shared resemblance of unmodified trochophore larva (Kocot *et al.* 2011), the Kryptozoa which claims an affinity between the Brachiopoda, Phoronida, and Nemertea (Dunn *et al.* 2008), the Lophophorata first proposed by Hyman (1959) (Brachiopoda, Phoronida, plus Bryzoa), and the Tetra-neuralia placing the Entoprocta with the Mollusca (Wanninger, 2009). However there has been little evidence supporting these proposals with many phylogenetic studies of the Lophotrochozoa failing to recover these groups (Struck & Fisse, 2008; Hejnol *et al.* 2009; Paps *et al.* 2009b; Hausdorf *et al.* 2010; Struck, 2014; Laumer *et al.* 2015; Kocot *et al.* 2016) suggesting they are artifacts.

However there are two strongly supported hypotheses for groupings of the lophotrochozoan phyla: the Trochozoa and Polyzoa. The Trochozoa was first suggested by Rouse (1999) which places the Brachiopoda, Phoronida, Mollusca, Annelida, and Nemertea together based on similarities between their primary trochophore larva, supported by Dunn *et al.* (2008); Paps *et al.* (2009b); Struck (2014); Laumer *et al.* (2015); Kocot *et al.* (2016) but with varying arrangement of these groups. The Polyzoa consists of the Entoprocta, Cycliophora, and Bryzoa recovered by Struck & Fisse, (2008); Hejnol *et al.* (2009); Paps *et al.* (2009b); Hausdorf *et al.* (2010); Kocot *et al.* (2016).

Since the Trochozoa and Polyzoa are the most highly supported hypotheses for lophotrochozoan phyla arrangement, the major studies recovering these groups are shown in [Figure 1.11](#) as an illustration of the current opinion of lophotrochozoan phylogeny.

Lack of molecular data has been a problem in assigning some of the smaller lophotrochozoan phyla falling outside the Trochozoa and Polyzoa (Paps *et al.* 2009b). However, even a recent phylogenomic scale analyses consisting of 32 lophotrochozoan transcriptomes (Kocot *et al.* 2016) has been unable to clarify the exact relationships within. The study from Kocot *et al.* (2016) took a vigorous approach to investigating the phylogeny of the Lophotrochozoa, employing 638 orthologous groups amongst 74 taxa with a strict respect for systematic biases. This resulted in eight differing phylogenetic scenarios for the Lophotrochozoa based on a variety of approaches to account for sources of phylogenetic error. Among these scenarios there was only consistent topological parity in the placement of the Mollusca, Brachiopoda, Phoronida, and Nemertea. The relationships of these groups broadly fit the positioning from a previous study (Paps *et al.* 2009b) [Figure 1.11 A] however the topology of the remaining lophotrochozoans is erratic depending on the type of dataset used (Kocot *et al.* 2016). See Figure 1.11 B & C for a reconstruction of two of their datasets: the full 74 taxa - 638 orthologous groups matrix and the dataset consisting of a restricted group of orthologs designed to maximize the reduction of systematic biases. A significant pattern nestled within the Kocot study is the viability of the proposed sister group to the Lophotrochozoa: the Platyzoa (Cavalier-Smith, 1998). The Platyzoa are only recovered in their entirety under one of the eight phylogenetic experiments and as methods are implemented to reduce systematic biases, support for the group begins to fall. The implication of these results is that the Platyzoa may be a phylogenetic artifact, and the protostomes affiliated with this group actually could belong in the Lophotrochozoa.



**Figure 1.11 A: Superphylum Lophotrochozoa Paps et al. (2009b)**  
 A dataset consisting of 96 taxa, 11 protein-coding genes, and 2 ribosomal genes.  
 Placed the chaetognaths within the Ecdysozoa - Platyzoa not recovered - Trochozoa & Polyzoa recovered.

**Figure 1.11 B: Superphylum Lophotrochozoa Kocot et al. (2016) Full Dataset**  
 A dataset consisting of 74 taxa, 32 lophotrochozoan transcriptomes, and 618 “orthologous groups”.  
 Platyzoa not recovered in this topology and only recovered in one out of eight experiments from this study.  
 Trochozoa & Polyzoa recovered.

**Figure 1.11 C: Superphylum Lophotrochozoa Kocot et al. (2016) Stringent Orthologs**  
 A dataset consisting of 74 taxa and 32 lophotrochozoan transcriptomes.  
 Reduced dataset than only includes the most stringent “orthologous groups”.  
 Platyzoa not recovered - Polyzoa not recovered - Trochozoa recovered.

On the topic of phylogenetically ambiguous lophotrochozoans, chapter 3 consists of an evolutionary investigation into one of the Animal Kingdoms first predators: the Chaetognatha. The chaetognaths are particularly fascinating, not just in regards to their ancient carnivore tendencies, but the complete confusion as to their phylogenetic affinity with a considerable number of conflicting studies trying to place them. Initially morphologists believed them to be deuterostomes because of their deuterostome-like development (Doncaster, 1902; Hyman, 1959; Ghirardelli, 1968 & 1981; Ducret 1978), however every molecular study places them within the protostomes but with little to no agreement (Telford & Holland, 1993; Papillon *et al.* 2003 & 2004; Matus *et al.* 2006; Marlétaz *et al.* 2006 & 2008; Paps *et al.* 2009b; Philippe *et al.* 2011b; Kocot *et al.* 2016). The protostome affinity of the chaetognaths is investigated in detail in chapter 3.

## 1.6 Thesis Aims

The aims of this thesis are broken down into their respective chapters.

### 1.6.1 Chapter 2

#### **Phylogenomics and the Case of the Rapidly Evolving Tardigrada**

- Generate a new ecdysozoan dataset on the backbone of the amalgamated Campbell *et al.* (2011) and deuterostome-pruned Philippe *et al.* (2011b) datasets with the supplementation of sixteen ecdysozoans from next generation sequencing experiments.
- Clarify the position of the tardigrades within the Ecdysozoa from phylogenetic reconstruction experiments of the new ecdysozoan super matrix.
- Investigate the presence of long branch attraction within the Ecdysozoa and test if it is influencing the grouping of the two fastest evolving members: the tardigrades and nematodes.
- Identify the point of origin of the Tardigrada using relaxed molecular clocks.

### 1.6.2 Chapter 3

#### **Chaetognatha: The Mosaic Metazoans**

- Reconstruct the phylogeny of the chaetognaths after mapping the orthologs of the *Parasagitta sp.* genome to the full Philippe *et al.* (2011b) dataset.
- Confirm that the chaetognaths are protostomes and pinpoint their much disputed phylogenetic position within the clade.
- Use the chaetognath fossil record and lineage characteristics of extant chaetognaths to encode their features into an existing morphological dataset

(Peterson & Ernisse, 2001). Generate the morphological phylogeny of the chaetognaths to better understand the living chaetognaths relationship with their fossil catalog.

- Divergence time estimation experiments: find the age of the chaetognaths using relaxed molecular clocks and apply the phylogenetic signal of both morphology and molecular data to a total evidence dating study in order to further understand the origins and evolutionary history of the Chaetognatha.

### **1.6.3 Chapter 4**

#### **Arthropod Terrestrialization**

- Investigate the timeline for Metazoan origins and the radiation of animal lineages using multi-model relaxed molecular clocks.
- Identify when the terrestrial arthropod subphyla (Hexapoda, Crustacea, Arachnida (class of chelicerates), and Myriapoda) colonized land.
- From these results infer the earliest known terrestrial ecosystem capable of supporting life.
- [Collaborators] Determine how the arthropods invaded land through ancestral character state reconstructions

### **1.6.4 Chapter 5**

#### **A Phylostratigraphic Study of Protein Family Evolution Across the Metazoa with Focus on the Protostomia**

- Expand the dataset of a “proof of concept” phylostratigraphic study of metazoan protein families (Pisani *et al.* 2013) with twenty-eight taxa from next generation sequencing experiments.

- Establish whether the findings of the initial study by Pisani *et al.* (2013) were an accurate portrayal of metazoan protein family evolution or whether results were impaired from a lack of taxa coverage and gene sampling.
- Generate protein families using the Markov Clustering Algorithm (Enright *et al.* 2002) and distribute these families amongst a metazoan supertree, the phylogeny of which is supported by a mixture of results from this thesis and published studies.
- Plot the rate of protein family acquisition across the metazoan supertree and highlight any nodes that experience a rate higher than the mean.
- Annotate the protein families pertaining to the nodes in the tree that experience a significant rate of protein family gain in order to garner functional insights into macroevolutionary adaptations spanning the timeframe of animal evolution.

## **1.6.5 Chapter 6**

### **Thesis Discussion**

- Apply the discoveries and knowledge from the four experimental chapters to discuss new insights into the complex evolution of the protostomes.
- Identify improvements for the methodologies used.
- Highlight opportunities for future work that could further expand our knowledge of phylogenomics and protostome evolution.

### 2.1 Introduction

#### 2.1.1 Phylum Tardigrada

The Tardigrada “slow walkers”, commonly referred to as water bears, are members of the Ecdysozoa, a clade sharing the common feature that all of its members undergo the process of moulting (i.e. ecdysis) (Aguinaldo *et al.* 1997).

Tardigrades are near microscopic (250-500  $\mu\text{m}$ ) invertebrate animals (Nelson, 2002) distinguished by features such as their segmented body divided into four sections (Gabriel, 2007), four pairs of stumpy legs attached to which are claws with varying arrangements depending on species (Guidetti & Bertolani, 2005), and well defined head structure. They possess reproductive, digestive, and nervous systems but lack respiratory and circulatory systems (Nelson, 2002). Currently there are over 1,000 known species (Degma, *et al.* 2016) broadly classified into two orders: the Heterotardigrada and Eutardigrada (Guidetti & Bertolani, 2005). The Heterotardigrada are distinguished from the Eutardigrada based on differences in the arrangement of their gonopore, anus, lack of “Malpighian tubules”, and pharynx structure (Guidetti & Bertolani, 2005). See [Figure 2.1](#) for images of tardigrade anatomy and internal structure.

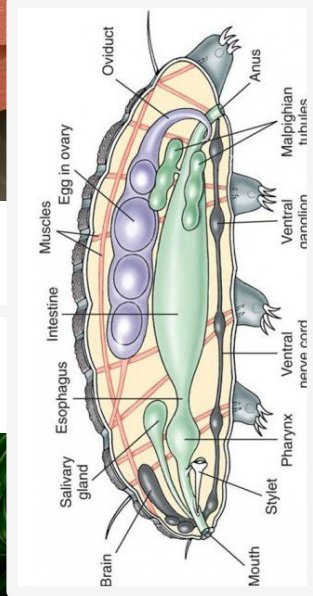




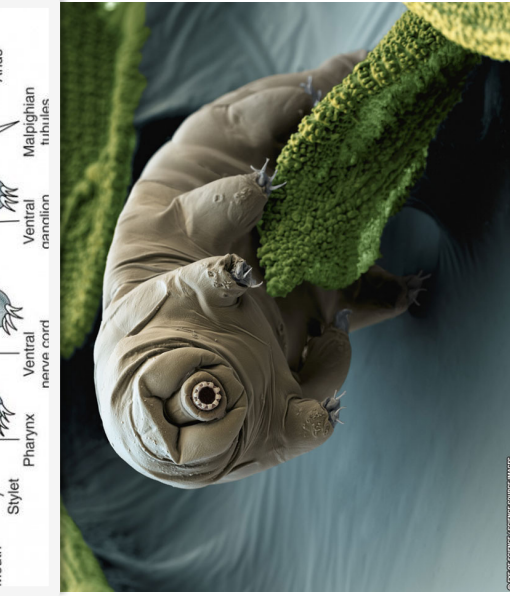
Science Photo Library / Power and Syred



Science Photo Library / Alamy



Eye of Science / Science Source



Eye of Science / Science Source



Eye of Science / Science Source

**Figure 2.1: Anatomy of the Tardigrades**

The tardigrades are commonly referred to as “water bears” or even “moss piglets” because of their appearance and marine and terrestrial habitats. Their most distinguishing features are their claws and distinct head which vary depending on species (Guidetti & Bertolani, 2005). Image source for internal anatomy: <https://netnature.wordpress.com/2015/07/31/tardigrados-calam-a-boca-da-criacao/>

The tardigrades are interesting creatures from an evolutionary standpoint as there is uncertainty surrounding their precise topological position amongst the Ecdysozoa, and scarce evidence of a timeline for their evolutionary radiation. While there have been numerous phylogenetic studies concentrating on the water bears (Lartillot & Philippe, 2008; Meusemann *et al.* 2010; Rota-Stabelli *et al.* 2010; Campbell *et al.* 2011; Borner *et al.* 2014; Smith & Ortega-Hernandez, 2014; Gross *et al.* 2015) there have been comparatively sparse divergence time estimation studies trying to identify when these creatures originated (Rota-Stabelli *et al.* 2013). This is somewhat surprising given the extent of molecular clock experiments involving their fellow ecdysozoans (Pisani *et al.* 2004; Regier *et al.* 2005; Rota-Stabelli *et al.* 2013; Wheat & Wahlberg, 2013; Misof *et al.* 2014) to name but a few, but may be a consequence of its disputed phylogenetic position, as a robust tree is an important criterion for an accurate relaxed clock analysis (Drummond *et al.* 2006).

### **2.1.2 Significance of the Tardigrades**

The most fascinating aspect of the tardigrades is their extreme ecological adaptability as they have been discovered in a large range of drastically different ecosystems across the planet. This includes freshwater (Nelson, 2000), marine - in all oceans both subtidal and at great depths as far as 4,690m (Nelson, 2002), terrestrial, (Guidetti *et al.* 1999), both Arctic and Antarctic (Pugh & McInnes, 1998 and Convey & McInnes, 2005), and geothermal (Nelson, 2000) environments. Indeed the tardigrade robustness is so impressive that they have been found to exist outside of all known ecosystems of the planet Earth itself: impressively surviving in the vacuum of space (Jönsson *et al.* 2008), the only animal known to endure such hostile conditions for life.

The implications of existing in the above conditions is that the tardigrades are incredibly thermostable, surviving the extreme heat of hot springs (Nelson, 2000) and the contrastingly extreme near absolute zero cold of outer space (Jönsson *et al.* 2008). In fact, experiments testing tardigrade resistance to temperatures have shown they can withstand a range between 151<sup>0</sup>C and -273<sup>0</sup>C (Rahm, 1921 & Becquerel, 1950). The ability to withstand these conditions is attributed to a survival technique known as cryptobiosis, a process whereby the animal desiccates itself through reversible anhydrobiosis creating a state in which all metabolic processes are shut down in response to pernicious environments (Clegg, 2001).

Additionally it is clear the water bears are resistant to both high pressure environments of the deep seas (Nelson, 2002) and the pressureless vacuum of space (Jönsson *et al.* 2008), while being exposed to high levels of solar radiation (deadly to animals) without the protection of a planet's atmosphere. Such an unusually strong constitution compared to other animals makes the water bears a subject of great interest to researchers from multiple scientific fields including the evolutionary (Gabriel *et al.* 2007), molecular (Schill *et al.* 2009), ecological (Nelson, 2002), medical (Hashimoto *et al.* 2016), and astrobiology disciplines (Horikawa *et al.* 2008 and Jönsson *et al.* 2008).

From the evolution viewpoint, the tardigrades are significant because none of their fellow ecdysozoans, or most of the Metazoa for that matter, share the same impressive durability to external environmental factors and adaptability to a multitude of challenging ecosystems (Nelson, 2002). This suggests that the tardigrades did not inherit their strong constitution from an ancestor but rather evolved these intense survival capabilities independently. Thus the evolution of the tardigrades from their point of origin becomes of great interest as one has to consider what were the

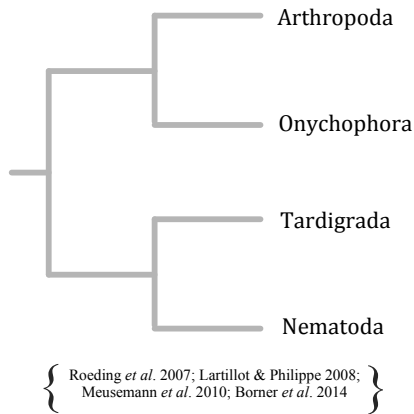
environmental circumstances the lineage encountered for such remarkable adaptations. Essentially the question becomes *why* such adaptations occurred in the tardigrade lineage and not other ecdysozoans.

The gene library, genetic networks, and metabolic pathways of the tardigrades are of particular interest to molecular biologists as genes novel in these species, or patterns of gene networks unique to the tardigrades, may help increase our understanding of how the tardigrade metabolism functions in extreme environments and explain *how* the tardigrades are able to survive a far wider range of hostile environments than most animals, particularly regarding the mechanisms of cryptobiosis. Clegg (2001) claims cryptobiosis could be activated by large concentrations of sugars involved in water retention such as trehalose and sucrose in addition to the expression of heat shock genes in response to external stimuli (harsh environmental conditions). Further studies on tardigrade cryptobiosis claim body size and the age of the creature influences its ability to survive the anhydrobiosis process, with larger older water bears lacking the energy reserves to rehydrate from a desiccated state (Jonsson & Rebecchi, 2002). Tardigrade sequencing projects (Hashimoto *et al.* 2016 & Koutsovoulos *et al.* 2016) have been useful in making the genes thought to be involved in the cryptobiosis process accessible. However, generating the genetic lexicon of some tardigrade species is only the start in a complicated process of identifying the genes, gene networks and metabolic pathways which may be responsible for tardigrade durability.

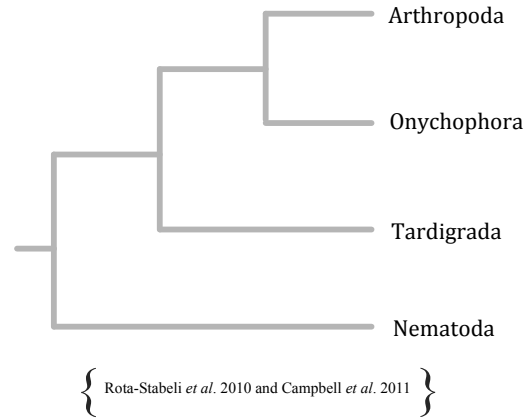
### 2.1.3 Conflicting Phylogenetic Theories

As discussed in chapter 1, the relationships within the Ecdysozoa are far from agreed upon. In the case of the tardigrades, there is uncertainty regarding their relationship with the nematodes, arthropods and onychophorans. Specifically, there is experimental support for a tardigrade and nematode grouping (Roeding *et al.* 2007; Lartillot & Philippe, 2008; Meusemann *et al.* 2010; Borner *et al.* 2014) [Figure 2.2 A], which contrasts against evidence of a sister relationship to the arthropods and onychophorans, forming a group called the Panarthropoda; (Rota-Stabelli *et al.* 2010 and Campbell *et al.* 2011) [Figure 2.2 B], both of which are in disagreement with morphological evidence suggesting they are a member of a group called the Tactopoda (Smith & Ortega-Hernandez, 2014 and Gross *et al.* 2015), which proposes an exclusive affinity between the Arthropoda and Tardigrada [Figure 2.2 C].

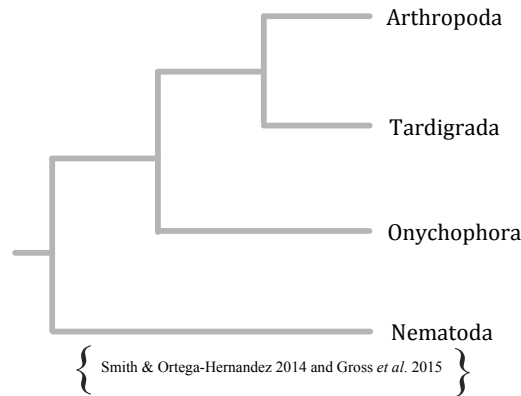
## A *Tardigrada & Nematoda*



## B *Panarthropoda*



## C *Tactopoda*



### Figure 2.2: Alternative Tardigrade Hypotheses

The three tardigrade topologies based on differing experimental support. The grouping of the tardigrades with the nematodes and the Panarthropoda are based on molecular datasets; mostly ESTs. The Tactopoda is supported entirely by morphological evidence.

Because of such phylogenetic uncertainty, and with molecular clocks being topology dependent (Drummond *et al.* 2006), there has been sparse investigation into the timing of the evolutionary radiation of the Tardigrada. Rota-Stabelli *et al.* (2013) points to a tardigrade divergence at the beginning of the Silurian, approximately 442 million years ago (MYA).

However, this divergence time estimation pertains to a Panarthropoda relationship and the tardigrade radiation of this dataset in particular varies depending on the type of data used and on clock model choice (Drummond *et al.* 2006).

Additionally, since the time of this study, gene sampling of the tardigrades has improved with new transcriptomic and genomic data (Borner *et al.* 2014 and Koutsovoulos *et al.* 2016) .

Possibly the strongest line of evidence for the sister grouping of the tardigrades with the arthropods and onychophorans comes from Campbell *et al.* (2011). Relying on an EST dataset, as NGS had not risen to prominence at the time, they generated a 255 protein dataset of 49,023 amino acids, introduced additional lines of molecular data for support, and ran phylogenetic signal dissection to identify LBA in their dataset.

Campbell's results support the Panarthropoda clade and demonstrate how the ever-present effects of LBA in datasets containing rapidly evolving taxa can bias results.

However, Campbell's dataset suffers from a lack of tardigrade sequence data and additionally relies on a very small number of miRNA phylogenetic markers to support the tardigrades placement. The robustness of miRNA as phylogenetic markers has since come into question as it seems miRNA loss is a more common occurrence than initially thought (Dunn, 2014).

More recently, an approach involving the sequencing of the *Echiniscus testudo* transcriptome produced an ecdysozoan phylogeny supporting the grouping of tardigrades and nematodes but with a notable degree of uncertainty. Building a dataset composed of 63 taxa, 189 genes and 24,429 amino acid sites, Borner *et al.* (2014) were unable to clarify the Tardigrada position within the Ecdysozoa due to conflicting results of their phylogenetic experiments. Concerned with systematic errors associated with rapidly evolving taxa, their dataset was partitioned into all sites, intermediate, fast, and slow sites.

All datasets returned the same tardigrade-nematode grouping with the exception of the slowest evolving partition that places the tardigrades sister to the arthropods and

onychophorans. The nematodes themselves are widely known to introduce LBA in to a dataset due to their rapid rate of evolution (Phillipe *et al.* 2005).

Borner *et al.* (2014) rejected the findings of their slowest evolving partition citing a lower posterior probability support of 0.8 for the tardigrades placement to that of their other partitions (1.0), and thus claimed that their dataset was not inflicted with a LBA bias.

However if one concludes that the majority of the phylogenetic signal in their dataset nests within the fastest, most saturated, sites then one cannot eliminate the possibility of the presence of a systematic bias such as LBA existing concurrently.

Unfortunately the modus of LBA is that it is positively misleading and will tend to falsely inflate support values due to its nature of grouping taxa based on their higher substitution rates (Philippe *et al.* 2005). Consequently, assuming its presence in the data based on the phylogeny of their slowest evolving partition, one would expect highly supported values for artificial groupings.

From a morphological standpoint, evidence for the positioning of the Tardigrada points singularly to the Tactopoda (Arthropoda + Tardigrada) (Smith & Ortega-Hernandez, 2014 and Gross *et al.* 2015). A cladistic study based on panarthropod head segmentation of *Hallucigenia sparsa*, a fossilized lobopodian from the Burgess Shale formation (Smith & Ortega-Hernandez, 2014) revealed that it has strong morphological similarity to the jaws teeth and claws of extant onychophorans. The eventual conclusion of their cladistics experiments was the proposition of a Tactopoda relationship.

Gross *et al.* (2015) provided an in-depth study of morphology of the Tardigrada concerning their nervous system. They claim the long thought synamorphly between



tardigrade and arthropod circumbuccal rings does not exist due to their differing patterns of development and innervation.

Contrary, the segmented ganglia, neuron arrangement, presence and orientation of leg nerves, axon development in the trunk, and specificity of adjacent segment linkage, novel to the tardigrades and arthropods (Harzsch, 2006; Whittington, 2006; Ungerer *et al.* 2011), all of which are either characters evolving from the immediate tardigrade-arthropod ancestor, or all lost in onychophorans and nematodes, or are homoplasys of convergent evolution. As such they conclude that evidence for tardigrades and arthropods sharing a nearest common ancestor is clear.

What complicates the matter is the presence of unique tardigrade-onychophoran traits such as the formation of brain neuropils, direction of axon growth, pattern of leg development, and timing of lateral nerve formation (Mayer *et al.* 2009 & 2010). This evidence on its own may point to a tardigrade – onychophoran sister grouping to the exclusion of the arthropoda but when considered with the other results of the study the most parsimonious scenario suggests a series of trait losing events in some of the lineages being discussed. Based on this morphological assessment one can imagine a scenario of a shared arthropod, tardigrade, onychophoran ancestor (Panarthropoda) after which either the onychophorans or arthropods lost several of these unique aforementioned traits over time. In addition, there does not seem to be morphological evidence for a Tardigrada - Nematoda grouping, perhaps further indicating it may be a long branch attraction artifact from molecular datasets.

#### **2.1.4 Long Branch Attraction within the Ecdysozoa**

Discussing the tardigrade phylogenetic studies not only brings up the ongoing molecules versus morphology debate, but highlights the presence of a suspected systematic bias that has been ingrained in ecdysozoan lineages since molecular evolutionary biologists started investigating this biodiverse superphylum; long branch attraction (Telford & Copley, 2005). See **Introduction 1.2.6** for a full description of this phenomenon.

The suspected source of LBA within the Ecdysozoa stems from two rapidly evolving groups: the Tardigrada and Nematoida (Nematoda + Nemetamorpha). Lineages with an unusually higher rate of site substitution can often be grouped together based on this rate of change as opposed to their true phylogenetic affinity in contemporary reconstruction studies (Felsenstein, 1978). The respective long branched nematodes and tardigrades compared to their ecdysozoan counterparts in conjunction with conflicting studies in which a LBA presence is suspected in these lineages (Campbell *et al.* 2011; Borner *et al.* 2014; Gross *et al.* 2015) has generated debate as to the correct phylogeny of these groups that is still in ongoing today.

#### **2.1.5 Dating the Tardigrada Origins with Newly Sequenced Taxa**

With the phylogeny of the tardigrades under dispute, and only a single dating study centered on one of the phylogenetic hypotheses (Rota-Stabelli *et al.* 2013), an approach was taken to clarify the relationships of the ecdysozoans: the Tardigrada, Arthropoda, Onychophora, and Nematoida by adding newly sequenced taxa to most of these groups; *Milnesium tardigradum*, *Hypsibius dujardini*, *Echiniscus testudo*,

*Scutigera coleoptrata*, *Polydesmus angustus*, *Symphylella vulgaris*, *Glomeridesmus sp.*, *Oniscidea sp.*, and *Pycnogonium littorale*; respectively.

An inclusive approach was taken to molecular dating with each of the three tardigrade hypotheses chosen for molecular clock experiments following a comparative analysis of the resulting origin dates. Divergence time estimation was previously explained in section **1.3.6 The Molecular Clock**, below is a summary of the important parameters required in an accurate dating analysis.

- High quality molecular data (typically from NGS experiments).
- Well taxonomically sampled dataset for the groups of interest.
- Accurate calibration points anchored by robust fossil dating.
- Correct identification and assignment of fossils to groups.
- A relaxed clock approach.
- A root prior that is an accurate reflection of the origins of the entire group studied.
- Testing and comparing the results from multiple evolutionary models.
- Preferably, a concentrated analysis avoiding the inclusion of many distantly related groups as this can stretch the abilities of the models imposed.

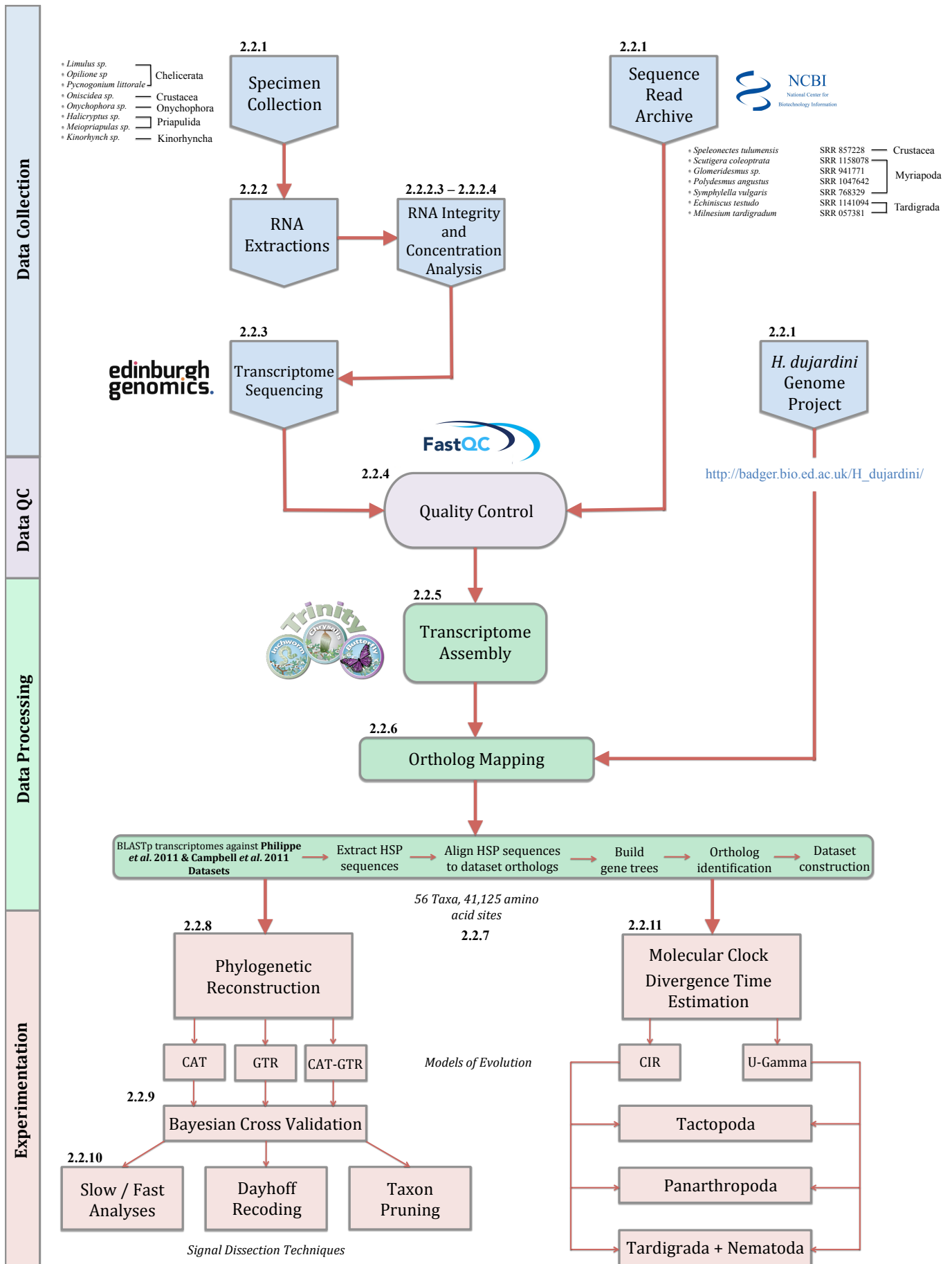
### **2.1.6 Aims of this Study**

The goal of this chapter is to apply sixteen newly sequenced taxa (see **Materials and Methods 2.2.1** for a full description) to two amalgamated existing ecdysozoan datasets (Campbell *et al.* 2011 & Philippe *et al.* 2011b) and recover the phylogeny of the Ecdysozoa with a focus on the tardigrade topology. Additionally, we wish to clarify whether the ecdysozoans are prone to the systematic error of LBA by

investigating the true phylogenetic nature of the tardigrades and their affinity with the nematodes. The principle method used for examining the presence of LBA is the implementation of a series of signal dissection techniques (discussed in section **1.3.2 Signal Dissection**) on the reconstructed phylogeny. Finally the origins of the tardigrades are dated using multi-model relaxed molecular clocks under each of the three proposed topologies: the Panarthropoda (Rota-Stabelli *et al.* 2010 and Campbell *et al.* 2011), Tactopoda (Smith & Ortega-Hernandez, 2014 and Gross *et al.* 2015), and Tardigrada plus Nematoda (Roeding *et al.* 2007; Lartillot & Philippe, 2008; Meusemann *et al.* 2010; Borner *et al.* 2014) to ensure a robust dating analysis, the inference of which is not restricted to a single phylogenetic hypothesis.

## **2.2 Materials and Methods**

The materials and methods detailed within this Chapter have been summarized in a flowchart [**Figure 2.3**]. Each step includes a number corresponding to the subsection of the materials and methods that describes the process. The flowchart is divided into four sections: data collection, data quality control, data processing, and experimentation to allow ease of navigation.



**Figure 2.3: Flowchart Detailing the Materials and Methods of Chapter 2**

The flowchart consists of four subsections: data collection, data QC, data processing, and experimentation. All methodologies used in this chapter are covered from specimen collection to divergence time estimation experiments. All methodologies are numbered to correspond to sections detailing them within the materials and methods.

### 2.2.1 Specimen Collection

This study centers on the origin and evolution of the tardigrades. As the Tardigrada are members of the Ecdysozoa (Aguinaldo *et al.* 1997), it was pertinent to collect genomic level data for their ecdysozoan sisters as well as the tardigrades themselves to ensure a good representation of taxa for this clade consisting of high quality molecular libraries.

#### 2.2.1.1 Pycnogonid (*Pycnogonium littorale*)

**Figure 2.4 A**

*Phylum Arthropoda / Subphylum Chelicerata / Class Pycnogonida / Order Pantopoda*

A colony of pycnogonids were collected from the Strangford Lough area in County Down with the aid of Dr. Julia Sigwart and her team in Queen's University Belfast Marine Laboratory. Specimen search and collections were carried out from 6:30am - 7:00pm for four days in November of 2011. Strangford Lough was chosen based on a previous study (Roberts, 1981) that had identified pycnogonid habitats in the seabed of its costal and bay areas.

November was chosen as the ideal time for searches due to the convenient low seasonal tide and workable light levels at sunrise of that time of year. This allowed us to wade out further into the bay and scavenge the seabed for the known habitats of the pycnogonids, mainly in and under the surface of rocks, seaweed, and kelp.

These objects were collected in to buckets and brought to the marine laboratory, roughly ten minutes away, and placed into tanks that were supplied with fresh seawater directly pumped in from the lough. This ensured that any specimens collected were living in an environment that was mimicking their habitat and thus maximized survival rates. For the rest of the day, every item collected was placed under a light microscope and pycnogonids were searched for. Any specimens found

were placed into their own tank containing a continuous supply of seawater pumped in from the lough, awaiting classification. Over the course of the four days, eight specimens were collected and identified as *Pycnogonum littorale* (classified based on the expertise of member of Queens University Belfast Marine Laboratory), three of which were stored in RNAlater (stabilization agent developed by QIAGEN) immediately. This was a preventative measure to ensure we would have DNA and RNA stable samples in the case of the live samples dying during transport as DNA and RNA degrade after death, the latter occurring at a rapid rate. The rest were transported to the National University of Ireland Maynooth in coolers containing seawater from Strangford Lough. They were then stored in a filtered tank in a cold room at a temperature similar to that of seawater.

Previous attempts at sequencing pycnogonids had failed due to large amounts of contaminant DNA in the samples. We believe this occurred as the most tissue plentiful parts (ideal for DNA & RNA extractions) of the pycnogonids anatomy are their legs, but unfortunately the legs also contains their guts. As such, previous extraction attempts resulted in sequencing the last meal of the pycnogonids as opposed to their genome. Our approach was to starve the specimens and monitor their status on a daily basis for three weeks, (the water being replaced by seawater from beaches on the coast of Dublin) this reduced the risk of extracting foreign DNA. Once these three weeks had elapsed, all specimens were submerged in RNAlater and stored in -80<sup>0</sup>C freezers awaiting extractions.

#### **2.2.1.2 Opilione (sp.)**

**Figure 2.4 B**

*Phylum Arthropoda / Subphylum Chelicerata / Class Arachnida / Order Opiliones*

Collecting opilione specimens was a much more straightforward task. A nest of opiliones was located in the local Kildare area. Identification of the specimens was carried out by laboratory post-doc at the time, Omar Rota-Stabelli. Specimens were stored in boxes before being transferred to vials of RNAlater and then stored in the -80°C freezers.

#### **2.2.1.3 Limulus (sp.)**

**Figure 2.4 C**

*Phylum Arthropoda / Subphylum Chelicerata / Class Xiphosura / Order Xiphosurida*

The horseshoe crab samples were sourced and sequenced by the University of Bristol.

#### **2.2.1.4 Oniscidea (sp.)**

**Figure 2.4 D**

*Phylum Arthropoda / Subphylum Crustacea / Class Malacostraca / Order Oniscidea*

Oniscidea specimens were collected and extracted by Omar Rota-Stabelli and sent for sequencing.

#### **2.2.1.5 Onychophora (*Epiperipatus* sp.)**

**Figure 2.4 E**

*Phylum Onychophora / Class Udeonychophora / Order Euonychophora*

Onychophora specimens were ordered from a website: ([www.exotic-pets.co.uk](http://www.exotic-pets.co.uk)).

#### **2.2.1.6 Halicyrtus (sp.)**

**Figure 2.4 F**

*Phylum Priapulida / Class Halicyrtomorpha / Order Halicyrtomorphida*

*Halicyrtus* sp. was sourced by Jakob Vinther of the University of Bristol's School of Biological Sciences and sequenced in-house.



### 2.2.1.7 *Meiopriapulas* (*sp.*)

Figure 2.4 G

*Phylum Priapulida / Class Meiopriapulomorpha / Order Meiopriapulomorphida*

*Meiopriapulas sp.* was also sourced by Jakob Vinther of the University of Bristol's School of Biological Sciences and sequenced in-house.

### 2.2.1.8 *Kinorhynch* (*sp.*)

Figure 2.4 H

*Phylum Kinorhyncha / Orders Cyclorhagida or Homalorhagida*

The kinorhynch was supplied by James Flaming of the University of Bristol paleobiology group. Extraction and sequencing of this sample was done in-house.

### 2.2.1.9 *Hypsibius dujardini*

Figure 2.4 I

*Phylum Tardigrada / Class Eutardigrada / Order Parachaela*

The assembled genome of *Hypsibius dujardini* was downloaded from its project page: [http://badger.bio.ed.ac.uk/H\\_dujardini/](http://badger.bio.ed.ac.uk/H_dujardini/). The genome has been screened through quality control and the genetic library within has been predicated by (Koutsovoulos *et al.* 2016) to a high standard. This meant that the *H. dujardini* genome did not need to pass through further assembly or quality control steps.

# A



**Figure 2.2.1 A: Pycnogonid**

- Common name: sea spider
- Non-terrestrial chelicerate
- Grow in size in deeper waters
- 1mm – 90cm

<http://www.discoverlife.org/z/q/?search=pycnogonida>

# B



**Figure 2.2.1 B: Opilione**

- Common name: harvestman / daddy long legs
- Leg span comparatively long
- 4mm – 4cm

[http://www.burkeseum.org/sites/default/files/styles/adaptive/adaptive-image/public/28-rupestre\\_1.jpg?itok=0pmedY1f](http://www.burkeseum.org/sites/default/files/styles/adaptive/adaptive-image/public/28-rupestre_1.jpg?itok=0pmedY1f)

# C



**Figure 2.2.1 C: Limulus**

- Common name: horseshoe crab
- Non-terrestrial chelicerate
- Crustacean-like appearance, body protected by a carapace
- 7cm – 60cm

<http://previews.123r.com/images/tonobalaguer005/tonobalaguer00500248/6987186-atlantic-horseshoe-crab.jpeg>

# D

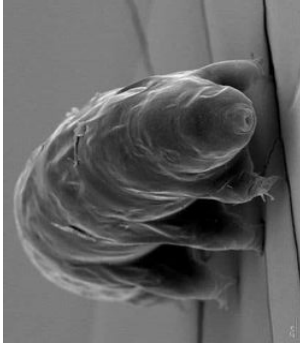


**Figure 2.2.1 D: Oniscidea**

- Common name: woodlouse
- 7 pairs of legs
- Large thorax consisting of 7 tough over-lapping plates
- 1cm – 3cm

[http://www.annalbase.uni-goettingen.de/animalbaseimage/Armadillidium-vulgare\\_01.jpg](http://www.annalbase.uni-goettingen.de/animalbaseimage/Armadillidium-vulgare_01.jpg)

# I



**Figure 2.2.1 I: Tardigrade**

- Common name: water bear
- *Hypsibius dujardini*, along with three other tardigrades, are the focus of this study
- Eight legged, segmented body,
- Microscopic – 0.5mm
- Extreme ecological adaptability

[http://badger.bio.ed.ac.uk/H\\_dujardini/images/H\\_dujardini.jpg](http://badger.bio.ed.ac.uk/H_dujardini/images/H_dujardini.jpg)  
[http://media.eol.org/content/2013/08/23/2361018\\_380\\_360.jpg](http://media.eol.org/content/2013/08/23/2361018_380_360.jpg)

# E



**Figure 2.2.1 E: Onychophoran**

- Common name: velvet worm
- Varying amount of appendages depending on species
- Hydrostatic skeleton
- 5mm – 20cm

[http://www.onychophora.com/images/fig\\_2.png](http://www.onychophora.com/images/fig_2.png)

# F

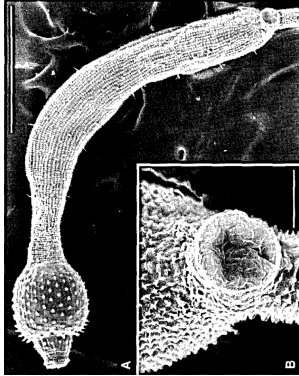


**Figure 2.2.1 F: Haliertyptus**

- Common name: penis worm
- Worm-like appearance
- Pharynx is lined with teeth
- 2mm – 39cm

<http://cdn.e.photoshelter.com/img-get2/10000kPhleKOY2yQ?fit=1000x750/DSC7618.jpg>

# G



**Figure 2.2.1 G: Meiopriapulas**

- Common name: penis worm
- Nervous system circumnavigates the pharynx
- 2mm – 39cm

(Todaro & Shirley 2003)

# H



**Figure 2.2.1 H: Kinorhynch**

- Common name: mud dragon
- Limbless, spiny body
- Less than 1mm in length
- Can survive at great ocean depth down to 8 kilometers.

<https://s-media-cache-ak0.pinimg.com/736x/67/a7/bd/67a7bd7855d2eb20e1a03e771b1b10.jpg>

## Figure 2.4: Sequenced Taxa [A - H] and Focus of this Study [I]

A visual representation of ecdysozoans sequenced for this study.

An image of the tardigrade *H. dujardini* is also included as the tardigrades are the focus of this study.

Sources for each image are provided.

### 2.2.1.10 Taxa Downloaded from the SRA

The Sequence Read Archive (SRA) is the largest online resource for NGS data (Leinonen *et al.* 2011) It is rapidly becoming the leading resource for phylogenomic studies and therefore was an ideal candidate for adding newly sequenced taxa to an established dataset of orthologs (Philippe *et al.* 2011b). Seven raw transcriptomes were downloaded from the SRA. These data provided a decent sampling of ecdysozoans with a focus on the Tardigrada (*Echiniscus testudo* and *Milnesium tardigradum*) but also covering the previously poorly sampled Myriapoda (*Scutigera coleoptrata*, *Glomeridesmus sp.*, *Polydesmus angustus*, and *Symphylella vulgaris*) and Crustacea (*Speleonectes tulumensis*). See [Table 2.1](#) for full details on these taxa.

**Table 2.1: Tardigrade Study: Transcriptomes Downloaded from the SRA**

The seven taxa downloaded from the SRA for the tardigrade study.

The SRA number and taxonomic information for all transcriptomes are provided.

**Table 2.1: Tardigrade Study Transcriptomes Downloaded From The SRA**

	Transcriptome	SRA Number	Phylum	Subphylum	Class	Order
1	<i>Speleonectes tulumensis</i>	SRR857228	Arthropoda	Crustacea	Remipedia	Nectipoda
2	<i>Scutigera coleoptrata</i>	SRR1158078	Arthropoda	Myriapoda	Chilopoda	Scutigeromorpha
3	<i>Glomeridesmus</i>	SRR941771	Arthropoda	Myriapoda	Diplopoda	Glomeridesmida
4	<i>Polydesmus angustus</i>	SRR1047642	Arthropoda	Myriapoda	Diplopoda	Polydesmida
5	<i>Symphylella vulgaris</i>	SRR768329	Arthropoda	Myriapoda	Symphyla	Symphylemida
6	<i>Echiniscus testudo</i>	SRR1141094	Tardigrada	-	Heterotardigrada	Echiniscoidea
7	<i>Milnesium tardigradum</i>	SRR057381	Tardigrada	-	Eutardigrada	Apochela

### 2.2.2 DNA & RNA Extractions

DNA and RNA extractions of the pycnogonids & opiliones were carried out by Eoin Mulvihill of NUIMs Nematode Genetics laboratory and by myself. Eoin had a large amount of experience with invertebrate DNA & RNA extractions and so was the perfect candidate for directing our efforts. Omar Rota-Stabelli carried out RNA extractions of the oniscidea and the onychophoran. RNA extraction of the horseshoe crabs, kinorhynch and priapulids were conducted at the University of Bristol. The following methodologies also apply to the molecular libraries used in this study that were sourced from other NGS projects. The *H. dujardini* data is from a genomic library and all data downloaded from the SRA are of transcriptomic origin. The full protocols for both DNA and RNA extractions can be found in [Appendices 2.2.2](#).

#### 2.2.2.1 DNA Concentration and Purity Analysis

DNA concentration levels were identified by measuring the absorbance of the solution at 260nm using the Nanodrop [[Supplementary Material 2.1](#)]. Qiagen extraction protocols state that any sample returning values falling within the range of 0.1 - 1.0 absorbance contain adequate concentration for sequencing.

The purity of the extracted DNA sample was determined by calculating the ratio of absorbance at 260nm to the absorbance at 280nm;  $A_{260}/A_{280}$  for protein contaminants and  $A_{260}/A_{230}$  for phenol and organic contaminants.

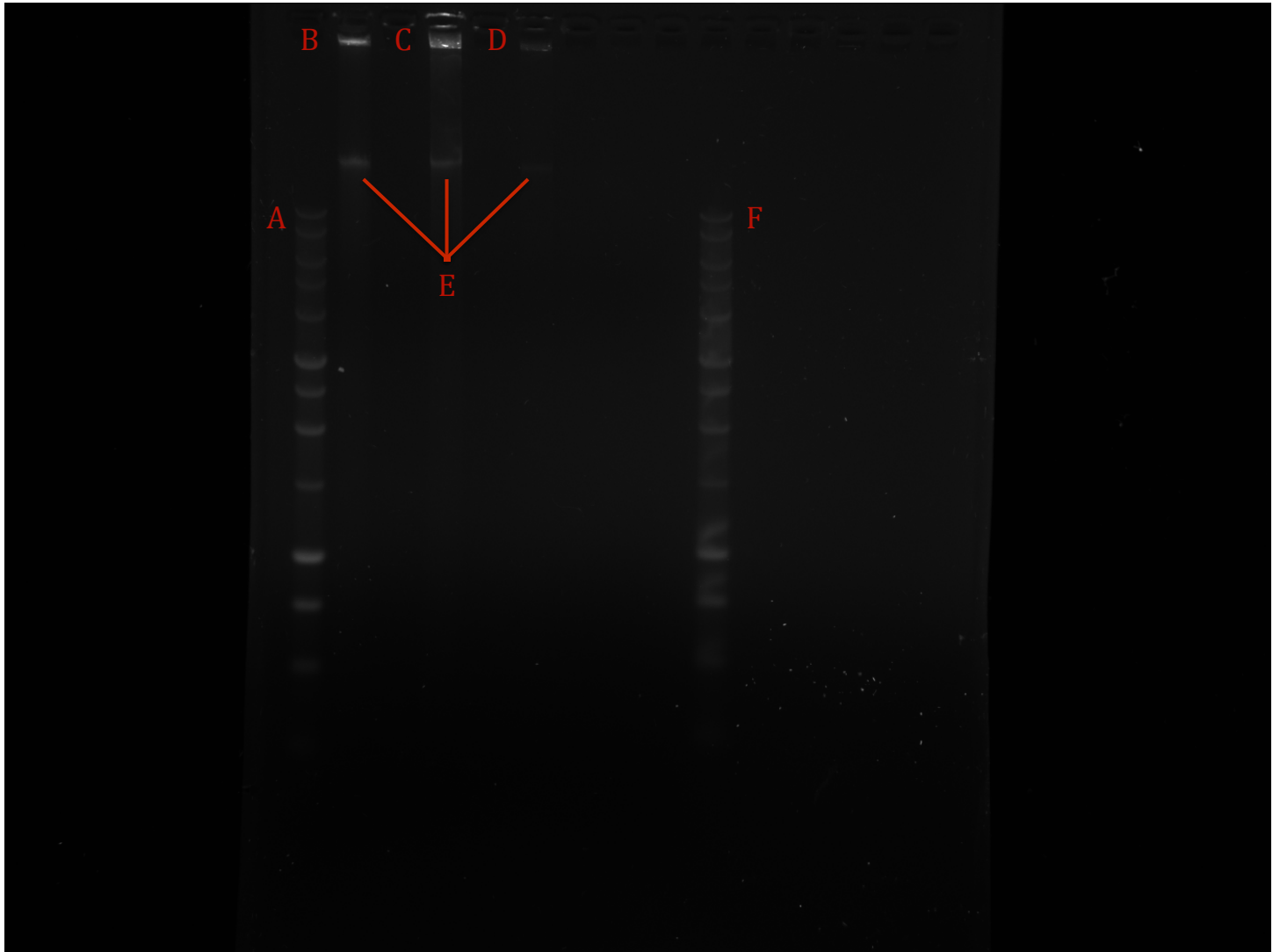
Pure DNA has an  $A_{260}/A_{280}$  ratio of 1.7 - 1.9 and an  $A_{260}/A_{230}$  ratio of 2.0 - 2.2 (<https://www.qiagen.com/ie/resources/resourcedetail?id=97640bc9-e4fe-4c4b-83f6-ac7ca4181597&lang=en>).

#### 2.2.2.2 DNA Integrity Analysis

DNA integrity was rated by gel electrophoresis. See [Figure 2.5](#) for the *Opilione sp.* sample and [Supplementary Material 2.1](#) for full results. Gel electrophoresis is a method of visualizing and separating gDNA in a solution, making it possible to gauge its integrity as it separates DNA based on its length (usually in kbp). Therefore it is a good indication of checking if gDNA is entirely intact or broken into parts.

After migration, isolated bands on the gel represent DNA strands of a single length. The brightness of a band is a loose indication of concentration (better quantified by nanodrop), and long blurry bands are representative of DNA of many sizes, in this case an indication of genomic DNA degradation.

DNA lengths are gauged with the use of ladders of known molecular length placed in either ends of the gel band. An ideal result of gDNA extraction is a single bright band at the end of the gel representing a single mass of intact gDNA.



**Figure 2.5: Gel Electrophoresis of the *Opilione sp.* gDNA**

**A:** molecular ladder of known masses. **B:** 100ng *Opilione sp.* gDNA. **C:** 200ng *Opilione sp.* gDNA. **D:** 50ng *Opilione sp.* gDNA. **E:** *Opilione sp.* mitochondrial DNA. **F:** molecular ladder of known masses. The integrity of the opilione gDNA was analysed at three different concentrations (50ng, 100ng, and 200ng). Clear isolated bands in these lanes indicate successful extractions of gDNA. There are no signs of RNA contamination as this would appear as bands further down the gel representing the smaller 18S and 24S RNA subunits.

### 2.2.2.3 RNA Extractions RNA Concentration and Purity Analysis

Similarly to the DNA prep, RNA concentration and purity was rated using the nanodrop, the only minor difference being that the absorbance value for protein contaminant free RNA ( $A_{260}/A_{280}$ ) is 2.0 as opposed to 1.7 - 1.9 for DNA [Supplementary Material 2.1].

### 2.2.2.4 RNA Integrity Analysis

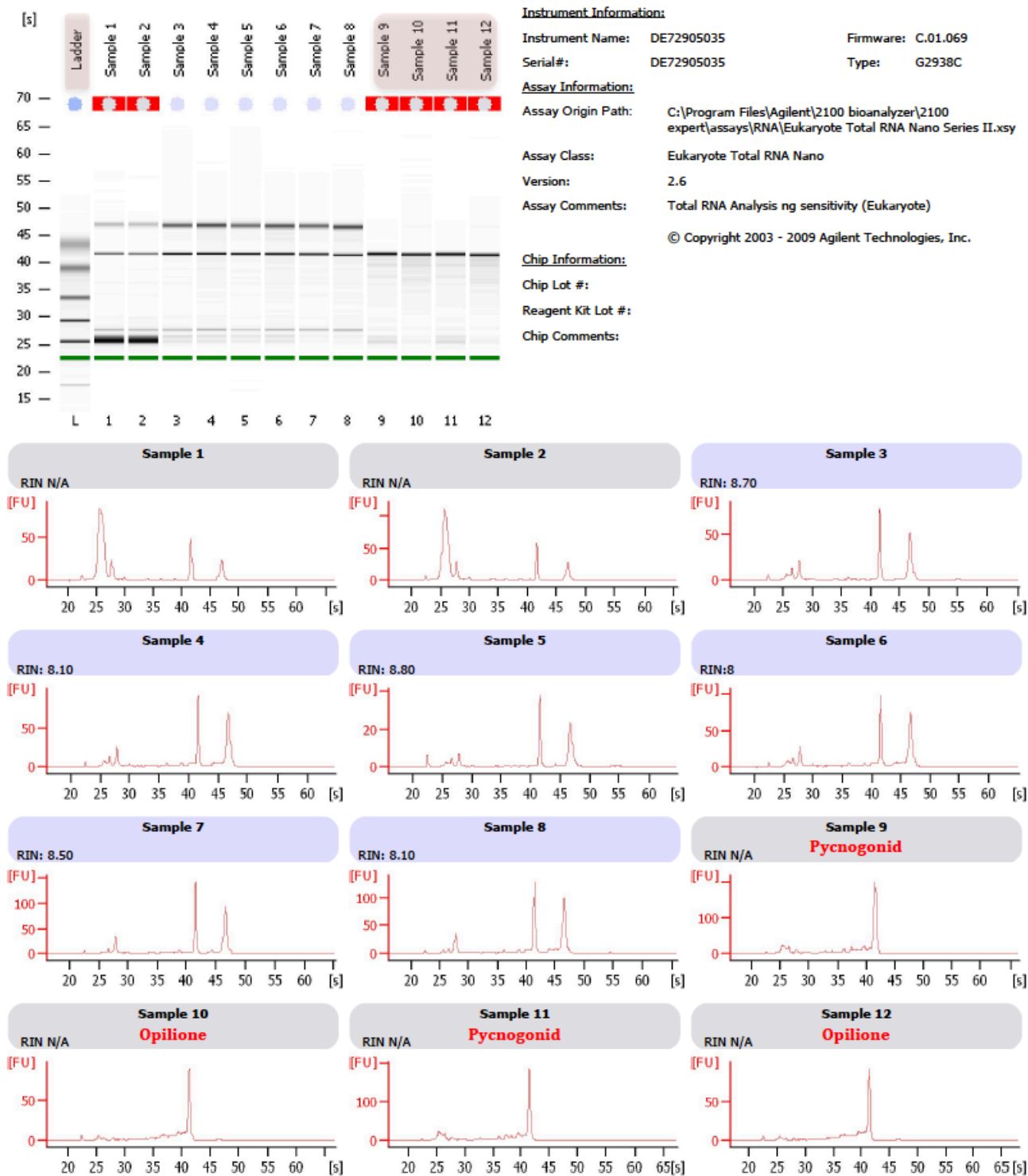
RNA integrity checks were carried out using a bioanalyzer. The bioanalyzer works as an automated form of electrophoresis, separating the RNA in the sample based on size. Typically, for quality integrity samples, we see two bands of RNA in large concentrations: 18S and 28S ribosomal RNA, as they are the largest RNA in the cells of eukaryotes and thus a good indication of integrity (Schroeder *et al.* 2006). These are represented as spikes on a graph in the bioanalyzers output [Figure 2.6 A]. A spike at only one or at multiple bands is often a sign of RNA degradation, as it would point to the likelihood that one of the major subunits had broken into smaller fragments.

However a study on insect RNA gel electrophoresis by Winnebeck *et al.* (2009) showed that the use of heat denaturing electrophoresis (such as the one adopted in these methods) on insect specimens results in a single peak (representing only one of the ribosomal subunits) for integral RNA instead of two. This is caused by the unusual rapid cleavage of 28S RNA near the center of the molecule under denaturing conditions (40-60<sup>0</sup>C) into fragments similar in size to the 18S subunit.

Although the study focused on insects, we propose that a similar phenomenon occurred with the chelicerate and crustacean samples as they displayed the same

results from the bioanalyzer (see [Figure 2.6 B](#) for the pycnogonid and opilione samples displaying this phenomenon) and passed further integrity QC tests at the Edinburgh Genomics sequencing center. This phenomenon may be found in the Arthropoda as a whole as opposed to just the Insecta. For full bioanalyzer results see [Supplementary Material 2.1](#).

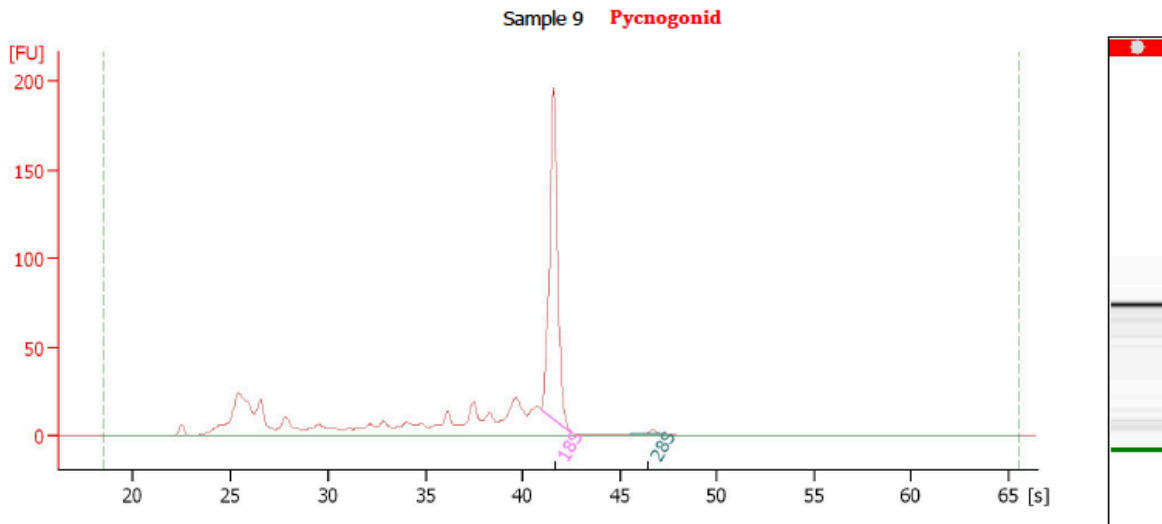




**Figure 2.6: Bioanalyzer Readings**

**Figure 2.6 A: Bioanalyzer Readings for Pycnogonid and Opilione RNA Extractions**

Both samples were split into two, with the pycnogonid RNA in lanes 9 & 11 and opilione RNA in lanes 10 & 12. Lane 1 consists of a ladder of differing known molecular masses, allowing the identification of the 24S and 18S RNA positions on the gel for the extracted RNA samples. Lanes 2-8 consist of RNA samples unrelated to this project. The lanes containing pycnogonid and opilione RNA display only a single band on the gel and peak on the graph.

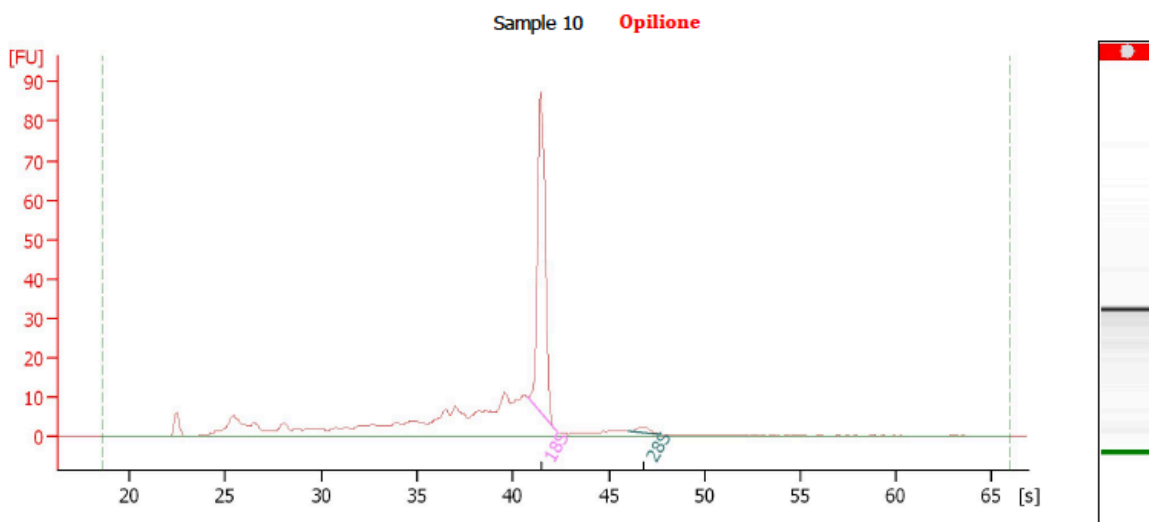


**Overall Results for sample 9 : Sample 9**

RNA Area:	712.2	RNA Integrity Number (RIN):	N/A (B.02.07)
RNA Concentration:	548 ng/ $\mu$ l	Result Flagging Color:	<span style="background-color: #cccccc; border: 1px solid black; display: inline-block; width: 20px; height: 10px;"></span>
rRNA Ratio [28s / 18s]:	0.0	Result Flagging Label:	RIN N/A

**Fragment table for sample 9 : Sample 9**

Name	Start Time [s]	End Time [s]	Area	% of total Area
18S	40.92	42.48	221.9	31.2
28S	45.52	47.42	2.9	0.4



**Overall Results for sample 10 : Sample 10**

RNA Area:	303.6	RNA Integrity Number (RIN):	N/A (B.02.07)
RNA Concentration:	234 ng/ $\mu$ l	Result Flagging Color:	<span style="background-color: #cccccc; border: 1px solid black; display: inline-block; width: 20px; height: 10px;"></span>
rRNA Ratio [28s / 18s]:	0.0	Result Flagging Label:	RIN N/A

**Fragment table for sample 10 : Sample 10**

Name	Start Time [s]	End Time [s]	Area	% of total Area
18S	40.80	42.23	87.2	28.7
28S	46.05	47.68	2.3	0.8

**Figure 2.6 B: Pycnogonid and Opilione Bioanalyzer Peaks**

The 18S peak is clearly visible for both pycnogonid and opilione samples. However both are missing the 28S peak. This usually indicates that the integrity of the samples has been compromised but the samples passed further RNA integrity tests at the sequencing center.

### 2.2.3 Genome and Transcriptome Sequencing

All gDNA and RNA samples were sequenced by Illumina SOLEXA at Edinburgh Genomics (<https://genomics.ed.ac.uk/services/sequencing>). Illumina technology is a form of NGS that creates much shorter reads than the previous Sanger methods. Genomic DNA samples were sequenced to a read length of 100 bp (pair end) and the RNA samples were sequenced to a read length of 80 and 60 (pair end).

The extracted DNA was broken into single stranded pieces of a specified length known as reads. These reads were attached via adapters to a flowcell, a unique surface that facilitates the attachment of millions of different reads on their own unique spot. This setup is the key to the speed of genome sequencing as it allows millions of reads to be sequenced at once and separates NGS from the more antiquated Sanger method. The reads were amplified, after which the original reads are washed away, leaving the cloned products.

Illumina's novel method of sequencing by synthesis followed. Chemically treated nucleotides were then added one cycle at a time. Each of the four nucleotides were altered to contain a unique fluorescently labeled reversible terminator that emitted a frequency (colour) when it hybridized to its complementary nucleotide fixed to the flowcell spot. These altered nucleotides have a secondary function in preventing more than one base addition per cycle. The frequency emitted from each spot was recorded by the machine for each cycle, leaving a digital recording of the composition of each read.

The SOLEXA machine measured the frequency emitted from each spot per cycle and called the appropriate base. A confidence score was assigned to each read which states how confident the machine is that the correct base was called based on the

intensity of the signal. A more in-depth description of NGS by Illumina SOLEXA can be found in [Supplementary Material 1.1](#).

#### 2.2.4 Data Quality Control

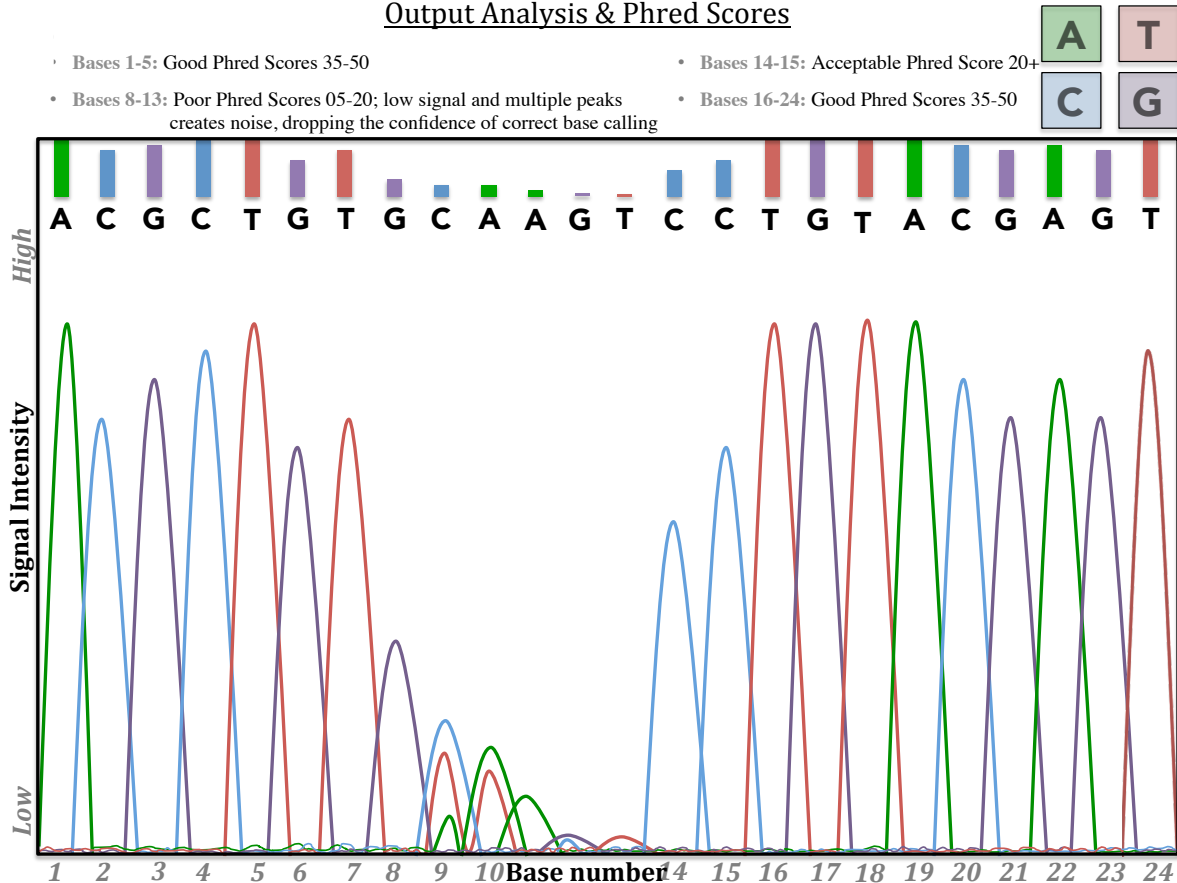
Once the data was retrieved from the sequencing center it was important to check the quality of the gDNA and RNA sequences. FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) was used to gauge the quality of our newly sequenced taxa in addition to all libraries downloaded from the SRA.

FastQC checks the raw sequence data and rates its quality based on a number of metrics such as Phred score, per base content, per sequence GC content, N content, sequence length distribution, adapter content and sequence duplication levels.

The *Phred score* is arguably the most important metric as it is a direct rating of the accuracy of the sequencing (Ewing *et al.* 1998). It rates the quality of the data based on a 0 - 50 scale. It does this by measuring the peak (the specific frequency of which corresponds to a particular base; A, C, G, T/U - the flurophone excitation of dNTPs) of each base called on the chromatogram and assigns a quality score to it that represents the probability that the base was called correctly [[Figure 2.7](#)].

For an full quality control report of the taxa sequenced for this study and the other chapters including data downloaded from the SRA see [Supplementary Material 2.2](#).

## Output Analysis & Phred Scores



**Figure 2.7: Chromatogram Visualization and Corresponding Phred Scores**

An illustration of the output from a chromatogram. The X-axis represents individual base calls and the Y-axis the signal intensity. Each base is identified by a unique frequency (colour), the peak of which is a measure of how clear that signal is. The higher the peak for a base the higher the Phred score and thus the more confident one can be in the correct base being called. Phred scores are encoded into the sequence headers for each read in the raw sequence data.

Note that chromatograms are not a feature of NGS machines, its inclusion is for illustrative purposes in describing Phred scores.

A Phred score of 50 indicates that the probability of the base having been called incorrectly is 1:100,000 (99.999% accuracy), of 40 indicates that the probability of the base having been called incorrectly is 1:10,000 (99.99% accuracy), and of 30 indicates that the probability of the base having been called incorrectly is 1:1,000 (99.9% accuracy) and so on (<http://www.phrap.com/phred/#qualityscores>). Generally, a mean Phred score across the entire library above 30 for raw sequence data is considered good quality, above 40 is excellent quality, and close to 50 is exceptional.

FastQC looks at quality scores per base for each of the reads and the mean quality of the entire library (*per sequence quality score*).

*Per base sequence content* measures the distribution of the bases A, C, G, and T/U across all base positions. An uneven distribution of any of these bases (usually seen as a spike in the percentage base in question for a particular base position) suggests a bias towards them in the sequencing step. Such an error would mean that the libraries generated would be compromised and the sequencer would require calibration.

*Per sequence GC content* of a library displays the distribution of GC across all the data. Since the GC percentage of an organism does not fluctuate, one expects a normal distribution of GC. This is an important check in data QC as a non-normal distribution of GC content suggests another organism has contaminated the library, representing an additional GC percentage distribution.

*Per base N content* recounts the number of uncalled bases in the library, the larger the number of N's in a library the more unmatched bases and the poorer the assembly.

The *sequence length distribution* simply measures the length of the reads in the library. Most sequencing projects specify constant read lengths for sequences, in our case 60bp and 80bp for transcriptomes and 100bp for genomes. Sequence length distribution can highlight any reads of unexpected length, signaling a potential fault with the sequencer.

The *Adapter content* test checks to see if any adapters in the sequencing process were left in the library. High adapter content in a sequence library can bias the assembly process as they can be mistaken for genuine parts of the sequences, creating false paths in the DeBruijn graphs. Adapters can be trimmed using software such as Ea-Utills (<https://code.google.com/archive/p/ea-utils/>).

*Sequence duplication levels* can be an issue or a sign of quality depending on the library. For RNA samples, large numbers of the same transcripts for genes that are highly expressed in the specimen prior to extraction are expected. The nature of transcriptomics means that many genes will be overly represented in a sample and some not at all based on gene expression levels. Subsequently, duplication levels in an RNA library are to be expected. For genomic libraries one ideally wants an equal representation of every gene in the genome, so high duplication levels in a genomic library may be representative of a sequencing bias.

After all newly sequenced data was passed for quality control they were input into an assembly pipeline.

### **2.2.5 Transcriptome Assembly and Translation**

Transcriptome assembly was achieved through the Trinity package (Grabherr *et al.* 2011) and peptides were predicted from the transcripts using TransDecoder (Haas *et al.* 2013). The commonly used gauge for assembly contiguity is the N50 statistic. The N50 value for an assembled transcriptome states that half of the assembled transcripts are the length of the given value, thus the higher the N50 statistic the more contiguous the assembly (Miller *et al.* 2010). For pair-end raw sequence files (where both strands have been replicated) the following command was used:

```
$ Trinity -seqType fq --JM 50G -single [raw_sequence_file] --run_as_paired --CPU 5
```

Single-end sequence files were assembled using the command below:

```
$ Trinity --seqType fq --JM 50G --single [raw_sequence_file] --CPU 5
```

The N50 statistics for each assembled transcriptome were generated through the *TrinityStats.pl* script, part of the Trinity package (Grabherr *et al.* 2011).

```
$ TrinityStats.pl [assembled_transcriptome]
```

Finally proteins were predicted from the assembled transcripts using Transdecoder (Haas *et al.* 2013). To further ensure genuine proteins were being predicted from the assembled transcripts, all putative proteins were cross-referenced with the Uniref90 (The Uniprot Consortium, 2008) and Pfam (Bateman *et al.* 2004) protein databases with only putative proteins scoring significant matches to these databases being kept.

```
$ blastp -query [assembled_transcriptome] -db uniprot_sprot.fasta -max_target_seqs
```

```
1 -outfmt 6 -evaluate 1e-10 -num_threads 10 > uniref90.out
```

```
$ hmmscan --cpu 5 --domtblout pfam.domtblout Pfam-A.hmm
```

```
[assembled_transcriptome]
```

```
$ Transdecoder.Predict -t [assembled_transcriptome] --retain_blastp_hits
```

```
uniref90.out --retain_pfam_hits pfam.domtblout
```

Assembly statistics for all taxa in this study including total number of transcripts, proteins identified from these transcripts, and N50 stats can be found in [Table 2.2](#).

For information as to which molecular libraries were pair-end or single-end see [Supplementary Material 2.2](#).



**Table 2.2: Tardigrada Study: Assembled & Translated Transcripts**

All newly sequenced taxa used in this study. The phred score, number of transcripts, number of proteins, and N50 statistics are provided. The transcripts and N50 assembly statistics for the kinorhynch were not available.

**Table 2.2: Tardigrada Study Assembled & Translated Transcripts**

<b>Sequenced Libraries</b>					
Taxa	Source	Phred Score	Transcripts	N50 Statistics	Proteins
Pycnogonid	in-house	39	87,838	1,765	26,668
Opilione	in-house	38	134,694	1,709	30,942
Limulus	in-house	37	117,946	1,181	30,282
Oniscidea	in-house	39	6,906	363	1,677
Onychophora	in-house	39	55,375	799	17,269
Kinorhynch	in-house	37	N/A	N/A	3,961
Halicryptus	in-house	37	64,406	1,896	29,057
Meiopriapulas	in-house	37	111,893	1,522	39,254
<b>Libraries Downloaded from the SRA</b>					
Taxa	Source	Phred Score	Transcripts	N50 Statistics	Proteins
Speleonectes	SRR857228	15	2,850	774	970
Scutigera	SRR1158078	37	228,504	421	43,674
Glomeridesmus	SRR941771	39	80,196	467	25,952
Polydesmus	SRR1047642	17	13,444	745	5,998
Symphylella	SRR768329	24-31	34,703	524	11,309
Echiniscus	SRR1141094	19	13,221	790	8,282
Milnesium	SRR057381	23-26	28,958	1,242	18,759

### 2.2.6 Ortholog Mapping

Orthologs can be defined as the same genes in different species as a direct result of a shared common ancestor (Sonnhammer & Koonin, 2002). They are a useful marker in estimating phylogenetic relationships and are regarded as the backbone of modern day molecular evolution datasets. The main risks to ortholog addition to datasets are paralogs, gene products of duplication events that are not necessarily representative of a shared common ancestor between studied taxa. As such, inclusion of paralogs in a dataset often results in homoplasy (Fitch, 2000 and Koonin, 2005). A robust and reliable method of ortholog identification in the newly sequenced species and mapping of these orthologs to a dataset is important for the integrity of phylogenomic studies. Such methods are provided herein.

The dataset of choice, to which the orthologs from the newly sequenced species are mapped, should ideally be published in a peer-reviewed journal, have a rich taxon sampling covering the groups of interest, and consist of a generous number of genes. For the purposes of this tardigrade study a thoughtful selection of the ecdysozoan members from the Philippe *et al.* (2011b) dataset in conjunction with supplementation of gene numbers from the Campbell *et al.* (2011) dataset provided a good backbone for this study of the tardigrades.

The protein libraries for each of the newly sequenced taxa from in-house experiments and external sources: *H. dujardini*, *E. testudo*, *M. tardigradum*, *P. littorale*, *Opiliones sp.*, *Limulus sp.*, *Oniscus sp.*, *Onychophoran sp.*, *Halicryptus sp.*, *Meiopriapulas sp.*, *Kinorhynch*, *S. tulumensis*, *S. coleoptrata*, *Glomeridesmus sp.*, *P. angustus*, and *S. vulgaris* were compared to every protein sequence in the newly amalgamated Philippe *et al.* (2011b) plus Campbell *et al.* (2011) dataset using BLASTp with the implementation of a strict E. value ( $1E^{-10}$ ) cut off. This involved comparing hundreds

of query files to a database so the script *create\_BLAST.pl* [[Supplementary Material 2.3](#)] was used to automate the process. A BLAST command was written for each query file in the directory and the output was written in tabulated format.

```
$ perl create_BLAST.pl
>> $ blastp -query [aligned_orthologs_file] -db [species.prot] -out
[aligned_ortholog_file_species.prot] -evaluate 1e-10 -outfmt 6
```

The output from the BLAST was parsed using the script *parse\_HSPs.py* [[Supplementary Material 2.4](#)]. This script searched through the files and identified only the top high scoring pairs (HSPs) for each BLAST and wrote them to a file designated *dataset\_protein-new\_taxa.HSPs*.

```
$ for i in *.ali [species]; do python parse_HSPs.py $i >>$i.HSPs; done
```

Another script, *retrieve\_fasta.py* [[Supplementary Material 2.5](#)], stored these HSPs and searched for them in the new taxa file (translated transcriptome). Whenever the script found a sequence it wrote it out to a file named after its sequence header.

```
$ for i in *.HSPs; do python retrieve_fasta.py [species].prot
```

Each of these sequences was a prospective ortholog from the newly sequenced taxa for one of the proteins in the dataset. Note that at this stage there may not necessarily be a prospective ortholog for every protein in the dataset. These prospective orthologs were aligned to their respective protein sequence using a muscle profile alignment (Edgar 2004). To automate this process the script *make\_muscle.pl* was used [[Supplementary Material 2.6](#)].

```
$ perl make_muscle.pl
>> $ muscle -profile -in1 [aligned_orthologs_file] -in2 [putative_ortholog] -out
[aligned_ortholog+putative_ortholog.fa]
```

At this stage of the process there was on occasion more than one candidate ortholog for a particular protein in the dataset. In such a case a phylogenetic tree was generated using the PhyML package (Guindon *et al.* 2010) and the ortholog was chosen based on its phylogenetic position and branch length. See [Figure 2.8](#) for some hypothetical scenarios illustrated to explain this process. This is also an important step in the process for when there is only a single candidate as we can check and remove any new ortholog that forms a long branch in the dataset. The benchmark of three standard deviations was usually considered for a prospective ortholog to be removed. It is important to remove these sequences particularly when studying taxa that have been mired with longbranch attraction. In terms of phylogenetic position, any prospective ortholog falling outside the Ecdysozoa was dumped since all molecular studies place them within that phylum. Given the disagreement amongst tardigrade phylogenetic studies, no one hypothesis was given preference over another in terms of prospective ortholog placement which was determined solely by branch length.

The files were converted to phylip format using *Fasta2Phylip.pl* [[Supplementary Material 2.7](#)].

```
$ for i in *.fa; do perl Fasta2Phylip.pl $i; done
```

A task list of PhyML runs was generated using a script adapted from *make\_muscle.pl*.

```
$ make_phyml_tasks.pl
```

```
>> $ phyml -i [aligned_ortholog+putative_ortholog.phy] -d aa -q -s best -b 0
```

Initially trees generated by PhyML were checked manually but with the increasing number of taxa being added to various datasets this became too long and laborious. The process was automated using the *check\_branch\_lengths.pl* [[Supplementary Material 2.8](#)], a script which removes all putative orthologs whose branch length is two standard deviations or more than the mean length of the original dataset

orthologs. It then informs the user which files contain more than one putative ortholog under three SD. From this point the correct ortholog was chosen based on the example given in [Figure 2.8](#).

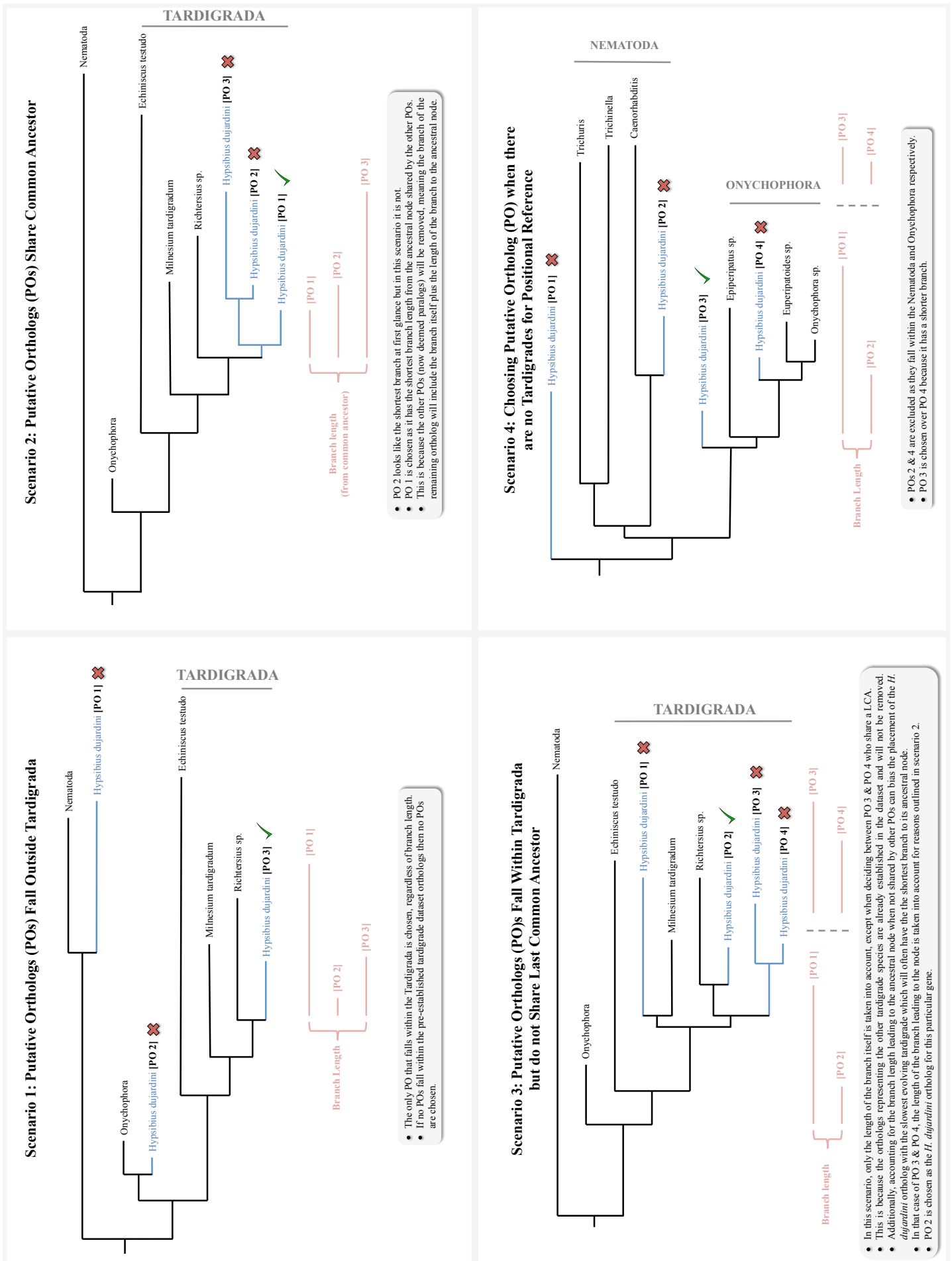
```
$ perl check_branch_lengths.pl
```

After the orthologs from the newly sequenced taxa were mapped to the amalgamated Philippe *et al.* (2011b) & Campbell *et al.* (2011) dataset, the sequences were “cleaned” using Gblocks (Castresana, 2000).

```
$ for i in *.phy; do Gblocks $i -t=p -b2=75 -b3=5 -b4=5 -b5=h >>$i-cleaned.phy
```

Gblocks identifies poorly aligned positions in alignments and removes them improving the quality of the phylogenetic signal within. Stringent settings were applied to the Gblocks cleaning procedure with a 75% conservation rate amongst alignments (-b2=75), a maximum number of contiguous non conserved positions of 5 (-b3=5), the minimum length of a block set to 5 (-b4=5), and gap positions allowed with half (-b5=h). This was important in ensuring a quality superalignment when all ortholog MSAs were concatenated using FASconCAT (Kück & Longo, 2014). Alignments were concatenated with nexus block and relaxed phylip formats.

```
$ FASconCAT_v.10 -n block -p relaxed
```



**Figure 2.8: Paralog Removal from Putative Orthologs**  
 Note that these scenarios are for descriptive purposes and not a common occurrence.

### 2.2.7 Dataset Summary

The tardigrade dataset was built on the backbone of the Philippe *et al.* (2011b) and Campbell *et al.* (2011) datasets which were edited down to mostly ecdysozoan taxa. Using the ortholog identification process outlined above, the tardigrades *Milnesium tardigradum* (SRR057381), *Hypsibius dujardini* ([http://badger.bio.ed.ac.uk/H\\_dujardini/home/download](http://badger.bio.ed.ac.uk/H_dujardini/home/download)), and *Echiniscus testudo* (SRR1141094) were added. To further supplement the Ecdysozoa, the newly sequenced *Meiopriapulas Sp.* (in-house), *Halicryptus Sp.* (in-house), *Kinorynch sp.* (in-house), *Scutigera coleopteran* (SRR1158078), *Polydesmus angustus* (SRR1047642), *Symphylella vulgaris* (SRR768329), *Glomeridesmus sp.* (SRR941771), *Speleonectes tulumensis* (SRR857228), *Oniscidea sp.* (in-house), *Opilione Sp.* (in-house), *Limulus Sp.* (in-house) and *Pycnogonium littorale* (in-house), were also included. Note that “in-house” refers to the specimen being sequenced by either Edinburgh Genomics or the University of Bristol. This generated a dataset of 56 taxa and 41,125 amino acid characters.

### 2.2.8 Phylogenetic Reconstruction

Phylogenetic trees were reconstructed using Phylobayes under the CAT, GTR, and CAT-GTR models (Lartillot *et al.* 2009). Two independent Hidden Markov Chains were run for ~5,000 generations with a burn-in rate of 25% and deemed converged once the maximum difference between the two chains was below 0.2. Below this threshold the posterior probability support for each node of the tree is significant (Lartillot *et al.* 2009). Note: the command line argument to run the CAT-GTR model

is rather obtusely “-GTR-CAT”. This point has been made to prevent confusion when documenting commands further on, particularly in the signal dissection subsections.

```
$ pb -d superalignment_gb75.phy -[model] -s -f superalignment_gb75-[model]-  
chain1
```

```
$ pb -d superalignment_gb75.phy -[model] -s -f superalignment_gb75-[model]-  
chain2
```

```
$ bpcomp -x 25 2 superalignment_gb75-[model]-chain1 superalignment_gb75-  
[model]-chain2 -c 0.1
```

### 2.2.9 Model Testing: Bayesian Cross Validation

Different models of evolution will often be more suitable for certain datasets over others therefore it is necessary to investigate which model fits the data the best. A Bayesian cross validation (BCV) (Lartillot *et al.* 2009) is one such method of choosing the most appropriate model for the dataset in question. BCV operates on the mantra of randomly separating the data into two unequal parts: the learning set and the test set. The model parameters are estimated on the learning set which is used to test the likelihood of the test set. This likelihood score measures how well the test set is predicted by the model in question. The dataset splitting procedure was repeated many times and the log likelihood score for each test set was averaged.

The dataset was jackknifed twice in an effort to limit the burden on computational resources. The seqboot package (Felsenstein, 1991) halved the dataset twice to generate a statistically representative dataset 25% the size of the original.

```
$ seqboot -d superalignment_gb75.phy -j b -% r -r 100 >> 50%_superalignment.phy
```

```
$ seqboot -d 50%_superalignment.phy -j b -% r -r 100 >> 25%_superalignment.phy
```



A number of replicate datasets were generated using the `cvrep` module of Phylobayes (Lartillot *et al.* 2009). These replicates represented the learning and test sets. The number of reps chosen was ten, meaning that the learning sets made up 9/10<sup>th</sup> of the dataset with the remaining 1/10<sup>th</sup> representing the test set. These replicates were run under a ten-fold BCV.

```
$ cvrep -nrep 10 -nfold 10 25%_superalignment.phy cv
```

The CAT, GTR, and CAT-GTR models were all run under these replicate learning sets. An MCMC chain was run under the respective phylogeny generated by each model from **Materials and Methods 2.2.8** with a burn in of 100.

```
$ pb -d dataset_[0-9]_learn.ali -T [models_concensus_tree.tre] -x 1 100  
[model]_dataset_[0-9]_learn.ali
```

The cross-validated log likelihood scores were calculated for each replicate. The likelihood of each test set was averaged over the parameter values estimated by Phylobayes on the corresponding learning sets. The log of the resulting average likelihood was written out for each replicate file.

```
$ readcv -nrep 10 -x 100 [model]_dataset
```

Summary statistics were then used to compare the suitability of the models between one other. In this case the suitability of the CAT and CAT-GTR models are compared to the suitability of the GTR model.

```
$ sumcv -nrep 10 CAT GTR-CAT GTR cvb
```

The model with the highest positive score is the best suited to the data, alternatively if the two compared models both return mean negative scores compared to the reference model then the reference model itself is the most suitable (Posada & Buckley, 2004).

### 2.2.10 Tardigrade Dataset: Signal Dissection

The phylogenetic signal within the dataset was tested under the three main methods of signal dissection: the slow / fast technique (Brinkmann & Philippe, 1999), Dayhoff recoding (Dayhoff *et al.* 1968), and taxon pruning (Aguinaldo *et al.* 1997).

#### 2.2.10.1 Slow / Fast Analysis

The dataset was divided into a series of monophyletic groups [Table 2.3] that were encoded in the end of the nexus format of the file.

PAUP (Swofford, 2002) calculated the substitution rate of each character (individual amino acid) as the sum of the numbers of steps for that corresponding position within its monophyletic group (Brinkmann & Philippe, 1999). All characters were then sorted in to categories of fastest evolving sites, ranging in intervals of 10%. These defined groups of characters were encoded in to the end of the nexus file of the full dataset. PAUP (Swofford, 2002) was used once again to generate sub datasets that only included the fastest percentage of characters of interest or conversely, can exclude these characters to produce slower evolving datasets with the fastest sites being incrementally removed in 10% intervals. The fastest 20%, 30%, and 40% characters were converted into datasets for phylogenetic reconstruction, as were the slowest 80%, 70%, and 60% characters.

Phylobayes (Lartillot *et al.* 2009) constructed phylogenies based on these incrementally rapidly and slowly evolving sub datasets under the best-fitting model for the data (previously tested in **Materials & Methods 2.2.9**).

*\$ pb -d [20-30-40]pc\_fastest.phy -GTR-CAT -s -f [20-30-40]pc\_fastest-CATGTR-  
chain1*

*\$ pb -d [20-30-40]pc\_fastest.phy -GTR-CAT -s -f [20-30-40]pc\_fastest-CATGTR-  
chain2*

*\$ pb -d [80-70-60]pc\_fastest.phy -GTR-CAT -s -f [80-70-60]pc\_fastest-CATGTR-  
chain1*

*\$ pb -d [80-70-60]pc\_fastest.phy -GTR-CAT -s -f [80-70-60]pc\_fastest-CATGTR-  
chain2*

The slow, fast, and original datasets were then compared to see if the most rapidly evolving sites are producing differing results to that of the other datasets.

**Table 2.3: Tardigrada Slow / Fast Dataset**

The dataset was divided into 8 monophyletic groups listed at the top of the table. Character numbers and P-scores are presented for the full dataset and the three fastest and slowest character datasets.

**Table 2.3: Tardigrada Slow / Fast Dataset**

	Original Dataset	Low Signal Datasets			High Signal Datasets		
		20% Fastest	30% Fastest	40% Fastest	80% Slowest	70% Slowest	60% Slowest
<b>Characters</b>	41,125	8,225	12,338	16,450	32,900	28,789	24,675
<b>Rate</b>	0 - 27	27 - 7	27 - 5	27 - 3	0 - 7	0 - 5	0 - 3

**Monophyletic Groups**

Chelicerata - Myriapoda - Mandubilata - Tardigrada - Cycloneurelia - Nematoida - Mollusca - Annelida

### 2.2.10.2 Dayhoff Recoding

The tardigrade dataset was recoded by three different substitution models: Dayhoff-6 recodes {A,G,P,S,T}, {D,E,N,Q}, {H,K,R}, {F,Y,W}, {I,L,M,V}, and {C} into six single characters.

```
$ pb -d superalignment_gb75.py -GTR-CAT -s -f -recode dayhoff6
```

```
superalignment_gb75-CATGTR-dayhoff6-chain1
```

```
$ pb -d superalignment_gb75.py -GTR-CAT -s -f -recode dayhoff6
```

```
superalignment_gb75-CATGTR-dayhoff6-chain2
```

Dayhoff-4 recodes {A,G,P,S,T}, {D,E,N,Q}, {H,K,R}, {F,Y,W,I,L,M,V}, and {C = ?} into four single characters.

```
$ pb -d superalignment_gb75.py -GTR-CAT -s -f -recode dayhoff4
```

```
superalignment_gb75-CATGTR-dayhoff4-chain1
```

```
$ pb -d superalignment_gb75.py -GTR-CAT -s -f -recode dayhoff4
```

```
superalignment_gb75-CATGTR-dayhoff4-chain2
```

Dayhoff-HP recodes {A,C,F,G,I,L,M,V,W}, {D,E,H,K,N,P,Q,R,S,T,Y} into two single characters.

```
$ pb -d superalignment_gb75.py -GTR-CAT -s -f -recode hp superalignment_gb75-
```

```
CATGTR-hp-chain1
```

```
$ pb -d superalignment_gb75.py -GTR-CAT -s -f -recode hp superalignment_gb75-
```

```
CATGTR-hp-chain2
```

The phylogeny of the recoded datasets were run under two parallel chains, with a 25% burn in, using the best fitting model for each dataset (CAT-GTR) until converged.

### 2.2.10.3 Taxon Pruning

The fastest evolving clades in the tardigrade dataset based on branch lengths were the Nematoida followed by the Tardigrada. The sole nemetamorph, *Spinochordedes sp.* was kept in the dataset while all but one of the nematodes were removed: *Trichuris sp.*, *Trichinella sp.*, *Pristionchus sp.*, *Caenorhabditis sp.*, *Brugia sp.*, and *Ascaris sp.*, leaving *Xiphinema sp.*, which has the shortest branch in the group. *E. testudo* was removed from the tardigrade clade as its branch was almost twice as long as the rest of group.

The phylogenetic trees for the taxon-pruned datasets were reconstructed using Phylobayes (Lartillot *et al.* 2009) under the CAT-GTR model.

```
$ pb -d pruned_superalignment_gb75.phy -GTR-CAT -s -f
pruned_superalignment_gb75-chain1
$ pb -d pruned_superalignment_gb75.phy -GTR-CAT -s -f
pruned_superalignment_gb75-chain2
```

### 2.2.11 Divergence Time Estimation

Divergence time estimation was completed using Phylobayes (Lartillot *et al.* 2009). The models of evolution chosen were the auto-correlated CIR and un-correlated Gamma. Chain correlation was measured through the readdiv function. Chains were deemed converged when the difference between them was less than 10MY for every corresponding node in the chronogram. It was essential to constrain divergence time estimations with calibrated bounds using the fossil record. The calibrations were carefully chosen with the aid of the robust fossil calibration recommendations from (Benton *et al.* 2015), these calibrations are displayed in [Table 2.4](#). The clocks were

run under relaxed settings with a soft maximum age of 650 MY for the root (deuterostome-protostome split) and a standard deviation of 20MY.

```
$ pb -d superalignment_gb75.phy -T CATGTR_concensus.tre -r outgroup -cal
calibrations -[model] -rp 650 20 superalignment_gb75-[model]-chain1
```

```
$ pb -d superalignment_gb75.phy -T CATGTR_concensus.tre -r outgroup -cal
calibrations -[model] -rp 650 20 superalignment_gb75-[model]-chain2
```

```
$ readdiv -x 500 superalignment_gb75-[model]-chain-1
```

```
$ readdiv -x 500 superalignment_gb75-[model]-chain-2
```

**Table 2.4: Tardigrade Dataset: Molecular Clock Calibrations**

The 27 calibration points used in the divergence time estimation study sourced from the fossil record and with guidance from (Benton *et al.* 2015).

**Table 2.4: Tardigrade Dataset Molecular Clock Calibrations**

Taxa		Bounds (MYA)		Taxa		Bounds (MYA)	
1	Human - Danio	444.9	- 420.7	15	Daphnia - Anoplodact	636.1	- 514
2	Aplysia - Crassostrea	636.1	- 534	16	Tubifex - Capitella	636.1	- 476.5
3	Human - Gallus	332.9	- 318	17	Mytilus - Loligo	549	- 532
4	Gallus - Taeniopygia	86	- 66	18	Mytilus - Tubifex	636.1	- 549
5	Human - Loxodonta	164.6	- 61.6	19	Epiperipatus - Daphnia	636.1	- 528.82
6	Human - Mus	164.6	- 61.6	20	Priapulid - Daphnia	636.1	- 528.82
7	Human - Xenopus	351	- 337	21	Priapulid - Meiopriapulid	636.1	- 528.82
8	Human - Leucoraja	468.4	- 420.7	22	Anoplodact - Acanthoscur	636.1	- 497
9	Petromyzon - Human	636.1	- 457.5	23	Scutigera - Strigamia	636.1	- 413
10	Ciona - Human	636.1	- 514	24	Scutigera - Glomerides	636.1	- 419
11	Saccogloss - Ptychodera	636.1	- 504.5	25	Rhodnius - Gryllus	414	- 267
12	Strongyloc - Patiria	636.1	- 480	26	Nasonia - Tribolium	414	- 307
13	Strongyloc - Saccogloss	636.1	- 515.5	27	Nasonia - Folsomia	636.1	- 395
14	Daphnia - Gryllus	636.1	- 523				

## 2.3 Results

### 2.3.1 Tardigrade Phylogeny

The 56 taxa 41,125 character phylogenomic tardigrade dataset containing new genomic & transcriptomic information for *H. dujardini*, *M. tardigradum*, *E. testudo*, *Halicryptus*, *Meiopriapulas*, *Kinorynch*, *S. coleopteran*, *P. angustus*, *S. vulgaris*, *Glomeridesmus*, *S. tulemensis*, *Oniscidea Sp.*, and *P. littorale* returned with contrary results depending on the model used.

#### 2.3.1.1 Tardigrade Phylogeny under the CAT Model

The CAT model was run as two independent chains for 4,000 generations before converging with a max\_diff below 0.2 and returned a Tardigrada plus Onychophora phylogeny with a posterior probability (PP) of 0.87 [Figure 2.9]. The Mandibulata (Pancrustacea + Myriapoda) was recovered (PP 1.0), rejecting the Myriochelata (Myriapoda + Chelicerata) and in agreement with (Rota-Stabelli *et al.* 2011; Misof *et al.* 2014; Borner *et al.* 2014). Concerning the Chelicerata, the marine Xiphosura and land-based Acari form an unconventional clade while the pycnogonids were outgroup to the chelicerates as expected but are also sister to the opiliones. The Diplopoda were recovered as a paraphyletic group and the Remipedia were positioned as the oldest pancrustacean.

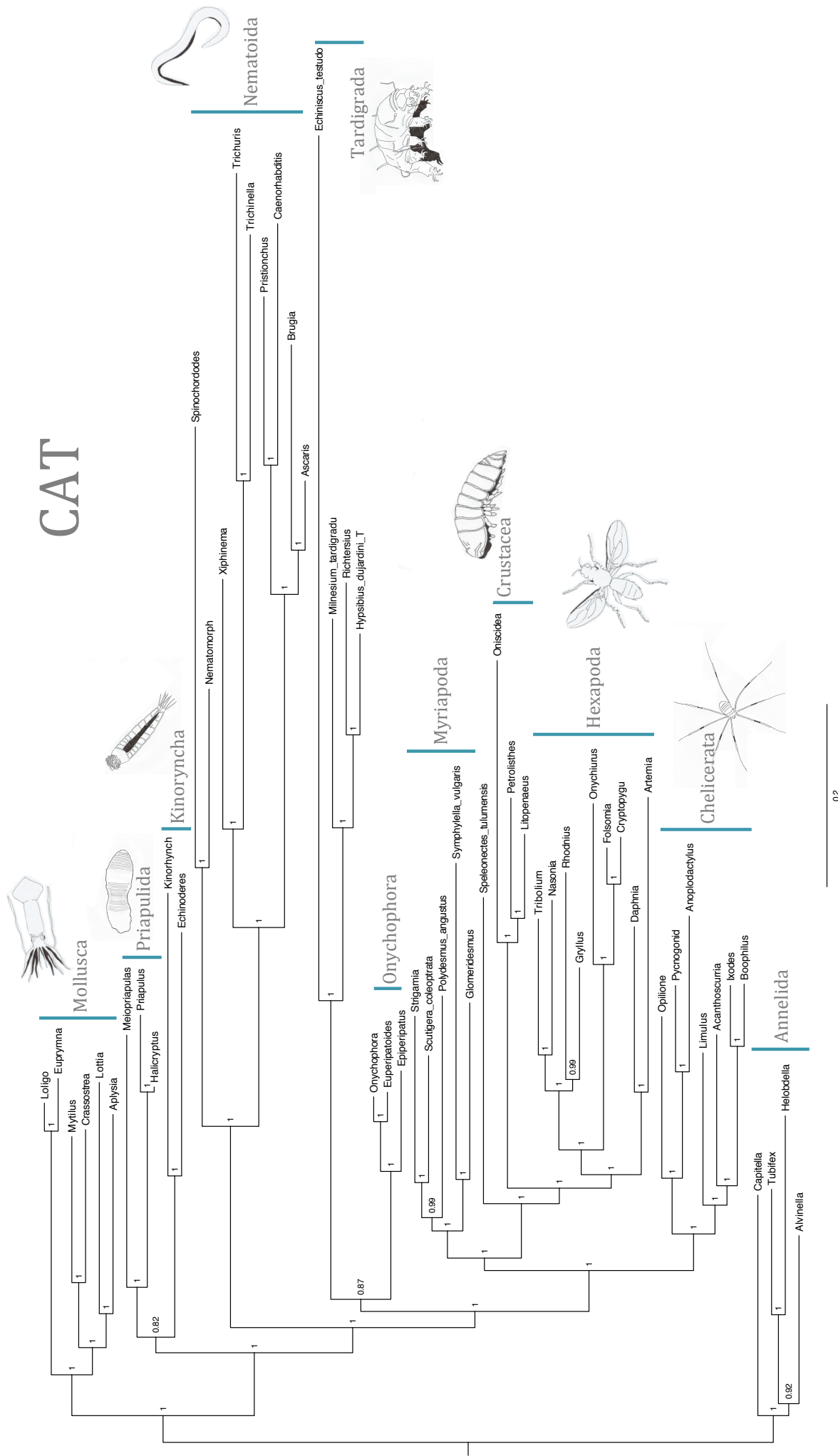
### **2.3.1.2 Tardigrade Phylogeny under the GTR Model**

The GTR model grouped the Tardigrada with the Nematoida with a PP of 0.92 [Figure 2.10]. In stark contrast to the CAT model, the Myriochelata was recovered over the Mandibulata in agreement with (Friedrich & Tautz, 1995; Cook *et al.* 2001; Pisani *et al.* 2004) but with poor support (PP = 0.58). The topology of the Chelicerata mirrored that of the CAT model, but differed in regards to the myriapods with the chilopods and diplopods well distinguished from one another. The Remipedia fall as outgroup to the Hexapoda as part of the well established Pancrustacea.

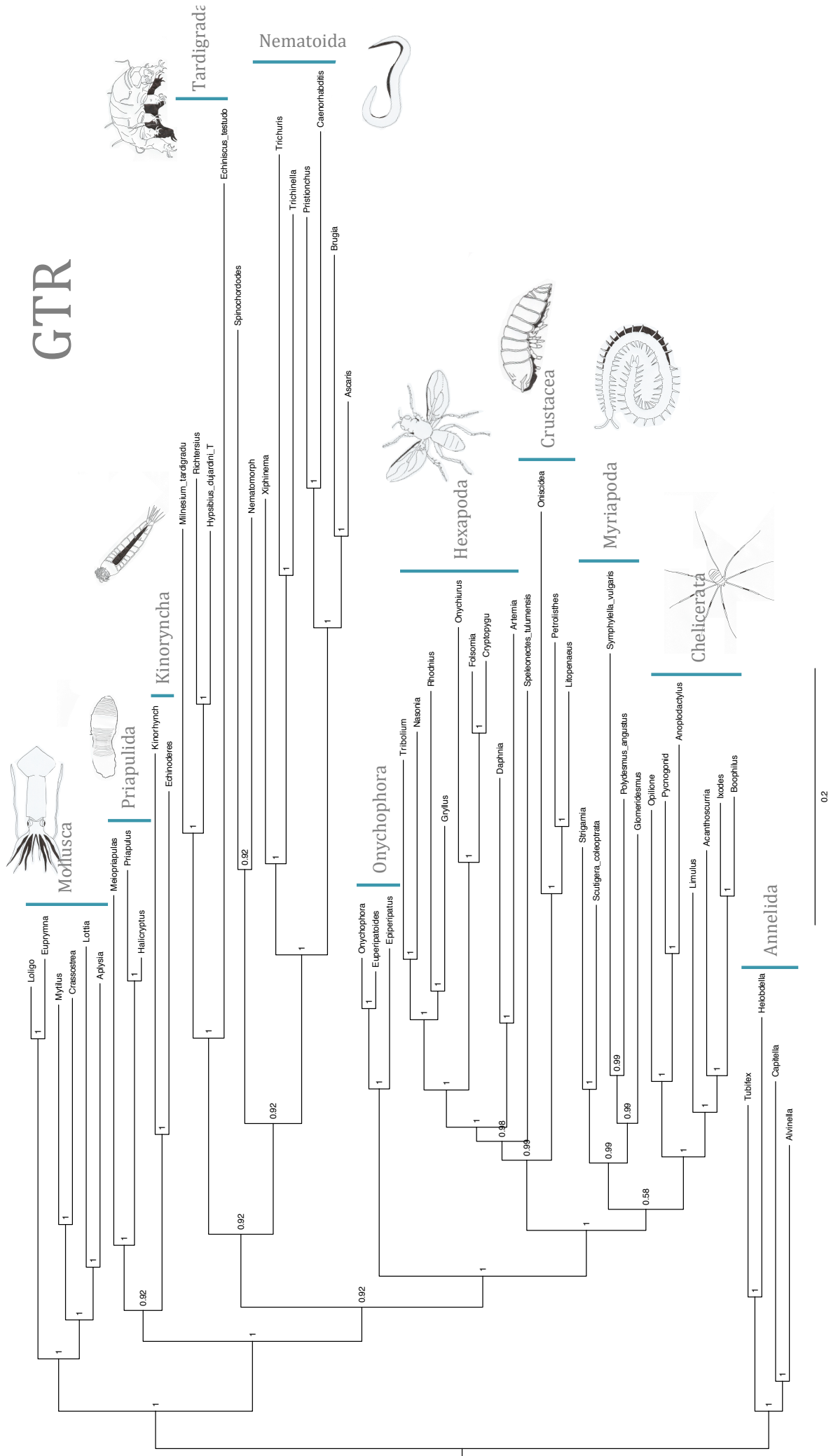
### **2.3.1.3 Tardigrade Phylogeny under the CAT-GTR Model**

The Tardigrada were grouped sister to the Nematoida for the CAT-GTR model [Figure 2.11] with a PP support of 0.98, in agreement with (Lartillot & Philippe, 2008; Meusemann *et al.* 2010; Borner *et al.* 2014). The Mandibulata was again recovered over the Myriochelata. The Chelicerata were found under the same unconventional topology as the other models. There was no clear separation between chilopods and diplopods with the Myriapoda and the Remipedia was recovered outgroup to the Hexapoda.

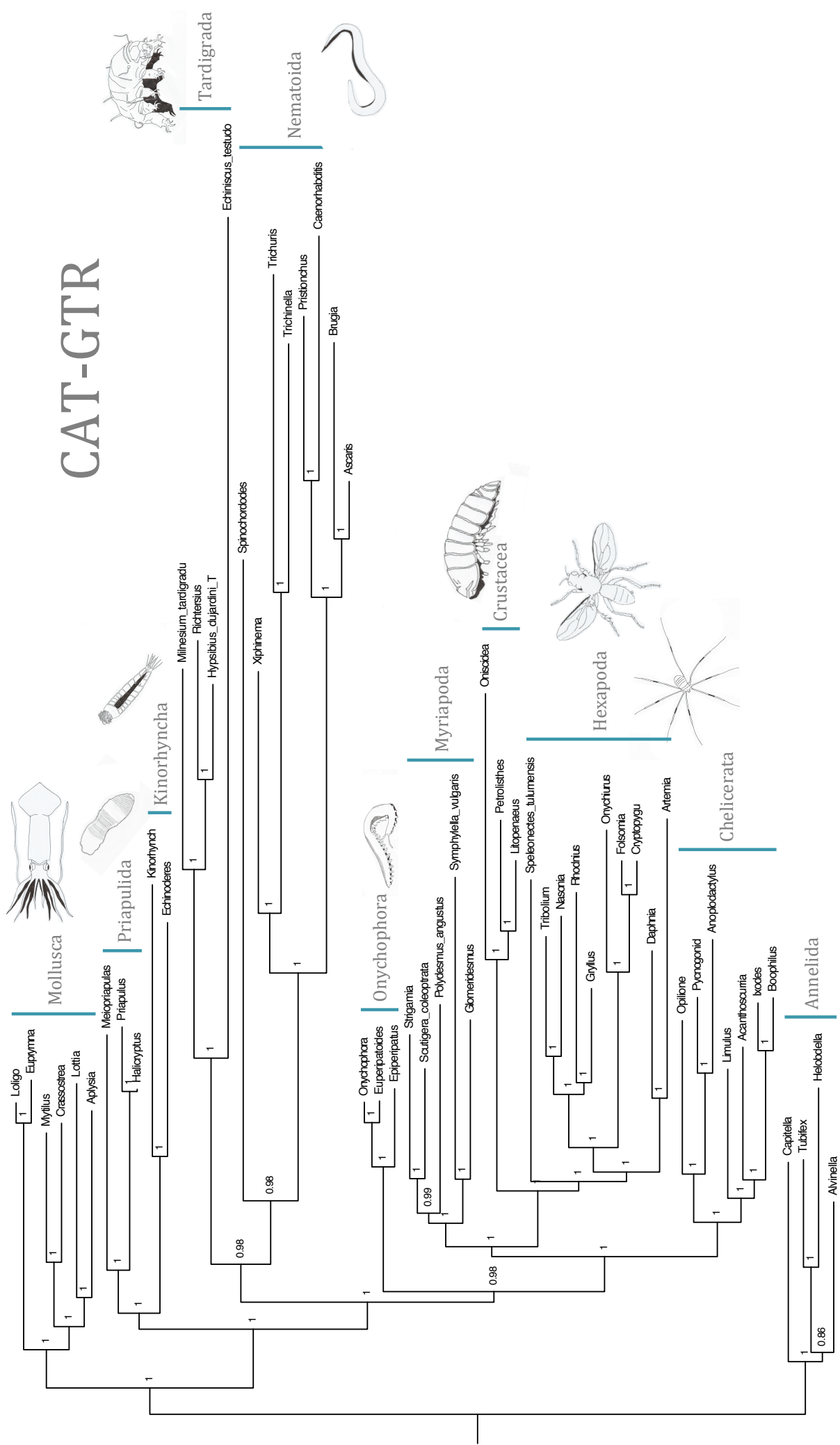




**Figure 2.9: Tardigrade CAT Phylogeny**  
 The Tardigrada are grouped with the Onychophora with a PP of 0.87



**Figure 2.10: Tardigrade GTR Phylogeny**  
 The Tardigrada are grouped with the Nematoida with a PP of 0.92



**Figure 2.11: Tardigrade CAT-GTR Phylogeny**  
 The Tardigrada are grouped with the Nematoida with a PP of 0.98

### 2.3.1.4 Best Fitting Model for the Data

Given the differing phylogenetic results from the models used, it was pertinent to test which model was the most suitable for the dataset via a Bayesian Cross Validation [Table 2.5]. Results suggest that the CAT-GTR is the most suitable; closely followed by the CAT model, with the GTR model being statistically rejected. However, this is with questionably certainty as the standard deviation between the CAT-GTR and CAT scores overlap leaving the lesser possibility that the CAT model could be the best fitting for the data.

Despite the ambiguity between the CAT and CAT-GTR models the GTR model can be rejected.

**Table 2.5: Tardigrade Dataset: BCV**

A Bayesian cross validation of the CAT, GTR, & CAT-GTR models show that the CAT-GTR model is the best fitting model for the data. However, high SD amongst the CAT and CAT-GTR models shows that the CAT model could be equally suitable. The GTR model is rejected.

**Table 2.5: Tardigrade Dataset BCV**

Model	Reference Model	Mean Score	Standard Deviation	Times Model is Most Suitable
CAT-GTR	GTR	949.26	+/- 922.903	4
CAT	GTR	801	+/- 1259.42	6

### 2.3.1.5 Tardigrade Slow / Fast Analysis

The slow / fast technique ranks the characters of the dataset by their rate of evolution and sorts them into categories, or bins, of rate evolution (Brinkmann & Philippe, 1999). Two approaches were taken: analyzing the phylogenetic signal emanating from datasets consisting of the 20%, 30%, and 40% fastest characters and contrarily the datasets for which these fastest sites were stripped away: comprising of the 80%, 70%, and 60% slowest characters. The contrasting positioning of the Tardigrada under these datasets can highlight a LBA influence.


The three datasets made up of the fastest characters all returned the Tardigrada plus Nematoida hypothesis, but with falling levels of posterior probability support: (20% fastest characters = 1.0, 30% fastest characters = 0.67, 40% fastest characters = 0.51) [Supplementary Material 2.9]. It is unclear whether the drop in confidence is due to loss of phylogenetic signal in the larger datasets or because the fastest characters are causing the high level of PP support due to the positively misleading effects of LBA. To clarify, these datasets are all subsets of the full dataset for which the two independent MCMC chains reached full convergence. All datasets were run for the same number of iterations meaning that subsets experiencing difficulty in converging was most likely due to loss of phylogenetic signal in the saturated sites as opposed to not allowing the chains enough time to converge.

The datasets for which the fastest 20 - 40% characters were stripped away recovered the same Tardigrada plus Nematoida grouping with a similar rate of falling support: (60% slowest characters = 0.96, 70% slowest characters = 0.75, 80% slowest characters = 0.5) [Supplementary Material 2.9].

**Table 2.6: Tardigrade Slow / Fast Results**

All datasets return the Tardigrada as a sister group to the Nematoda. Most datasets had convergence issues with the exception being the 20% fastest. All trees were reconstructed under the CAT-GTR model.

**Table 2.6: Tardigrade Slow / Fast Results**

<i>Stable Character Reconstructions</i>				<i>Character Saturation</i> 
<b>Dataset</b>	<b>Tardigrade Position</b>	<b>PP</b>	<b>Convergence</b>	
80% slowest	sister with Nematoda	0.5	1	
70% slowest	sister with Nematoda	0.75	1	
60% slowest	sister with Nematoda	0.96	1	
<i>Saturated Character Reconstructions</i>				
<b>Dataset</b>	<b>Tardigrade Position</b>	<b>PP</b>	<b>Convergence</b>	
40% fastest	sister with Nematoda	0.51	1	
30% fastest	sister with Nematoda	0.67	0.57	
20% fastest	sister with Nematoda	1	0.26	

### 2.3.1.6 Dayhoff Recoding

The three Dayhoff recoding models (Dayhoff 4, 6, & HP) were run under CAT-GTR, the two independent chains were deemed converged when they reached a max\_diff below 0.2 after roughly 1,500 generations. A tardigrade plus onychophoran grouping was recovered under all recoding strategies with very strong support (PP = 0.98) [Figure 2.12]. The models returned broadly similar results with minor discrepancies regarding the positioning of *S. tulumensis* and some chelicerates. The Dayhoff recoded dataset mirrored the phylogeny of the CAT-GTR model with the only exception being the recovering of Tardigrada + Onychophora over Tardigrada + Nematoda.

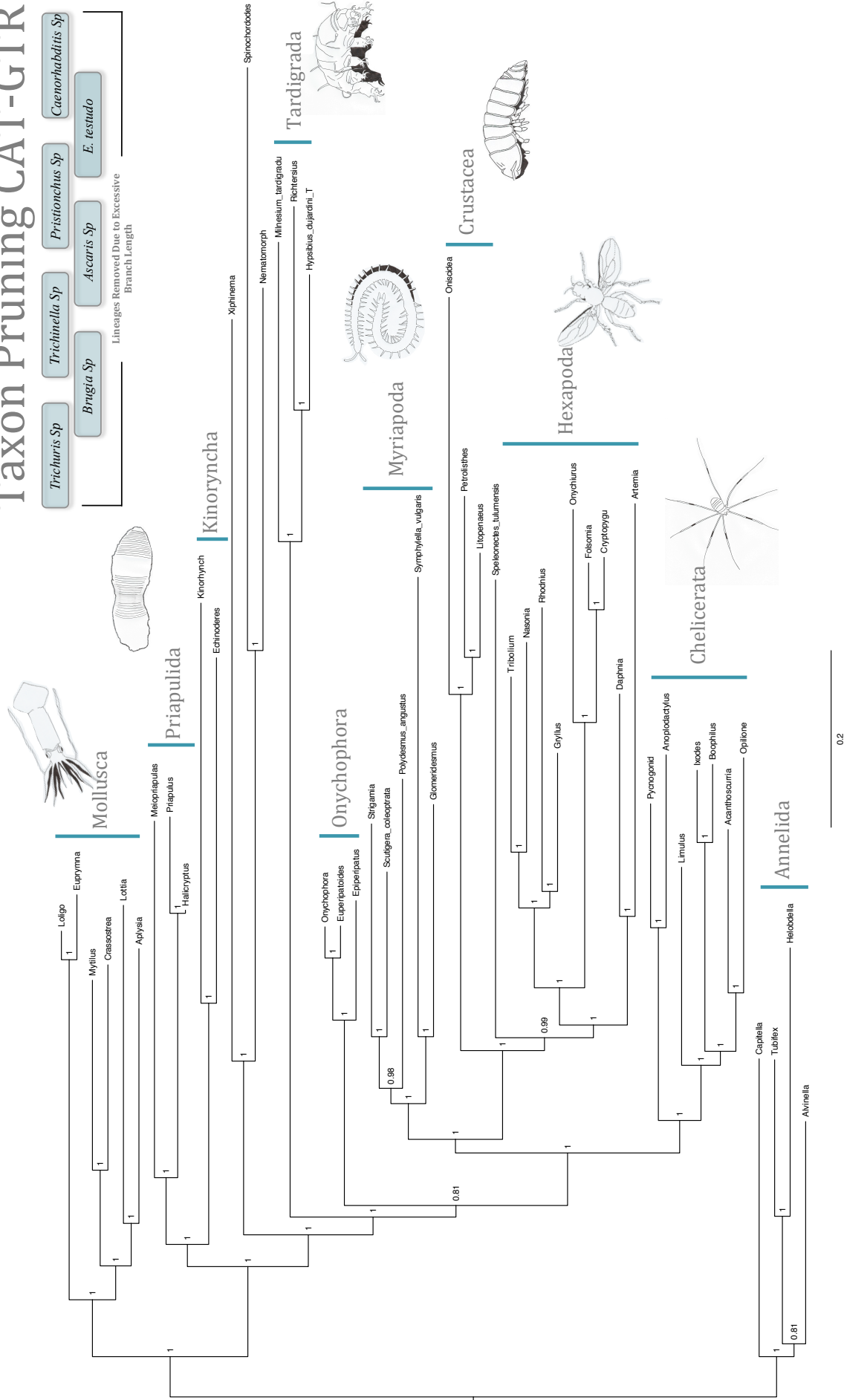


### 2.3.1.7 Taxon Pruning

Removing the fastest evolving lineages from the dataset: *Trichuris Sp.*, *Trichinella Sp.*, *Pristionchus Sp.*, *Caenorhabditis Sp.*, *Brugia Sp.*, *Ascaris Sp.*, and *E. Testudo* altered the recovered phylogeny under the CAT-GTR model, finding support for the Panarthropoda [**Figure 2.13**].



# Taxon Pruning CAT-GTR



**Figure 2.13: Tardigrade Taxon Pruning CAT-GTR**  
The Panarthropoda are recovered with a PP of 1.0

## 2.3.2 Tardigrade Divergence Time Estimation

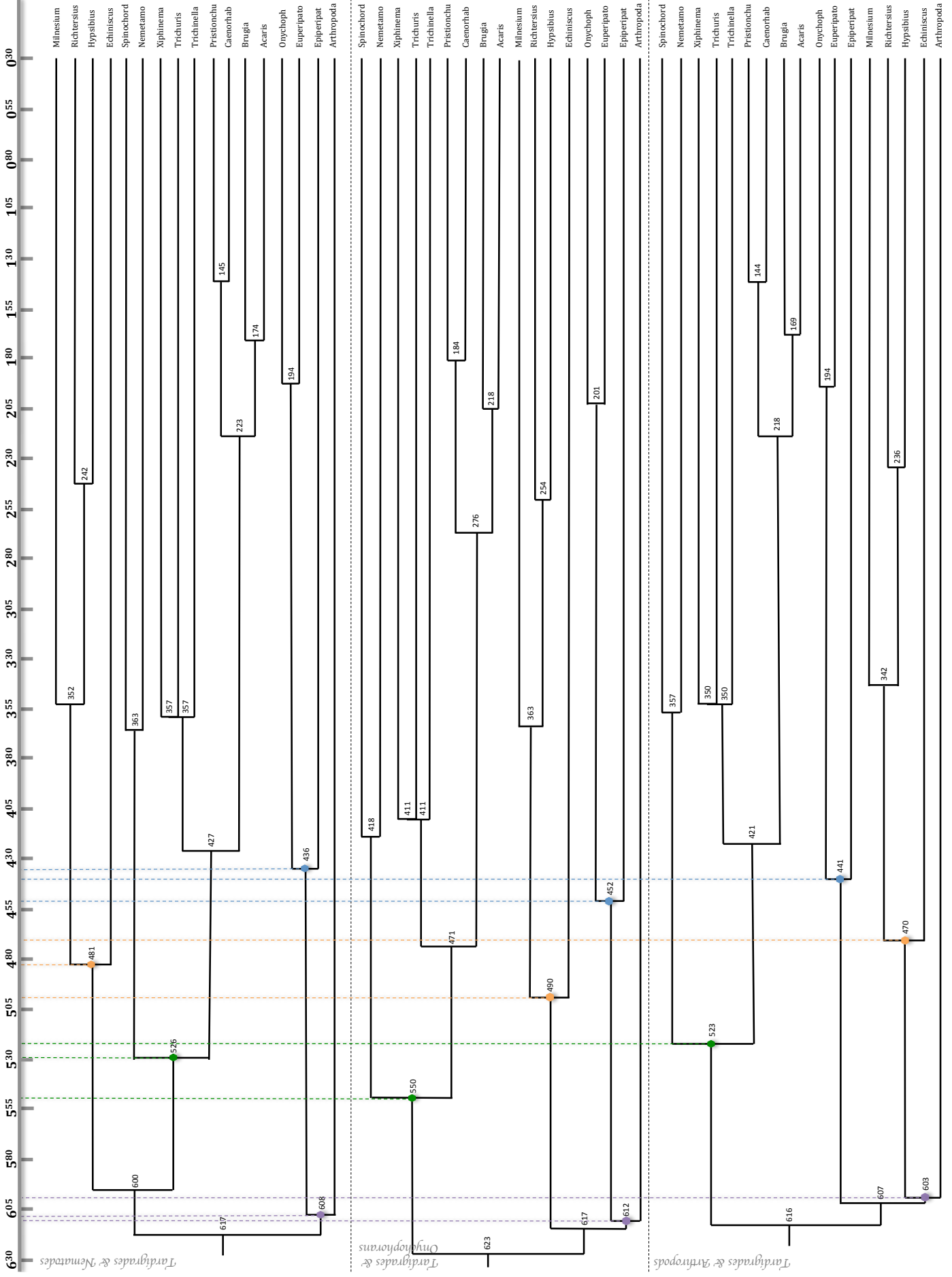
### 2.3.2.1 Dating the Alternative Hypotheses

The long disputed phylogeny of the Tardigrada has prevented in depth dating studies because of the need for an uncontested topology in the molecular clock process. We decided to take the approach of dating the Tardigrada under the three competing hypotheses; Tardigrada + Nematoida (T+N), Tardigrada + Onychophora (T+O), and Tardigrada + Arthropoda (T+A), under the CIR and U-GAMMA models and compare the divergence dates.

Interestingly we see only minor differences in divergence dates for the Tardigrada under the separate hypothesis, regardless of model, with the tardigrades diverging mostly in the Lower to Mid Ordovician: 490 (T+O), 481 (T+N), and 470 (T+A) MYA under the auto correlated CIR model and slightly earlier in the Mid Ordovician under the U-Gamma model: 470 (T+O), 464 (T+N), and 465 (T+A) MYA

[[Figures 2.14 A & B](#)].

# CIR Model



**Figure 2.14: Tardigrade Molecular Clocks**  
**Figure 2.14 A: Dating the Topologies under the CIR Model**

Discrepancies in divergence dates between the alternative hypotheses are illustrated with a focus on the Onychophora (blue), Tardigrada (orange), Nematoida (green), and arthropod-onychophoran / arthropod-tardigrade ancestor (purple). The scale is in millions of years, from left to right, ancient to recent.



**Table 2.7: Summary of Ecdysozoan divergence dates under the CIR and U-Gamma models**

Divergence dates for each of the three tardigrade hypotheses supported by scientific evidence are presented. Divergence time estimations for all topologies were performed under the autocorrelated CIR model and uncorrelated U-Gamma Model. The full set of chronograms can be found in [Supplementary Material 2.10](#)

**Table 2.7: Summary of Ecdysozoan Divergence Dates under the CIR and U-GAMMA Models**

Topological Hypothesis	Node	Node Age (Millions of Years)	
		<i>Auto-Correlated CIR Model</i>	<i>Un-Correlated GAMMA Model</i>
<i>Tardigrada + Nematoida</i>	Tardigrada Origins	481	464
	Nematoida Origins	526	445
	Onychophora Origins	436	269
	Arthropod-Onychophoran Split	608	610
<i>Tardigrada + Onychophora</i>	Tardigrada Origins	490	470
	Nematoida Origins	550	456
	Onychophora Origins	452	271
	Arthropod-Onychophoran Split	612	606
<i>Tardigrada + Arthropoda</i>	Tardigrada Origins	470	465
	Nematoida Origins	523	454
	Onychophora Origins	441	267
	Arthropod-Tardigrade Split	603	607
<i>Mean Dates</i>	Tardigrada Origins	480	466
	Nematoida Origins	533	452
	Onychophora Origins	443	269
	Arthropod-Onychophoran Split	610	608

## 2.4 Discussion

### 2.4.1 Assessing De-Novo Assembly Methods

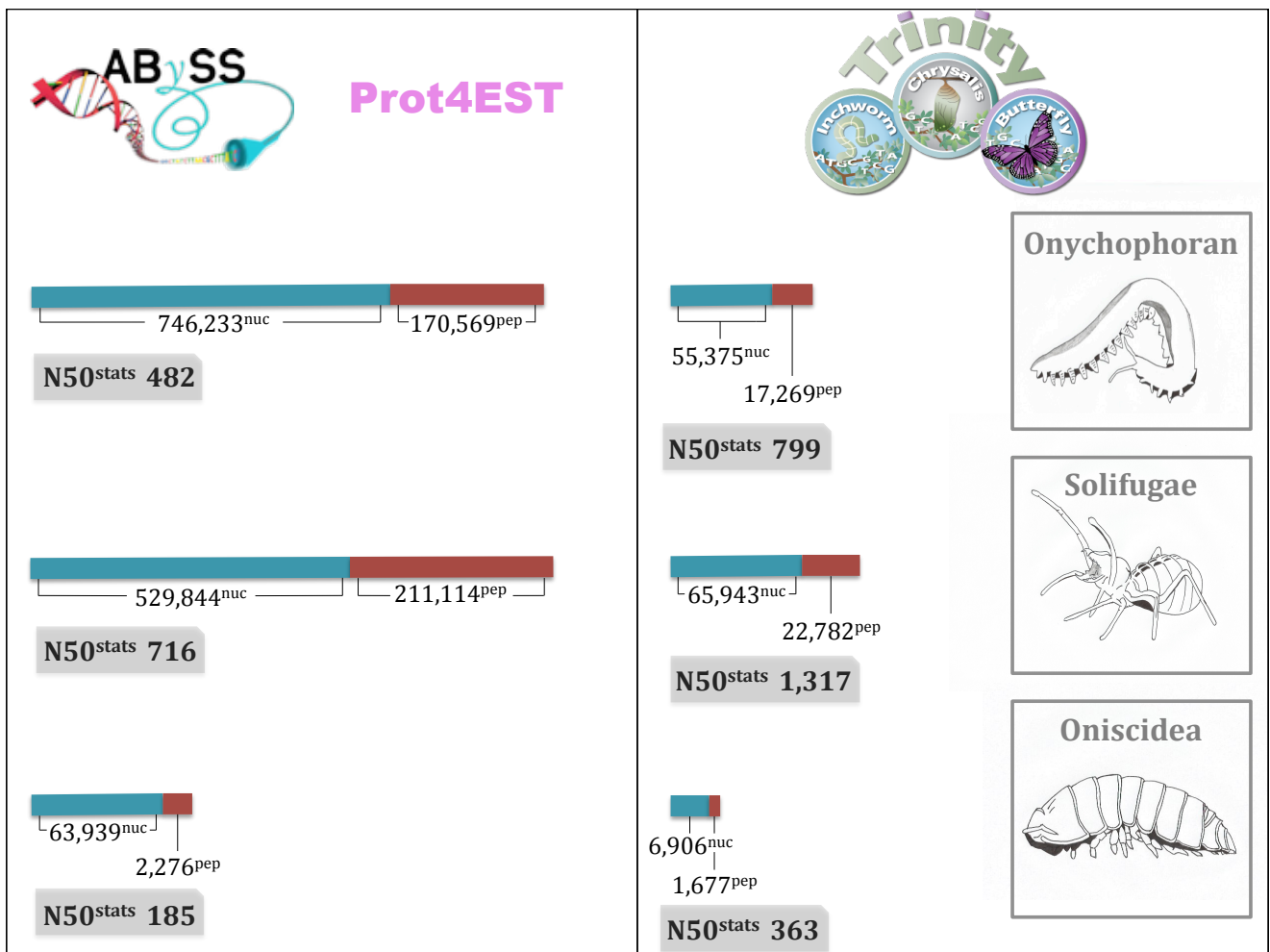
Initially de-novo transcriptome assembly was achieved through the ABySS and Trans-ABySS packages (Simpson *et al.* 2009 and Robertson *et al.* 2010), with translation of transcripts facilitated by Prot4EST (Wasmuth & Blaxter, 2004). However with the release of subsequent software; Trinity and TransDecoder (Grabherr *et al.* 2011 and Haas *et al.* 2013) along with the following studies comparing the two open source assembly software (Zhao *et al.* 2011 and Clarke *et al.* 2013) the decision was made to test the newly sequenced libraries against both.

Both assembly methods take the de Bruijn graph approach of assembling short reads but the important difference is the k-mer strategy imposed. Trinity chooses a single optimal k-mer while ABySS gives the user the option to run numerous assemblies with varying k-mer lengths to generate a mosaic assembly of multiple k-mer lengths. The issue with a multi k-mer approach to de Bruijn graph assemblies is that the entire process must be run many times depending on read length, taking a large amount of time and computational resources. This resource-limiting factor is multiplied for each species sequenced.

In agreement with Zhao *et al.* (2011) and Clarke *et al.* (2013), the Trinity and TransDecoder packages proved to be the more reliable, streamline, least computationally exhaustive in addition to producing the most contiguous assemblies and translated amino acid sequences [Figure 2.15].

ABySS and Prot4EST generated a greater quantity of sequences unanimously across all three taxa tested. However, the N50 stats show that they also produce the least contiguous assemblies that are on average 45% smaller than the N50 stats from

Trinity. We chose quality over quantity, valuing a more contiguous assembly over a more fractured assembly resulting in more sequences.



**Figure 2.15: A Comparison of ABySS & Prot4EST against Trinity & Transdecoder**

The cyan bar corresponds to number of assembled nucleotides while the red bar corresponds to the number of translated proteins from the nucleotide sequences. The comparison between methods was centered on three taxa: one from each of the Onychophora, Chelicerata, and Crustacea for balance. The N50 stats are a gauge of contiguity of the assembly the value of which refers to the length of half the assembled sequences being of equal or greater length than the value of N50.

### 2.4.2 Ortholog Mapping

Methods for identifying homology are long debated both in terms of terminology (Fitch, 2000) and methodology. Our methods start with a homology search using BLAST as an initial similarity search to identify prospective orthologs for newly sequenced taxa in relation to a dataset of hundreds of homologous genes. Often this is not enough however, as paralogous sequences are sometimes too similar to the original genes for BLAST to distinguish between orthology and paralogy (particularly concerning in-paralogs), even while implementing strict E. value cut-offs. It is common for a BLAST search to result in more than one putative ortholog for a particular gene in the dataset. To this point it is necessary to take homology searches a step further. This was done by building trees of the gene datasets using PhyML (Guindon *et al.* 2010), a useful balance of complex and quick phylogenetic reconstruction. This allows the evaluation of the putative orthologs for a particular gene on two further levels: placement in the tree and branch length. In relation to placement, a putative ortholog that is not recovered well within its evolutionary group is discarded. For example, a putative ortholog of the tardigrade *E. testudo* for a particular gene would be discarded if it fell outside of the Ecdysozoa as all molecular studies have placed them within this superphylum. However, one must be careful not to be too exact in this editing process to the point of bias, such as choosing the orientation of the tardigrades within the Ecdysozoa. This must be at the discretion of the tree reconstruction software.

The second metric, branch length, comes in to play once there are multiple sequences in plausible locations of the tree, Oftentimes these are clustered together, but not always. Since paralogs are copies, it follows that they will likely undergo more change over time than the original gene (Hurles, 2004). Therefore paralogs should



have a longer branch length than the real ortholog that we wish to add to the dataset. So the final step in our homology search is to keep the putative ortholog with the shortest branch length and discard the rest.

### **2.4.3 Phylogenomic Datasets**

The backbone of our dataset is an amalgamation of (Philippe *et al.* 2011b and Campbell *et al.* 2011) with a focus on ecdysozoan taxa. Sensibly, genes are primarily chosen for their slow rate of evolution across a multitude of lineages as this prevents character saturation and loss of phylogenetic signal. Signal saturation cannot be entirely avoided however as the addition of groups with a comparatively higher rate of evolution such as the Tardigrada or Nematoda will break this rule and introduce orthologs with a higher rate of change to these slowly evolving genes. There are no universally slowly evolving gene catalogs for all types of lineage. The best one can do is avoid evolutionary distant groups and rapidly evolving taxa. This is not always possible when working in the realms of deep node phylogenetics or when studying rapidly evolving lineages. This study concerns both, which is why further methodological steps were required to reduce the effects of systematic errors (see **Materials and Methods 2.2.10**).

The choice of gene for these foundation datasets can sometimes be arbitrary however. These datasets were built before the widespread availability of NGS technology so are mostly made up of ESTs. A consequence of this is that choice of gene is limited to its coverage from EST experiments for the taxa of interest. Full molecular libraries were not available at the time. As a consequence, useful gene candidates for the dataset could have been excluded because of such limiting factors. We are only presently

moving into a window where full molecular libraries are becoming available and such stochastic errors are no longer a problem. It will be important to readdress the foundations of these datasets in the future and supplement not only the taxa coverage, but to add to the number of phylogenetically useful genes in order to strive to build the most complete evolutionary scenarios possible with the current technology at our disposal.

#### **2.4.4 Phylogenomics and Systematic Error**

Phylogenomics has unequivocally broadened the horizon for molecular evolution studies, allowing researchers to study the full genomic library of virtually any living species on the planet. The key weaknesses of molecular evolution this new technology addresses are stochastic errors such as poor taxon sampling and insufficiently sized datasets that are not a reflective subsample of the evolutionary history of the species studied. These benefits have allowed us to investigate the more obscure animals such as the Chaetognatha (chapter 3) and clarify disputed topological placements such as the Tardigrada by resolving ambiguity through the addition of more molecular data and taxa.

While phylogenomics has been important in resolving such issues it would be a naïve assumption to conclude it solves all the problems molecular evolution studies face today, the most prominent of which are systematic biases that elude the benefits of adding information to a dataset, discussed in section **1.2.6 Stochastic and Systematic Error**. The principle form of systematic error accounted for in this chapter is LBA because of the rapidly evolving nature of the tardigrades and nematodes and their correspondingly long branches in the tree compared to their sister ecdysozoans.

It is clear based on the experimentation of this chapter that phylogenomic datasets are highly susceptible to systematic errors such as LBA and heterotachy when concerning rapidly evolving lineages and poorly fitting models of evolution respectively, regardless of how much data is collected and tested. However we have also shown that taking steps to nullify such errors can be beneficial in minimizing their effects. Signal dissection measures were successful in identifying LBA and uncovering the true phylogeny of fast evolving taxa while model testing revealed that the site-heterogeneous model CAT-GTR was the best fit for the data.

In light of these results we would recommend such methods accounting for systematic errors become part of the standard operating procedure for large-scale phylogenomic studies involving rapidly evolving groups.

#### **2.4.5 The Importance of Model Testing**

In addition to testing site-heterogeneous models of evolution in efforts to avoid heterotachy, a multi-model approach is necessary in order to account for the highly variable nature of molecular datasets. Past phylogenetic studies have come into question for their use of antiquated evolutionary models (Telford & Holland, 1993; Papillon *et al.* 2003 & 2004; Matus *et al.* 2006). Often this is understandable due to the lack of sophisticated models at the time after all one can only use what was available. But there have been cases where studies have arbitrarily chosen a model without testing its suitability against others. Without an objective method for identifying what model tested best fits the data, our experiments would have been mired in uncertainty due to the differing phylogenies generated by the CAT, GTR, and CAT-GTR models. Therefore a necessary step in phylogenetic reconstruction should be a critical approach to model choice, in this case a Bayesian cross validation. As molecular evolution models continue to develop and strive to increase their efficiency in copying nature, the models used in this study will also age and diminish in comparative sophistication. It is important to remain grounded in our phylogenetic estimates, as they are a best appraisal of incredibly complex scenarios with the tools, data, and models at our current disposal.

#### **2.4.6 The Uncertainty of the Molecular Clock**

Molecular clocks are far from infallible and their misuse can result in bizarre divergence time estimations; see Graur & Martin (2004) for a tongue in cheek yet important discussion on such lackadaisical clock methods. Below are some aspects of clocks that one must be wary of when planning divergence time estimation experiments.

##### **2.4.6.1 Gaps in the Fossil Record**

The obvious weakness of the fossil record is of course its incompleteness. Fossilization by its very nature is a rare occurrence, more so for the ancient soft-bodied metazoans (Lipps & Signor, 1992 and Conway-Morris, 1993). There are various conditions required for such an event. First and foremost the specimen needs to expire in sedimentary conditions such as sand or silt, often found on shorelines or seabed. These sediments, when deposited from their suspension, can form the basis of sedimentary rock over millions of years, becoming an excellent preservative for the specimen. Secondly because of the rapid degradation of biological tissue post-mortem, the fossilization event needs to occur quickly often in a natural disaster such as flooding (Raup *et al.* 1978). This also prevents the specimen from getting damaged or destroyed by predators. Anaerobic or low oxygen environments also aid in the prevention in decay, and in order for preservation post fossilization the specimen should be protected from exposure to the elements. Finally the harder the composition of the specimen's body plan, the better the chance for fossilization. This is not a preservation "advantage" which the protostomes possess.

Because a combination of such exceptional conditions must be met for the animal to be preserved, it is statistically improbable that the oldest known fossil for a particular clade is representative of that clade's origin. Instead fossils are treated as minimum age constraints in divergence time estimation. A minimum age constraint for a particular clade is simply a limit as to how young the clade can be dated. For example, if we know a fossil for a certain clade exists in the Cambrian, the origin of said clade could not be post-Cambrian, assuming the fossil has been correctly ascribed and is a member of the crown group.

A mathematical method has been designed to address the holes in the fossil record, prior probability distributions estimate factors such as the chance of fossil preservation and discovery based on a variety of metrics such as geographical fossil preservation biases; fossilization is often restricted to particular environments such as areas of heavy sedimentation like river beds and shores, chance of discovery; locating intact fossils is usually limited to known paleontological sites which exhibit magnificently unusual levels of preservation (Wilkinson *et al.* 2011).

#### **2.4.6.2 Fossil Identification**

The correct interpretation of clade and lineage defining features in a fossil is an essential step in the process of divergence time estimation. In paleontology this is known as fossil ascribing (Raup *et al.* 1978). Mistakes in this process can lead to erroneous calibration points as incorrectly identified fossils can be appointed to the wrong clades causing them to be constrained by false dates. Donoghue & Benton (2007) remark that the difficulty of accurate fossil identification is compounded by our reliance on exceptionally preserved fossils. Damaged or missing parts of the

specimen can cause defining characteristics to be lost making ascribing the fossil impossible. This generates uncertainty in the accurate assignment of the sample to a particular group. Furthermore, even if the fossil is well preserved, the ancient nature of the specimen can create morphological disparity between itself and its extant relatives who may have lost certain ancestral character features over long periods of time.

#### **2.4.6.3 Crown and Stem Groups**

A longstanding complication in fossil calibrations of molecular clocks is the constraint of a fossil as a minimum age to a lineage or group without being sure of a crown group relationship. A crown group consists of a number of closely related lineages and their last common ancestor, it also includes all that ancestor's descendants. It is essentially a group displaying the full evolutionary radiation of a particular ancestor, making them monophyletic. Conversely, a stem group is an incomplete picture of the lineage radiation from an ancestor, a paraphyly composed of an ancestor, its radiating lineages, but minus its living descendants. Stem groups have undergone extinction events, creating a missing link between themselves and their living crown group cousins. Assigning fossils that are a more accurate representation of a stem group than a crown group to extant lineages in a divergence time estimation study will result in a large disparity between the dating of the fossils and molecules. An example of this issue is found in chapter 3.

## 2.5 Conclusions

While the best fitting model for the data (CAT-GTR) recovered the Tardigrada grouping with the Nematoida [Figure 2.11], in agreement with (Lartillot & Philippe, 2008; Meusemann *et al.* 2010; Borner *et al.* 2014). Lesser fitting models tested, CAT [Figure 2.9] and GTR [Figure 2.10], returned with a tardigrade and onychophoran grouping and a tardigrade plus nematoides topology respectively. Additionally the BCV was not conclusive in its assessment that CAT-GTR was more suitable than the CAT model for this dataset [Table 2.5]. We found no experimental evidence for the sister affinity of tardigrades to the arthropods (Smith & Ortega-Hernandez, 2014 and Gross *et al.* 2015).

However the results from our robust signal dissection experiments of rapidly evolving characters discovered evidence that LBA is likely influencing the grouping of the Tardigrada and Nematoida.

Removing the longest branched tardigrade, *E. testudo*, along with many of the long branched nematodes, recovered a Tardigrada plus Onychophora grouping [Figure 2.13], while a Dayhoff recoding, which attempts to nullify the deleterious effects of over-saturated sites, returned a Panarthropod grouping [Figure 2.12] both under the CAT-GTR model. Interestingly, the topological position of the other Ecdysozoan clades, reconstructed under the original dataset, remain broadly constant under the Dayhoff Recoding and taxon pruning strategies.

While all slow / fast technique subsets returned the same tardigrade plus nematoides topology, the posterior probability of this grouping consistently falls when excluding faster characters, meaning as we remove saturation confidence of this grouping systematically begins to fall. In addition to the evidence presented from the Dayhoff



recoding and taxon pruning, these findings unequivocally highlight the deleterious influence of low signal, saturated characters in rapidly evolving lineages, causing the grouping of the two fastest evolving clades of animals within the Ecdysozoa to be artificially grouped together based on their high rate of evolution as opposed to a true affinity of sharing a LCA. There is enough evidence from these experiments to confidently say the tardigrade nematode grouping in this dataset is an artifact of LBA.

Taking all aspects of phylogenetic evidence in to account, we can conclude that there is no molecular evidence for the Tactopoda, therefore the shared morphological characteristics unique to the tardigrades and arthropods must be due to a loss of such traits in the Onychophora. The Tardigrada - Nematoda grouping is an artifact of LBA based on the results of the signal dissection experiments but there is still uncertainty in relation to the tardigrades relationship with the onychophorans. The taxon pruning experiment construes a Panarthropod relationship in agreement with Rota-Stabelli *et al.* (2010) and Campbell *et al.* (2011) with high PP support (1.0) but the Dayhoff recoding strategies place the tardigrades as a sister group with the onychophorans (PP = 0.98) as does the CAT model. This is the first line of molecular evidence suggesting such a grouping but has been suggested by morphologists (Mayer *et al.* 2009 & 2010). While we cannot conclusively determine which of these two scenarios reveal the true affinity of the Tardigrada, we can conclude that the Tardigrada, Onychophora, and Arthropoda groups share a common ancestor, to the exclusion of the Nematoda.

Our phylogenetic reconstruction of the Arthropoda also recovers the Mandibulata (Rota-Stabelli *et al.* 2011; Misof *et al.* 2014; Borner *et al.* 2014) over the Myriochelata (Friedrich & Tautz, 1995; Cook *et al.* 2001; Pisani *et al.* 2004) but raises uncertainties in the Myriapoda with a paraphyletic Diplopoda. A different

sampling of myriapods produced a contrasting topology (Lozano-Fernandez *et al.* 2016): a monophyletic Chilopoda and Diplopoda, but lacking any sampling from the Symphylella. We recommend studying phylogenies with a full sampling of myriapods in future studies to clarify this issue.

Due to the initial uncertainty of the position of the Tardigrada an all-encompassing approach was necessary in addressing their origins. All three tardigrade topological hypotheses were tested: Tardigrada+Nematoida (T+N). Panarthropoda (T+O), and Tactopoda (T+A). Unfortunately we did not foresee a tardigrade - onychophoran grouping during the experimental design stage as up until now there had been no molecular evidence suggesting such an affinity.

The origin of the tardigrades was most likely in the Lower to Mid Ordovician 480 - 466 MYA, with only a single of the six scenarios (three topologies under two models) dating them slightly older, in the very Late Cambrian, 490 MYA [Table 2.7]. This estimation is roughly 38 - 24 MY older than the previous most notable study of tardigrade origins (Rota-Stabelli *et al.* 2013) which places them in the Silurian, approximately 442 MYA. The interesting aspect of this topologically comprehensive experimental method of divergence time estimation is that we see very little difference in the timing of the clades regardless of topology. For example, the Tardigrada see a  $\Delta$  20 MY under the CIR model and  $\Delta$  6 MY under the U-GAMMA model across the varying topologies, the other clades span even smaller differences.

While the divergence date of the Tardigrada is similar under the two models tested, we see large discrepancies between the dating of the Onychophora and Nematoida under the auto-correlated CIR and uncorrelated GAMMA reconstructions. The mean difference in divergence dates for the Onychophora and Nematoida under the two models are 174 MYA and 81 MYA respectively. With these deviations in

mind, without topological reference, we place a vague origin for the Onychophora sometime between the Silurian and Permian, and an Early Cambrian to Mid Ordovician origin for the Nematoida.

Despite this ambiguity, we can confidently conclude a more concise timing for the Tardigrada divergence: Late Cambrian to Mid-Lower Ordovician. If one considers that we found no evidence for a T+A grouping and consider the T+N affinity to be an artifact of LBA, the Tardigrada origins can be narrowed down even further to a 20 MY time scale running between the Late Cambrian to Mid Ordovician [Table 2.7].

The implications of these divergence time estimations is that we can place the origins of the tardigrades in the Cambrian explosion which fits the narrative of Enright *et al.* (2011) claiming that while the origins of animals originated pre-Cambrian, most phylum level crown group animals radiated in the Cambrian by taking advantage of geological changes such as ocean redox and with the aid of pre-formed gene networks developed during the macroevolutionary lag between the Ediacaran and Cambrian. The explanation for such radiation and rapid diversification of lineages could have been driven by the increased number of predators appearing in the sea availing of the increased oxygen levels that promoted their metabolic needs (Sperling *et al.* 2013) resulting in a wide array of morphological adaptations (see chapter 3 for a study of such predators). Furthermore the deep date for the Ecdyszoa origins in our dataset are broadly in line with Enright *et al.* (2011) Ediacaran divergence dates, laying credence to the pre-Cambrian origins of animals.

Moving closer towards the tips of molecular time trees, our tardigrade divergence dates also agree with the evolutionary divergence of the Ecdyszoa from Rota-Stabelli *et al.* (2013) conforming to their estimates that the Tardigrada are some of the youngest members of the superphylum.

### 3.1 Introduction

#### 3.1.1 Chaetognatha: Ancient Predators

Chaetognatha “bristle jaws”, are a small group of planktonic, carnivorous, marine worms that range in size from 2mm to 120mm (Bieri, 1959). Their most noticeable characteristics are their grasping spines emanating from the side of their head, sharp array of teeth (both essential to their predatory lifestyle) in conjunction with lateral, dorsal, and ventral fins – the number of which varies among species (Tokioka, 1965 and Feigenbaum & Maris, 1984). All known species are hermapharditic and they make up a large percentage of the oceans plankton (Bieri, 1959).

Due to their ancient catalog of fossils (Schram, 1973; Chen, 2002; Doguzhaeva *et al.* 2002; Vannier *et al.* 2007) in addition to their carnivorous characteristics, the Chaetognatha are considered to be some of the first predatory animals.

The timing of the chaetognath origins is of particular significance as there is evidence that shows low oxygen levels in the sea during the Cambrian had a direct influence on inhibiting the number and diversity of carnivorous animals (Sperling *et al.* 2013). Essentially this makes the Chaetognatha candidates for the some of the earliest predators to appear in the Animal Kingdom as the fossil evidence proves their existence in this time period. It is not only possible that they were the first predators in the ocean, but for a period of time they could have been on top of the food chain

because of the lack of competition. Considering that contemporary chaetognaths consist of much of the oceans plankton (Bieri, 1959 and Parsons, 1988), which is basically food for larger animals such as whales, the last 500 million years would have seen them fall drastically down the food chain in a major evolutionary shift of the oceans foodweb.

The importance of the ancient Chaetognatha grows when one considers the further claim of Sperling *et al.* (2013), that such predators were the driving force for the Cambrian explosion, specifically the radiation of diverse lineages and eclectic disparate body plans. This is an agreement with the explanation for the Cambrian radiation by Erwin *et al.* (2011). This major evolutionary event coincided with an ocean redox that raised the oxygen levels in the seas, but Sperling argues such environmental changes, while essential in facilitating the event by promoting carnivorous lifestyles, lack the driving force to incur such disparity and diversification of animals lineages that radiated from the explosion. Instead they summarise that the appearance of such animal body plans in a small space of time was the result of a predator – prey arms race, occurring in a non-conventional food web, that promoted morphological innovation.

With the importance of these animals outlined, it is not surprising that there have been many evolution studies on these ancient predators.

Similarly to the Tardigrada, the phylogeny of the Chaetognatha is also ambiguous, with early morphological studies classifying them as deuterostomes (Hyman, 1959; Ghirardelli, 1968 & 1981; Ducret, 1978) and later molecular studies placing them within the Protostomia but with a wide array of disagreement (Telford & Holland, 1993; Papillon *et al.* 2003 & 2004; Matus *et al.* 2006; Marlétaz *et al.* 2006 & 2008; Paps *et al.* 2009b; Philippe *et al.* 2011b).

### 3.1.2 Deuterostomes or Protostomes?

The phylogenetic affinity of the chaetognaths is infamously ambiguous despite half a dozen studies that have tried to place them. This problem stems from their eclectic possession of both deuterostome and protostome traits, brought to the attention of the scientific community by the earliest morphological studies (Doncaster, 1902).

The chitinous spines of the chaetognaths are similar to the chitinous jaws of rotifers (lophotrochozoans), and the presence of a ventral nerve cord suggests a protostome affiliation (Nielson, 2001). However, their tripartite body plan with a post anal tail, and radial intermediate cleavage (Matus, 2006), along with what is seen to be their most morphologically defining characteristic; the formation of the anus from the blastopore, all point to deuterostome origins.

Since the advent of molecular data, a range of studies using various lines of molecular evidence (Telford & Holland, 1993; Papillon *et al.* 2003 & 2004; Marlétaz *et al.* 2006 & 2008; Matus *et al.* 2006; Paps *et al.* 2009b; Philippe *et al.* 2011b) [Figure 3.1] have categorized the chaetognaths as protostomes yet none have been able to robustly place them within the clade.

### 3.1.3 Competing Phylogenetic Hypotheses

The initial molecular study of the chaetognaths was conducted by Telford & Holland (1993) using an 18s ribosomal dataset from *Saggita elegans*. Their results, although important at the time as the first piece of molecular evidence suggesting that chaetognaths were not deuterostomes, suffers from stochastic and systematic errors: a small dataset, outdated phylogenetic reconstruction methods and poorly fitting evolutionary models.

As such, they were unable to precisely place the chaetognaths but concluded that they either form a sister group to the coelomate protostomes or lie outside the coelomates [Figure 3.1 A].

Telford & Holland also concluded that the chaetognaths many deuterostome-like characteristics may not be synapomorphies (traits novel to an ancestor and its descendants) (Zelditch *et al.* 1995) rather pleisomorphies (an ancestral trait state) (Olmstead, 1995) or homoplastic apomorphies (novel traits derived in the extant lineage that seem to have ancestral origins but do not) (Wägele, 1996).

A phylogenetic analysis of mitochondrial datasets indicates that the chaetognaths may be lophotrochozoans [Figure 3.1 B]. The analyses, centered on the sequencing of *Spadella cephaloptera*'s mitochondrial genome, also showed that the chaetognath mito genome has experienced similar gene loss to several lophotrochozoans that are otherwise conserved amongst the Metazoa (Papillon *et al.* 2004).

However, the exact positioning of the chaetognaths within this super phylum could not be fixed.

The uncertainty of their location in the phylogeny could be attributed to the overreliance on mitochondrial data, which is particularly susceptible to compositional heterogeneity on two levels: strand asymmetry and the homoplastic clustering of taxa because of guanine and cytosine deficiencies (Rota-Stabelli & Telford, 2008).

The placement of the chaetognaths within the Lophotrochozoa, based on mitochondrial data, is contrary to the phylogeny obtained through a previous study based on a six hox gene dataset extracted from the same species of chaetognath (Papillon *et al.* 2003) that was in agreement with a pre deuterostome protostome split divergence (Telford & Holland, 1993).

A later study (Marlétaz *et al.* 2006) using a dataset derived from 11,526 ESTs of *Spadella cephaloptera* and ribosomal proteins, found the chaetognaths to be basal protostomes [Figure 3.1 C]. Both maximum likelihood and Bayesian inference reconstruction methods were used and steps were taken to reduce the influence of both compositional heterogeneity and LBA.

Marlétaz *et al.* found a gene in the chaetognaths, Guanidinoacetate N-methyltransferase, which is present in non-protostome Bilateria such as the deuterostomes and cnidarians. It is unclear whether this is indicative of the chaetognaths true phylogenetic affinity or a feature of their mosaic nature. Further studies (Marlétaz *et al.* 2008) using a wider phyla scope, taxon sampling and updated inference methods (site heterogeneous CAT model) further supported the proposal for the Chaetognatha as the sister to all other protostomes.

A fourth hypothesis was proposed by Matus *et al.* (2006). Using small and large subunit rRNAs and mitochondrial genomes in conjunction with roughly 5,000 ESTs sequenced from *Flaccisagitta enflata* they conclude that the Chaetognaths may be a sister group to the Lophotrochozoa [Figure 3.1 D].

Paps *et al.* (2009b) published a study of 13 orthologs across 90 taxa in order to address the complex Bilaterian phylogeny which alluded to an Ecdysozoan origin of the Chaetognatha, placing them sister to the Onychophora [Figure 3.1 E].

However, the authors themselves claim that the clustering of chaetognaths with the other Ecdysozoans in their dataset (nematodes, arthropods, and onychophorans) is most likely due to LBA, which has obscured ecdysozoans relationships and phylogenetic reconstruction methods in general (Bergsten, 2005).



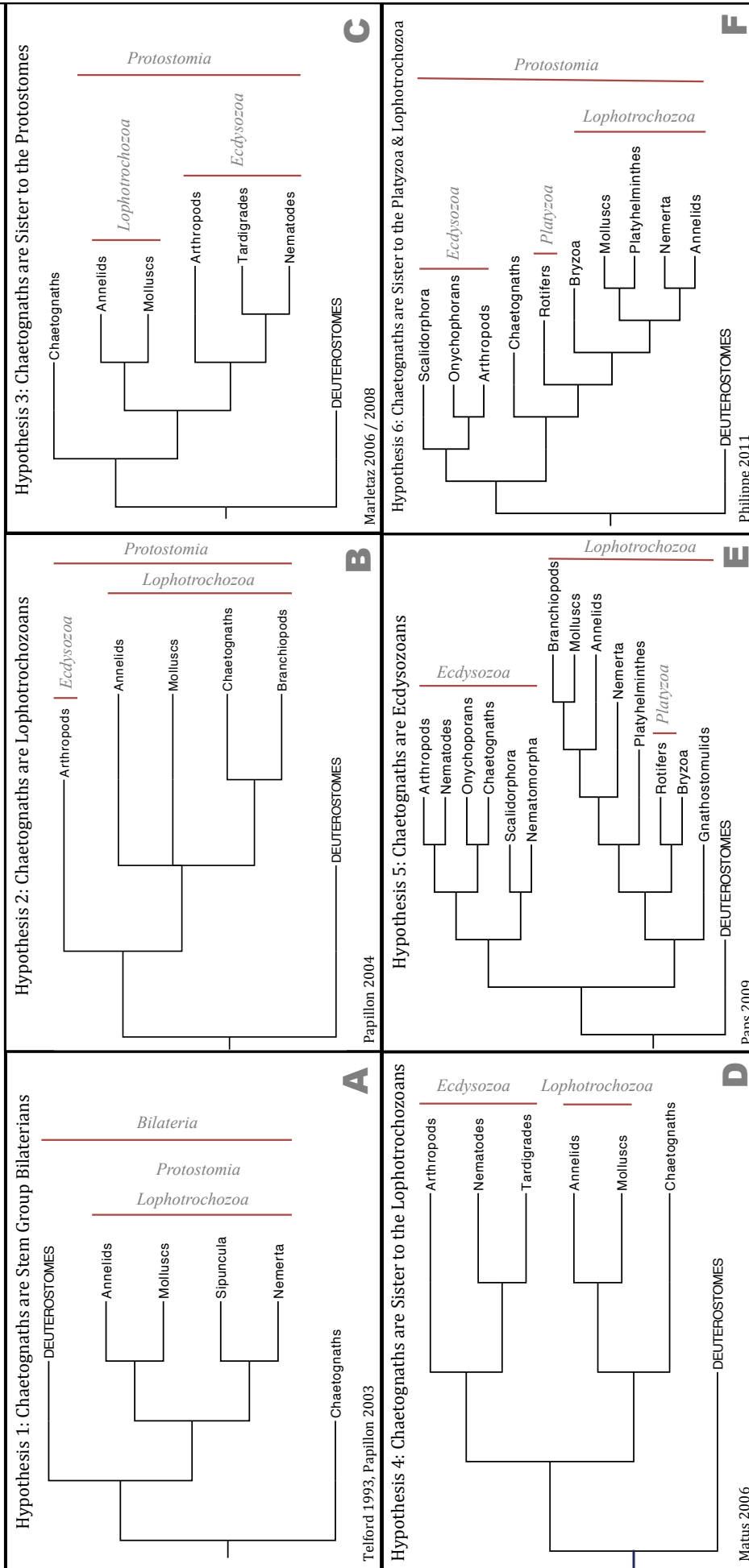
One of the more recent studies of chaetognath relationships consisted of three different molecular sources – a 66 taxa, 197 gene phylogenomic dataset with additional mitochondrial, and miRNA evidence (Philippe *et al.* 2011b).

These findings were similar to that of Marlétaz *et al.* (2006 & 2008), classifying the chaetognaths as the sister group to the Lophotrochozoa but also to the platyzoan rotifers [**Figure 3.1 F**].

Finally a very recent phylogenomic study of the Lophotrochozoa from Kocot *et al.* (2016) found the chaetognaths to be sister to all lophotrochozoans but without the existence of the Platyzoa. The Kocot group experimented with their dataset, generating eight differing phylogenies depending on the type and degree of systematic bias being accounted for. Only a small number of recovered groups throughout the eight phylogenetic experiments remained broadly topologically consistent: the Mollusca, Brachiopoda, and most notably for this study: the Chaetognatha.

These studies have highlighted the need for the use of expansive yet balanced datasets while stressing the importance of reducing the pitfalls of phylogenetic reconstruction: signal saturation, avoidance of compositional biases and choice of the most realistic evolutionary model for the data.

# Previous Hypotheses of Chaetognath Phylogeny from Molecular Datasets



**Figure 3.1: Chaetognath Phylogenies from Molecular Studies**  
 The placement of the chaetognaths is still unclear even after a range of molecular studies ranging near twenty years. Morphological studies predating the molecular reconstructions considered the chaetognaths Deuterostomes due to their deuterostome-like development.

### 3.1.4 The Chaetognath Fossil Record

The fossil record of the chaetognaths is poor, perhaps due to their benthic nature and body composition not suited to conditions for fossilization (Lipps & Signor, 1992). However, there are a small number of acceptably well ascribed fossils known [Figure 3.2].

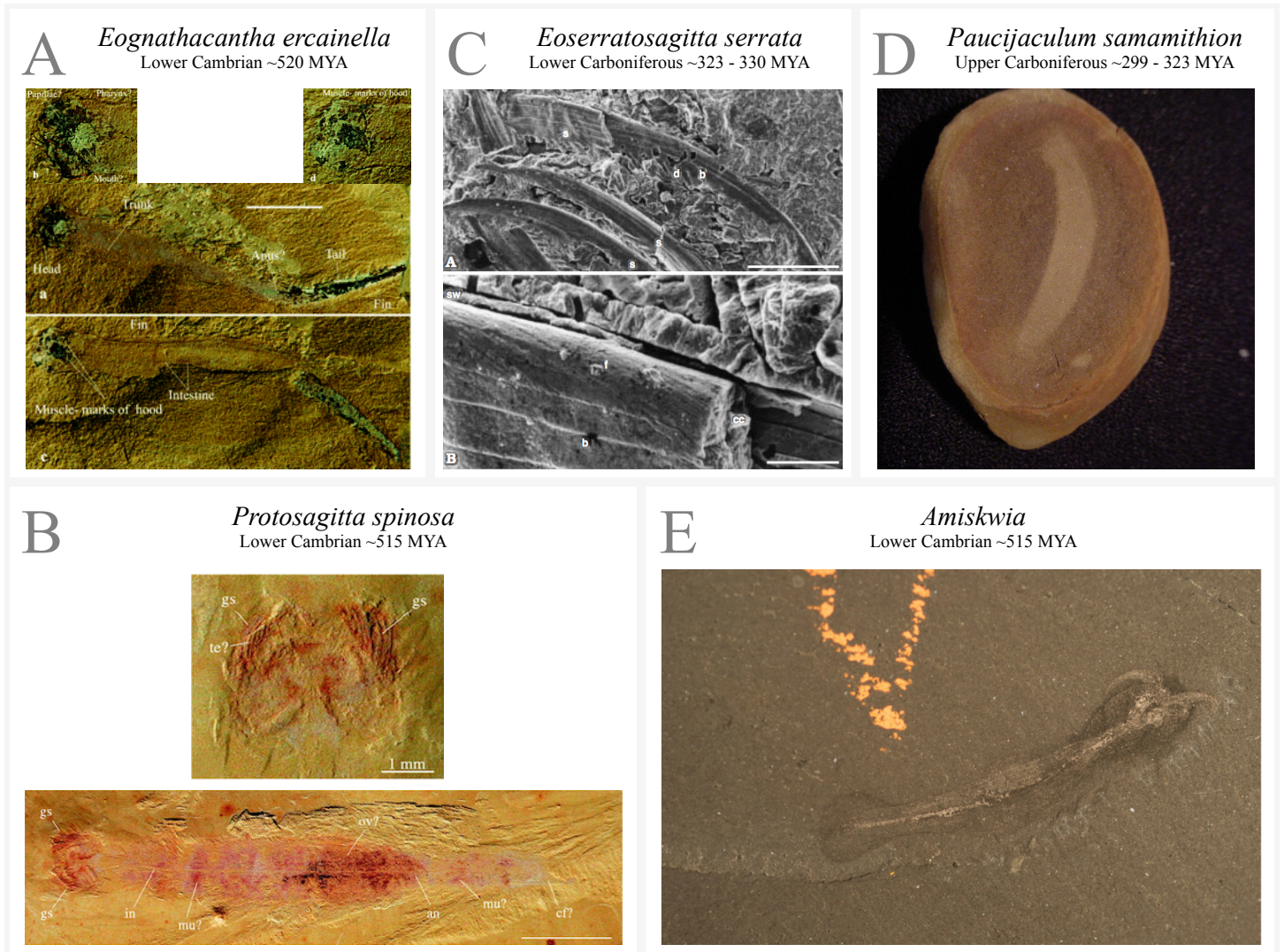
The oldest known chaetognath fossil is *Eognathacantha ercainella*, dated approximately 520 million years ago (MYA) in the Lower Cambrian (Chen, 2002). This fossil was found in the Maotianshan Shale Haiko of Kunming South China, and its degradation has made it difficult to identify chaetognath features. Yet its grasping spines are definitely evident [Figure 3.2 A].

The most preserved chaetognath fossil was discovered in the Chengjiang biota Yuanshan formation, China. Similar in age to that of *E. ercainella*, *Protosagitta spinosa* has been dated as roughly 515 MYA (Vannier *et al.* 2007) [Figure 3.2 B].

Younger chaetognath fossils have been found dating back to the Carboniferous period, *Eoserratosagitta serrata* (Doguzhaeva *et al.* 2002), dated to the Lower Carboniferous, possesses clear grasping spines [Figure 3.2 C] and *Paucijaculum samamithion* (Schram, 1973) is believed to be from the Upper Carboniferous [Figure 3.2 D].

A taxon that has been at stages suggested to belong to the Chaetognatha is the iconic Burgess Shale fossil *Amiskwia* (515 MYA) (Conway-Morris, 1977). *Amiskwia*'s morphological peculiarities have made it a problematic taxon to ascribe. Despite having a similar body plan to the chaetognaths, including dorsal, ventral and lateral fins, they lack the characteristic grasping spines, possess antennae which are not found in fossil chaetognaths and its anus is placed at the end of tail, different to that of extant chaetognaths [Figure 3.2 E].

Tokioka (1965) offered an explanation for these discrepancies with traditional chaetognath features claiming *Amiskwia* is a member of a separate class, the extinct Apherogomorpha, to that of the rest of the extant chaetognaths who are placed in Sagittoidea.



**Figure 3.2: The Fossil Record of the Chaetognatha and Amiskwia**

Discovered in archeological sites from China to the USA, from 299 to 520 million years old, the most characteristic morphological element is their grasping spines. Image Sources: **A:** Chen & Huang (2002) **B:** Doguzhaeva *et al.* (2002) **C:** Schram (1973) **D:** Vannier *et al.* (2007) **E:** Conway-Morris (1977).

### 3.1.5 Lineage Characteristics and Ascribing the Amiskwia Fossil

Re-visualizing a detailed description of chaetognath lineage characteristics (Tokioka, 1965), enabled the possibility of ascribing the most preserved chaetognath fossils and the ambiguous *Amwiskwia* [Figure 3.3]. The accuracy of these interpretations depends on the correct polarisation of character-state changes across the Chaetognatha by Tokioka (1965).

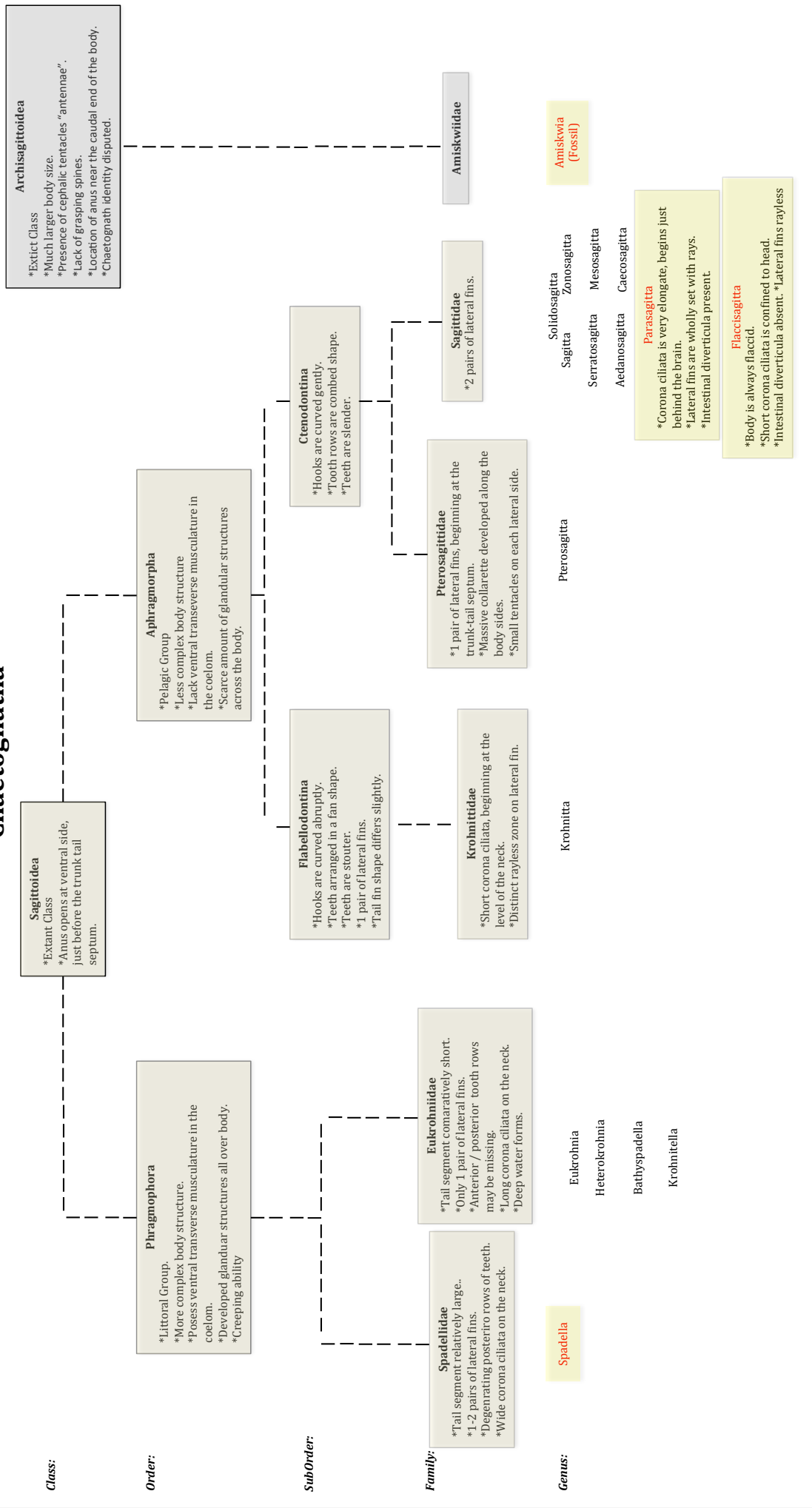
Several lineage-defining characteristics were identified in *P. spinosa* and *E. ercainella* making it possible to assign them to a specific family of Chaetognatha including:

- The location of the anus (An) is before the end of the tail, a defining feature of the extant chaetognath class Sagittoidea.
- Evidence of musculature in the coelom (Mu), indicating that it falls under the Phragmophora order of chaetognaths with more complex body structures.
- A thin tail segment (Ts) is indicative of a member of the Eukrohniidae family.
- A single pair of lateral fins (Lf) also points to the Eukrohniidae.

While some members of the neighbouring Spadellidae family under the same order also have only a single pair of lateral fins, the thinning tail segment clearly indicates *P. spinosa* should be placed in the Eukrohniidae.

Similarly, one of the defining characteristics of the Pterosagittidae is also a single pair of lateral fins but again *P. spinosa* lacks the other defining features of this family including the simpler body structure, larger tail size, and antenna [Supplementary Material 3.1].

# Chaetognatha



**Figure 3.3: Lineage Characteristics of the Chaetognatha**

An illustration of Tokioka's (1965) characterization of the Chaetognatha. These polarizing features were useful in classifying the ascribed chaetognath fossils and the controversial *Amiskwia*.

### 3.1.6 Aims of this Study

The purpose of this study is to clarify the phylogenetic position of the Chaetognatha with the supplementation of the *Parasagitta sp.* genome to a large dataset of protostomes and deuterostomes (Philippe *et al.* 2011b), additionally to clarify their affinity with their fossil catalog, in particular the disputed *Amiskwia* (Conway-Morris, 1977), and finally to date the origins of the chaetognath clade.

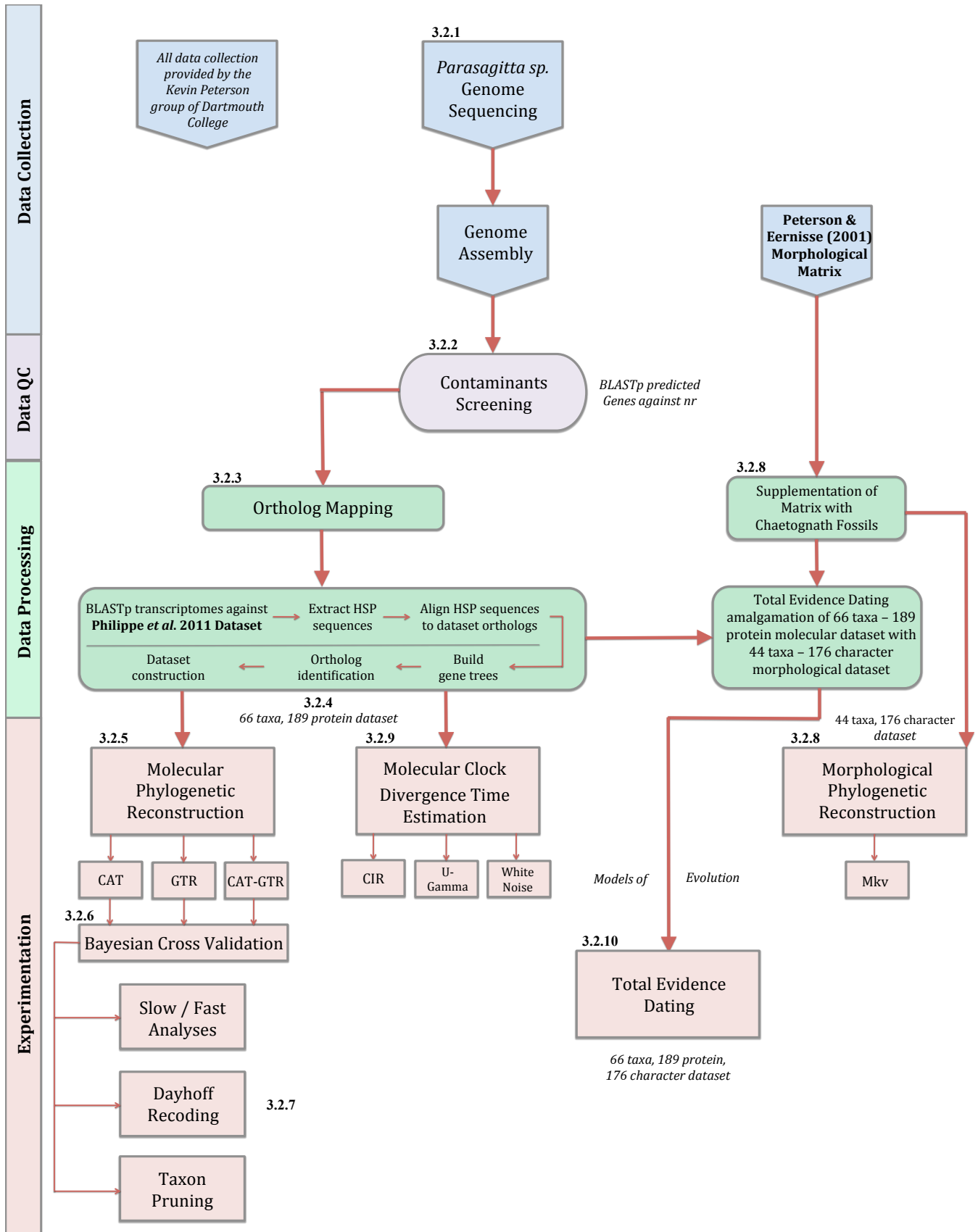
A series of phylogenetic reconstruction experiments were used to resolve the topological placement of the chaetognaths. The relationships between the extant chaetognaths and their most preserved fossils (*Amiskwia* (contested) & *P. spinosa*) (Conway-Morris, 1977 and Vannier *et al.* 2007) were defined by encoding the physical characteristics of both into a large morphological matrix of bilaterians (Peterson & Ernisse, 2001) and a phylogeny was built using this information.

Furthermore, the origin of the chaetognaths is of particular interest because of the old age of their fossils (Vannier *et al.* 2007) and carnivorous morphological characteristics (Bieri, 1959; Tokioka, 1965; Feigenbaum & Maris, 1984) making them some of the first predators in the Animal Kingdom. The Philippe *et al.* (2011b) dataset, enhanced to include chaetognath molecular libraries, was subject to divergence time estimation experiments using molecular clocks in order to find the origins of this clade. In addition to using fossil evidence to specify bounds for molecular clocks (Sanderson, 1996; Thorne *et al.* 1998; Yang & Donoghue, 2016), the ascribed characteristics of ancient chaetognaths from the enhanced morphological dataset (Peterson & Ernisse, 2001) were applied to a total evidence study (Ronquist *et al.* 2012a) in conjunction with the molecular dataset in order to further specify the origins of the Chaetognatha using multiple lines of phylogenetic signal.

## **3.2 Materials and Methods**

A flowchart has been designed to summarize the material and methods used in this chapter [Figure 3.4]. Each numbered step corresponds to a subsection within the materials and methods, describing the particular methodology in detail throughout the distinct processes of data collection, data quality control, data processing, and experimentation.





**Figure 3.4: Flowchart Detailing the Materials and Methods of Chapter 3**

The flowchart consists of four subsections: data collection, data QC, data processing, and experimentation. All methodologies are numbered to correspond to sections detailing them within the materials and methods.

### 3.2.1 Specimen Collection, gDNA Extraction, Sequencing and Assembly

The *Parasagitta sp.* genome was provided by Professor Kevin Peterson and colleagues of Dartmouth College's Biological and Earth Sciences Department. This included the gene predicted file and the translated proteome. Descriptions of the DNA extraction process and sequencing of gDNA important to the preparation of the chaetognath genomic libraries can be found in **Materials and Methods 2.2.2** and **Introduction 1.4.3** respectively.

### 3.2.2 Data Quality Control - Contaminant Screening

As outlined, the *Parasagitta sp.* genome entered the materials and methods pipeline of this project pre-assembled. The raw sequence files were not available thus quality control measurement such as phred scores (encoded into the sequence headers of the raw reads) could not be determined. However it was still possible to perform some quality control measures such as contaminant screening. While FastQC was used to check for contaminant DNA and RNA in all other newly sequenced taxa (see **Material and Methods 2.2.5**) a different approach had to be taken for the assembled and translated chaetognath genome. Instead the *Parasagitta sp.* proteome was compared against the NCBI non-redundant protein database (nr) (Pruitt *et al.* 2005) using BLAST (Altschul *et al.* 1990).

```
$ blastp -query Parasagitta_Proteome -db [nr.00 - nr.66] -out screened-[nr.00 -  
nr.66] -evalue 1e-10 -outfmt 6
```

All *Parasagitta sp.* sequences with a top HSP of a non-protostome or deuterostome species were removed. This particular cut-off was chosen because of the uncertainty of the chaetognath bilaterian affiliation to the Deuterostomia and Protostomia

(Hyman, 1959; Ghirardelli, 1968 & 1981; Ducret, 1978 versus Telford & Holland, 1993; Papillon *et al.* 2003 & 2004; Matus *et al.* 2006; Marlétaz *et al.* 2006 & 2008; Paps *et al.* 2009b; Philippe *et al.* 2011b; Kocot *et al.* 2016).

### 3.2.3 Ortholog Mapping

The ortholog mapping process followed the same protocol as **Materials and Methods 2.2.6** with some slight modifications. Instead of identifying the orthologs from multiple newly sequenced transcriptomes and mapping them onto corresponding orthologs of the Campbell *et al.* (2011) and Philippe *et al.* (2011b) datasets (taxonomically pruned for ecdysozoan relevance). The orthologs were identified from a single taxon of genomic source (*Parasagitta sp.*) and mapped to the full Philippe *et al.* (2011b) dataset. This dataset consists of a large number of deuterostome and protostome taxa. A quality sampling of both groups ensured certainty when placing the Chaetognatha. The superalignment was cleaned using the same Gblock (Castresana, 2000) settings as **Materials and Methods 2.2.6**.

### 3.2.4 Dataset Summary

105 orthologs from the *Parasagitta sp.* genome were mapped to the Philippe dataset (Philippe *et al.* 2011b) by way of the ortholog mapping procedure previously outlined in **Materials and Methods 2.2.6**. This created a new dataset of 66 taxa and 189 genes [Table 3.1].

**Table 3.1: Chaetognath Dataset**

Distribution of orthologs of all 66 taxa in the 189 gene dataset. The newly sequenced *Parasagitta sp.* suffered from foreign DNA contamination explaining a coverage of only 105 out of 189 orthologs from a genomic source.

**Table 3.1: Chaetognath Dataset - 66 Taxa, 189 Genes**

Group	Taxa	Orthologs	Group	Taxa	Orthologs	Group	Taxa	Orthologs	
<i>Porifera</i>	Amphimedo	185	<i>Agnatha</i>	Petromyzon	172	<i>Mollusca</i>	Aplysia	188	
	Leucetta	141		Danio	177		Crassostr	175	
	Oscarella	94		Gallus	174		Euprymna	164	
	Suberites	137		Leucoraja	150		Lottia	181	
<i>Placazoa</i>	Trichoplax	185	<i>Gnathostomata</i>	Xenopus	179		Mytilus	176	
<i>Cnidaria</i>	Acropora	174			Ciona	185	<i>Annelida</i>	Alvinella	164
	Anemonia	158		<i>Tunicata</i>	Brugia	117		Capitella	184
	Cyanea	92				Halocynthi		110	Helobdell
	Hydra	187			Molgula	177		Pomatocero	112
	Hydractini	152	<i>Chaetognatha</i>		Flaccisagitta	24		Tubifex	169
	Nemastoste	188		<b>Parasagitta</b>	<b>105</b>	Echinoder		95	
<i>Xenacoelomorpha</i>	Meara	93		Spadella	112	<i>Scalidorpha</i>	Priapulul	123	
	Nemertoder	111	<i>Platyzoa</i>	Brachionu	181	<i>Onychophora</i>	Euperipato	116	
	Xenoturbel	130			Philodina	124		Acanthoscu	150
	<i>Hemichordata</i>	Ptychodera	114	<i>Bryzoa</i>	Bugula	117	<i>Chelicerata</i>	Anoplodac	120
Saccoglos		188			Cristatell	108			Ixodes
<i>Echinodermata</i>	Holothuria	127	<i>Nemerta</i>	Carinoma	97	<i>Myriapoda</i>	Scutigera	99	
	Patiria	167			Cerebratul		90		Daphnia
	<i>Acoelomorpha</i>	Strongyloc	188	<i>Entoprocta</i>	Symbion	117	<i>Pancrustacea</i>	Gryllus	181
Convolutri		42			Pedicelli	128			Litopenae
Isodiametr		76	<i>Platyhelminthes</i>	Paraplanoc	84			Onychiuru	178
Symsagitti		122			Macrostrom	106			Rhodnius

### 3.2.5 Phylogenetic Reconstruction

Phylogenetic trees were reconstructed using Phylobayes under the CAT, GTR, and CAT-GTR models (Lartillot *et al.* 2009). Two independent chains were run for ~5,000 generations under the CAT and GTR models and for ~2,000 generations under the more computationally demanding CAT-GTR model. A burn in rate of 10% was applied meaning the first 500 (CAT, GTR) / 200 trees (CAT-GTR) were excluded. A lower burn of 10% was chosen instead of the 25% used in the tardigrade dataset because this data had more taxa and took longer for chains to finish a cycle.

```
$ pb -d superalignment_gb75.phy -[model] -s -f superalignment_gb75-[model]-  
chain1
```

```
$ pb -d superalignment_gb75.phy -[model] -s -f superalignment_gb75-[model]-  
chain2
```

```
$ bpcomp -x 10 2 superalignment_gb75-[model]-chain1 superalignment_gb75-  
[model]-chain2 -c 0.1
```

### **3.2.6 Model Testing: Bayesian Cross Validation**

The model testing for the chaetognath dataset followed the same procedure as the BCV for the tardigrade dataset (see **Materials and Methods 2.2.9**). Seqboot (Felsenstein, 1991) reduced the dataset size down to 25% of its original size, as a BCV is a computationally intensive. The reduced dataset was tested with the CAT, GTR, and CAT-GTR models under ten replicates, making nine learning sets and one test set. Cross validated log likelihood scores were calculated for each replicate. The likelihood of the test set was averaged over the parameter values estimated by Phylobayes on the corresponding learning sets. The log of the resulting average likelihood was written out for each replicate file. Summary statistics were then used to compare the suitability of the models between one other. In this case the suitability of the CAT and GTR models are compared to the suitability of the CAT-GTR model. The model with the highest positive score is the best suited to the data, alternatively if the two compared models both return mean negative scores compared to the reference model then the reference model itself is the most suitable (Posada & Buckley, 2004).

### 3.2.7 Chaetognath Dataset: Signal Dissection

A robust approach to phylogenetic reconstruction was necessary to ensure confidence in the placement of the Chaetognatha given the large amount of disagreement on its topological position. Signal dissection protocols applied in chapter 2 (**Materials and Methods 2.2.10**) were also used in the chaetognath experiments: slow / fast analysis (Brinkmann & Philippe, 1999), Dayhoff recoding (Dayhoff *et al.* 1968), and taxon pruning (Aguinaldo *et al.* 1997).

#### 3.2.7.1 Slow / Fast Analysis

The Chaetognath dataset was broken into twelve monophyletic groups and the characters within were ranked based on their rate of change [Table 3.2]. The chaetognath dataset was divided into the same incremental percentages of fastest and slowest evolving sites as the tardigrade dataset. For a full description of the slow / fast technique see **Materials and Methods 2.2.10.1**. The phylogeny of each subset was reconstructed under the CAT-GTR model as a Bayesian cross validation identified it as the best fitting model of the three tested for the data (see **Results 3.3.1.4**).

**Table 3.2: Chaetognath Slow / Fast Dataset**

A description of the groups used for the slow / fast experiment for the chaetognath dataset. The 21,187 character dataset was divided into bins of the 20, 30, & 40 % fastest evolving characters and contrastingly the 80, 70, 60 % slowest evolving characters based on their substitution rate.

<b>Table 3.2: Chaetognath Slow / Fast Dataset</b>							
	Original Dataset	Low Signal Datasets			High Signal Datasets		
		20% Fastest	30% Fastest	40% Fastest	80% Slowest	70% Slowest	60% Slowest
<b>Characters</b>	21,187	4,237	6,356	8,475	16,950	14,831	12,712
Rate	0 - 48	48 - 16	48 - 11	48 - 7	0 - 16	0 - 11	0 - 7
<b>Monophyletic Groups</b>							
Ecdysozoa - Tunicata+Gnathostomata+Agnatha - Deuterostomia - Lophotrochozoa+Platyzoa Echinodermata+Hemichordata+Xenacoelomorpha - Cnidaria - Porifera							

### 3.2.7.2 Dayhoff Recoding

The chaetognath dataset was recoded with three different Dayhoff recoding recipes: Dayhoff-6, Dayhoff-4, and Dayhoff-HP. For a full description of the Dayhoff recoding strategies see **Materials and Methods 2.2.10.2**. A phylogeny for each of the three datasets was built using the CAT-GTR model.

### 3.2.7.3 Taxon Pruning

The Acoelomorpha are the fastest evolving clade in the chaetognath datasets by a considerable margin, *Isodiametr sp.*, *Symsagitti sp.* and *Convolutri sp.* were pruned from this clade. Other long branched taxa removed included *Meara sp.* of the Xenacoelomorpha, *Philodina sp.* of the Platyzoa, and *Macrostrom sp.* of the Lophotrochozoa. The phylogenetic trees for the taxon-pruned datasets were reconstructed using Phylobayes under the CAT-GTR model.

### 3.2.8 Morphological Dataset

The disparity between molecular clock divergence dates of the extant chaetognaths and the chaetognath fossils (see **Results 3.3.2**) suggests there is a stem lineage relationship as opposed to an encompassing crown group. Many molecular experiments are restricted with their usage of fossil data for evolution studies; solely using the fossil record as calibration points for clocks (Yang & Donoghue, 2016). Looking at the chaetognaths from a morphological point of view can garner further insight into the relationships between the fossils and extant species.

The chaetognath fossils were encoded into the morphological matrix from Kevin Peterson of Dartmouth College (Peterson & Ernisse, 2001) based on the ascribing expertise of Jakob Vinther of the University of Bristol and with the aid of the known lineage characteristics of the chaetognaths outlined by (Tokioka, 1965). The morphological phylogeny was reconstructed with Mr. Bayes (Ronquist *et al.* 2012b) using the Mkv model under default settings. See [Supplementary Material 3.2](#) for the full matrix and encoded nexus commands.

### 3.2.9 Divergence Time Estimation

The molecular clock function of the Phylobayes package (Lartillot *et al.* 2009) was used for divergence time estimation experiments. Initially two models of evolution were chosen: the correlated CIR and uncorrelated Gamma (U-GAMMA). However, given the surprising results from the molecular clock experiments (**Results 3.3.2.1**) it was advisable to run the experiment under another model, white-noise, for further comparisons in order to establish whether there were biases in the previously used models skewing the divergence time estimations. Each clock was run under relaxed settings, a soft maximum age for the root of the tree (833 MYA) with a standard deviation equal to the mean (552.8). Since a different set of taxa comprises the chaetognath dataset compared to that of the tardigrade dataset (mainly a greater number of lophotrochozoans and deuterostomes) a new set of calibration points was required to for divergence bounds of the lineages studied. See [Table 3.3](#) for these calibration points. Chains were considered converged when the difference between their corresponding nodes was no larger than 10 MY.



```
$ pb -d superalignment_gb75.phy -T CATGTR_concensus.tre -r outgroup -cal
calibrations -[model] -rp 833 552.8 superalignment_gb75-[model]-chain1
```

```
$ pb -d superalignment_gb75.phy -T CATGTR_concensus.tre -r outgroup -cal
calibrations -[model] -rp 833 552.8 superalignment_gb75-[model]-chain2
```

```
$ readdiv -x 300 superalignment_gb75-[model]-chain-1
```

```
$ readdiv -x 300 superalignment_gb75-[model]-chain-2
```

**Table 3.3: Chaetognath Study Molecular Clock Calibrations**

The 17 calibration points used in the divergence time estimation including their upper and lower bounds. Reference points were taken from Benton *et al.* (2015).

**Table 3.3: Chaetognath Study Molecular Clock Calibrations**

	<i>Taxa</i>		<i>Bounds (MYA)</i>			<i>Taxa</i>		<i>Bounds (MYA)</i>					
1	Hydractini	-	Ixodes	636.1	-	552.85	10	Philodina	-	Ixodes	636.1	-	552.85
2	Hydractini	-	Acropora	636.1	-	529	11	Euprymna	-	Aplysia	549	-	534
3	Meara	-	Ixodes	636.1	-	552.85	12	Euprymna	-	Alvinella	636.1	-	552.85
4	Xenoturbel	-	Branchios	636.1	-	514	13	Pomatocero	-	Alvinella	636.1	-	-1
5	Saccoglos	-	Holothuria	636.1	-	515.5	14	Priapulid	-	Ixodes	636.1	-	528.82
6	Saccoglos	-	Ptychodera	636.1	-	504.5	15	Euperipato	-	Ixodes	636.1	-	528.82
7	Strongyloc	-	Holothuria	549	-	509	16	Litopenae	-	Ixodes	636.1	-	514
8	Petromyzon	-	Branchios	636.1	-	514	17	Leucetta	-	Suberites	713	-	-1
9	Leucoraja	-	Danio	457.5	-	420.7							

### 3.2.10 Total Evidence Dating

The total evidence dating dataset was generated by combining the molecular dataset for the phylogenetic and divergence time estimation experiments and the morphological dataset, which was useful in placing the chaetognath fossils. Concatenating these datasets generated a TED matrix of 68 taxa and 21,397 characters [Supplementary Material 3.3]. The best-preserved chaetognath fossil, *P. spinosa* (Vannier *et al.* 2007) and the disputed putative chaetognath fossil *Amiskwia*

*sp.* (Conway-Morris, 1977) were included in the analysis. The rooted tree used in the analysis mirrored the topology of the phylogeny generated by the best fitting model for the dataset by way of a BCV (CAT-GTR). All taxa positions were fixed based on this tree topology with the exception of the fossils, which were allowed to freely move positions based solely on the information encoded within the TED dataset. As per TED recommendations (Ronquist *et al.* 2012a), three separate experiments were setup: a non-clock (no parameters), strict clock (parameters and bounds), and IGR (divergence time estimation model, relevant to both morphology and molecules).

Mr. Bayes (Ronquist *et al.* 2012b) was used to run the TED experiments. The two independent MCMC chains of Mr. Bayes TED runs require a large number of generations, partially because of the large “warm up time” required for the chains, but also because of the relatively diverse form of data types used in the study in comparison to standard divergence time estimation studies which adds to the complexity of the analysis. As a consequence, for a dataset of this size a generation number of 50,000,000 was required as implemented by Ronquist *et al.’s* (2012) TED study on the Hymenoptera. The TED dataset was run under two differing scenarios: an un-calibrated non-clock analysis where the data is examined without settings or models of evolution (the result of which allows one to estimate the rate prior and approximate rate of substitutions across the tree) and a relaxed clock IGR model analysis with defined fossil calibrations (all fossils assigned to nodes with the exception of the chaetognath fossils), a model variance of 37.12, whose rate prior and substitution rate needed to be calculated by the former clock settings. See [Supplementary Material 3.3](#) for the full nexus commands.

## 3.3 Results

### 3.3.1 Chaetognath Phylogeny

#### 3.3.1.1 CAT Model

The CAT model placed the chaetognaths within the protostomes as sister group to all other lophotrochozoans (PP 0.97), diverging before most of the other phyla within the Protostomia [Figure 3.5]. Two MCMC chains were run in parallel for over 8,400 generations before converging with a maximum difference of 0.2, well within acceptable levels of maximum posterior probability discrepancies based on the Phylobayes manual (Lartillot *et al.* 2009).

The newly sequenced *Parasagitta sp.* is grouped along side *Flaccisagitta sp.* (PP = 1.0) with *Spadella sp.* as outgroup of extant chaetognaths with strong support (PP = 0.99).

Interestingly the Mandibulata is not recovered within the Arthropoda instead we see the grouping of the chelicerates with the myriapods: the Myriochelata with a strong PP of 1.0.

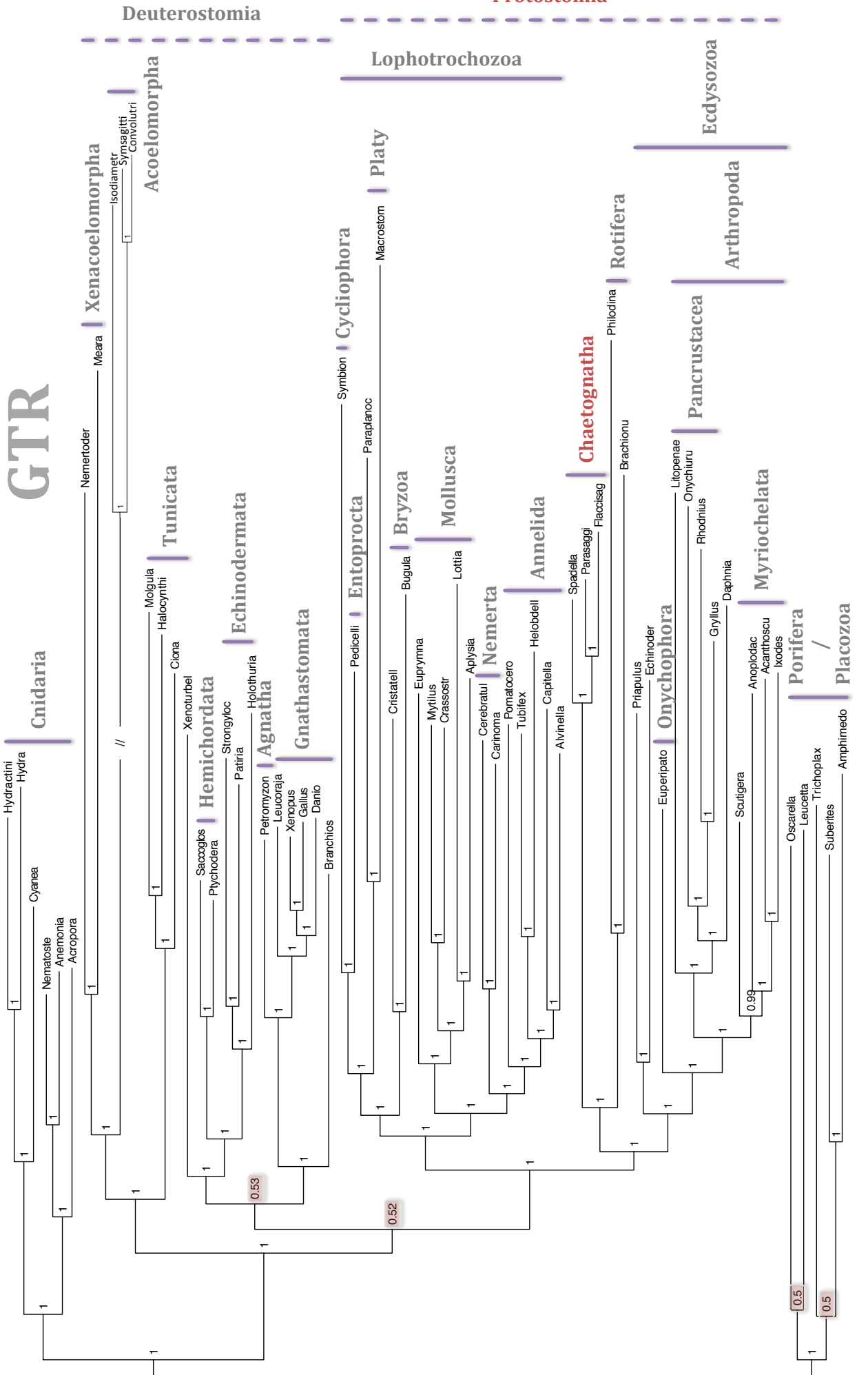
The most uncertain placement is found within the Annelida where *Capitella sp.* and *Alvinella sp.* are grouped with a PP of 0.5, low PP scores are also seen in the deep nodes amongst the Porifera.



### 3.3.1.2 GTR Model

The GTR model experienced difficulty converging on a definitive phylogeny, with the major discrepancies between the chains revolving around the nodes of the Deuterostomia ancestor and the ancestral node of the deuterostome / protostome divergence with poor PPs of 0.53 and 0.52 respectively.

The GTR model returned with dramatically different results to the other models: grouping the Chaetognatha with the Rotifera, which in turn is sister to the Ecdysozoa [Figure 3.6]. Since the chains could not sufficiently converge, with a maximum difference between the two of 1.0, these results have been rejected.



0.2

Figure 3.6: Chaetognatha GTR Phylogeny

The GTR model places the Chaetognatha within the protostomes with a posterior probability of 0.97. The chaetognaths are grouped with the Rotifera with a PP of 1.0. Platyzoa not recovered

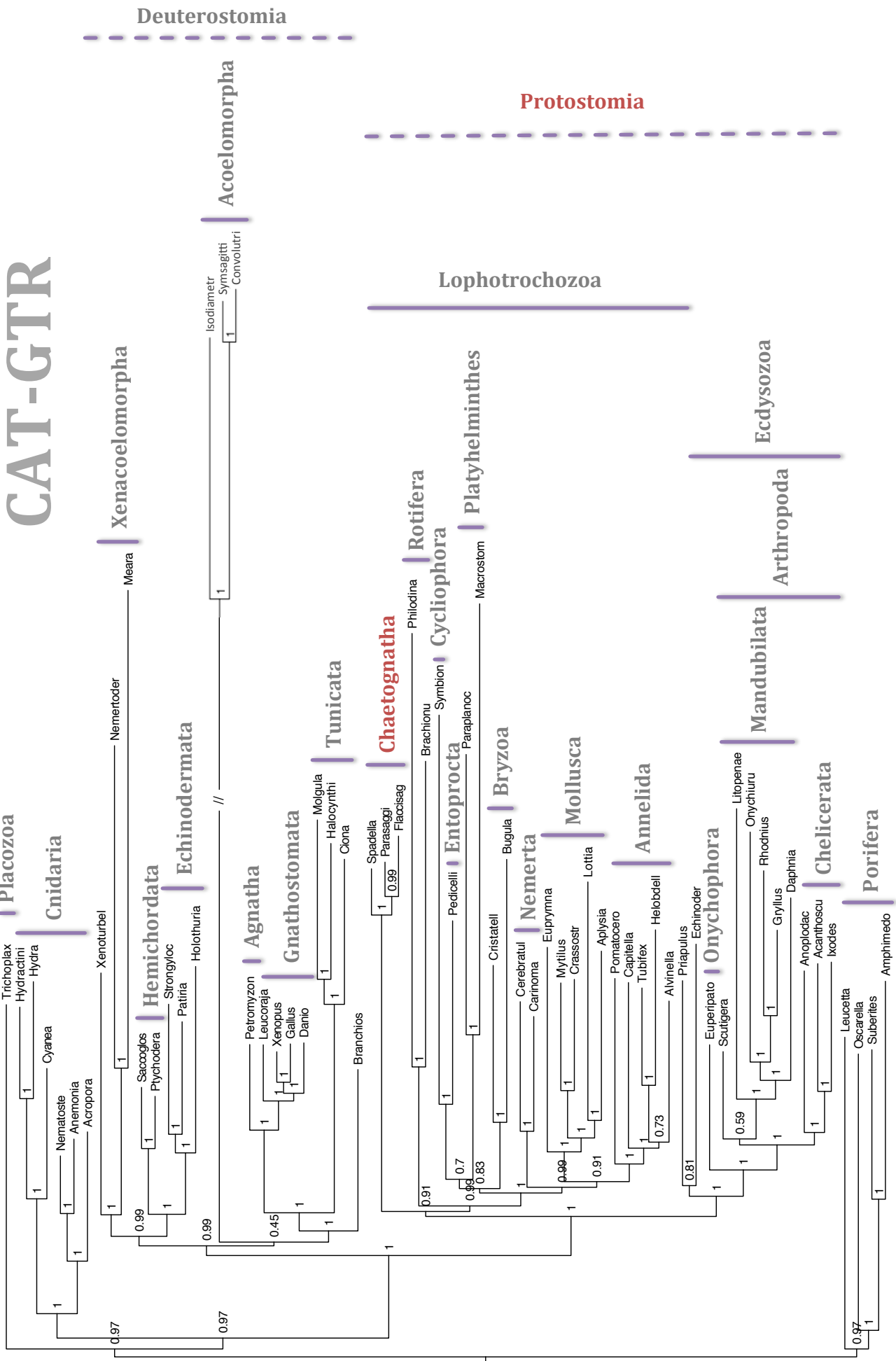
### 3.3.1.3 CAT-GTR Model

The CAT-GTR model returned with the same phylogeny for the Chaetognatha as the CAT model: nested within the Protostomia as basal lophotrochozoans, with a PP of 0.9 [Figure 3.7]. Again the Platyzoa, a monophyly of the Rotifera, Cycliophora, and Platyhelminthes is not recovered.

The CAT-GTR model returns the Mandibulata and not the Myriochelata inside the Arthropoda albeit with low PP support (0.59). The other main difference between the models is found amongst the lophotrochozoans: the CAT-GTR model does not group the Annelids with the Nemerta and instead has the former sister to the Mollusca.

Due to computational limitations, the MCM chains could only be run for roughly half the generations of the CAT model, just under 4,400. Convergence issues across the CAT-GTR generated tree are seen in the Mandibulata (PP = 0.5) and the Acoelomorpha (PP = 0.46), however we see much stronger support in the deeper nodes compared to the CAT model particularly among the Cnidaria (PP = 1.0), Placozoa (PP = 0.97), and Porifera (PP = 0.97). It is entirely possible we would have seen higher support values for the Mandibulata if we had the computational resources required to run the chains for longer.

# CAT-GTR



0.4

**Figure 3.7: Chaetognatha CAT-GTR Phylogeny**  
 The Chaetognatha are protostomes, basal lophotrochozoans.  
 The maximum difference between the two chains was 0.4  
 and the placement of the Chaetognatha is with a posterior probability of 0.9



### 3.3.1.4 Best Fitting Model for the Dataset

The results of the Bayesian Cross Validation showed that the CAT-GTR model was clearly the best fit for the data, followed by the CAT model, with GTR being the poorest fit by a significant margin [Table 3.4]. Consequently, the CAT and GTR topologies are rejected and thus we conclude the CAT-GTR hypothesis: the Chaetognatha were the first lophotrochozoans to diverge.

**Table 3.4: Chaetognath Dataset: BCV**

The CAT and GTR models are a statistically poor choice compared to that of the CAT-GTR model for analyses with this dataset. The SD spread does not conflict with the confidence of the mean score values.

**Table 3.4: Chaetognath Dataset BCV**

Models	Reference Model	Mean Score	Standard Deviation	Times Model is Most Suitable
CAT	CAT-GTR	-151.44	+/- 52.171	0
GTR	CAT-GTR	-946.6	+/-71.615	0

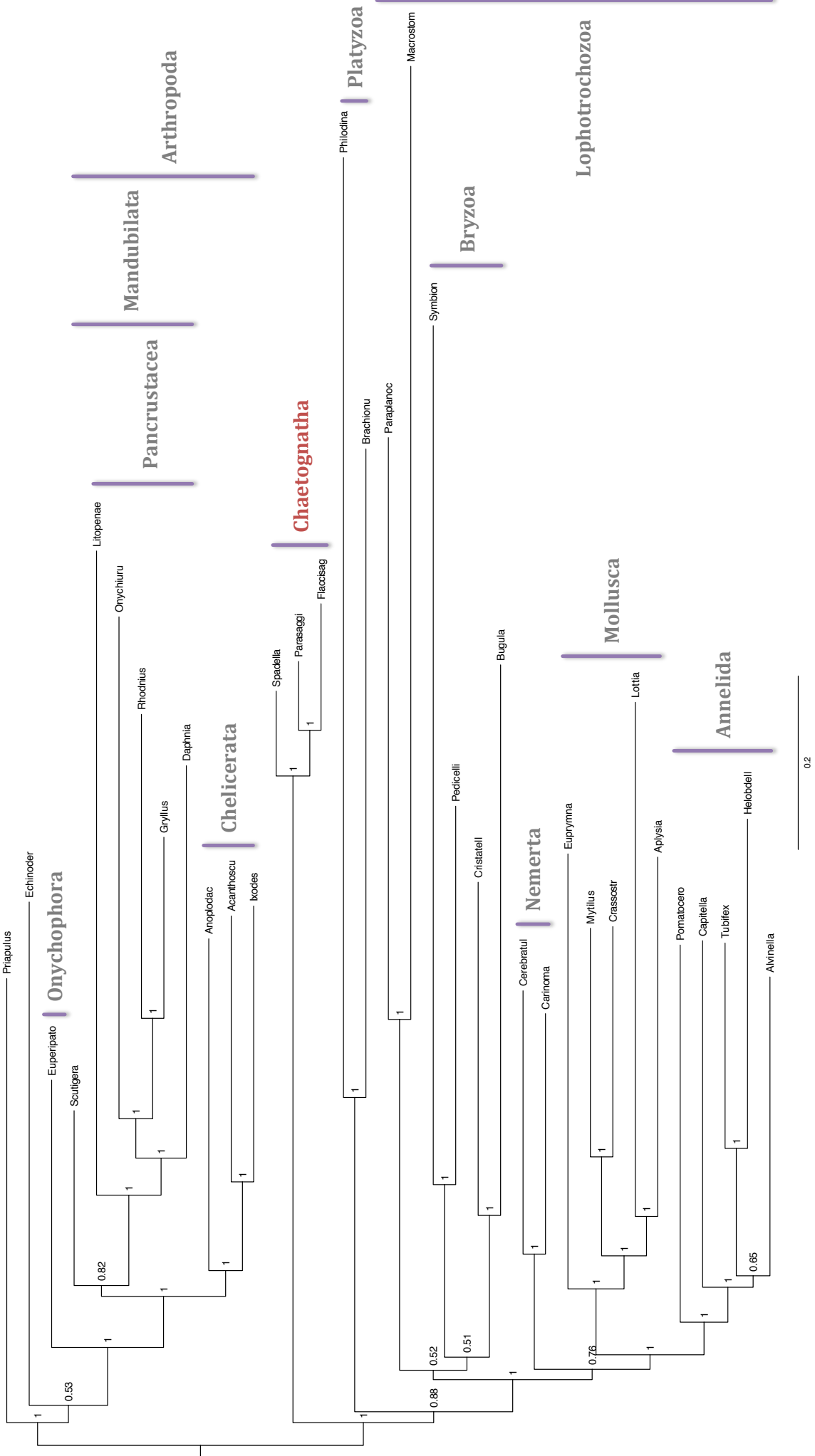
### 3.3.1.5 Protostome Phylogeny

Cautious of the dangers of deep node phylogenetics, and given that all tested models placed the chaetognaths in the protostomes, it was important to reconstruct a more focused dataset containing just the protostomes. This would reveal if the deep nodes in the tree were skewing the data. Generally speaking, the deeper the nodes in phylogenetic reconstruction (i.e. the older the taxa being studied) the more time for the DNA / peptide code to change, causing site saturation, thus skewing or even masking the underlying phylogenetic signal.

Phylogenetic reconstruction was run under the CAT-GTR model since it was deemed best fitting by the BCV. Two independent chains were run for 4,500 generations returning a max difference between the two of 0.2.

The phylogeny of the Chaetognatha does not change to that of the CAT-GTR tree in the smaller protostome dataset (PP 1.0), in fact we see additional support for the Mandibulata over the Myriochelata with a PP value of 0.8 [Figure 3.8].

# Protostomia CAT-GTR



**Figure 3.8: Protostome CAT-GTR Phylogeny**  
 A phylogenetic reconstruction of just the protostomes in the dataset.  
 The Chaetognath relationship remains unchanged seeing full PP support of 1.0

### 3.1.3.6 Slow / Fast Analyses


Previous studies on the chaetognaths have been suspected of being influenced by long branch attraction (LBA) (Paps *et al.* 2009b) therefore we felt it was pertinent to take a closer look at the phylogenetic signal in the dataset via the slow / fast method (Brinkmann & Philippe, 1999).

The slow / fast technique allows one to divide the characters of a dataset into different categories of evolution rate. The dataset was divided into saturated character reconstructions: 20%, 30%, and 40% fastest characters and compared to the contrastingly stable character reconstructions: 80%, 70%, and 60% slowest characters. All datasets were reconstructed under the best fitting model for the data: CAT-GTR.

**Table 3.5: Chaetognath Slow / Fast Results**

The subset of characters, chaetognath position, posterior probability (PP) of that node, and the convergence of the two independent chains are provided.

<b>Table 3.5: Chaetognath Slow / Fast Results</b>			
<i>Stable Character Reconstructions</i>			
<b>Dataset</b>	<b>Chaetognath Position</b>	<b>PP</b>	<b>Convergence</b>
80% slowest	sister to the ecdysozoa	0.82	0.4
70% slowest	sister to all other lophotrochozoans	0.43	0.26
60% slowest	sister to all other lophotrochozoans	0.57	0.29
<i>Saturated Character Reconstructions</i>			
<b>Dataset</b>	<b>Chaetognath Position</b>	<b>PP</b>	<b>Convergence</b>
40% fastest	sister, with Acoelomorpha. to all other protostomes	0.51	1
30% fastest	sister to all other protostomes	0.5	1
20% fastest	sister to all other lophotrochozoans	0.71	0.46



Character Saturation

### 3.3.1.7 Dayhoff Recoding

The dataset was recoded under the three standard Dayhoff recipes for amino acid biochemical similarity: Dayhoff 4, 6, and HP.

The dataset recoded with the Dayhoff 4 & 6 substitution matrices returned the same Chaetognath positioning as the CAT-GTR model. Topology was consistent with the CAT-GTR tree with small exceptions such as the arrangements of the Annelida, Mollusca, Bryzoa, and Gnathostomata, however the position of these clades remained constant. The major difference between the original and recoded datasets was the arrangements of the Arthropoda, finding a sister grouping of the Pancrustacea with the Chelicerata and with Myriapoda as sister group [[Supplementary Material 3.4](#)].

The HP recoded version of the dataset experienced convergence issues but recovered the rotifers as outgroup to the sister grouping of the Chaetognatha and Lophotrochozoa. The placement of the Chaetognatha suffers from low support however (PP = 0.42). Other notable differences to the original CAT-GTR topology include the grouping of the Acoelomorpha with the Echinodermata albeit also with a low posterior probability (0.52) [[Supplementary Material 3.4](#)].

### **3.3.1.8 Taxon Pruning**

The longest branched lineages, particularly the entire Acoelomorpha, were removed to see if reducing rapidly evolving taxa changed the topology of the tree under the CAT-GTR model. Drastic alterations in clade topology could be an indication of LBA influence in the dataset.

The taxon pruning experiment recovered the Chaetognatha as basal lophotrochozoans with a PP of 1.0, agreeing with the best fitting model for the data (CAT-GTR). This suggests that the longest branched lineages in the dataset were not influencing the positioning of the Chaetognatha [[Supplementary Material 3.5](#)].

### **3.3.1.9 Morphological Phylogeny**

The adapted dataset from Peterson to include chaetognath characteristics and fossils was run under the Mkv model in Mr. Bayes (Ronquist *et al.* 2012b). Results show a stem lineage relationship of the chaetognath fossils to that of extant chaetognaths [[Figure 3.9](#)].



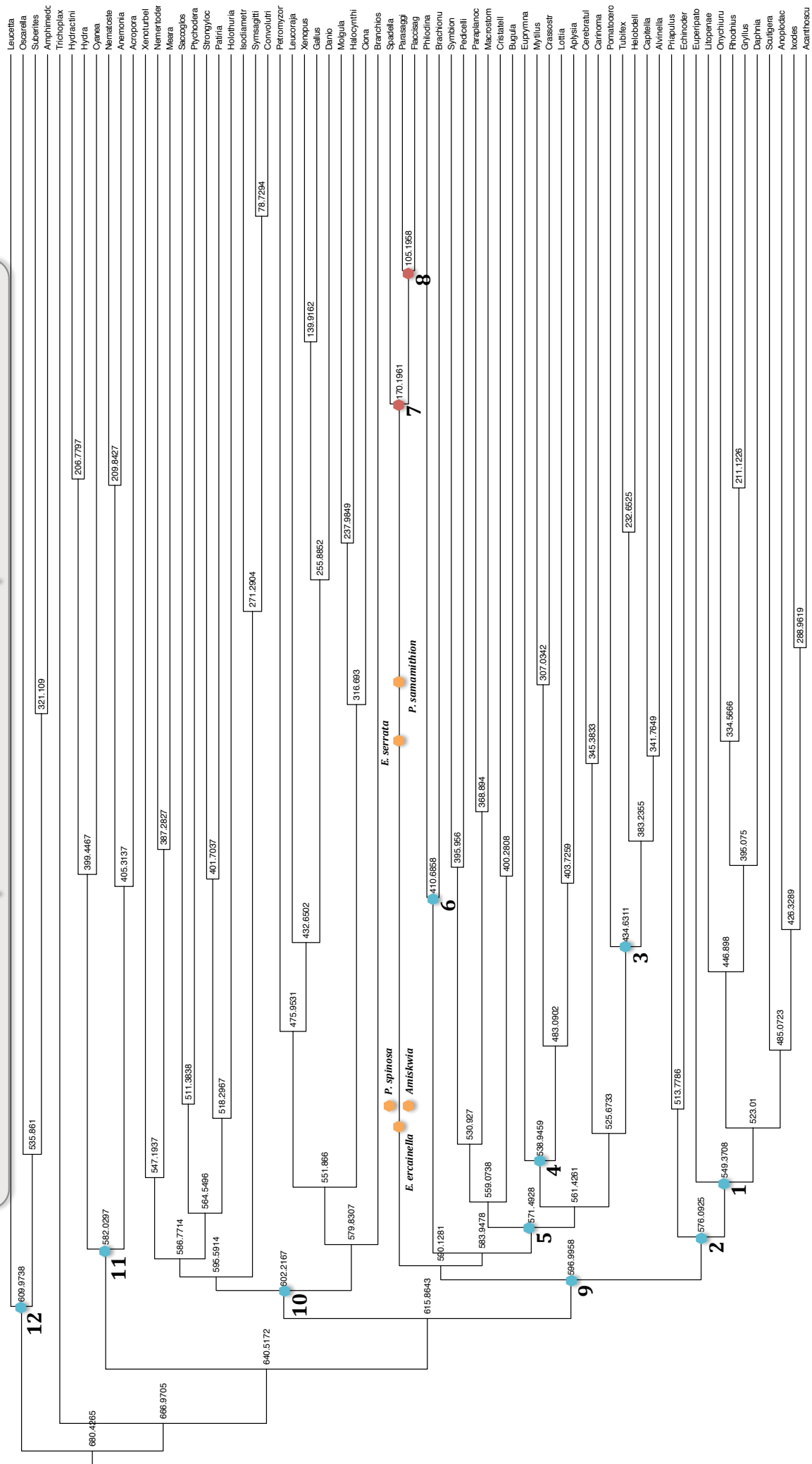
### 3.3.2 Chaetognath Divergence Time Estimation

#### 3.3.2.1 Chaetognatha Origins Under Clock Models

The Chaetognatha split from the rest of the Lophotrochozoa in the Ediacaran, 590 MYA [Figure 3.10]. Interestingly, extant chaetognaths are much younger with *Spadella* splitting from *Parasagitta* and *Flaccisagitta* in the Mid-Jurassic just 170 - 203 MYA and the latter group emanating from the Lower Cretaceous 105 - 139 MYA. This is in stark contrast to the chaetognath fossils found in geological time periods ranging from the Lower Cambrian (*E. ercainella*) to the Upper Carboniferous (*P. sammithion*). The three models of evolution used, CIR, UGAMMA, and white noise, returned broadly similar results with the largest discrepancy falling between the CIR and UGAMMA Annelida origins (Delta 110 MYA). In general, the CIR model generated younger dates across the nodes of the tree.



Group	$\Delta$ U-gamma	$\Delta$ wht nse	Group	$\Delta$ U-gamma	$\Delta$ wht nse	Group	$\Delta$ U-gamma	$\Delta$ wht nse
1 Arthropoda	-3 MY	+13 MY	5 Lophotroch	+5 MY	+12 MY	9 Protostomia	-4 MY	+20 MY
2 Ecdysozoa	0 MY	+16 MY	6 Platyzoa	+76 MY	+56 MY	10 Deuterostomia	+9 MY	+12 MY
3 Annelida	+110 MY	+9 MY	7 Chaetognatha	+17 MY	+33 MY	11 Cnidaria	+64 MY	-6 MY
4 Mollusca	-1 MY	+1 MY	8 Para - Flac	+34 MY	+24 MY	12 Porifera	+70 MY	+43 MY



**Fig 3.10: Chaetognath Molecular Clock**

The results from the U-gamma and white noise models compared to CIR in the table. Major groups are highlighted 1-12 with the chaetognaths in red. The chaetognath fossils have been included in the timeline as orange markers.

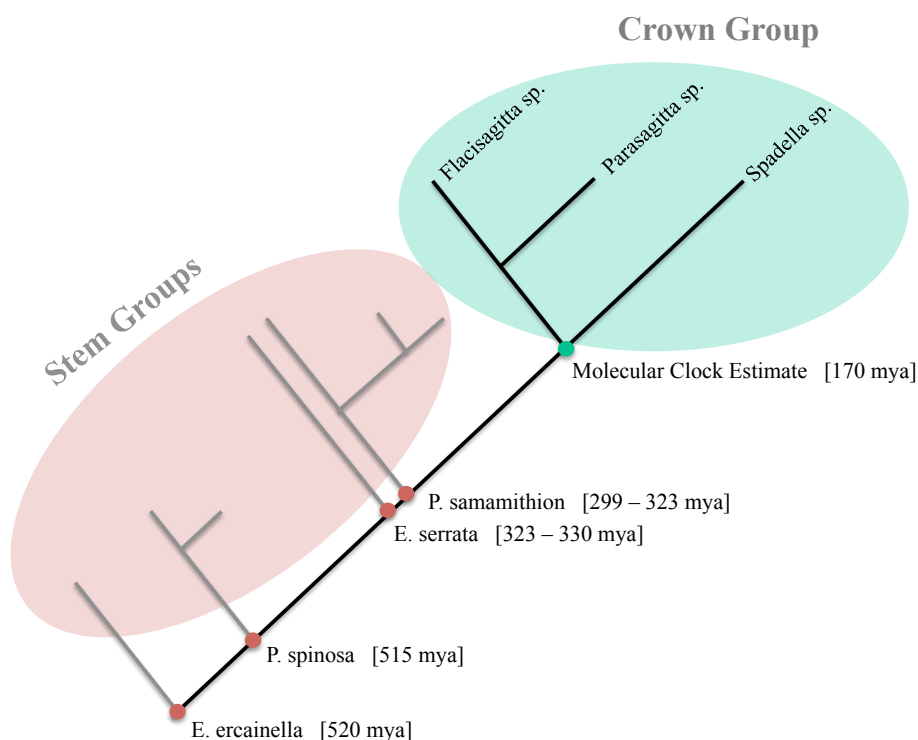
### **3.3.2.2 Total Evidence Dating**

The TED analyses proved not to be feasible due to the large number of MCMC generations required for convergence, 50,000,000 as recommended by Ronquist *et al.* (2012a), and the intense computational resources required. Time estimations based on a small number of generations suggest that it would take many years to match the generation quota required to reach that of the robustness and integrity of the Ronquist Hymenoptera study. This is an unfortunate case of resource limitations sometimes encountered when running complex analyses with phylogenomic sized datasets.

## 3.4 Discussion

### 3.4.1 Disparity Between Rocks and Clocks

It is clear, from every molecular clock experiment undertaken in this study, that there is a large disparity between the dating of the extant chaetognaths, represented by molecular data, and the extinct fossils. This disagreement ranges from 130 MY (extant lineages – *P. samamithion*) to 350 MY (extant lineages – *E. ercainella*). An explanation for such differing time estimations from representatives of the same clade of animals is the existence of several extinct stem groups for which the fossils are a much closer representative of [Figure 3.11].



**Figure 3.11: Crown Group Lineages and Stem Group Fossils**

The disparity between rocks and clocks suggests the fossils represent extinct chaetognath stem groups. The youngest fossil representative for extant lineages (the flacisagitta – parasagitta – spadella ancestor) has not been discovered or may not exist.

### 3.4.2 Total Evidence Dating

The importance of collaboration between morphological and molecular divergence time estimations is becoming more apparent after each evolutionary study. Many pioneering scientists in the field advocate such an approach (dos Reis *et al.* 2016 and Yang & Donoghue, 2016) that exploits the strengths of both data sources while minimizing their weaknesses. From a molecular point of view this involves the application of fossils, not just as minimal time constraints, but taking advantage of the evolutionary signal found within their morphological characteristics and including them with the molecular matrix. Inclusive methods such as this, which aim to maximize the phylogenetic signal available via multiple morphological and molecular data sources, have been proposed by (Ronquist *et al.* 2012a). Total evidence dating (TED) can achieve this approach and has a secondary use in its ability to apply morphological information in assigning uncertain fossils to clades under a fixed topology. TED would have been incredibly useful in our study of Chaetognatha divergence time estimation as it may have reduced some of the ambiguity of the affinity of *Amiskwia* and addressed the disparity between the fossil ages and the dating of the extant chaetognaths lineages.

While these novel methods hold promise, the computational resources required for the joining of large-scale morphological matrices and phylogenomic sized datasets is considerable. Early estimations of the TED dataset we generated, consisting of the 66 taxa, 21,187 character chaetognath dataset and the Peterson 42 taxa 166 character morphological dataset, indicated a three year processing time with the resources available. Such large-scale tests are only within reach to those that have formidable computational power, but moving forward, with the exciting rate of

technological advancement, such integrative genomic-scale methods will soon be possible for all.

### 3.4.3 Disagreement Amongst Signal Dissection Experiments

As seen in **Results 3.3.1.6**, there was disagreement between the outcome of the slow / fast experiments and other signal dissection techniques. Most of the slowest subsets returned the same chaetognath positioning as the CAT-GTR model, however the subset with the 20% fastest characters removed did not. This may be in relation to how each of them operate vis-à-vis disruption of phylogenetic signal. All three techniques aim to root out saturated characters in the dataset with the cost of losing some phylogenetic signal, but they do so in different ways.

Taxon pruning (Aguinaldo *et al.* 1997) simply removes long branched taxa in the dataset in efforts to prevent them from drawing other similarly long branches to them. The underlying phylogenetic signal is disrupted but remains ordered, i.e. the concatenated genes in the dataset remain intact and are not changed or mixed up. Similarly the Dayhoff recoding preserves the order of the characters and just substitutes them for simplified versions based on their composition. The Dayhoff method (Dayhoff *et al.* 1968) uses various recipes of differing simplicity, each at the cost of more phylogenetic signal. The slow fast technique however not only disrupts the phylogenetic signal of the dataset but it also disrupts the order of the characters too. This may be part of the reason for the differing results.

The slow / fast technique (Brinkmann & Philippe, 1999) works on the principle of ranking characters in a dataset by their rates of change and then excluding certain groups with a similar rate of change from datasets to see how the phylogenetic tree is

reconstructed without them. Common approaches are to remove the fastest evolving characters to see how the tree looks with less saturated characters or inversely, removal of the most stable characters to see how what the phylogeny looks like under conditions strongly susceptible to stochastic errors. Both of these methods however completely reorder the composition of the concatenated genes in the dataset to the point where their biological relevancy is arguable. This level of character disruption may create too much noise which could explain why only one of six signal dissection experiments reached that of acceptable convergence of independent Markov chains.

### 3.5 Conclusions

Based on the results from the evolutionary model that fits the data best: the CAT-GTR phylogenetic reconstruction [Figure 3.7], in conjunction with supporting evidence from the dataset focused on just the Protostomia [Figure 3.8], we conclude that the chaetognaths are sister to the Lophotrochozoa making them some of the most ancient protostomes.

The results from the signal dissection experiments mostly correlated with our proposed phylogenetic scenario with the Dayhoff recoding and taxon pruning experiments in agreement with the CAT-GTR phylogeny. However some of the signal dissection experiments conflict with the rest of our results. The Chaetognaths, along with the Ecdysozoa, were the earliest protostome groups to diverge likely explaining their puzzling mosaic features and perhaps explaining as to why they have a deuterostome-like development that has confused evolutionary biologists for years. The placement of the Chaetognatha correlates with two of the numerous previous studies (Matus, 2006 and Kocot *et al.* 2016), rejecting the other chaetognath topologies: stem group bilaterians (Telford *et al.* 1993 and Papillon *et al.* 2003), young members of the Lophotrochozoa (Papillon *et al.* 2004), sister to both the lophotrochozoans and ecdysozoans (Martelaz *et al.* 2006 & 2008), sister group to the lophotrochozoans and platyzoans (Philippe *et al.* 2011b), and placement within the ecdysozoans (Paps *et al.* 2009b). Furthermore, a monophyletic Platyzoa was not recovered as only two of its proposed members shared a last common ancestor: the rotifers and cycliophorans. However the poor sampling of platyzoans taxa in the dataset make it difficult to conclusively rule out the existence of the group.

The difficult evolutionary question to resolve is how the chaetognaths possess deuterostome-like traits. As each of the chaetognaths deuterostome-like characteristics are taken into account: the tripartite body plan, post anal tail, radial intermediate cleavage, and formation of the anus from the blastopore (Matus, 2006) it becomes less and less likely that they are all individual products of homoplastic apomorphies manifesting themselves in a staggeringly coincidental amount of convergent evolution from opposing sides of the bilaterian tree. Therefore the most parsimonious explanation is that the lophotrochozoan ancestor possessed these traits while all diverging lophotrochozoan lineages lost them with the exception of the oldest: the chaetognaths who retained these ancestral lophotrochozoan pleisomorphic characteristics. Under this scenario the protostome ancestor would have had developmental patterns and characteristics similar to that of deuterostomes but these traits were lost after the divergence of the chaetognaths right before the emergence of the remaining lophotrochozoans. However, when the phylogeny of the entire Protostomia is taken into consideration then these traits must have also been lost in the Ecdysozoa but independently to that of the trait loss in the Lophotrochozoa for this scenario to hold true. Particularly since molecular clock experiments date the origins of the Ecdysozoa deeper in time to that of both their sister protostomes and the chaetognath fossils (**Results 3.3.2.1**). This is provided that the origin of extant chaetognaths does not pre-dates the fossils, a fair assumption given the result of our molecular clock experiments (**Results 3.3.2.1**).

A series of deuterostome-like apomorphic trait gains seems unlikely to have occurred along the chaetognath lineage but separate independent losses of these exact characteristics in the lophotrochozoan and ecdysozoan lineages are also unlikely.



The answer to this question is important, not just for our understanding of chaetognath evolution but to our understanding of the protostome ancestor. If the first scenario holds true then the protostome ancestor was much more similar to contemporary deuterostomes than previously thought, mirroring deuterostome developmental and body plan characteristics. If the latter scenario is correct then the evolution of the chaetognaths becomes more bizarre as they would have effectively accrued a series of deuterostome defining characteristics soon after their divergence, made evident by their fossil record (Chen & Huang, 2002), followed by roughly half a billion years of stagnant evolution given the remarkable morphological similarity of extant chaetognaths to their 500 MYA fossils (Chen & Huang, 2002 and Doguzhaeva *et al.* 2002). The position of the chaetognaths within the Protostomia, and the known chaetognath fossils are offset by the surprisingly young origin estimates for living chaetognaths. The fossil record places the minimum origin of the Chaetognatha in the Cambrian, 520 MYA (Chen & Huang, 2002). However our divergence time estimations using new phylogenomic data dates the origin of extant chaetognaths much later, between 170 and 203 MYA [Figure 3.10]. We conclude that the disparity between these dates is due to the fossils being representative of extinct stem group chaetognaths and not the living crown group represented by the molecular data [Figure 3.9]. This would mean that the Chaetognath have been part of a near total extinction event explaining not only the disparity between the extant lineages and the oldest fossil but the long chaetognath branch in the lophotrochozoan tree. The above scenario suggests the chaetognaths have been witness to a very eventful evolutionary history. 520 MYA they were some of the first predators to appear in the oceans, with very little competition due to the low oxygen levels not suitable for competing carnivorous lifestyles (Sperling *et al.* 2013). The redox of the oceans led to an

increase in predation and a major change in the foodweb with increased predators roaming the seas. The influx of predator – prey scenarios led to diversification in body plans in attempts to adapt to a much more dangerous environment which is seen in the sudden radiation and diversification of animal lineages post Cambrian.

It is difficult to speculate on the circumstances of the extinction events of many of the chaetognath lineages. With the morphology of extant chaetognaths being virtually identical to their ancestors (Szaniawski, 2005), indicating few morphological adaptations over the last 500 million years, we can compare the very different ecosystems of the ancient chaetognaths to the living members to form some idea as to how these extinction event may have happened. Ancient chaetognaths were some of the first predators in the ocean, with very little competition, therefore they were likely top of the food chain in the Cambrian when the radiation of animal lineages known today was in its infancy. Hundreds of millions of years of evolutionary adaptations across the tree of life later and the contemporary chaetognaths find themselves in a drastically different and more dangerous environment. The oceans are now filled with predators far more numerous and dangerous to the morphologically preserved chaetognaths, inverting their position in the food chain has them serving as the chief component of plankton – the common food source for many animals in the sea (Bieri, 1959). Such an inversion of predation roles may explain why many of the chaetognath lineages are now extinct. Given this observation perhaps the question should not be why certain chaetognaths went extinct, but considering how little they have changed over the last 500 million years, why the extant chaetognath lineages exist today? These results may be an early indication that the chaetognaths, the primary component of plankton essential to the oceans foodweb, are under threat of extinction in the near future.

### 4.1 Introduction

#### 4.1.1 Arthropod Terrestrialization

Terrestrialization, the colonization of land from sea, was a major evolutionary event with lasting impact on the life of this planet, changing the landscape of contemporary terrain ecosystems and the carbon cycle (Kenrick *et al.* 2012).

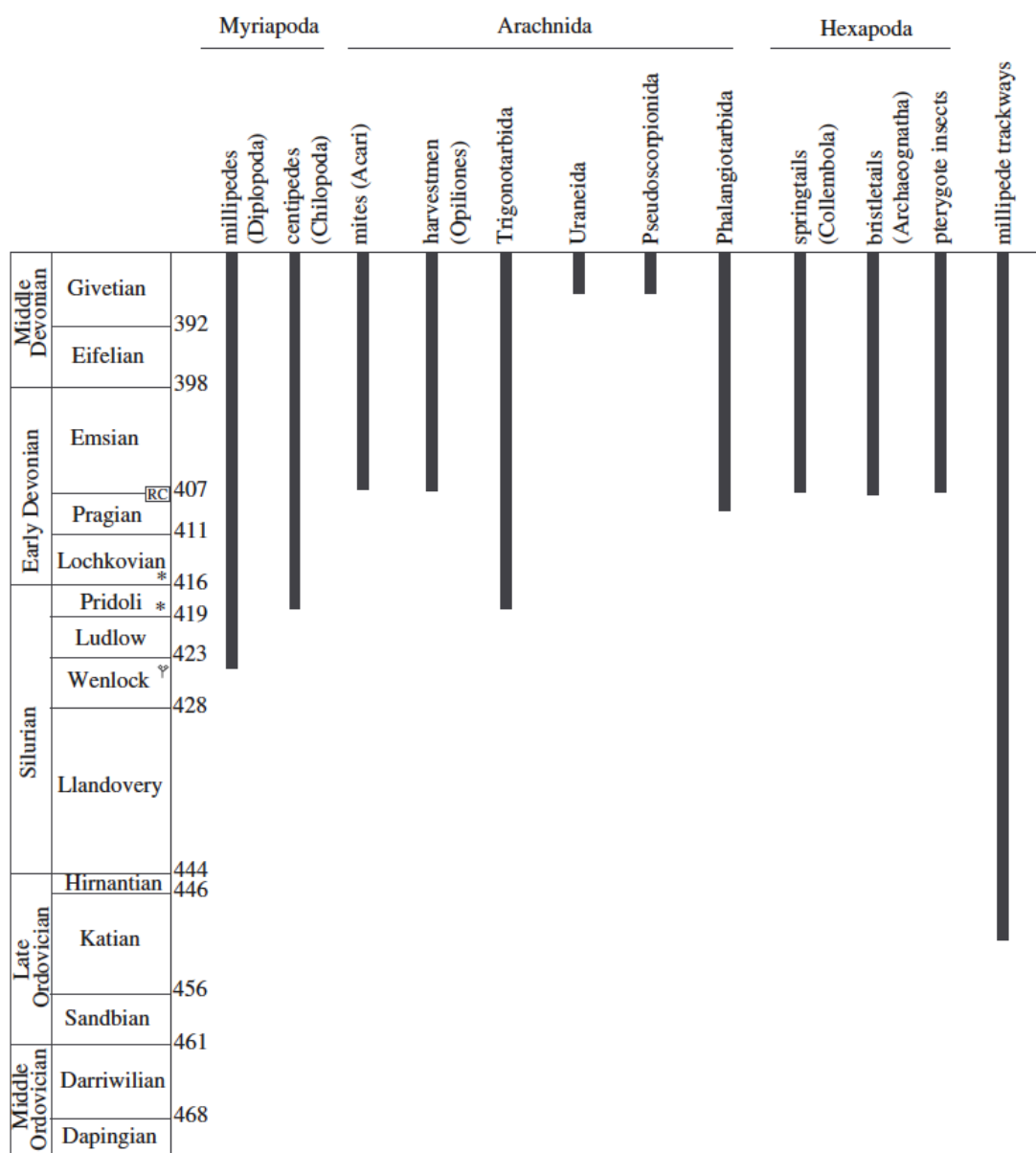
The first animals to colonize land, based on current fossil evidence, were ecdysozoans, specifically myriapods (Wilson & Anderson, 2004). The most biodiverse clade of ecdysozoans; the Arthropoda, are one of the more studied terrestrial phyla as not only are they some of the earliest animals to attain this feat, but three of its major subphyla the Hexapoda, Myriapoda, and Arachnida seemed to have colonized land independently on at least three Paleozoic occasions (Little, 1983) numerous times in more contemporary time periods (isopods & amphipods of the Crustacea), and perhaps even multiple times in the same lineage (Rota-Stabelli *et al.* 2013). This makes their journey particularly interesting as animals with shared morphological characteristics and similar genetic blueprints garnered the ability to transverse drastic environmental changes, overcoming the obstacles caused by the disparity between water and air such as the different physical properties of these media, sensory reception, locomotion, gas exchange, increased exposure to ultraviolet radiation, and change in mating patterns, *independently* (Dunlop *et al.* 2013) and not directly by virtue of inheritance from a single common ancestor, as it itself was

marine based (Malloof *et al.* 2010). Were these impressive accomplishments of evolution reliant on the timing of geological and atmospheric factors such as a rise in oxygen levels roughly 400 - 280 MYA (Berner, 1999). Conceivably, they were conditioning on the colonization of land by plants, laying the groundwork for the well-established food web of terrestrial ecosystems, but see (Rota-Stabelli *et al.* 2013). It is possible they were drastic efforts driven by the avoidance of marine based predators, Dunlop *et al.* (2013) suggest Horseshoe crabs laid their eggs on safe, previously non-colonized, shorelines to protect their young from such dangers. Perhaps it was the product of a series of genetic and morphological changes giving rise to land adaptations such as alterations in olfactory receptor genes to facilitate the binding of airborne odorants (Niimura & Nei, 2005 and Vieira *et al.* 2011) and the development of breathing apparatus trachea. Ostensibly it is probable that such a complex shift in habitat was facilitated by a mixture of these factors, the timing of the previous giving rise to the formation of the latter.

The primary source of evidence for arthropod terrestrialization is the fossil record: physical preservation of animals nested in a particular geological time period, allowing us to set a minimum age constraint on their existence. With regards to colonization, we are interested in the arthropods first appearance on land more so than their first appearance in the fossil record. Present evidence points to the Myriapoda being the first to make the transition, *Pneumodesmus newmani* was discovered in the Cowie Formation Scotland, emanating from the Ludlow Series of the Silurian, 426 MYA (Wilson & Anderson, 2004). Slightly younger is a trigonotarbid of the Arachnida, prominent member of the chelicerates, found in Shropshire England, of the Pridoli Series of the very late Silurian, 419 MYA (Jeram *et al.* 1990). However, to

make matters more interesting and undoubtedly more convoluted, there is evidence of myriapod-like trackways as far back as the Lower Cambrian, 530 MYA (MacNaughton *et al.* 2002). This extends the possibility of arthropod terrestrialization deeper into time, the delimitation of which depends on how credible one finds the ascribing of these trackways.

A convenient summary of the terrestrial arthropod fossil record is provided by Kenrick *et al.* (2012) [Figure 4.1].



**Figure 4.1: The Fossil Record of Terrestrial Arthropods**  
Stratigraphic record of terrestrial arthropod groups through the Middle Ordovician - Middle Devonian (Kenrick *et al.* 2012).

#### 4.1.2 Terrestrialization: A Complex Timeline

There have been a number of studies attempting to comprehensively determine a timeline for arthropod terrestrialization many of which have encountered difficulties due to restrictive dataset size (Regier *et al.* 2005), poor taxon sampling (Pisani *et al.* 2004), conflicting phylogenetic hypotheses; mainly Myriochelata vs. Mandibulata (Friedrich & Tautz, 1995; Cook *et al.* 2001; Pisani *et al.* 2004 vs. Rota-Stabelli *et al.* 2011; Misof *et al.* 2014; Borner *et al.* 2014), the incompleteness of the fossil record (Benton *et al.* 2015), the correct designation of fossils to taxonomic groups (Donoghue & Benton, 2007 and Inoue *et al.* 2010), the varying nature of molecular clocks (Rota-Stabelli *et al.* 2013), and lastly the complexity and uncertain numeracy of the terrestrialization events themselves.

The fossil record is a crucial piece of evidence in the terrestrialization puzzle, however it alone is not enough. A major hurdle facing modern colonization studies is the incomplete nature of the fossil record that, although providing us with minimum age constraints for species origins or evolutionary events, cannot be guaranteed as an accurate reflection of the origin of these events. This is because, as previously discussed, fossilization is heavily dependent on unique paleontological factors such as clay sedimentation, (Ager, 1981) that are found in such famous archeological sites as the Burgess Shale (Middle Cambrian), the Wenlock Series of Herefordshire (Middle Silurian), and the Rhynie Chert (Lower Devonian). The global distribution of these known paleontological sites is erratic and, concerning arthropod fossils, almost entirely located in Eumerica (Kenrick *et al.* 2012), making the catalog of fossils to study from even more limited. Not only that, but ancient arthropods were soft bodied in nature (Conway-Morris, 1993), highly susceptible to erosion and disintegration over hundreds of millions of years, making preservation even less likely than their

robust skeletal cousins on the deuterostome side of the animal tree. Consequently one has to consider how much time has elapsed between the terrestrialization of the first arthropods and first terrestrial arthropod to encounter the suitable and somewhat exceptionally unlikely conditions for preservation, furthermore, we rely on these fossils not only existing but being discovered.

This window of uncertainty is addressed by molecular clock studies that endeavor to fill this void of knowledge, resulting in fluctuating levels of reliability (Graur & Martin, 2004 and Kumar, 2005). The variability of dating studies can be attributed to a multitude of reasons such as the correct ascribing of fossils, how one defines bounds based on the fossil record, the application of these calibrations to the appropriate clades, dataset size (gene & taxa numbers), suitability of taxon coverage, the chosen phylogeny, strict versus relaxed clock methods, and the models of evolution applied to the analyses. Naturally these complexities have created considerable disparity between terrestrialization studies.

#### **4.1.3 Aims of this Study**

With a number of newly sequenced taxa pivotal to the terrestrial puzzle becoming available the time is right to readdress the timeline. Genomic level data for a series of myriapods (*Glomeridesmus* sp. SRR941771, *Lithobius forficatus* SRR1159752, *Polyxenus lagurus* SRR1048056, *Prostemmiulus* sp. SRR945439, *Scutigera coleoptrata* SRR1158078), terrestrial crustaceans (*Oniscidea* sp. in-house, *Speleonectes tulumensis* SRR857228), and chelicerates (*Pycnogonium littorale* in-house) allowed us to generate a dataset with sufficient taxa and gene coverage to run a phylogenetic reconstruction and terrestrialization dating experiments to further clarify

when animal life on land began. This study involves reconstructing the phylogeny of the Arthropoda to confirm the well-established relationships of their subphyla and to identify the non-colonized sister groups to the terrestrial lineages. Following this, a divergence time estimation analysis was necessary to reassess Arthropoda origins and terrestrial timelines with new taxa coverage, and finally an ancestral character state reconstruction on the sister group of the Hexapoda, whether the Branchiopoda or Remipedia (decided by the phylogenetic reconstruction), to clarify their poorly understood path to land.

My role in this study was to assemble the new transcriptomic data and identify their orthologs for the dataset. Following this I was involved with the divergence time estimation study. The phylogenetic reconstruction and ancestral character state reconstruction are not included in this thesis as I was not involved. Details of these experiments can be found in (Lozano-Fernandez *et al.* 2016)



## 4.2 Materials and Methods

### 4.2.1 Generation of Molecular Libraries

Several specimens collected for studies in the previous chapters were also used for this study of terrestrialization. These were *Oniscus sp.*, *P. littorale*, *Limulus sp.*, and *Epiperipatus sp.* For information on how these specimens were collected see **Materials and Methods 2.2.1**. The RNA extraction for these specimens was carried out in chapter 2 and full protocols can be found in the **Appendices**. The molecular libraries for these taxa were generated in the Edinburgh Genomics Sequencing Centre (see section **Materials and Methods 2.2.3**). To supplement these data, nine ecdysozoans were downloaded from the SRA: *Glomeridesmus* (SRR941771), *S. vulgaris* (SRR768329), *S. coleoprata* (SRR1158078), *P. lagurus* (SRR1048056), *Prostemmiulus* (SRR945439), *L. forficatus* (SRR1159752), *P. angustus* (SRR1047642), *M. tardigradum* (SRR057381), and *E. testudo* (SRR1141094) in conjunction with *H. dujardini* which was sourced from ([http://badger.bio.ed.ac.uk/H\\_dujardini/home/download](http://badger.bio.ed.ac.uk/H_dujardini/home/download)). All data was funneled through a quality control pipeline [**Materials and Methods 2.2.4**].

## 4.2.2 Transcriptome Assembly and Translation

The clean molecular libraries for the fourteen ecdysozoans were assembled and translated to proteins using Trinity (Grabherr *et al.* 2011) and Transdecoder (Haas *et al.* 2013). **Table 4.1** describes the assembly statistics for these taxa. **Materials and Methods 2.2.5** describes the process in full.

**Table 4.1: Terrestrialization Study: Assembled and Translated Transcripts**

Phred, scores, N-50 statistics, transcripts, and translated proteins for the fourteen ecdysozoans.

Sequenced Libraries					
Taxa	Source	Phred Score	Transcripts	N50 Statistics	Proteins
Oniscidea	in-house	39	6,906	363	1,677
Pycnogonid	in-house	39	87,838	1,765	26,668
Limulus	in-house	37	117,946	1,181	30,282
Onychophora	in-house	39	55,375	799	17,269
Glomeridesmus	SRR941771	39	80,196	467	25,952
Symphylella	SRR768329	24-31	34,703	524	11,309
Scutigera	SRR1158078	37	228,504	421	43,674
Polyxenus	SRR1048056	29	9,792	407	1,763
Prostemmiulus	SRR945439	39	41,181	355	5,849
Lithobius	SRR1159752	38	63,999	227	1,571
Polydesmus	SRR1047642	17	13,444	745	5,998
Milnesium	SRR057381	23-26	28,958	1,242	18,759
Echiniscus	SRR1141094	19	13,221	790	8,282
Hypsibius	<a href="http://badger.bio.ed">badger.bio.ed</a>	-	14,421	-	12,729

## 4.2.3 Ortholog Mapping

Ortholog identification and mapping protocols follow **Materials and Methods 2.2.6**.

The fourteen taxa were mapped to the Campbell *et al.* (2011) dataset as it was deemed an ideal candidate for studies on arthropod evolution. This generated a supermatrix of 30 species represented by 246 proteins.

#### 4.2.4 Divergence Time Estimation

Divergence time estimation was performed using Phylobayes (Latillot *et al.* 2009) on a fixed topology. Two alternative relaxed molecular clock models were used: the auto correlated CIR model and the Uncorrelated Gamma Multipliers model (UGAMMA). The tree was rooted on the Deuterostomia-Protostomia split. A set of twenty-four calibrations [Table 4.2] was used based on divergence estimates from Labandera (2005); Shear (1991); Strother *et al.* (2001), with a root prior defined using a Gamma distribution of mean 636MY and standard deviation of 30 MY.

```
$ pb -d terrestrialization_dataset.phy -T CATGTR_concensus.tre -r outgroup -cal
```

```
calibrations -[model] -rp 636 30 d terrestrialization_dataset-[model]-chain1
```

```
$ pb -d d terrestrialization_dataset.phy -T CATGTR_concensus.tre -r outgroup -cal
```

```
calibrations -[model] -rp 636 30 d terrestrialization_dataset-[model]-chain2
```

```
$ readdiv -x 500 d terrestrialization_dataset-[model]-chain-1
```

```
$ readdiv -x d terrestrialization_dataset-[model]-chain-2
```

**Table 4.2: Terrestrialization Study Molecular Clock Calibrations**

The 23 calibration points used in the clock experiments.

**Table 4.2: Terrestrialization Study Molecular Clock Calibrations**

<i>Taxa</i>		<i>Bounds (MYA)</i>		<i>Taxa</i>		<i>Bounds (MYA)</i>	
1	Ixodes - Acanthos	636.1	- 410	13	Daphnia - Anoplodact	636.1	- 514
2	Homo - Danio	444.9	- 420.7	14	Aplysia - Capitella	636.1	- 550.25
3	Aplysia - Lottia	636.1	- 534	15	Epiperipatus - Daphnia	636.1	- 528.82
4	Homo - Gallus	332.9	- 318	16	Priapulid - Daphnia	636.1	- 528.25
5	Gallus - Taeniopygia	86	- 66	17	Anoplodact - Acanthos	636.1	- 497
6	Homo - Mus	164.6	- 61.6	18	Scutigera - Strigamia	636.1	- 413
7	Homo - Xenopus	351	- 337	19	Scutigera - Glomerides	636.1	- 419
8	Homo - Leucoraja	468.4	- 420.7	20	Rhodnius - Gryllus	414	- 267
9	Petromyzon - Homo	636.1	- 457.5	21	Nasonia - Trilobolium	414	- 307
10	Ciona - Homo	636.1	- 514	22	Nasonia - Onychiurus	636.1	- 395
11	Strongyloc - Saccoglossus	636.1	- 515.5	23	Onisdidea - Litopenaeus	636.1	- 358.5
12	Daphnia - Gryllus	636.1	- 523	<i>(Labandera, 2005; Shear, 1991; and Strother et al, 2001)</i>			

## 4.3 Results

### 4.3.1 Divergence Time Estimation

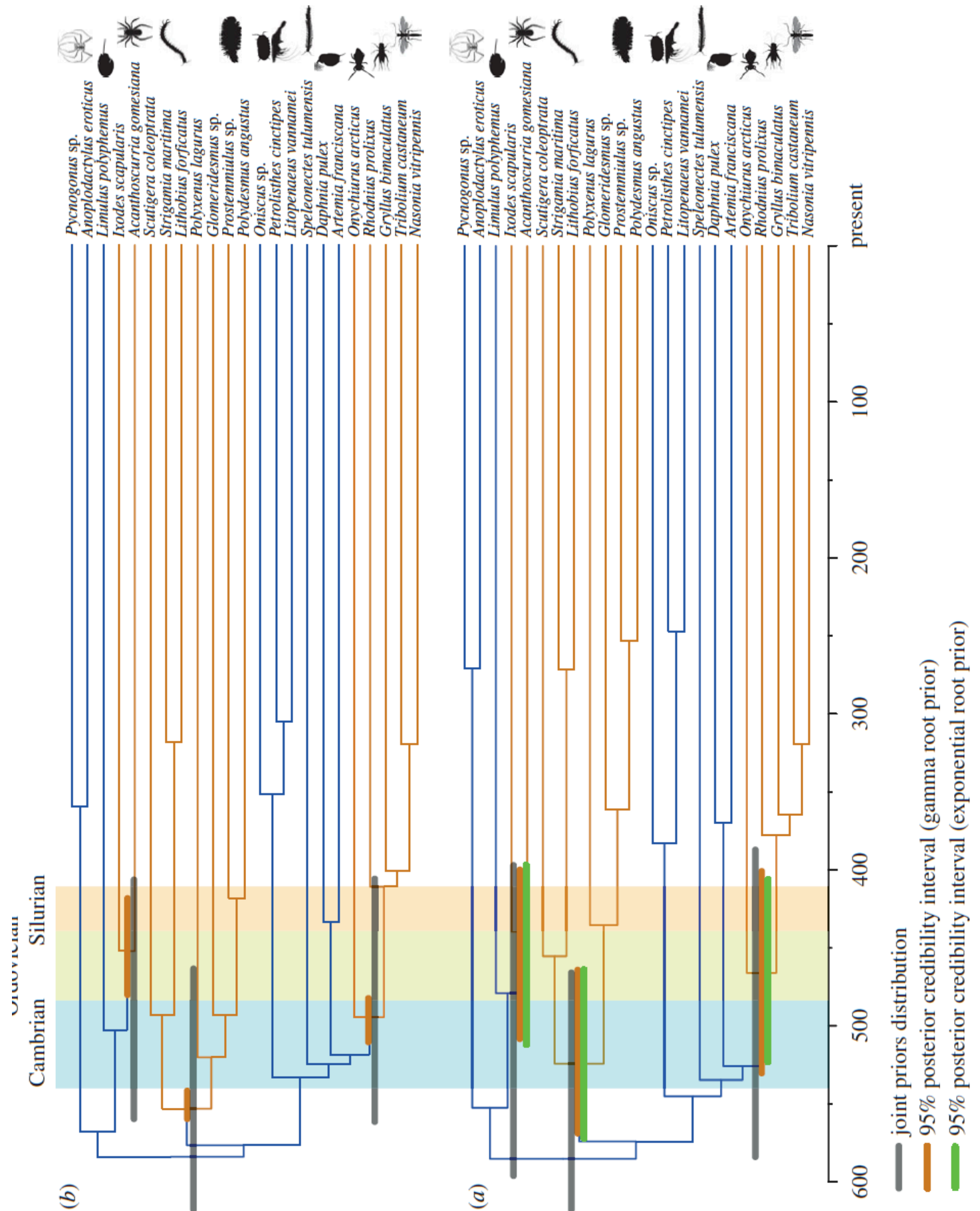
Molecular divergence time estimations are summarized in [Figure 4.2](#), with a more succinct explanation in [Table 4.3](#), and a detailed composition of each experiment in [Supplementary Material 4.1](#).

The 95 % credibility intervals (CI) for the CIR model were more concise than the U-GAMMA CI, however, regarding nodes under Paleozoic terrestrialization events, all CIR CI values fell under the U-GAMMA CIs. The two models did differ somewhat significantly on several aspects, with the CIR model resulting in deeper divergence times. The U-GAMMA model was in agreement with the fossil record of an Upper Cambrian to Silurian origin of the three terrestrial arthropod lineages while the CIR model placed the origin of the Myriapoda pre-Cambrian. The CIR model points to an Ordovician invasion of land for the Arachnida while U-GAMMA estimates the same event occurred more recently, in the Silurian. The Hexapoda diverged from their Pancrustacean ancestor in the Ordovician according to U-GAMMA while CIR offers a contrary, Cambrian origins.

**Table 4.3: Terrestrialization Study: Divergence Time Estimation**

Divergence time estimations for the arthropods under the U-GAMMA and CIR models.

Clade	U-GAMMA		CIR	
	Mean Age	95% CI	Mean Age	95% CI
Myriapoda	528	463 - 568	558	544 - 572
Chilopoda	457	408 - 526	490	452 - 511
Diplopoda	439	317 - 537	519	486 - 541
Hexapoda	468	407 - 512	499	394 - 431
Arachnida	440	397 - 518	460	413 - 493



**Figure 4.2: Molecular Clock Results re-viewing the Terrestrial Timeline**

**(A) Divergence times obtained under the CIR auto correlated, relaxed, molecular clock model**

**(B) Divergence times obtained using the Uncorrelated Gamma Multipliers model**

In both cases, nodes in the tree represent average divergence times estimated using the root prior with 636 Ma mean and 30 Ma SD. Brown bars represent 95% credibility intervals from the considered analysis. Grey bars represent the joint priors (for the considered nodes and analyses). Green bars in Figure 2B indicate 95% credibility intervals obtained using the exponential prior of average 636 Ma.

## 4.4 Discussion

The power of this approach was the comparative and multifaceted aspect of the analyses. Instead of looking at isolated lineages of arthropod terrestrialization, contemporary sequencing technology allowed us supplement taxa deficient clades, such as the Myriapoda, to address the terrestrialization question armed with more information previous studies lacked (Pisani *et al.* 2004). In conjunction with the phylogenetic analysis, an ancestral environmental reconstruction was carried out by Mark Puttick of the University of Bristol's paleobiology group, which proved useful in understanding terrestrialization from a differing facet; revealing the path the Hexapoda took to land colonization as opposed to the timeframe. The ancestral environment reconstruction estimates that the hexapods invaded land from a marine environment (with a probability of 0.84) as opposed to freshwater ( $p = 0.15$ ) or brackish ( $p = 0.002$ ) habitats. This is a further example of the benefits of molecular, morphological, and paleobiological working as one to find solutions to complex evolutionary questions as opposed to narrow independent studies.

Divergence dating studies are as varied as they are numerous, differing calibrations, clock models, phylogenetic interpretations, and widely disparate datasets make comparing the resulting conclusions problematic. With the absence of a protocol consensus, one can only endeavor to produce the most robust results possible by avoiding under-sampled datasets for the taxa of interest, application of a concrete reference topology, relaxed clock methods preventing lineage effects, a multi-model approach, and a well calibrated clock. Implementing these methods, we have summarized our divergence time estimations with model variation in mind.

## 4.5 Conclusions

Animals most likely originated in the Cryogenian and radiated close to the Upper Cambrian. The absence of pre-Cambrian fossils can perhaps be explained by a lack of animal diversity at the time, the soft bodied plan of the earliest animals proving hard to fossilize (Conway-Morris, 1993), the incomplete nature of the fossil record, and the geological biases of paleobiological sites (Kenrick *et al.* 2012).

In conclusion, based on the findings of this study, the contemporary terrain based arthropod lineages invaded land on at least three, if not four, separate occasions; the Hexapoda in the Ordovician, the Arachnida in the Silurian or the Ordovician, both broadly inline with the fossil record, with the Myriapoda colonizing land possibly twice, initially in the Cambrian [Figure 4.2]. This result is much earlier than the myriapod fossil record, however it fits the timescale for terrestrial myriapod-like tracks discovered 530 MYA (MacNaughton *et al.* 2002).

Unlike the origins of the subphyla to which they belong, the divergence of the diplopods and chilopods is in agreement with fossil evidence. Current theories see the Myriapoda as invading land once due to presence of terrain adapted breathing apparatus in both the chilopods and diplopods, but our divergence dating results point to a further possible explanation: independent colonization of land for each of these groups meaning the myriapods would have colonized land twice. This would signify that the development of the myriapod trachea would also have occurred independently in chilopods and diplopods in a classic example of convergent evolution, not unheard of in terrestrial evolution (Little, 1983). Morphological evidence gives some credit to this possibility (Dohle, 1998) with observations in structural differences in trachea



and positional disparity of spiracles suggesting the land-adapted breathing traits may not be the product of a LCA.

Additionally it seems, based on the resulting environmental reconstruction of the Branchiopoda, in accordance with the evidence for the marine based sister groups of the Myriapoda and the Arachnida; terrestrialization of early animals initiated at coastlines of the continents and advanced inwards. The long road to arthropod terrestrialization did not emanate from rivers or lakes, but from the oceans.

### 5.1 Introduction

#### 5.1.1 Preliminary Study

The concept for a large-scale study of protein family evolution originated from a preliminary “proof of concept” study (Pisani *et al.* 2013), which itself was inspired by a phylostratigraphic study from Domazet-Loso *et al.* (2007).

Despite focusing mostly on the Arthropoda, the experiment from Pisani *et al.* (2013) suffered from a lack of taxon sampling in three out of four of its subphyla: the Crustacea, Myriapoda, and Chelicerata and were only represented by EST datasets. Our approach was to expand on the experiment by supplementing the dataset with twenty-eight newly sequenced taxa (ten of which are the product of in-house sequencing experiments) from poorly sampled subphyla such as the Crustacea, Chelicerata, Myriapoda, and Tardigrada [[Supplementary Material 5.1](#)], augmenting the 21 taxa 389,994 protein dataset to 49 taxa and 847,640 proteins. This significant improvement in molecular information made it possible to reevaluate the results of the previous study that found no significant changes concerning new protein family acquisition traversing the history of the Metazoa, with the sole exception being a notable increase within the flying insects; the Diptera.

### 5.1.2 Protein Families

A protein family is a group of genes related by common ancestry (Wu *et al.* 2003). Protein families can be shared from speciation, gene duplication or even by lateral gene transfer events (Wu *et al.* 2003). Protein family members share similar sequences and functions and are considered homologous to one another (Dayhoff, 1976; Krause *et al.* 2005; and Demuth & Hahn, 2009). The protein family structure can be used to investigate how organisms have changed over time, whether through speciation events resulting in the formation of taxonomic groups, or changes at the molecular level during major environmental changes transitions as terrestrialization (Pisani *et al.* 2004 and Kenrick *et al.* 2012). For the purposes of this study the concept of *new protein families* are of interest, loosely referred to as orphan protein families. Here we define a new protein family as a network of proteins that did not exist in the last common ancestor, or in terms of a phylogenetic framework, a network of proteins that did not exist in the previous ancestral node in the tree.

The rate of new protein family acquisition throughout the history of the Metazoa can be investigated by applying the theory behind phylostratigraphy (Domazet-Loso *et al.* 2007) to a large database of high quality molecular data (847,640 proteins mostly from NGS sources) that have been grouped into families using the Markov Clustering Algorithm (MCL) (Enright *et al.* 2002). This makes it possible to view trends of significant protein family gains throughout metazoan history, when these gains occurred, and what groups were involved. This is in effort to further understand how a complex and diverse Animal Kingdom has diversified and radiated to the present day on a macroevolutionary level.

### 5.1.3 Aims of this Study

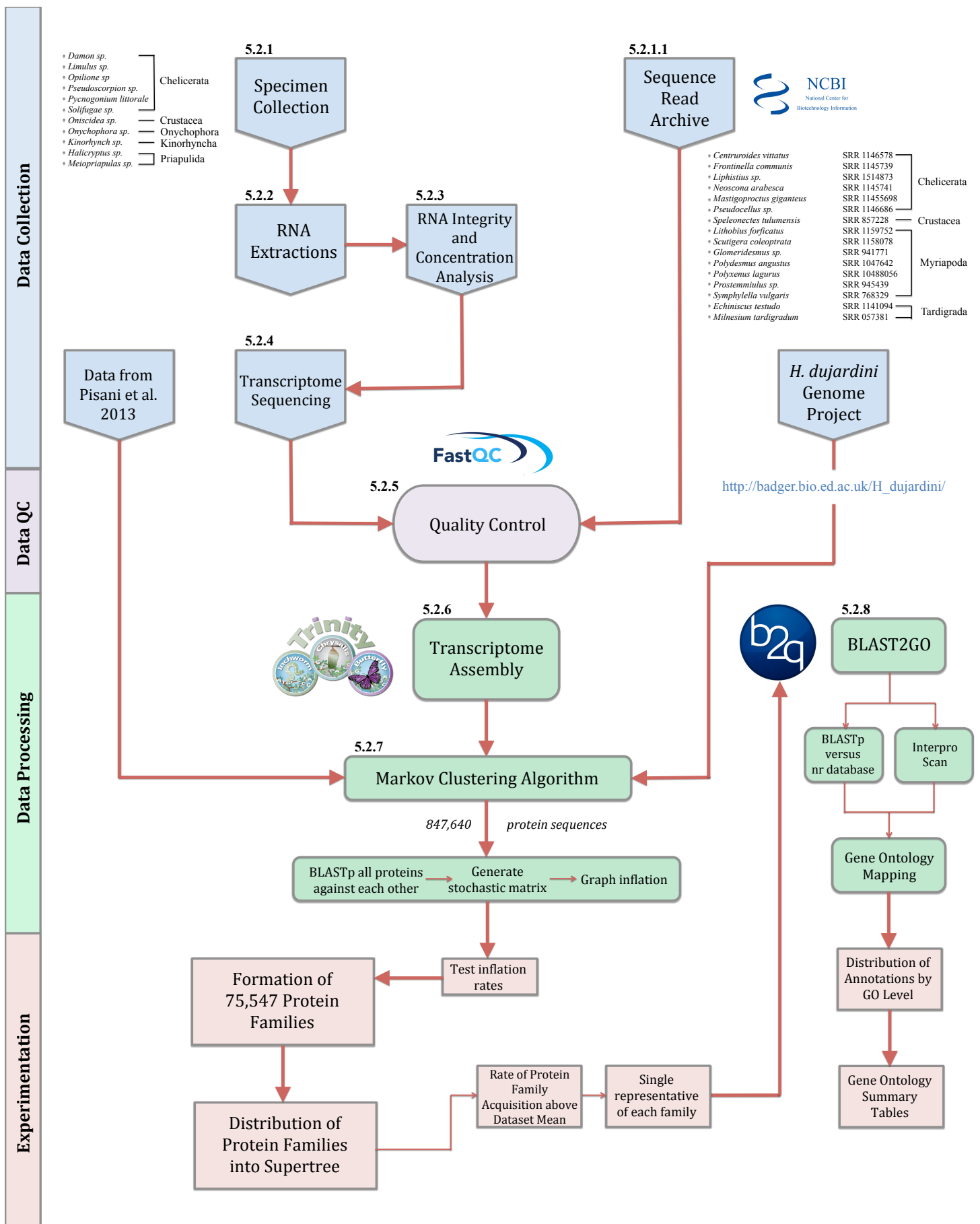
The aim of this chapter was to generate a new collection of protein families using the Markov Clustering Algorithm (MCL) (Enright *et al.* 2002) by amalgamating the data from the previous study (Pisani *et al.* 2013) and the twenty-eight newly sequenced taxa from NGS experiments. These protein families were distributed throughout a supertree with nodes experiencing a rate of protein family acquisition above the mean being of interest. Nodes that displayed significant gains in protein families were annotated using BLAST2GO (Conesa *et al.* 2005) in order to gain some functional understanding into the rate of protein family adaptation.

The hypothesis of this study is to review whether the results of the preliminary experiment (Pisani *et al.* 2013) are an accurate representation of the protein family trend spanning the Metazoa or whether it is a restricted view due to lack of information because of taxa restrictions in major clades. Advancements in NGS technology helped us to double the amount of information in our dataset by subsidizing these species barren ecdysozoan clades thus broadening the scope of the study and perhaps even allowing us to uncover previously unseen molecular trends during taxonomic formation or even during large-scale environmental changes such as sea versus atmospheric oxygen levels (Berner, 1999) or even ecosystem adaptations such as some of the first terrestrialization events (Wilson & Anderson, 2004 and Jeram *et al.* 1990).

## 5.2 Material and Methods

In accordance with chapters 2 and 3, a flowchart is provided to illustrate the materials and methods of this chapter [Figure 5.1]. This details the steps taken from preliminary data collection stage to the final experimentation stage. Many of the methods that were used in this chapter have already been detailed in chapter 2. To avoid repetition a concise summary of these methods is provided instead of a full description. These sections include:

5.2.1 Specimen Collection	see 2.2.1 for full description
5.2.2 DNA & RNA Extractions	see 2.2.2 for full description
5.2.3 Concentration and Integrity Analyses	see 2.2.2 for full description
5.2.4 Genome and Transcriptome Sequencing	see 2.2.3 for full description
5.2.5 Data Quality Control	see 2.2.4 for full description
5.2.6 Transcriptome Assembly and Translation	see 2.2.5 for full description



**Figure 5.1: Flowchart Detailing Materials and Methods of Chapter 5**

These steps detail how the new protein families were generated annotated and distributed across the metazoan supertree from specimen collection to MCL protein clustering and BLAST2GO annotation.

### 5.2.1 Specimen Collection

Many of the specimens used in this study were used in chapter 2. These were *P. littorale*, *Opilione sp.*, *Limulus sp.*, *Oniscidea sp.*, *Epiperipatus sp.*, *Halicryptus sp.*, *Meiopriapulas sp.*, and *Kinorhynch sp.* Information about the collection of these animals can be found in section **Materials and Methods 2.2.1**.


Additional specimens were sourced for this study and are outlined herein. *Pseudoscorpion sp.* were identified and collected by Karl-Hinrich Kielhorn & Jason Dunlop at the Leibniz Institute for Research on Evolution Biodiversity at the Humboldt University, Berlin. Two sun spider (*Galeodidae sp.*) samples were collected by Stuart Longhorn and stored in RNAlater vials at -80°C. The arachnid *Damon sp.* from the amblypygi order was sourced from an exotic animals website ([www.exotic-pets.co.uk](http://www.exotic-pets.co.uk)) and sent to the University of Bristol for extraction and sequencing. Information on the classification of the collected specimens can be found in **Table 5.1** and images in **Figure 5.2**.

**Table 5.1: Specimens Collected for Protein Family Study**

A heavy focus on Arthropods, particularly chelicerates was a deliberate choice to address subphyla that lacked data from the preliminary study.

**Table 5.1: Specimens Collected for Protein Family Study**

	Specimen	Phylum	Subphylum	Class	Order
1	<i>Pycnogonium littorale</i>	Arthropoda	Chelicerata	Pycnogonida	Pantopoda
2	<i>Opilione sp.</i>	Arthropoda	Chelicerata	Arachnida	Opiliones
3	<i>Galeodidae sp.</i>	Arthropoda	Chelicerata	Arachnida	Solifugae
4	<i>Damon sp.</i>	Arthropoda	Chelicerata	Arachnida	Amblypygi
5	<i>Pseudoscorpion sp.</i>	Arthropoda	Chelicerata	Arachnida	Pseudoscorpiones
6	<i>Limulus sp.</i>	Arthropoda	Chelicerata	Xiphosura	Xiphosurida
7	<i>Oniscidea sp.</i>	Arthropoda	Crustacea	Malacostraca	Oniscidea
8	<i>Epiperipatus sp.</i>	Onychophora	-	Udeonychophora	Euonychophora
9	<i>Halicryptus sp.</i>	Priapulida	-	Halicryptomorpha	Halicryptomorphida
10	<i>Meiopriapulas sp.</i>	Priapulida	-	Meiopriapulomorpha	Meiopriapulomorphida
11	<i>Kinorhynch sp.</i>	Kinorhyncha	-	Cyclorhagida or Homalorhagida	

<p><i>Pycnogonum littorale</i></p> 	<p><b>Pycnogonida</b></p> <ul style="list-style-type: none"> <li>• Common name: sea spider</li> <li>• Non-terrestrial chelicerate</li> <li>• Grow in size in deeper waters</li> <li>• 1mm – 90cm</li> </ul> <p><a href="http://www.discoverlife.org/20/q/search-pycnogonida">http://www.discoverlife.org/20/q/search-pycnogonida</a></p>
<p><i>Opilione sp.</i></p> 	<p><b>Opiliones</b></p> <ul style="list-style-type: none"> <li>• Common name: harvestman / daddy long legs</li> <li>• Leg span comparatively long</li> <li>• 4mm – 4cm</li> </ul> <p><a href="http://www.burkhemuseum.org/sites/default/files/styles/adaptive/adaptive-image-public/28-rapestre_1.jpg?itok=0npdy1f">http://www.burkhemuseum.org/sites/default/files/styles/adaptive/adaptive-image-public/28-rapestre_1.jpg?itok=0npdy1f</a></p>
<p><i>Pseudoscorpion sp.</i></p> 	<p><b>Pseudoscorpion</b></p> <ul style="list-style-type: none"> <li>• Common name: false scorpion</li> <li>• Characterized by their long pincers</li> <li>• 2mm – 12mm</li> </ul> <p><a href="http://somebingsawlingmyhair.com/wp-content/uploads/2007/09/pseudoscorpion_sharpfocus5888k.jpg">http://somebingsawlingmyhair.com/wp-content/uploads/2007/09/pseudoscorpion_sharpfocus5888k.jpg</a></p>
<p><i>Galeodidae sp.</i></p> 	<p><b>Solifugae</b></p> <ul style="list-style-type: none"> <li>• Common name: camel / sun spider</li> <li>• Appears to have 5 pairs of legs but the front pair are actually pedipalps</li> <li>• 5cm – 15cm</li> </ul> <p><a href="https://upload.wikimedia.org/wikipedia/commons/1/1d/Unidentified_solifugae_Umanac_district_MP_India.jpg">https://upload.wikimedia.org/wikipedia/commons/1/1d/Unidentified_solifugae_Umanac_district_MP_India.jpg</a></p>
<p><i>Damon sp.</i></p> 	<p><b>Amblypygi</b></p> <ul style="list-style-type: none"> <li>• Common name: tailless whip scorpion</li> <li>• Lack a tail which distinguishes them from the closely related Thelyphronida</li> <li>• 5cm – 70cm</li> </ul> <p><a href="http://www.panarthropoda.de/sub/galerie/pictures/physnus_whicief01.jpg">http://www.panarthropoda.de/sub/galerie/pictures/physnus_whicief01.jpg</a></p>
<p><i>Limulus sp.</i></p> 	<p><b>Xiphosura</b></p> <ul style="list-style-type: none"> <li>• Common name: horseshoe crab</li> <li>• Non-terrestrial chelicerate</li> <li>• Crustacean-like appearance, body protected by a carapace</li> <li>• 7cm – 60cm</li> </ul> <p><a href="http://previews.123rf.com/images/tonobalaguert1005/tonobalaguert100500248/6987186-atlantic-horseshoes-cnb.jpeg">http://previews.123rf.com/images/tonobalaguert1005/tonobalaguert100500248/6987186-atlantic-horseshoes-cnb.jpeg</a></p>
<p><i>Oniscidea sp.</i></p> 	<p><b>Oniscidea</b></p> <ul style="list-style-type: none"> <li>• Common name: woodlouse</li> <li>• 7 pairs of legs</li> <li>• Large thorax consisting of 7 tough over-lapping plates</li> <li>• 1cm – 3cm</li> </ul> <p><a href="http://www.animalbase.uni-goettingen.de/animalbaseimage/Armadillidium-vulgare_01.jpg">http://www.animalbase.uni-goettingen.de/animalbaseimage/Armadillidium-vulgare_01.jpg</a></p>
<p><i>Epiperipatus sp.</i></p> 	<p><b>Figure 2.2.1 H: Onychophora</b></p> <ul style="list-style-type: none"> <li>• Common name: velvet worm</li> <li>• Varying amount of appendages depending on species</li> <li>• Hydrostatic skeleton</li> <li>• 5mm– 20cm</li> </ul> <p><a href="http://www.onychophora.com/images/fig_2.png">http://www.onychophora.com/images/fig_2.png</a></p>
<p><i>Halicryptus sp. &amp; Meiopriapulas sp.</i></p> 	<p><b>Priapulida</b></p> <ul style="list-style-type: none"> <li>• Common name: penis worm</li> <li>• Worm-like appearance</li> <li>• Pharynx is lined with teeth</li> <li>• 2mm– 39cm</li> </ul> <p>(Todaro &amp; Shirley 2003)</p> <p><a href="http://cdn.e-photoshelter.com/img-qa2/10000089h6kKOY2xQ/1000x750/DSC7618.jpg">http://cdn.e-photoshelter.com/img-qa2/10000089h6kKOY2xQ/1000x750/DSC7618.jpg</a></p>
<p><i>Kinorhynch sp.</i></p> 	<p><b>Kinorhyncha</b></p> <ul style="list-style-type: none"> <li>• Common name: mud dragon</li> <li>• Limbless, spiny body</li> <li>• Less than 1mm in length</li> <li>• Can survive at great ocean depth down to 8 kilometers.</li> </ul> <p><a href="https://s-media-cache-ak0.pinimg.com/736x/67/a7/bd/67a7bd78555db20e103e7f1b10.jpg">https://s-media-cache-ak0.pinimg.com/736x/67/a7/bd/67a7bd78555db20e103e7f1b10.jpg</a></p>

**Figure 5.2: Specimens Collected and Sequenced for Chapter 5**  
 Images and brief descriptions of the specimens whose molecular libraries have been generated for the protein family evolution study.  
 The url for each image source is provided.



### 5.2.1.1 Taxa Downloaded from the Sequence Read Archive

The SRA was an excellent source of data to further supplement the preliminary study (Pisani *et al.* 2013) with NGS data, particularly in areas of sparse taxon coverage such as the Chelicerata, Myriapoda, and Tardigrada. Transcriptomes for the following taxa were downloaded and entered into the quality control and data assembly pipelines.

**Table 5.2: Transcriptomes Downloaded from the SRA**

The sixteen transcriptomes sourced from the SRA used in this study. Together with the specimens collected and sequenced from in-house experiments, and the *H. dujardini* genome project made twenty-eight taxa to supplement the preliminary study.

**Table 5.2: Transcriptomes Downloaded From The SRA**

	Transcriptome	SRA Number	Phylum	Subphylum	Class	Order
1	Centruroides vittatus	SRR1146578	Arthropoda	Chelicerata	Arachnida	Scorpiones
2	Frontinella communis	SRR1145739	Arthropoda	Chelicerata	Arachnida	Araneae
3	Liphistius sp.	SRR1514873	Arthropoda	Chelicerata	Arachnida	Araneae
4	Neoscona arabesca	SRR1145741	Arthropoda	Chelicerata	Arachnida	Araneae
5	Mastigoproctus giganteus	SRR1145698	Arthropoda	Chelicerata	Arachnida	Thelyphonida
6	Pseudocellus sp.	SRR1146686	Arthropoda	Chelicerata	Arachnida	Ricinulei
7	Speleonectes tulumensis	SRR857228	Arthropoda	Crustacea	Remipedia	Nectipoda
8	Lithobius forficatus	SRR1159752	Arthropoda	Myriapoda	Chilopoda	Lithobiomorpha
9	Scutigera coleoptrata	SRR1158078	Arthropoda	Myriapoda	Chilopoda	Scutigermorpha
10	Glomeridesmus	SRR941771	Arthropoda	Myriapoda	Diplopoda	Glomeridesmida
11	Polydesmus angustus	SRR1047642	Arthropoda	Myriapoda	Diplopoda	Polydesmida
12	Polyxenus lagurus	SRR1048056	Arthropoda	Myriapoda	Diplopoda	Polyxenida
13	Prostemmiulus	SRR945439	Arthropoda	Myriapoda	Diplopoda	Stemmiulida
14	Symphylella vulgaris	SRR768329	Arthropoda	Myriapoda	Symphyla	Symphylemida
15	Echiniscus testudo	SRR1141094	Tardigrada	-	Heterotardigrada	Echiniscoidea
16	Milnesium tardigradum	SRR057381	Tardigrada	-	Eutardigrada	Apocheila

In addition to sourcing specimens from in-house sequencing experiments and the SRA, the genome of *Hypsibius dujardini* was downloaded from ([http://badger.bio.ed.ac.uk/H\\_dujardini/home/download](http://badger.bio.ed.ac.uk/H_dujardini/home/download)). For further information on the source of these taxa see **Supplementary Material 2.2**.

### 5.2.2 DNA & RNA Extractions

DNA extractions of the pycnogonids, opiliones, solifugae, and RNA extractions of the pycnogonids, opiliones, solifugae, and pseudoscorpions were carried out by Eoin Mulvihill of NUIMs Nematode Genetics laboratory and by myself. Eoin had a large amount of experience with invertebrate DNA & RNA extractions and so was the perfect candidate for directing our efforts. Omar Rota-Stabelli carried out RNA extractions of the oniscidea and the onychophoran. RNA extraction of the amblypygi, horseshoe crabs, kinorhynch and priapulids were conducted at the University of Bristol. Protocols for DNA and RNA extractions can be found in the **Appendices**.

### 5.2.3 DNA & RNA Concentration and Integrity Analysis

DNA concentration levels were identified by measuring the absorbance of the solution at 260nm using the Nanodrop [[Supplementary Material 2.1](#)]. Qiagen extraction protocols state that any sample returning values falling within the range of 0.1 - 1.0 absorbance contain adequate concentration for sequencing.

The purity of the extracted DNA sample was determined by calculating the ratio of absorbance at 260nm to the absorbance at 280nm;  $A_{260}/A_{280}$  for protein contaminants and  $A_{260}/A_{230}$  for phenol and organic contaminants.

DNA integrity was rated by gel electrophoresis. See [Supplementary Material 2.1](#) for full results. A current was run through the gel and after migration, isolated bands on the gel represented DNA strands of a single length. The brightness of a band is a loose indication of concentration (better quantified by nanodrop), and long blurry bands are representative of DNA of many sizes, in this case an indication of genomic DNA

degradation. DNA lengths were gauged with the use of ladders of known molecular length placed in either ends of the gel band.

Similarly to the DNA prep, RNA concentration and purity was rated using the nanodrop, the only minor difference being that the absorbance value for protein contaminant free RNA ( $A_{260}/A_{280}$ ) is 2.0 as opposed to 1.7 - 1.9 for DNA [**Supplementary Material 2.1**].

RNA integrity checks were carried out using a bioanalyzer. A sample with intact RNA will return two peaks on the graph representing the largest RNA subunits 18S and 28S. A spike at only one or at multiple bands is usually a sign of RNA degradation, as it would point to the likelihood that one of the major subunits had broken into smaller fragments. However the samples encountered a phenomenon described by Winnebeck *et al.* (2009), displaying a single peak for intact RNA samples. See **Materials and Methods 2.2.2.4** for a discussion on this topic. For full bioanalyzer results see **Supplementary Material 2.1**.

#### **5.2.4 Genome and Transcriptome Sequencing**

All gDNA and RNA samples were sequenced by Illumina Solexa at Edinburgh Genomics (<https://genomics.ed.ac.uk/services/sequencing>). A description of the NGS process with Illumina Ssolexa technology can be found in **Materials and Methods 2.2.3**. A step by step guide to NGS by Illumina Solexa is detailed in **Supplementary Material 1.1**.

### 5.2.5 Data Quality Control

Once the data was retrieved from the sequencing center it was important to check the quality of the gDNA and RNA sequences. FastQC (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) was used to gauge the quality of our newly sequenced taxa in addition to all libraries downloaded from the SRA.

FastQC checks the raw sequence data and rates its quality based on a number of metrics such as Phred score, per base content, per sequence GC content, N content, sequence length distribution, adapter content and sequence duplication levels.

See **Table 5.3** for a brief description of data quality of the sequenced taxa and SRA sourced taxa from this study and **Supplementary Material 2.2** for a more in-depth report.

### 5.2.6 Transcriptome Assembly and Translation

Transcriptome assembly was achieved through the Trinity package (Grabherr *et al.* 2011) and amino acid sequences were predicted from the transcripts using TransDecoder (Haas *et al.* 2013). The commonly used gauge for assembly contiguity is the N50 statistic. The N50 value for an assembled transcriptome states that half of the assembled transcripts are the length of the given value, thus the higher the N50 statistic the more contiguous the assembly (Miller *et al.* 2010). Assembly statistics including total number of transcripts, proteins identified from these transcripts, and N50 values can be found in **Table 5.3**.

**Table 5.3: Protein Families Study Assembled and Translated Transcripts**

The table is divided into libraries generated from in-house sequence experiments and libraries downloaded from the SRA. There is a stark contrast between the Phred scores and N50 stats between the two sets of libraries. This suggests the libraries generated in-house are of greater quality than those downloaded from the SRA.

**Table 5.3: Protein Families Study Assembled & Translated Transcripts**

Sequenced Libraries					
Taxa	Source	Phred Score	Transcripts	N50 Statistics	Proteins
Pycnogonid	in-house	39	87,838	1,765	26,668
Opilione	in-house	38	134,694	1,709	30,942
Pseudoscorpion	in-house	39	86,196	874	24,142
Solifugae	in-house	38	65,943	1,317	22,782
Amblypygi	in-house	37	87,015	1,554	24,564
Limulus	in-house	37	117,946	1,181	30,282
Oniscidea	in-house	39	6,906	363	1,677
Onychophora	in-house	39	55,375	799	17,269
Halicryptus	in-house	37	64,406	1,896	29,057
Meiopriapulas	in-house	37	111,893	1,522	39,254
Libraries Downloaded from the SRA					
Taxa	Source	Phred Score	Transcripts	N50 Statistics	Proteins
Centruroides	SRR1146578	14-26-38	46,919	282	2,050
Frontinella	SRR1145739	37	160,683	330	16,241
Liphistius	SRR15114873	39	40,822	245	1,244
Neoscona	SRR1145741	38	175,980	349	29,147
Mastigoproctus	SRR1145698	37	120,987	712	32,648
Pseudocellus	SRR1146686	37	81,107	291	4,438
Speleonectes	SRR857228	15	2,850	774	970
Lithobius	SRR1159752	38	63,999	227	1,571
Scutigera	SRR1158078	37	228,504	421	43,674
Glomeridesmus	SRR941771	39	80,196	467	25,952
Polydesmus	SRR1047642	17	13,444	745	5,998
Polyxenus	SRR1048056	29	9,792	407	1,763
Prostemmiulus	SRR945439	39	41,181	355	5,849
Symphylella	SRR768329	24-31	34,703	524	11,309
Echiniscus	SRR1141094	19	13,221	790	8,282
Milnesium	SRR057381	23-26	28,958	1,242	18,759

### 5.2.7 MCL: Protein Family Generation

The first step in establishing MCL gene families is with BLAST (Altschul, 1990). With the addition of proteins from the twenty-eight newly sequenced taxa, the dataset from Pisani *et al.* (2013) grew from 389,994 sequences to 847,640. All sequences were compared and ranked against each other based on their E. value, with a minimum threshold of  $1E^{-10}$ . This all versus all ranking system allows for the creation of a stochastic matrix, the backbone of MCL experiments.

```
$ cat [Pisani et al. (2013) dataset] [all taxa from Table 5.2.6] >> MCL_dataset.fa
$ blastall -i MCL_dataset.fa -d MCL_dataset.fa -out MCL_blastall.out -evalue 1e-10
                                     -outfmt 6
```

A BLAST all matrix of 847,640 x 847,640 protein sequences produced 155 million HSPs in tabulated format. The tabulated output was then converted into abc format in which each line is tabulated to denote only the reference node (1<sup>st</sup> tab), the node it is being compared to (2<sup>nd</sup> tab) and the E. value / edge weight that scores their similarity (11<sup>th</sup> tab).

```
$ cut -f 1,2,11 MCL_blastall.out > MCL_blastall.abc
```

The E. values were formatted to correspond to weighted edges between the nodes compared, by converting their expectation probabilities using negated logarithms, and an undirected graph was generated. This was achieved using the `mcxload` module inbuilt into the MCL package.

```
$ mcxload -abc MCL_blastall.abc --stream-mirror --stream-neg-log10 -stream-tf
'ceil(200)' -o MCL_blastall.mci -write-tab MCL_blastall.tab
```

The “stream mirror” option ensures an undirected network, the “stream-neg-log10” command converts the E. values to  $\log_{10}$  while negating negative values, and

“ceil(200)’ sets any E. value below  $1E^{-200}$  as the maximum edge weight of 200 - i.e. the best possible “match” (Enright *et al.* 2002).

This generates two output files: .tab and .mci. The .tab file assigns each sequence from the .abc file to a unique number identifier, the .mci file uses the unique number identifier as the nodes in the graph to greatly speed up the process. Writing out large sequence headers, particularly from sequencing experiments, greatly slows down the MCL process.

MCL then generated random walks in this stochastic matrix, using “edge weights” to define similarity between protein nodes, creating flow paths of high similarity. The inflation option alters the granularity of the graph, with lower values generating more coarse clusterings and larger ones forming crisp disparity between groups of nodes. With the variable nature of molecular datasets in mind, the inflation rate was set to a broad set of values 1.2, 2, 4, 6, & 8 as per MCL guidelines (Enright *et al.* 2002).

```
$ mcl MCL_blastall.mci -I 1.2 >> MCL-1.2.out
```

```
$ mcl MCL_blastall.mci -I 2 >> MCL-2.out
```

```
$ mcl MCL_blastall.mci -I 4 >> MCL-4.out
```

```
$ mcl MCL_blastall.mci -I 6 >> MCL-6.out
```

```
$ mcl MCL_blastall.mci -I 8 >> MCL-8.out
```

An inflation value of 8 produced the most defined clusterings and more families numerically than the other inflation experiments [Table 5.4]. Additionally, protein families with only a single member were removed, leaving 75,547 protein families remaining for delegation to the 49 taxa and 48 internal nodes of the tree, as the very definition of a family is that it consists of more than one member (sequence).

**Table 5.4: Influence of Inflation Rate on a 847,637 x 847,637 Protein Clustering Matrix**  
 An inflation rate of 8 produced the most defined clusterings. This was the same rate used in the preliminary study (Pisani *et al.* 2013)

<i>Inflation Rate</i>	<i>Total Protein Families</i>	<i>Families With At Least Two Members</i>
<i>I 1.2</i>	124,787	33,863
<i>I 2</i>	143,796	52,859
<i>I 4</i>	157,151	65,986
<i>I 6</i>	163,673	72,019
<i>I 8</i>	167,673	75,547

Once the Markov clustering was complete the next step was to translate the clustered unique number identifiers back to their sequence headers. This was achieved using the inbuilt `mexdump` module, a script that writes all the sequence headers comprising each protein family to file, putting each family on a single line, the protein families were ordered in descending number of sequence members.

```
$ mexdump -icl MCL-8.out -tabr MCL_blastall.tab -o dump.MCL-I-8
```

A shell script, `extract_family_headers.sh`, [Supplementary Material 5.2] was used to take all the headers of each protein family from the `.dump` file and write them to their own individual files, with the title of the file corresponding to the number designated to each family of the MCL output; 0 for the first and largest protein family, 1 for the second largest, ... , 75,546 for the smallest.

```
$ sh extract_family_headers.sh
```

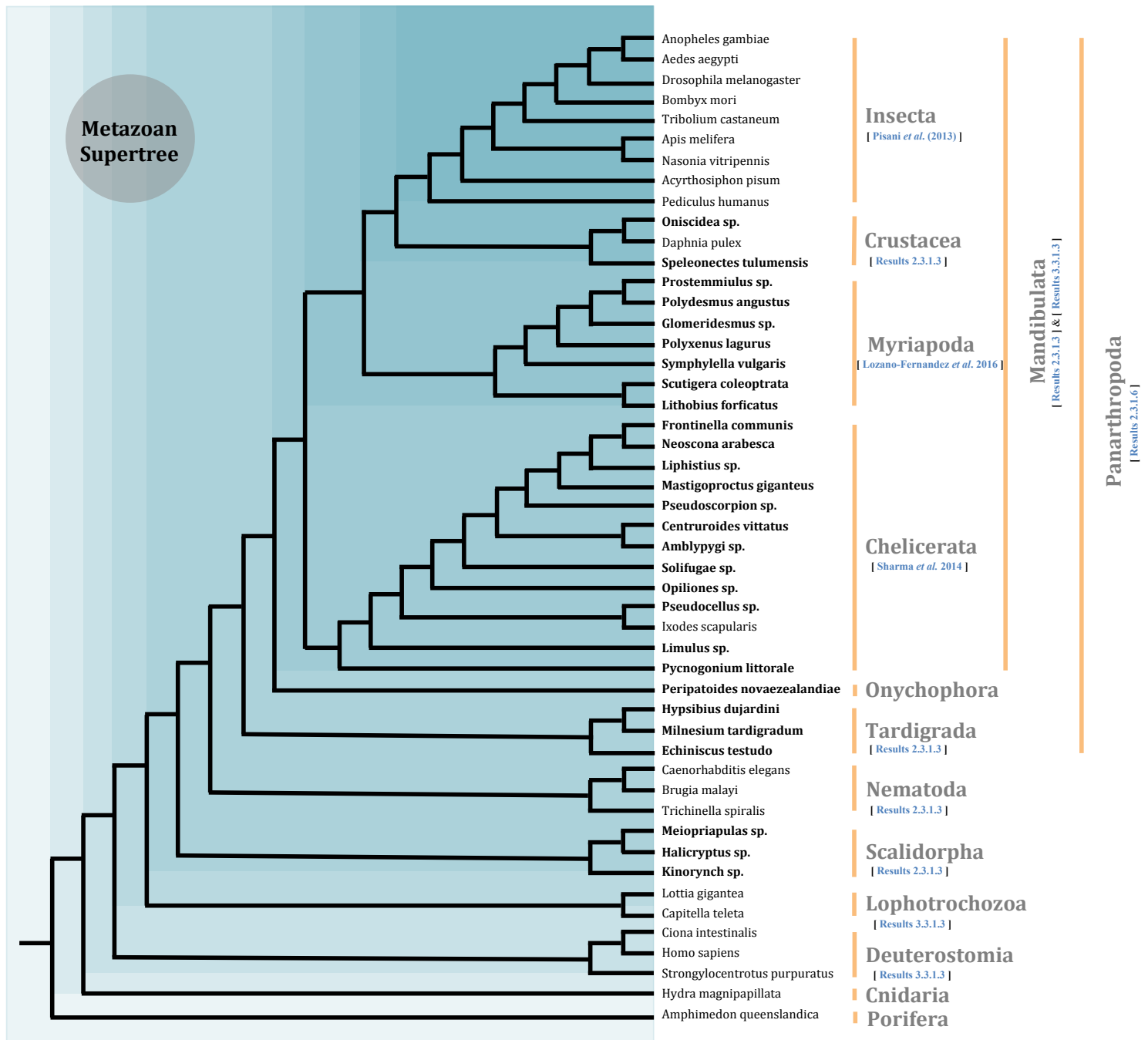
The sequence headers in each protein file were rearranged so that there was a single header per line. Then the sequences for each header for all protein families were unified using `match_fasta.py` [Supplementary Material 5.3] resulting in a designated fasta file for each protein family.

```
$ for i in *.txt; do python match_fasta.py MCL_database.fa $i; done
```



### 5.2.8 Distribution of Protein Families Across a Metazoan Supertree

Once the protein families had been generated the next step was to distribute the families throughout the metazoan clades and lineages. A topological framework was required in order to do this accurately. Building a new full sized phylogenetic dataset of all the metazoan taxa in this study would have been too laborious and time consuming for the scope of this project as the workload required to identify and map orthologs of all 49 members, in addition to the computational resources and timescale required for independent Phylobayes MCMC chains (Lartillot *et al.* 2009) to converge for multiple models (CAT, GTR, & CAT-GTR) would be far too large. Furthermore the molecular libraries for the taxa varied in size and quality mainly because many members from the initial study (Pisani *et al.* 2013) were sourced from EST and not NGS experiments resulting in a fragmented coverage of proteins. Consequently there would be no guarantee of decent ortholog coverage for all forty-nine representatives in the dataset. Finally, for the purposes of this study, a phylogenetic representation of the species (taking in to account branch lengths) was not necessary as only the cladistic relationships of the taxa was important when plotting the rate of protein family gain within nodes and lineages, therefore a cladogram would suffice. With this in mind the most efficient option was to construct a metazoan supertree pieced together from the phylogenetic results of the previous chapters and information from the literature. This supertree contained taxa from the Porifera, Cnidaria, and the Bilateria (mostly sampling the Protostomia) meaning the majority of metazoan groups were included with the exception of the Placozoa. See [Figure 5.3](#) for an illustration of the supertree and the justification for the topology.



**Figure 5.3: Supertree Metazoa**

A cladogram representing the topological assortment of the forty-nine metazoan lineages studied. New taxa generated from NGS experiments are in bold. Sources for the internal and external relationships of groups are provided in blue.

The *assign\_families\_to\_clades.py* script [Supplementary Material 5.4] parsed through the newly formed protein family files and designated them into one of the forty-nine lineages or forty-eight nodes in the tree based on the species composition of each protein family.

```
$ for i in *.txt.fa; do python assign_families_to_clades_v5.py $i; done
```

This involved counting the species in every sequence header of a particular protein family and storing it in a list. This list of species, for the protein family in question, was the identifier for which clade the family should be designated to. For example, a family containing *Amblypygi sp.*, *Opilione sp.*, and *F. communis* would be designated to the Arachnida. However, if the family contained *Amblypygi sp.*, *Opilione sp.*, *F. communis*, and *A. aegypti* then it would be assigned to the Arthropoda etc.

The prerequisite to assigning a protein family to a particular node of the tree was that it contained at least one sequence belonging to the oldest member of that node, this accounted for the incomplete nature of transcriptomic libraries and gene loss over time. The data was scaled, and a rate of protein family acquisition calculated, by dividing the number of families of each node by the total number of sequences used in the dataset. The preliminary study (Pisani *et al.* 2013) used the gene number of each genome to normalise the data but this was not possible for the expanded study. Given that many of the specimens sequenced are novel and from transcriptomic sources, often little is known about their genetic catalog. Transcriptome sizes are not an accurate reflection of the gene library of a genome as they are only representative of the genes that are expressed just prior to the mRNA extraction procedure (Chomczynski & Sacchi, 1987 and Wang *et al.* 2009). As such, they do not have a record for genes that are “turned off” or whose expression levels are so low that the QIAGENs extraction procedures cannot detect them. In addition, it is long known that

the idea of one gene producing a single protein product is a fallacy as many protein-coding genes are capable of producing multiple protein products (Marcotte *et al.* 1999). The final reason for not using the numbers of sequenced transcripts from NGS experiments as a substitute for the number of genes in an organisms genome is that transcripts can consist of multiple isoforms of the same gene and transcriptome libraries will contain multiple copies of the same transcripts depending on the level of expression. This greatly inflates libraries of transcriptomic nature making them poor representatives of the species genomic catalog.

Instead the protein families were normalized by dividing them by the total number of proteins clustered by MCL consisting of at least two sequences: 75,547.

### 5.2.9 BLAST2GO: Annotating Protein Families

Annotation of the new protein families was achieved using BLAST2GO (Conesa *et al.* 2005). BLAST2GO uses various gene ontology databases unified in a single consortium (Ashburner *et al.* 2000) as a method to robustly annotate the sequences making up the protein families of interest. It reaches this goal through a number of steps presented in the flowchart below [Figure 5.4]. Initially the BLAST package (Altschul, 1990) was used to compare the protein families of interest against the NCBI non-redundant protein database (nr). BLAST2GO only selects putative protein sequences that have a high quality match to a known protein sequence in the nr database. This ensures that only “real” proteins are funneled into the annotation pipeline.

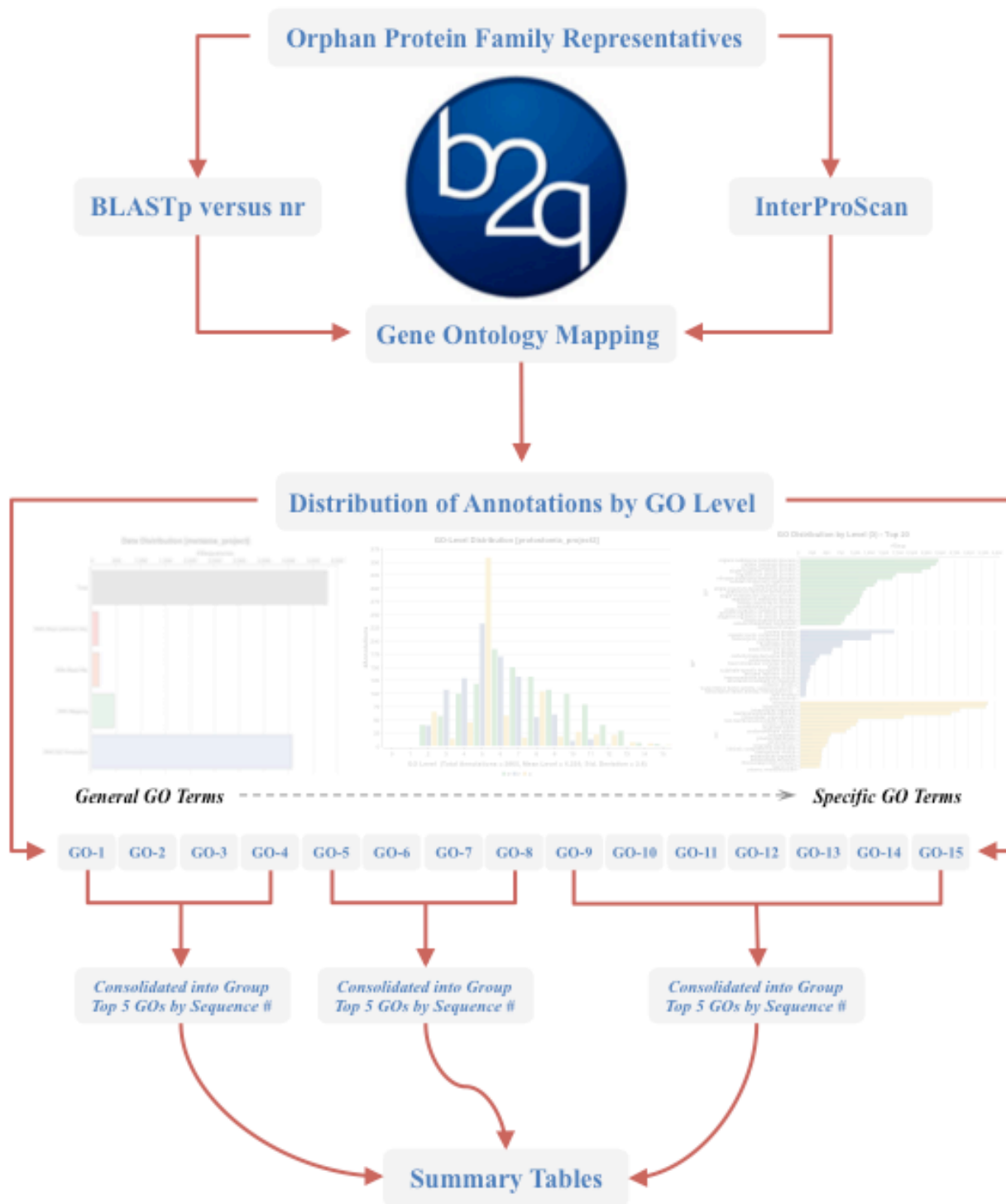
```
$ blastp -query [clade] -db nr.[00 - 66] -out [clade]-nr.[00 - 66] -evaluate 1e-10  
-outfmt 6
```

Following this, InterProScan (Quevillon *et al.* 2005) identified real protein domains within our families. InterProScan assigns known protein signatures to confirmed protein sequences with the aid of a large consortium of protein databases including Pfam, PROSITE, HAMAP, and PRINTS to name just a few. These commands were carried out through the BLAST2GO interface.

The BLAST and protein fingerprints results were then mapped to functional information in the GO database and assigned confidence ratings. Finally BLAST2GO (Conesa *et al.* 2005) sorted all the information gathered into an annotation table revolving around biological processes, molecular function, and cellular component categories. Due to limited computational resources and time, only the eight groups in **Figure 5.5** that displayed significant gains of new protein families compared to the mean were annotated, as the ever-growing nr database is enormous (composed of 70,427,238 proteins as of September 2016 <https://www.ncbi.nlm.nih.gov/refseq/>). This left six groups that had an above average rate of protein family acquisition without functional information. The eight taxonomic groups that were annotated are labeled in **Figure 5.6**. To increase the efficiency of the annotation of these families we took advantage of the fact that members belonging to the same protein families share similar functions (Dayhoff, 1976; Krause *et al.* 2005; and Demuth & Hahn, 2009). Therefore a single representative of each family should be a sufficient summary of each family as a whole. For example concerning the new metazoan protein families, this meant comparing a single representative of each family (4,771) instead of every sequence in the 4,771 protein families (343,295). This greatly sped up the annotation process.

At this stage it may be useful to write a short reminder on GO levels and their affects on results distribution. GO levels work in orders of specificity and are relative to each

dataset. Some of the annotated protein families returned up to twelve levels of GO information, which is twelve orders of specificity, which is twelve ways of distributing *the same functional information*. Others have returned shallower levels of GO, it depends on the number of sequences being annotated, their BLAST scores, coverage of InterProScan in relation to the sequences and the overlapping coverage between the nr and protein fingerprint databases. The nature of these levels of specificity means that shallower levels of GO distribution will be more concentrated, i.e. there will be larger clusters of proteins sharing the same broad function. As one moves up the levels of specificity these clusters will break up as proteins are sorted into more specific categories of functions. With this in mind one can consider the large amount of information that is generated and the difficulty of describing functional distribution. Does one take a broad or specific approach to describing these results? Which will describe the data the best? In order for these results to be biologically informative a balance must be struck. Too broad and the functional adaptations of the different groups of protein families will homogenize, too specific and it becomes difficult to infer meaning to a highly specified set of functions with a shallow distribution. An approach was taken to consolidate similar GO levels into single groups and from there select the five functions that represented the most sequences and put them in a summary table [Figure 5.4]. This prevented an overload of information from dozens of output graphs from BLAST2GO while simultaneously presenting the most relevant functional adaptations for each of the eight groups at multiple levels of specificity.



**Figure 5.4: Flowchart for Protein Family Annotation**

This flowchart displays the annotation procedure for eight of the groups that displayed a significant rate of orphan protein family acquisition. The differing levels of GO graphs (BLAST2GOs output) are consolidated into three groups of similar specificity. The higher the GO level the more specific its description of the functions. GO functions were ranked in order of number of sequences with those functions. The top 5 functions for each of the consolidated GO groups are summarised in **Results Tables 5.5 - 5.12** for biological process and molecular function categories.

## 5.3 Results

### 5.3.1 Data Quality and Assembly Statistics

Raw data statistics, quality control results from FASTQC, and assembly results for in-house sequencing projects and transcriptomes downloaded from the SRA can be found in [Supplementary Material 2.2](#).

### 5.3.2 MCL Protein Clustering

The Markov Cluster Algorithm produced 167,673 protein families (75,547 with two or more members) from 847,637 sequences, 457,646 of which were the product of new NGS sequencing experiments. An inflation value of 8 was chosen as this produced the most defined clusterings and more families numerically than the other inflation experiments [[Table 5.4](#)]. This inflation value was the same as the preliminary study (Pisani *et al.* 2013), which is useful for experimental consistency when comparing the two sets of results.

### 5.3.3 Protein Families

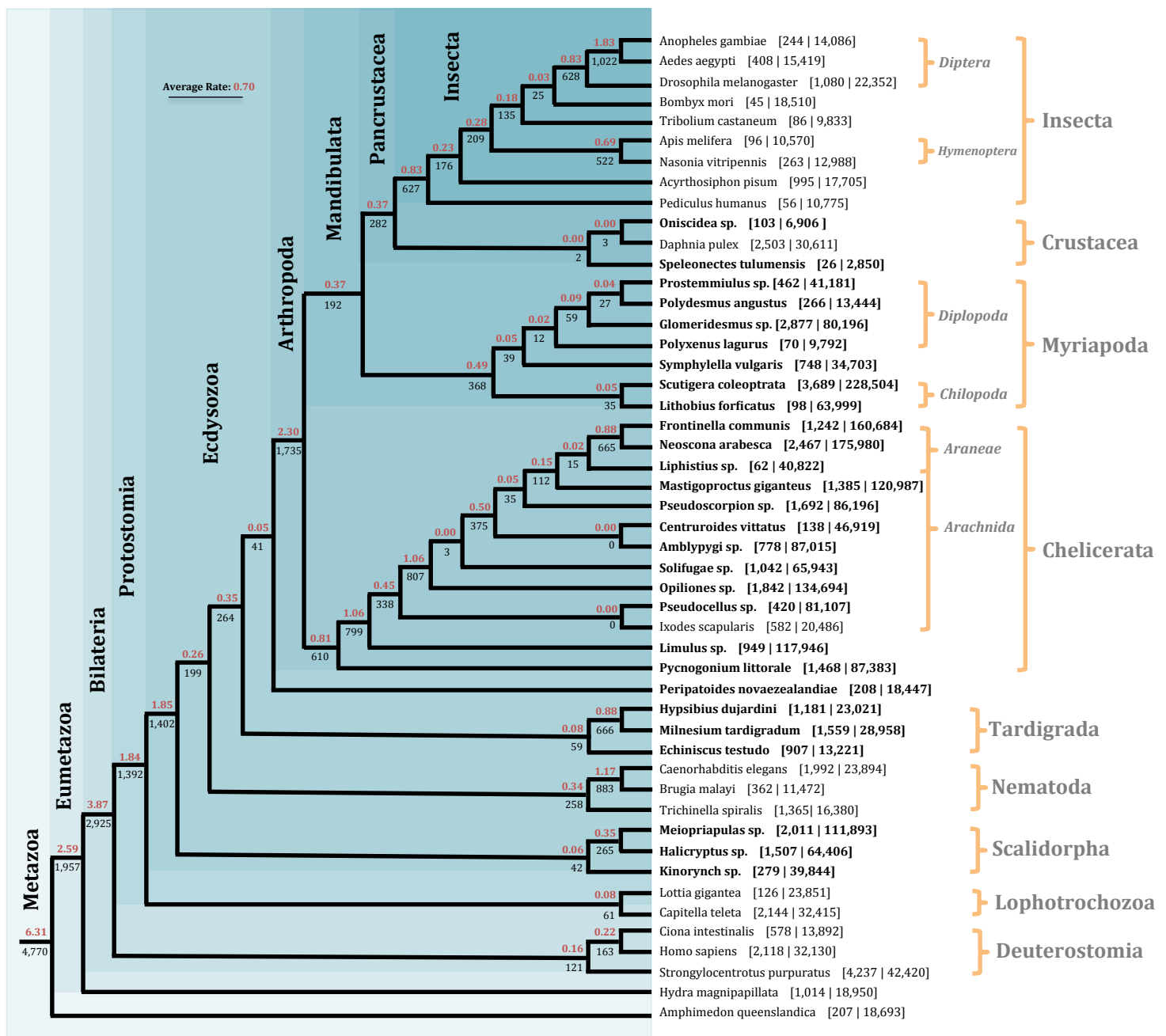
The 75,547 protein families were distributed amongst the taxa and ancestral nodes of the radiating lineages on the basis of the species composition of the protein families. An expanded cladogram from the preliminary study was used [[Supplementary Material 5.1](#)]. The justification for topology choice for the majorly expanded (sub)phyla is based on (Sharma *et al.* 2014) for the Chelicerata, (Lozano-Fernandez *et al.* 2016) for the Myriapoda, [Results 2.3.1.3](#) for the Crustacea, Tardigrada, and



Scalidorpha, **Results 2.3.1.3** & **Results 3.3.1.3** for the Mandubilata, and **Results 2.3.1.6** for the Panarthropoda. **Figure 5.5** depicts the number of orphan gene families for each node (number below node) and a scaled value for rate of protein family acquisition (red number above node).

The average number of orphan protein families per internal node in the tree (ancestor) was calculated as 527. This was the total number of orphan families spanning all internal nodes (25,290) divided by the number of internal nodes (48). The data was scaled by dividing the number of orphan families for each node by the total number of clustered sequences (consisting of at least two members) from MCL: 75,547. This follows that the mean rate of orphan protein family acquisition for the internal nodes of the tree is 0.7. Values above the mean were considered significant.

The numbers beside each taxon in square brackets represents the number of orphan genes in the family from which that lineage belongs. The second number in brackets represents the number of genes in the genome for taxa used in the initial study or the transcriptome size for newly sequenced taxa. Gene numbers are difficult to source for de-novo sequenced taxa as usually little is known about their genomic lexicon.



**Figure 5.5: Distribution of Metazoan Protein Families**

A metazoan supertree consisting of the taxa from the preliminary study and the twenty-eight newly sequenced taxa. Justification for the topology can be found in [Figure 5.3](#). The numbers in black below each node are the number of protein families. The numbers in red above each node are the scaled values for rate of protein family acquisition. The numbers beside each taxon in square brackets represents the number of orphan genes in the family from which that lineage belongs. The second number in brackets represents the number of genes in the genome for taxa used in the initial study or the transcriptome size for newly sequenced taxa. Newly sequenced taxa added to the study are in bold. Significant gains are seen in 14 groups.

### 5.3.4 Rate of New Protein Family Acquisition

As per **Figure 5.5**, the number of protein families for each node in the tree was scaled by dividing them by the total number of grouped proteins from the MCL clustering. The rate of new protein family gain was plotted against all forty-eight nodes in the study in order to identify any significant gains [**Figure 5.6**].

Fourteen out of the forty-eight nodes (29%) saw a rate of protein family acquisition above the mean, out of these fourteen groups eight experienced a dramatic level of gain. All fourteen groups are highlighted in black on the chart below, with a summary of the functional information for the eight groups with the most dramatic rise in new protein families. This functional data was obtained from **Results 5.3.5**.



### 5.3.5 Protein Family Annotation

Fourteen groups displayed a rate of protein family acquisition that deviated significantly from the mean. We consider these groups to have an unusual rate of protein family gain in relation to the other thirty-four groups (nodes in the supertree) included in the analysis. These groups were the Metazoa, the Eumetazoa, the Bilateria, the Protostomia, the Ecdysozoa, the ancestor of the Secernentea and Chromadorea (which belong in the Nematoda), the Eutardigrada, Arthropoda, Chelicerata, Xiphosura and Arachnida ancestor, two nodes within the spiders, Hexapoda, and the ancestor of *Anopheles* and *Aedes* (members of the Diptera). Due to a lack of resources not all of these groups could be annotated, instead the eight groups that saw protein family gains far greater than the mean were chosen.

Below are the functional statistics for the protein families of nodes that experienced a largely significant rate of gain. The gene ontologies are broken down into three categories: biological process (BP - green), metabolic function (MF - blue), and cellular location (CL - yellow). The results have been summarised in tables. See [Supplementary Material 5.5](#) for all the GO graphs for each of the eight groups.

## Metazoa

**Table 5.5: Prominent Functions of Metazoa Protein Families**

The table is divided into three levels of gene ontology specificity [GO 1-4], [GO 5-8], and [GO 9-15]. Both biological process and molecular function data are presented under these three levels of specificity. The functional specificity for the Metazoa ranged 12 levels.

Gene Ontology (GO) Levels: Annotation Distribution									
General GO								Specific GO	
GO 1	GO 2	GO 3	GO 4	GO 5	GO 6	GO 7	GO 8	GO 9 - GO 15	
Biological Process									
metabolic & cellular processes			gene expression				protein ubiquitination		
regulation of cellular processes			cellular protein metabolic process				RNA splicing		
transport			cellular protein modification				calcium ion homeostasis		
system development			neuron generation				peptidyl-lysine acetylation		
cell communication			nucleic acid transcription				regulation of cellular pH		
Molecular Function									
protein binding			DNA & RNA regulation				coupled ATPase activity		
compound binding			ATPase activity				hydrogen transportation		
transferase activity			transcription & RNA regulation				RNA polymerase II regulatory region		
nucleic acid binding			nucleoside-triphosphatase activity				RNA polymerase II promoter		
protein dimerization			nucleoside binding				multiple helicase activities		
Metazoan GO Levels: 12					Protein Family Annotation Percentage 83.3%				

The protein families originating in the Metazoa had roles in metabolic and cellular processes including the regulation of both. These protein families were also involved in most aspects of the gene to protein process: gene expression, transcription, and protein modification. Some of the metazoan protein families that have seen a significant rate of gain also take part in neuron generation. This is of particular interest as neuron networks are a defining characteristic of most animals (Caveney *et al.* 2006; Moroz, 2009; Jekely, 2013). From a general perspective, the Metazoa saw an increase in proteins involved in the formation and development of cellular operating systems, the only group to see such an increase.

## Eumetazoa

**Table 5.6: Prominent Functions of Eumetazoa Protein Families**

The table is divided into three levels of gene ontology specificity [GO 1-4], [GO 5-8], and [GO 9-15]. Both biological process and molecular function data are presented under these three levels of specificity. The functional specificity for the Eumetazoa ranged 10 levels.

Gene Ontology (GO) Levels: Annotation Distribution									
General GO								Specific GO	
GO 1	GO 2	GO 3	GO 4	GO 5	GO 6	GO 7	GO 8	GO 9 - GO 15	
Biological Function									
metabolic & cellular processes				nucleic acid-template transcription				regulation of cysteine endopeptidase	
biological regulation				regulation of nucleic acid-template transcription				calcium ion transport	
response to stimulus				protein phosphorylation				T cell apoptotic process	
localization				generation of neurons				neuron-neuro synaptic transmission	
signalling				divalent metal ion transport				regulation of pH	
Molecular Function									
catalytic activity				serine / threonine protein kinase activity				RNA polymerase II regulatory region	
binding				nucleoside-triphosphatase activity				ATPase activity coupled	
transferase activity				transcription regulatory region				ATP-dependent helicase	
transporter activity				ATPase activity				potassium & hydrogen transporter	
oxidoreductase				cation transmembrane transportation				cGMP activated cation channel	
Eumetazoan GO Levels: 10					Protein Family Annotation Percentage 10.9%				

The Eumetazoa protein families share a similar functional distribution to the Metazoa with some considerable exceptions. Most notably a gain in immune system proteins and it seems protein families originating in the Eumetazoa had a further focus on neuron development. This may have been an essential aid in adaptability and variability of the radiating animals lineages from the Eumetazoa. However these functions only cover 10.9 % of protein families for these groups and with such limited coverage it is unclear if they are representative of the rest of the families.

## Bilateria

**Table 5.7: Prominent Functions of Bilateria Protein Families**

The table is divided into three levels of gene ontology specificity [GO 1-4], [GO 5-8], and [GO 9-15]. Both biological process and molecular function data are presented under these three levels of specificity. The functional specificity for the Bilateria ranged 11 levels.

Gene Ontology (GO) Levels: Annotation Distribution										
General GO				→						Specific GO
GO 1	GO 2	GO 3	GO 4	GO 5	GO 6	GO 7	GO 8	GO 9 - GO 15		
Biological Process										
metabolic & cellular process			protein modification by conjugation				protein polyubiquitination			
biological regulation			protein catabolic process				peptidyl-lysine acetylation			
response to stimulus			peptidyl-lysine modification				regulation of cytosolic calcium			
biogenesis			regulation of protein kinase				iron ion homeostasis			
developmental process			DNA repair via homologous recombination				positive regulation of protein kinase C			
Molecular Function										
protein & ion binding			transcription regulatory regions				DNA-dependent ATPase activity			
organic compound binding			RNA polymerase II regulatory region				ATP-dependent ATPase activity			
transferase activity			ATPase activity				copper ion transporter			
hydrolase activity			motor activity				5'-3' DNA helicase activity			
small molecular binding			metal ion transporter				RNA-dependent ATPase activity			
Bilaterian GO Levels: 11					Protein Family Annotation Percentage 24.8%					

The bilaterian families also suffered from poor annotation coverage with only a quarter being designated functions. The most significant trend amongst the biological processes is the modification of proteins. These would be important processes for functional modifications for a multitude of proteins. The most dominant type of molecular functions are based on proteins catalyzing energy driven reactions.



## Protostomia

**Table 5.8: Prominent Functions of Protostomia Protein Families**

The table is divided into three levels of gene ontology specificity [GO 1-4], [GO 5-8], and [GO 9-15]. Both biological process and molecular function data are presented under these three levels of specificity. The functional specificity for the Protostomia ranged 11 levels.

Gene Ontology (GO) Levels: Annotation Distribution									
General GO								Specific GO	
GO 1	GO 2	GO 3	GO 4	GO 5	GO 6	GO 7	GO 8	GO 9 - GO 15	
<b>Biological Process</b>									
cellular & metabolic processes				RNA biosynthetic process				protein polyubiquitination	
single-organism process				nucleic acid-templated transcription				peptidyl-lysine acetylation	
biological regulation				generation of neurons				cytidine deamination	
response to stimulus				cation transmembrane transport				neuromuscular synaptic transmission	
signalling				protein modification by conjugation				NAD biosynthetic process	
<b>Molecular Function</b>									
catalytic activity				ATP metabolic process				ATP dependent helicase activity	
signal transduction				protein dephosphorylation				DNA dependent ATPase activity	
transporter activity				cation channel activity				protein DNA loading ATPase	
receptor activity				DNA binding				four-way junction helicase activity	
small molecule binding				metal ion transmembrane transport				ATP dependent RNA helicase activity	
Protostome GO Levels: 11					Protein Family Annotation Percentage: 63.1%				

The Protostomia are set apart from the older animal groups with their notable adaptations in regards to environmental stimulus, receptor activity, and cell signaling. The trend of protein families influencing neurological processes continues from the Metazoa. Additionally there seems to have been a large influence of new protostome protein families involved in cellular energy reactions.

## Ecdysozoa

**Table 5.9: Prominent Functions of Ecdysozoa Protein Families**

The table is divided into three levels of gene ontology specificity [GO 1-4], [GO 5-8], and [GO 9-15]. Both biological process and molecular function data are presented under these three levels of specificity. The functional specificity for the Ecdysozoa ranged 10 levels.

Gene Ontology (GO) Levels: Annotation Distribution										
General GO				→						Specific GO
GO 1	GO 2	GO 3	GO 4	GO 5	GO 6	GO 7	GO 8	GO 9 - GO 15		
<b>Biological Process</b>										
metabolic & cellular processes			cellular protein modification				protein ubiquitination			
macromolecule metabolic processes			protein phosphorylation				peptidyl-tyrosine phosphorylation			
protein metabolic processes			neuron generation				RNA splicing			
cell communication			protein modification				activation of phospholipase C			
signal transduction			negative regulation of endopeptase				iron ion import			
<b>Molecular Function</b>										
nucleic acid, cation, nucleotide, binding				purine ribonucleoside binding			microtubule motor activity			
catalytic activity				protein kinase activity			ATPase activity coupled			
transporter activity				ATPase activity			histone acetyltransferase			
transferase activity				protein tyrosine phosphatase			type II deoxyribonuclease			
peptidase activity				helicase activity			hydrogen & calcium ion transporters			
Ecdysozoan GO Levels: 10					Protein Family Annotation Percentage 43.5%					

The protein families originating from the Ecdysozoa have very similar function to their preceding groups. These proteins have had roles in neuron development, protein modification, which is essential for functional adaptation, and energy interactions.

## Secernentea & Chromadorea Ancestor

**Table 5.10 Prominent Functions of the Secernentea & Chromadorea Ancestor Protein Families**

The table is divided into three levels of gene ontology specificity [GO 1-4], [GO 5-8], and [GO 9-15]. Both biological process and molecular function data are presented under these three levels of specificity. The functional specificity for the Secernentea & Chromadorea ancestor ranged 10 levels.

Gene Ontology (GO) Levels: Annotation Distribution										
General GO				→						Specific GO
GO 1	GO 2	GO 3	GO 4	GO 5	GO 6	GO 7	GO 8	GO 9 - GO 15		
<b>Biological Process</b>										
multicellular organism development				embryo development				protein ubiquitination		
macromolecule metabolic process				gene expression				synaptic transmission		
anatomical structure development				nematode larval development				ATP synthesis		
reproduction				cellular & DNA regulation				RNA splicing via transesterification		
biological regulation				ATP metabolic process				calcium ion transport		
<b>Molecular Function</b>										
nucleic acid & cation binding				various molecule binders				RNA polymerase II regulatory region		
substrate transmembrane transporters				ion & cation channel activity				hydrogen, potassium, sodium, calcium transporter		
hydrolase & transferase activity				cation transmembrane transporters				ATPase activity, coupled		
signal transducer activity				RNA polymerase II regulation				voltage-gated calcium channel		
transmembrane receptor activity				nucleoside-triphosphatase activity				microtubule motor activity		
Secernentea - Chromadorea Ancestor GO Levels: 10					Protein Family Annotation Percentage 74.7%					

These nematodes display similar molecular functions in their protein families to their ecdysozoan cousins but the notable differences are seen in their biological processes. One of the most significant processes for which the orphan protein families of the *Secernentea* and *Chromadorea* ancestor contribute to is nematode larval development. Following this, the top biological adaptations born from these families include embryo development, reproduction, locomotion, proteins influencing adult lifespan, endocytosis and lipid storage. These adaptive functions are very much developmentally themed.

## Arthropoda

**Table 5.11: Prominent Functions of Arthropoda Protein Families**

The table is divided into three levels of gene ontology specificity [GO 1-4], [GO 5-8], and [GO 9-15]. Both biological process and molecular function data are presented under these three levels of specificity. The functional specificity for the Arthropoda ranged 10 levels.

Gene Ontology (GO) Levels: Annotation Distribution									
General GO				→ Specific GO					
GO 1	GO 2	GO 3	GO 4	GO 5	GO 6	GO 7	GO 8	GO 9 - GO 15	
Biological Process									
metabolic & cellular processes			gene expression				protein ubiquitination		
biological regulation			RNA biosynthetic process				ATP synthesis		
regulation of biological processes			regulation of biosynthetic processes				peptidyl-lysine acetylation		
response to stimulus			nucleic acid-templated transcription				calcium ion transport		
cell signalling			cation transmembrane transport				histone phosphorylation		
Molecular Function									
nucleic acid & cation binding			zinc ion binding				ATPase activity coupled		
transferase activity			nucleoside-triphosphatase activity				hydrogen, calcium, potassium transporters		
protein, ion, carbohydrate binding			transcription regulation				ion & calcium channel activity		
signalling receptor activity			ATPase activity				ATP-dependent helicase activity		
chitin binding			helicase activity				RNA polymerase II distal enhancer		
Arthropod GO Levels: 10					Protein Family Annotation Percentage 35.7%				

The coverage for the Arthropod protein families was poor, with just 35.7% of representative sequences annotated. However, the most interesting functional gain we see amongst the Arthropoda is in chitin metabolic processes. This is of particular significance because chitin is a major component of the arthropod exoskeleton and considered one of the defining features of the phylum, see Rebers & Willis (2001) for supporting evidence for protein chitin receptors in arthropods. Other small gains are seen in DNA regulation, protein regulation and adaptation, oxidation-reduction processes, and signal transduction.

## Anopheles & Aedes Ancestor

**Table 5.12: Prominent Functions of Anopheles & Aedes Ancestor Protein Families**

The table is divided into three levels of gene ontology specificity [GO 1-4], [GO 5-8], and [GO 9-15]. Both biological process and molecular function data are presented under these three levels of specificity. The functional specificity for the Anopheles & Aedes Ancestor ranged 10 levels.

Gene Ontology (GO) Levels: Annotation Distribution									
General GO				Specific GO					
GO 1	GO 2	GO 3	GO 4	GO 5	GO 6	GO 7	GO 8	GO 9 - GO 15	
<b>Biological Process</b>									
macromolecule metabolic process			sensory perception to chemical stimulus				metal ion homeostasis		
protein macromolecule process			sensory perception of smell				iron ion transport		
biological regulation			gene expression				protein ubiquitination		
cellular response to stimulus			nucleic acid-templated transcription & regulation				tRNA splicing via endonucleolytic		
single organism signalling			sensory perception of taste				regulation of cAMP-dependent kinase		
<b>Molecular Function</b>									
organic & heterocyclic compound binding			element binding				microtubule motor activity		
hydrolase activity			cation transmembrane transport				potassium & sodium ion transport		
odorant binding			ATPase activity				ATPase activity coupled		
hydrolase activity			endopeptidase activity				DNA helicase activity		
signalling receptor activity			olfactory receptor activity				RNA polymerase II regulatory region		
Anopheles - Aedes Ancestor GO Levels: 10					Protein Family Annotation Percentage 53.1%				

The most prominent functions of the families originating in the *Anopheles – Aedes* ancestor are receptor based, involved in the binding of various molecules such as zinc, ion and even more complex molecules such as nucleic acids and odorants. Other significant gains in functions include olfactory receptors and endopeptase activity. The more broad biological processes metric returns similar results, significant numbers of sensory stimulus receptors, proteins involved in sensory perception, taste perception, signal transduction in addition to proteins involved in molecular regulation such as proteolysis, transcription regulation, and phosphorylation. Based on these annotations, it seems these dipterans have gained a significant number of sensory proteins compared to the rest of the flies as well as many of their insect and arthropod cousins.

## 5.4 Discussion

### 5.4.1 Adjusting the Balance of Sequenced Ecdysozoans

As discussed in chapter 1, before NGS became widely available, molecular data for non-model organisms was restricted to small gene dataset experiments. This lack of data has hindered our ability to reliably study large groups of animals such as the Protostomia due to poor taxon coverage and an insufficient quantity of molecular libraries. The Insecta is arguably the most studied group of the Ecdysozoa due to their economic impact on agriculture (Kevan *et al.* 1990; Robinson *et al.* 2004; Calderone, 2012), thus it is of no surprise that before high throughput sequencing data became available, the insects comprised the major source of arthropod sequence data.

One of the goals of this project was to contribute to the library of knowledge of protostome sequence data (specifically ecdysozoans) by supplementing under sampled clades with newly sequenced taxa using NGS technology. To this aim we have generated transcriptomic data for the Chelicerata: *Amblypygi sp.*, *Limulus sp.*, *Opilione Sp.*, *P. littorale*, *Galeodidae sp.*, the Crustacea: *Oniscidea sp.*, the Priapulida: *Halicryptus sp.*, *Meiopriapulid sp.*, and the Onychophora: *Epiperipatus sp* [Figure 5.2]. Furthermore we have availed of sixteen NGS projects whose raw data has been deposited in the SRA to further augment our protostome molecular library [Table 5.2]. The growth trend of the SRA is ever increasing, some thirty-fold since the commencement of this project [Figure 1.7] and because of its open accessibility it is an incredibly valuable resource to molecular biologists around the world, arguably the most significant molecular database since the inception of the NCBI (Pruitt *et al.* 2005) and the Ensembl genome browser (Filcek *et al.* 2012).

### 5.4.2 Data Quality of the SRA

A concerning discovery from our data analyses was the overall poor quality of raw sequence data downloaded from the SRA. This ranged from fixable errors such as reads containing sequence adapters to more detrimental errors such as poor phred scores, directly caused by inaccurate base calling during the sequencing process.

From the sixteen transcriptomes downloaded from the SRA, 37% contained adapters, 37% had an average phred score below an acceptable margin for accurate base calling, 12% tested positive for contaminants, and overall 87% were flagged for at least one form of quality control issue [[Supplementary Material 2.2](#)].

Adapters and foreign RNA contaminants can be removed from transcriptomic data with relative ease using adapter-trimming software and through BLAST comparisons to the NCBI's non-redundant protein database (nr) (Pruitt *et al.* 2005). However, below acceptable phred scores, due to inaccurate base calling during the sequencing process, is a much more complicated problem as the likelihood of bases being erroneously called is not insignificant and calls the reliability of the transcriptome into question.

What we cannot establish at this juncture is if the sixteen raw data projects, downloaded as part of this study and flagged for quality control issues, are a representative sample of the sequence quality of the SRA. The sample size is only a fraction of the 5,000 tera bases (Tb) currently residing in the database. Only an extensive large-scale quality control analysis of a statistically significant sample size, representative of the database, could address this question.

However, what we do know, by virtue of the failed QC tests, is that the SRA either has no quality control standards for data submission or these standards are not being

enforced frequently enough. The quality regulation of the SRA may be the next step in improving this essential resource for molecular biologists.

Finally, it is also important to highlight that many raw data projects submitted to the SRA, in fact most of the sixteen downloaded and used in this study, are the data source for published peer reviewed studies (Brewer & Bond, 2013; Borner *et al.* 2014; Fernandez *et al.* 2014; von Reumont *et al.* 2014), many from highly respected scientific institutions. The importance of sequence quality and reliability cannot be understated as if the underlying raw data is poor then the methods, results and conclusions of an experiment must be brought into question. Unfortunately it seems the quality of NGS data is not yet a talking point in the field of molecular biology, perhaps drowned out by the race to sequence as many taxa as possible in this new age of data acquisition. This is a worrying trend, with the ever growing SRA and use of its raw data in many collaborative projects around the world one wonders how long will it take for the reliability and consistency of this resource be addressed for the good of future molecular studies.



### 5.4.3 Future Improvements to Experimental Design

Although NGS experiments have greatly increased the usefulness of phylostratigraphic studies by giving us the opportunity to supplement taxon poor clades with large molecular libraries, there are still aspects of this study that can be improved upon. Mainly the use of genomic data instead of transcriptomic data. Transcriptomic libraries extracted from an organism and sequenced represent the catalogue of expressed genes just prior to extraction procedures. This is not representative of the organism's entire genomic library as it excludes genes not producing a mRNA product or genes that are very lowly expressed. Therefore transcriptomic libraries, while very informative, are not the full genomic picture. If each taxa in this study was represented by their entire catalog of genes it would be a much more complete study. The reasons for not using genomic libraries for our protein family study range from the increased costs to do so, computational resources (genome assembly is much more intense compared to that of the transcriptome assemblies), and the infancy of open-source gene prediction software such as AUGUSTUS (Stanke & Waack, 2003) which is difficult to configure for de-novo sequenced animals as complex as the Bilateria, furthermore gene prediction is not required for transcriptomes, as we know mRNA are gene products and thus there is no requirement to identify them from non-coding DNA.

Furthermore, as more newly sequenced taxa belonging to under-sampled phyla and subphyla become available, the addition of more data to other Protostomes in this study such as the Crustacea subphyla, Nematoda phyla, and in particular superphyla Lophotrochozoa and Deuterostomia would be useful in generating a complete study of the evolution of protostome protein families.

## 5.5 Conclusions

Our NGS approach to greatly increase the quality and scope of the preliminary study (Pisani *et al.* 2013), with the addition of twenty-eight taxa and 457,646 sequences, has revealed interesting biological information concerning the pattern of protein family origins in the Animal Kingdom that had been previously hidden because of the lack of data.

The acquisition of new protein families played a role in the formation of high-level taxonomic grouping. This is evident by significant gains in large animal defining clades such as the Metazoa, Eumetazoa, Bilateria, Protostomia, Ecdysozoa, Arthropoda, and Chelicerata. The two exceptions are the deuterostomes and lophotrochozoans, however this may be due to an under sampling of these groups. For this study the deuterostomes are only represented by three taxa: *H. sapiens*, *C. intestinalis*, and *S. purpuratus*. Comparing the results of this study to the preliminary (Pisani *et al.* 2013) it is reasonable to assume that sparse taxon sampling and restrictive EST datasets do have an influence on the results of phylostratigraphic experiments. We see insignificant to no gains in most of the internal node (younger) groups, meaning the rate of protein family acquisition towards the tips of the tree is mostly neutral (there are some exceptions to this). This suggests that the protein family networks required for these organisms were already developed at an earlier stage. Essentially, one can infer that new protein families arise in broad taxonomic groupings such as kingdoms, superphyla, phyla etc. and further adaptations within the radiating groups are the result of the tweaking and re-wiring of these families and their protein-protein interaction networks, specifically changes in cellular processes, regulatory functions, and developmental networks. This pattern falls in line with

Erwin *et al*'s (2011) description of the Cambrian explosion, where the molecular architecture of extant animal lineages may have been put in place long before their existence in the form of evolutionary acquisitions of their ancient ancestors. These new protein families in the ancient animal nodes could have acted as the blueprints, the adaptations, regulation, and re-wiring of which led to radiation of diverse lineages (see Knoll & Carroll 1999). Similarly the formation of new protein families were almost entirely restricted to broad level taxonomic groups, the radiating lineages did not for the most part acquire new protein families, instead it is conceivable they were altered to promote diversification.

As alluded to earlier, we do see a significant rate of new protein family gains in some nodes closer to the tips of the tree. Within the Nematoda, the *Secernentea* and *Chromadoreia* ancestor, within two nodes of the chelicerates, one of the tardigrade orders, and within the Diptera, the *Anopheles* and *Aedes* ancestor. Interestingly, the trend exclusively shared between the two annotated nodes (concerning nematodes and dipterans) is a more specific phenotypic gain in protein families, as opposed to the acquisitions seen in the older nodes which tend to relate to molecular processes such as proteins involved in gene expression and regulation, protein modification, cell signaling and ATP catalytic activity. The *Secernentea* and *Chromadoreia* ancestor has seen a clear gain of protein families involved in developmental processes such as embryo development and a form of larval development unique to nematodes. Concerning the ancestor of *Anopheles* and *Aedes*, the gain in functions such as perception of chemical stimulus, smell, and taste in addition to signaling receptor activity suggests a functional adaptation in sensory protein families.

There was no pattern of protein gain during the terrestrialization events within the arthropods. With the ancestral arthropod being marine based (Maloof *et al.* 2000),

there were independent land colonization events in the Hexapoda, Myriapoda and Arachnida (the chelicerate outgroups are marine based). With a neutral gain of newly generated protein families across most of these subphyla and class, there is no evidence that protein family acquisition played a role in this complicated process. However there was an influx of protein families within the chelicerates but not in the ancestral node considered relevant to land colonization (the terrestrial arachnid ancestor). Finally an above average rate of protein family generation was also found within the tardigrades concerning the Eutardigrada (*H. dujardini* and *M. tardigradum*). This is interesting as it is unclear as to why this occurred in one tardigrade order but not the other (Heterotardigrada). The essential differences between these groups are morphological characteristics such as arrangement of their gonopore, anus, “Malpighian tubules”, and pharynx structure (Guidetti & Bertolani, 2005) so it’s possible the increase in protein family gain amongst the Eutardigrada may have been to help facilitate these morphological adaptations.

Gene ontology is not only useful for annotating proteins; its structured universal vocabulary makes it a powerful and consistent tool for describing these functions. The ever-expanding GO database is powered by AmiGO 2 (<http://amigo.geneontology.org/amigo/landing>). AmiGO 2 is a search engine designed to link GO queries to functional descriptions identified by experimentation, information from protein fingerprint databases, similarity to other known GO terms, genomic context, species context, and a combination of these factors. AmiGO 2 also serves a database of the entire GO lexicon that can be browsed by GO number or term. We can avail of this database to gain further insight into the functions of the most common GO terms applied to protein families across all eight groups with a

significant rate of gain. There were 12 GO terms that were continually represented in the top five functions across all eight groups. A description of each is below including an inference in what a gain in protein families involved in such functions could have meant in the formation of high-level taxonomic groups.

*Catalytic activity* (GO:0003824)

A variety of enzymes possessing specific binding sites for complementary substrates in order to take run a multitude of cellular processes.

*ATP dependent activity* (GO:0008026)

This is a broad description of reactions that require ATP in order to process. These involved DNA & RNA binding, helicase and protease activity.

*ATPase activity, coupled* (GO:0042623)

A catalytic reaction that drives another reaction, specifically linked with ion transportation across cell membranes.

*Signal transduction* (GO:0007165)

A process whereby the signal changes the state or activity of a cell. Signal transduction can originate intra or extra cellularly and involves the reception of the signal followed by a downstream regulation of a cellular process such as transcription regulation.

*DNA helicase activity* (GO:0003678)

The action of unwinding DNA with the helicase enzyme. Activity with such an enzyme is almost certainly linked with transcription as this is the primary purpose for unwinding super coiled DNA.

*Ion transport* (GO:0006811)

Which may be linked to coupled ATPase activity, considers the involvement of protein families in moving ions pertaining to hydrogen, calcium, iron, zinc, and copper within or between cells.

*Ion channel activity* (GO:0005216)

One of the media by which the above ions are transported through a cell membrane. This function is associated with protein families which facilitate this process via diffusion.

*Microtubule motor activity* (GO:0003777)

Microtubules are found in the cytoplasm of the cell often referred to as the cellular skeleton. Protein activity influencing these structures tends to involve enzymes catalysing the movement often coupled with the hydrolysis of ATP.

*Protein modification* (GO:0036211)

Protein modifications take the form of covalent alteration of one or more amino acids in a protein. This includes the co-translational and post-translational stages. The most abundant form of modification amongst the eight groups was protein phosphorylation. Such modification can alter the functional role of the protein.

*Neuron generation* (GO:0048699)

A significant amount of neuron generation is seen in new protein families from the older nodes in the tree: Metazoa, Eumetazoa, and Ecdysozoa. This GO term speaks for itself, the formation of nerve cells and neuroblasts.

*Response to stimulus* (GO:0050896)

The detection of an internal or external stimulus which results in the change of activity of a cell whether it be gene expression, enzyme activity, pH regulation ect.

*Biological regulation* (GO:0065007)

A very broad term that attributes the regulation of any biological process at the molecular level. There is very little that one can infer from such a term.

The common theme amongst the most prominent annotations is that of every day cellular processes, particularly enzyme activity. An increase in such may have been needed for an organism in response to increased demands on metabolism as these processes could contribute to the increased workload required for the formation of the complex animal body plan for example, or development of new morphological features. The most notable GO term prominent amongst protein families is that of neuron generation. Protein families involving neuron formation saw an increase in most of the older nodes on four separate occasions, from the Metazoa to the Ecdysozoa with the exception of the Bilateria (which only had a 20% annotation rate). This is a clear sign of neurological development over time, an essential biological network found in contemporary animal systems.

### 6.1 Phylogenomics, an Important Step Forward

The improvements NGS and phylogenomic datasets have had on molecular evolution studies are evident in this thesis. The threat of stochastic errors in phylogenetic reconstructions have virtually been eradicated, assuming a thoughtful approach to taxon sampling for the clades of interest. Increasing the number of taxa in the preliminary phylostratigraphic study of orphan gene families (Pisani *et al.* 2013) with NGS libraries revealed a large amount of evolutionary information previously concealed by a lack of data. Sequencing the genome of the *Parasagitta sp.* was essential in the clarification of the chaetognaths placement within the Protostomia in addition to highlighting a near total extinction of the chaetognaths, revealing that extant chaetognaths are much younger than their oldest fossils, the direct descendent lineages of which no longer exist. Furthermore, NGS data enabled a more substantial investigation into evolutionary events that changes the landscape of how animals live on earth: terrestrialization.

However it is important to note that such technological advancement has introduced new problems such as data quality issues and accentuated others such as systematic errors, specifically LBA. When applied to rapidly evolving lineages such as the nematodes and tardigrades, one can generate conflicting phylogenies when the data is being inflicted by systematic errors such as LBA.



## 6.2 Experiment Chapter Summaries

### 6.2.1 Chapter 2

Chapter 2 demonstrated the strengths and weaknesses of phylogenomic datasets. Firstly it added more molecular data to the tardigrades than any preceding study. However we showed that the Tardigrada and Nematoda are highly susceptible to LBA error based on our signal dissection results. The addition of more data tends to accentuate positively misleading systematic artifacts and artificially inflate PP support values (Felsenstein 1978) making it difficult to follow the underlying phylogenetic signal. Ultimately results showed that the nematode - tardigrade grouping is most likely incorrect and there is no molecular evidence for an exclusive arthropod - tardigrade sister grouping. However, our phylogenomic experiments were unable to discern the exact position of the Tardigrada as there was strong evidence for both the Panarthropoda (Campbell *et al.* 2010 and Rota-Stabelli *et al.* 2011) and a previously unreported grouping of the tardigrades and onychophorans. In addition, further clarity concerning the arthropod subphyla as the Mandibulata was recovered over the Myriochelata and the origins of the Tardigrada was dated as 480 MYA.

### 6.2.2 Chapter 3

Sequencing the *Parasagitta sp.* genome revealed the internal chaetognath relationships: the *Parasagitta* and *Flaccisagitta* genii share a most recent common ancestor with the *Spadella* genus as the outgroup. Our phylogenomic analyses place the chaetognaths as basal lophotrochozoans. Molecular clock results show a large 350 MY disparity between extant chaetognaths and the fossils [Figure 3.10]. This

suggests a large extinction event wiped out the oldest chaetognath lineages, leaving only the crown group chaetognaths, indicating a major shift of the chaetognath position in the food web over the last 500 MY, from one of the first predators in the ocean to the extinction of many of its lineages and current role in the oceans food chain as the primary composite of plankton.

### **6.2.3 Chapter 4**

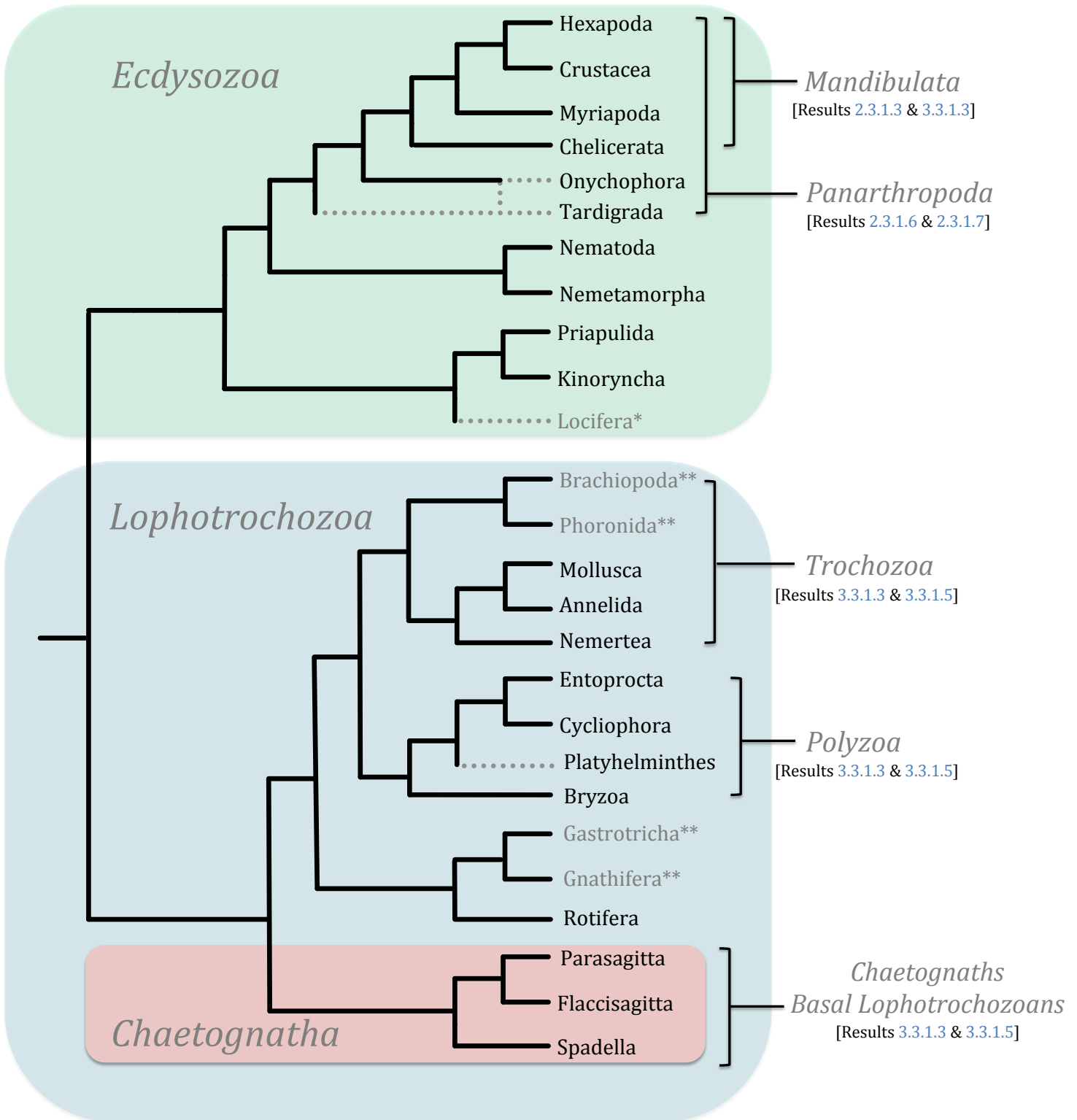
Possibly the most interesting development from the terrestrialization study is the evidence of an invertebrate land colonization event as early as the Cambrian. A Myriapod divergence time estimation of 528 MYA does not fit with their fossil record, the oldest known being from the Silurian 426 MYA (Wilson & Anderson 2004). However, our divergence time estimations do coincide with what appear to be myriapod trace fossils from the Cambrian (MacNaughton *et al.* 2002). It is important to mention that the molecular libraries for the myriapods were from NGS experiments, a further example of how this technology is influencing the field. Our findings also assigned the terrestrialization of the other arthropod subphyla in to geological time. The Hexapoda colonized land in the Ordovician and the Arachnids (not the chelicerates as the Pycnogonida - Xiphosura outgroup is marine based) in the Ordovician or Silurian. Furthermore, an ancestral character state reconstruction by Mark Puttick of the University of Bristol revealed the route arthropods took to land, via the oceans.

#### 6.2.4 Chapter 5

Chapter 5 involved a large-scale study of orphan protein families ranging across most of the Animal Kingdom, greatly augmenting a preliminary study (Pisani *et al.* 2013) with twenty-eight newly sequenced taxa, mostly ecdysozoans, consisting of over 400,000 proteins. In a clear demonstration of the benefits of NGS data to molecular evolution studies, our new dataset uncovered a series of significant orphan protein families in fourteen, most of which had been missed by the initial study. A pattern of orphan protein family acquisition was uncovered in the ancestral nodes of most high-level taxonomic animal groups [Figure 5.5 & Figure 5.6]. A neutral gain in families was seen emanating towards the crown group metazoans with notable exceptions within the nematodes, dipterans tardigrades, and chelicerates. Families within the first two of these groups experience a phenotypic growth in new protein families focusing on development and sensory perception respectively. Most new protein families in the ancient animal nodes revolved around routine cellular processes such as enzyme reactions, protein receptors, ion transport and in particular genetic regulation and protein modification. The latter functions are important influences in the adaptation of molecular processes. Interesting patterns include a unique gain in chitin related proteins in the Arthropoda, a substance that makes an important part of their exoskeleton. It also points to the possibility, at least from the point of view of protein families, that the unprecedented level of biodiversity amongst the arthropods is the result of alterations of existing biological systems as opposed to the generation of new ones. In addition to this there was a striking overall trend of continual gain of proteins involved in neuron generation for most of the ancient animal nodes. This is of particular significance as neurological systems are a keystone in animal biology (Moroz, 2009).

## 6.3 Suggested Alterations to Protostome Phylogeny

Based on our findings of the tardigrade and chaetognath studies, the protostome cladogram illustrated in chapter 1 [Figure 1.8] can now be updated to reflect the findings of this thesis [Figure 6.1]. Beginning with the Ecdysozoa, the interrelationships of the arthropod subphyla are resolved as the Mandibulata was consistently recovered over the Myriochelata when using the best fitting model for the data (Results 2.3.1.3 & Results 3.3.1.3). The Tardigrada are either a member of the Panarthropoda (Results 2.3.1.6) or sister to the Onychophora (Results 2.3.1.7). We have shown that the tardigrade - nematode affinity is a LBA artifact and found no support under any molecular scenario for the Tactopoda. Regarding the remaining protostomes: the Lophotrochozoa, the Trochozoa and Polyzoa topologies were recovered and remained stable under phylogenetic experiments in chapter 3, with the Mollusca, Annelida, Nemertea, and Entoprocta, Cycliophora, Bryzoa forming respective monophyletic groups (Results 3.3.1.3 & 3.3.1.5). None of the phylogenetic models returned a monophyly between the prospective platyzoan taxa that were sampled for this dataset: the Rotifera, Cycliophora, and Platyhelminthes; further suggesting this group is an artifact in agreement with Kocot *et al.* (2016) which consisted of the remaining “platyzoans”, the Gnathostomulida and Gastrotricha. The chaetognaths are basal lophotrochozoans, diverging before the aforementioned phyla, as evident by Results 3.3.1.3 & 3.3.1.5. Additional representative phyla, not available at the time of this study, have been added from the recent literature Yamasaki *et al.* (2015) for the Locifera and Kocot *et al.* (2016) for the Brachiopoda, Phoronida, Gastrotricha, and Gnathifera.



**Figure 6.1: Revised Protostome Relationships**

A protostome supertree representing the results of chapters 2 and 3 in conjunction with a further sampling of phyla from recent literature to complete the tree. Phyla from literature are represented in grey at the tips of the cladogram. The Kocot dataset chosen was the one designed to limit the influence of systematic biases; it was also the most congruous to our reconstruction of the Lophotrochozoa out of their eight phylogenetic experiments. Dotted branches construe topological uncertainty.

\* Yamasaki *et al.* (2015)

\*\*Kocot *et al.* (2016)

## 6.4 Reflections on the Cambrian Explosion

Results for each of the four experimental chapters knit together to broadly fit the Erwin *et al.* (2011) narrative for the Cambrian explosion. Independent divergence time estimations applying molecular clocks bound by the fossil record from the tardigrade, chaetognath, and terrestrialization studies place the origins of animals pre-Cambrian and chapter 4 even contains evidence of a Cambrian migration on to land for the Myriapoda (Lozano-Fernandez *et al.* 2016). The phylostratigraphic investigation into orphan protein families belonging to the Metazoa shows a pattern of acquisition in the large animal-defining clades such as the Metazoa, Eumetazoa, Bilateria, Protostomia, Ecdysozoa, and Arthropoda, followed by a neutral rate of novel protein family gain in the crown group metazoans. This suggests a system of evolutionary diversification where the blueprints for all protein families required for diversification were developed in the older nodes of the Metazoan tree pre-Cambrian. Therefore, the diversification and radiation of animal lineages in the Cambrian may have been facilitated by early predators such as the ancient chaetognaths whose threat prompted a rapid rate of morphological innovation. The protein families inherited from ancestral animals (the high-level taxonomic groups in the tree) were some of the tools needed to facilitate such adaptations and most likely achieved such a feat by tweaking the already existing networks.

## 6.5 Discussion of Thesis Findings

This thesis has demonstrated the benefits of applying next generation sequencing projects to deep node evolutionary questions. Phylogenomics has undoubtedly improved the availability of pertinent molecular data required to answer many phylogenetic conundrums, essentially removing the threat of stochastic error. However, a concerning widespread publication of poor quality data is quickly becoming a problem and these sequencing projects require strict quality control standards and practices to prevent the open-source market from getting flooded with unreliable sequence data. The projects within this thesis, particularly the work on the tardigrades, has also illustrated that phylogenomic-scale datasets are not the solution to all problems concerning molecular evolution. This is because systematic error becomes inflated with the addition of data (Felsenstein, 1978). As a consequence, improving taxon and gene coverage to a systematic problem only makes the error worse. To this end this work has demonstrated how to identify systematic errors such as long branch affinities between rapidly evolving groups through signal dissection measures and how to uncover the true phylogenetic signal beneath. It is important to note that just improving on the dataset quality is not a guarantee in the success of phylogenetic reconstructions, these data must be applied thoughtfully and intelligently based on the context of the situation (i.e. is one studying deep nodes or shallow nodes - the amount of character saturation in the molecular dataset tends to vary greatly between the two resulting in complications in discerning branch topologies in either event (Philippe *et al.* 2011a) or perhaps the taxa of interest is rapidly evolving - increasing the risk of systematic errors) and with multi-model approaches which should be tested for suitability.

A phylogenomic approach, while implementing signal dissection techniques, has revealed that the affinity between the tardigrades and nematodes is an artifact of LBA and that the tardigrades are more closely related to the onychophorans, sharing a common ancestor as either a sister group or as part of the Panarthropoda. Furthermore generating and applying genomic data to the Chaetognatha phylum established the inner relationships of the group with *Parasagitta sp.* and *Flaccisagitta sp.* sharing a monophyly with *Spadella sp.* more distantly related. Results confirm the chaetognaths as protostomes and further reveal them to be the oldest lophotrochozoans. The significance of this means that the protostome ancestor either possessed development and morphological characteristics more similar to that of deuterostomes, before the radiation of most extant protostome lineages (with the chaetognaths retaining these pleisomorphic traits), or that the chaetognaths have undergone a remarkable amount of convergent evolution in order to possess some key deuterostome traits (apomorphies) for an indeterminable reason.

Divergence time estimations from the projects making up this thesis all follow the same narrative of a pre-Cambrian origin of animals, opposing the traditional Cambrian explosion idea of animal life beginning in a burst of relatively immediate diversification and radiation of lineages (Conway-Morris, 2000). Instead we have delineated a Cryogenian period of animal origins with lineage radiation beginning in the Cambrian with even one of the earliest terrestrialization events occurring during this time (**Results 4.3.1**). The lack of physical evidence for animals existing pre-Cambrian can be explained by the incomplete nature of the fossil record in conjunction with the small possibility of the first soft-bodied metazoans suiting unusually specific fossilization conditions (Petrovich, 2001). Moreover the small number and geographically biased sampling of pre-Cambrian paleontological sites



(Peterson & Butterfield, 2005) make finding fossil remains of the first metazoans even more improbable.

Finally a macroevolutionary study of protein family acquisition spanning the Animal Kingdom has identified patterns of significant gains in their numbers and functionality in high-level taxonomic groupings (representing the ancestral node for large amounts of animals sharing certain defining traits) as opposed to a steady gain of these families over time distributed across the younger diversified lineages. Ergo, most protein families were formed in the older metazoan nodes such as the Bilateria, Protostomia, and Ecdysozoa, meaning that the protein networks existing in extant lineages today were developed hundreds of millions of years before these animals existed. We estimated that the relatively recent radial evolution of extant lineages was perhaps aided by the re-wiring and adaptations of these long existing protein networks (families) as opposed to the creation of new ones.

## **6.6 Future Work**

### **6.6.1 SRA Data Quality Standards**

The growth of the SRA from NGS projects, the source of which is often from published peer-reviewed research, in conjunction with its usefulness as a source of data for new molecular research means that it is an essential database for future phylogenomic studies. Accounting for the importance of this database for current and future molecular research I would suggest an implementation of QC standards for each submission, or at least a filter for raw sequence data that has been through QC.

### **6.6.2 Supplementing the Foundations of Phylogenomic Datasets**

It would be advisable to readdress the foundations of the datasets from which we build our phylogenomic studies to which we add the orthologs of newly sequenced taxa. Many of these datasets are pre-NGS era and are possibly limited in their gene number due to lack of available molecular libraries at the time. An expansion with more slowly evolving genes would be of benefit to the phylogenetic signal of the dataset and could prevent problems regarding missing data and large gaps in the MSA we generate.

### 6.6.3 Phylostratigraphic Investigations of Protein Families

As mentioned in the chapter 5 discussion section, there are a number of improvements that could be worked into future phylostratigraphic orphan protein family experiments that would be beneficial. First and foremost the application of genomes instead of transcriptomes would ensure a full genetic lexicon for each of the taxa and thus a complete proteome from which we could identify families via MCL (Enright 2002). Secondly the application of NGS data to the Deuterostomia and Lophotrochozoa would be a necessary expansion in order to see if the pattern of new protein family gain in ancient animal nodes remains constant. Next, further experiments on the influence of inflation rate on large phylogenomic matrices may result in more defined clusters as only 75,547 out of 167,673 of the proteins in the study clustered to an acceptable degree. Finally, additional time and resources would be beneficial in uncovering the function of the five groups above the mean rate of protein family acquisition [Figure 5.6] but were not inserted into the BLAST2GO annotation pipeline [Figure 5.4]. This may reveal more information as to how these younger groups have adapted over time.

### 6.6.4 Total Evidence Dating of the Chaetognatha

Further sequencing of the Chaetognatha would be ideal for further investigations into these predators. Presently we only have the genome of *Parasagitta sp.*, as *Spadella sp.* and *Flaccisagitta sp.* are sourced ESTs. Further divergence time estimations such as TED, which would include phylogenetic signal from the fossils may further clarify the origins of extant chaetognaths and their relation to the ambiguous *Amiskwia* (Conway-Morris 1977).

### 6.6.5 Chaetognaths and Lophotrochozoan Phylogenomic Dataset

The chapter 3 study on the chaetognaths is built on a dataset from Philippe *et al.* (2011b). However a recent dataset has been recently published from Kocot *et al.* (2016) which generated many new molecular libraries from NGS experiments in order to fully sample each phyla of the Lophotrochozoa including all grouped purported to belong to the Platyzoa (Cavalier-Smith, 1998). Unfortunately this data was released at the very late stages of this project and so could not be used. A further study mapping the *Parasagitta sp.* orthologs to this new dataset could be useful in re-confirming our placement of the Chaetognatha and in investigating the evolutionary timescale of the lophotrochozoan phyla.

## Bibliography

- Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., ... Venter, J. C. (2000). The genome sequence of *Drosophila melanogaster*. *Science*, *287*, 2185–2195.
- Ager, D. V. (1981). *The Nature of the Stratigraphical Record*. Macmillan Press.
- Aguinaldo, A.-M. A., Turbeville, J. M., Linford, L. S., Rivera, M. C., Garey, J. R., Raff, R. A., & Lake, J. A. (1997). Evidence for a clade of nematodes, arthropods and other molting animals. *Nature*, *387*, 489–493.
- Altaba, C. R. (2009). Universal artifacts affect the branching of phylogenetic trees, not universal scaling laws. *PLoS ONE*, *4*(2), 1–13.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology*, *215*(3), 403–410.
- Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402.
- Andrews, S., & Rothnagel, J. (2014). Emerging evidence for functional peptides encoded by short open reading frames. *Nature Reviews Genetics*, *15*(286), 193–204.
- Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J., Dehal, P., ... Brenner, S. (2002). Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science*, *297*(5585), 1301–1310.
- Aris-Brosou, S., & Yang, Z. (2003). Bayesian models of episodic evolution support a late Precambrian explosive diversification of the Metazoa. *Molecular Biology and Evolution*, *20*(12), 1947–1954.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., ... Sherlock, G. (2000). Gene Ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, *25*(1), 25–29.
- Ball, C., Dolinski, K., Dwight, S. S., Harris, M., Issel-Tarver, L., Kasarskis, A., ... Cherry, J. M. (2000). Integrating functional genomic information into the *Saccharomyces* genome database. *Nucleic Acids Research*, *28*(1), 77–80.
- Bapst, D. W. (2012). Paleotree: An R package for paleontological and phylogenetic analyses of evolution. *Methods in Ecology and Evolution*, *3*(5), 803–807.
- Basu, M. K., Carmel, L., Rogozin, I. B., & Koonin, E. V. (2008). Evolution of protein domain promiscuity in eukaryotes. *Genome Research*, *18*, 449–461.

- Bateman, A., Coin, L. J., Durbin, R., Finn, R., & Hollich, V. (2004). Pfam: The protein families database. *Nucleic Acids Research*, *32*, 138–141.
- Becquerel, P. (1960). La suspension de la vie au dessus de 1/20 K absolu par demagnetization adiabatique de l'alun de fer dans le vide les plus elève. *Comptes Rendus de l'Académie Des Sciences*, *231*, 261–263.
- Bennett, S. (2004). Solexa Ltd. *Pharmacogenomics*, *5*(4), 433–438.
- Benton, M. J., Donoghue, P. C. J., Asher, R. J., Friedman, M., Near, T. J., & Vinther, J. (2015). Constraints on the timescale of animal evolutionary history. *Palaeontologia Electronica*, 1–107.
- Bergsten, J. (2005). A review of long-branch attraction. *Cladistics*, *21*, 163–193.
- Berner, R. A. (1999). Atmospheric oxygen over Phanerozoic time. *Proceedings of the National Academy of Sciences*, *96*(20), 10955–10957.
- Bieri, R. (1959). The distribution of the planktonic Chaetognatha in the pacific and their relationship to the water masses. *Limnology and Oceanography*, *4*(1), 1–28.
- Blake, J. A., Eppig, J. T., Richardson, J. E., & Davisson, M. T. (2000). The mouse genome database (MGD): expanding genetic and genomic resources for the laboratory mouse. The Mouse Genome Database Group. *Nucleic Acids Research*, *28*(1), 108–111.
- Blattner, F. R., Plunkett III, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., ... Shao, Y. (1997). The complete genome sequence of Escherichia coli K-12. *Science*, *277*, 1453–1462.
- Bonfield, J. K., Smith, K. f, & Staden, R. (1995). A new DNA sequence assembly program. *Nucleic Acids Research*, *23*(24), 4992–4999. =
- Boore, J. L., Collins, T. M., Stanton, D., Daehler, L. L., & Brown, W. M. (1995). Deducing the pattern of arthropod phylogeny from mitochondrial DNA rearrangements. *Nature*, *376*, 163.
- Borner, J., Rehm, P., Schill, R. O., Ebersberger, I., & Burmester, T. (2014). A transcriptome approach to ecdysozoan phylogeny. *Molecular Phylogenetics and Evolution*, *80*, 79–87.
- Boto, L. (2014). Horizontal gene transfer in the acquisition of novel traits by metazoans. *Proceedings of the Royal Society B*, *8*(281), 1–8.
- Boudali, H., & Duga, J. B. (2005). A new Bayesian network approach to solve dynamic fault trees. *Reliability and Maintainability Symposium, Conference*, 451–455.

- Boussau, B., Walton, Z., Delgado, J. A., Collantes, F., Beani, L., Stewart, I. J., ... Huelsenbeck, J. P. (2014). Strepsiptera, phylogenomics and the long branch attraction problem. *PLoS ONE*, *9*(10), 1–9.
- Brewer, M., & Bond, J. E. (2013). Ordinal-level phylogenomics of the arthropod class Diplopoda (millipedes) based on an analysis of 221 nuclear protein-coding loci generated using next-generation sequence analyses. *PLoS ONE*, *13*(8), 1–15.
- Brinkmann, H., & Philippe, H. (1999). Archaea sister group of bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. *Molecular Biology and Evolution*, *16*(6), 817–825.
- Brinkmann, H., van der Giezen, M., Zhou, Y., Poncelin de Raucourt, G., & Philippe, H. (2005). An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Systematic Biology*, *54*(5), 743–757.
- Brohée, S., & van Helden, J. (2006). Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, *7*, 488.
- Brownlee, G. G., & Sanger, F. (1969). Chromatography of <sup>32</sup>P-labelled oligonucleotides on thin layers of DEAE-cellulose. *FEBS*, *11*, 395–399.
- Burgess, D. (2016). Gene Expression: A space for transcriptomics. *Nature Reviews Genetics*, *17*, 436–437.
- Calderone, N. W. (2012). Insect pollinated crops, insect pollinators and US agriculture: Trend analysis of aggregate data for the period 1992–2009. *PLoS ONE*, *7*(5).
- Campbell, L. I., Rota-Stabelli, O., Edgecombe, G. D., Marchioro, T., Longhorn, S. J., Telford, M. J., ... Pisani, D. (2011). MicroRNAs and phylogenomics resolve the relationships of Tardigrada and suggest that velvet worms are the sister group of Arthropoda. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(38), 15920–15924.
- Canard, B., & Sarfati, R. S. (1994). DNA polymerase fluorescent substrates with reversible 3'-tags. *Gene*, *148*(1), 1–6.
- Cannon, J. T., Vellutini, B. C., Smith, J., Ronquist, F., Jondelius, U., & Hejnol, A. (2016). Xenacoelomorpha is the sister group to Nephrozoa. *Nature*, *530*, 89–93.
- Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, *17*, 540–552.
- Cavalier-Smith, T. (1998). A revised six-kingdom system of life. *Biology Reviews*, *73*, 203–266.

- Caveney, S., Cladman, W., Verellen, L., & Donly, C. (2006). Ancestry of neuronal monoamine transporters in the Metazoa. *Journal of Experimental Biology*, 209, 4858–4868.
- Chen, J., & Huang, D. (2006). A possible lower Cambrian chaetognath (arrow worm). *Science*, 298(5591), 187.
- Chen, M.-H., Kuo, L., & Lewis, P. O. (2014). Bayesian phylogenetics: Methods, algorithms, and applications. *Champan and Hall/CRC Mathematical and Computational Biology*, 1–49.
- Clarke, K., Yang, Y., Marsh, R., Xie, L. L., & Zhang, K. K. (2013). Comparative analysis of de novo transcriptome assembly. *Science China Life Sciences*, 56(2), 156–162.
- Clegg, J. S. (2001). Cryptobiosis - a peculiar state of biological organization. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, 128(4), 613–624.
- Cohen, K., Finney, S., Gibbard P., & Fan, J. (2013) The ICS international chronostratigraphic chart. *Episodes*, 36(3), 199-204.
- Compeau, P. E. C., Pevzner, P. A., & Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, 29(11), 987–991.
- Conesa, A., Götz, S., García-gómez, J. M., Terol, J., Talón, M., Genómica, D., ... Valencia, U. P. De. (2005). Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 21(18), 3674–3676.
- Convey, P., & McInnes, S. J. (2005). Exceptional tardigrade-dominated ecosystems in Ellsworth Land, Antarctica. *Ecology*, 86(2), 519–527.
- Conway-Morris, S. (1993). The fossil record and the early evolution of the Metazoa. *Nature*, 361, 219–225.
- Conway-Morris, S. (2000). The Cambrian “explosion”: Slow-fuse or megatonnage? *Proceedings of the National Academy of Sciences of the United States of America*, 97(9), 4426–4429.
- Conway-Morris, S. (1977). A redescription of the Middle Cambrian worm *Amiskwia sagittiformis* from the Burgess Shale of British Columbia. *Paläontologische Zeitschrift*, 51(3), 271–287.
- Cook, C. E., Smith, M. L., Telford, M. J., Bastianello, A., & Akam, M. (2001). Hox genes and the phylogeny of the arthropods. *Current Biology*, 11(10), 759–763.
- Cox, J. C., Ingersoll, J. E., & Ross, S. A. (1985). A theory of the term structure of interest rates. *Econometrica*, 53, 385–408.



- Crick, F. H. C., Barnett, L., Brenner, S., & Watts-Tobin, R. J. (1961). General nature of the genetic code for proteins. *Nature*, *192*, 1227–1232.
- Dabert, M., Witalinski, W., Kazmierski, A., Olszanowski, Z., & Dabert, J. (2010). Molecular phylogeny of acariform mites (Acari, Arachnida): Strong conflict between phylogenetic signal and long-branch attraction artifacts. *Molecular Phylogenetics and Evolution*, *56*(1), 222–241.
- Darwin, C. (1859). On the origins of the species by means of natural selection, or the preservation of favoured races in the struggle for life.
- Darwin, C. (1837). Darwin's Notebook B, 36.
- Davidson, E., Peterson, J., & Cameron, A. (1995). Origin of bilaterian body plans: Evolution of developmental regulatory mechanisms. *Science*, *270*(5240), 1319–1326.
- Dayhoff M. O. (1968). Atlas of protein sequence and structure. *Silver Spring: National Biomedical Research Foundation*, 345–352.
- Dayhoff, M. O. (1976). Origin and evolution of protein superfamilies. *Fed Proc*, *35*, 2132–2138.
- Degma, P., Bertolani, R., & Guidetti, R. (2016). Actual checklist of Tardigrada species. *31st Edition (15-12-2016)*, *31*, 1–47.
- Dehal, P., & Boore, J. L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biology*, *3*(10), 1700–1708.
- Demuth, J. P., & Hahn, M. W. (2009). The life and death of gene families. *BioEssays*, *31*(1), 29–39.
- Doerr, D., Gronau, I., Moran, S., & Yavneh, I. (2012). Stochastic errors vs. modeling errors in distance based phylogenetic reconstructions. *Algorithms for Molecular Biology*, *7*(22), 1–16.
- Doguzhaeva, L. A., Mutvei, H., & Mapes, R. H. (2002). Chaetognath grasping spines from the Upper Mississippian of Arkansas (USA). *Acta Palaeontologica Polonica*, *47*(3), 421–430.
- Dohle, W. (1998). *Arthropod Relationships* (Vol. Chapter 23).
- Domazet-Lošo, T., Brajković, J., & Tautz, D. (2007). A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends in Genetics*, *23*(11), 533–539.
- Doncaster, L. (1902). On the development of Sagitta, with notes on the anatomy of the adult. *Journal of Cell Science*, *2*(46), 351–359.

- Donoghue, P. C. J., & Benton, M. J. (2007). Rocks and clocks: Calibrating the Tree of Life using fossils and molecules. *Trends in Ecology & Evolution*, 22(8), 424–431.
- dos Reis, M., Donoghue, P. C. J., & Yang, Z. (2016). Bayesian molecular clock dating of species divergences in the genomics era. *Nature Reviews Genetics*, 17, 71–80.
- Douady, C. J., Boucher, Y., Doolittle, W. F., & Douzery, E. J. P. (2003). Comparison of bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Molecular Biology and Evolution*, 20(2), 248–254.
- Drummond, A. J., Ho, S. Y. W., Phillips, M. J., & Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biology*, 4(5), 699–710.
- Drummond, A. J., Suchard, M. A., Xie, D., & Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, 29(8), 1969–1973.
- Ducret, F. (1978). Specific systematic structures of two chaetognaths (*Sagitta tasmanica* and *Eukrohnia hamata*) and phylogenetic implications. *Zoomorphology*, 91, 201–215.
- Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., ... Birney, E. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), 57–74.
- Dunn, C. W. (2014). Reconsidering the phylogenetic utility of miRNA in animals. *Proceedings of the National Academy of Sciences of the United States of America*, 111(35), 12576–12577.
- Dunn, C. W. (2013). Evolution: Out of the ocean. *Current Biology*, 23(6), 241–243.
- Dunn, C. W., Hejnol, A., Matus, D. Q., Pang, K., Browne, W. E., Smith, S. A., ... Giribet, G. (2008). Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, 452, 745–749.
- Dzidic, S., & Bedekovic, V. (2003). Horizontal gene transfer-emerging multidrug resistance. *Acta Pharmacologica Sinica*, 24(6), 519–526.
- Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5), 1792–1797.
- Edwards, A. W. F., & Cavalli-Sforza, L. L. (1964). Reconstruction of evolutionary trees. *Phenetic and Phylogenetic Classification*, 6(6), 67–76.
- Edwards, A. W. F., & Cavalli-Sforza, L. L. (1963). The reconstruction of evolution. *Annals of Human Genetics*, 27, 105–106.

- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1), 1–26.
- Ekblom, R., & Galindo, J. (2010). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, 107(1), 1–15.
- Enright, A. J., Dongen, S. Van, & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7), 1575–1584.
- Erwin, D. H., Laflamme, M., Tweedt, S. M., Sperling, E. A., Pisani, D., & Peterson, K. J. (2011). The Cambrian Conundrum: Early Divergence and Later Ecological Success in the Early History of Animals. *Science*, 334(6059), 1091–1097.
- Erwin, D., & Valentine, J. (2012). The Cambrian explosion: the construction of animal biodiversity. *Roberts Publishing*.
- Ewing, B., Hillier, L., Wendl, M. C., & Green, P. (1998). Base-calling of automated sequencer traces using phred. *Genome Research*, 8, 175–185.
- Eyre-Walker, A., & Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nature Reviews Genetics*, 8, 610–618.
- Fan, J., Farman, M., & Gijbels, I. (1998). Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society Series B*, 60(3), 591–608.
- Fayers, S. R., & Trewin, N. H. (2002). A new crustacean from the Early Devonian Rhynie chert, Aberdeenshire, Scotland. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, 93, 355–382.
- Felsenfeld, G., & Miles, H. T. (1967). The physical and chemical properties of nucleic acids. *Annual Review of Biochemistry*, 36(1), 407–448.
- Felsenstein, J. (2004). Inferring phylogenies. *Sunderland, Massachusetts: Sinauer*.
- Felsenstein, J. (1976). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, 27(4), 401–410.
- Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39(783–791).
- Felsenstein, J. (2005). PHYLIP (Phylogeny Inference Package) version 3.69. *Dept of Genome Sciences, University of Washington, Seattle*.
- Fernandez, R., Hormiga, G., & Giribet, G. (2014). Phylogenomic analysis of spiders reveals nonmonophyly of orb weavers. *Current Biology*, 3, 1772–1777.
- Fitch, W. M. (2000). Homology a personal view on some of the problems. *Trends in Genetics*, 16(5), 227–231.

- Fitch, W. M., & Margoliash, E. (1967). Construction of phylogenetic trees. *Science*, 155(3760), 279–284.
- Fleagle, J. G. (2013). Primate adaptation and evolution. *Academic Press*.
- Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-silva, D., ... Searle, S. M. J. (2012). Ensembl 2012. *Nucleic Acids Research*, 40, 84–90.
- Flintoft, L. (2013). Isoform diversity revealed. *Nature Reviews Genetics*, 14(369).
- Floudas, D., Binder, M., Riley, R., Barrie, K., & Blanchette, R. A. (2012). The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. *Science*, 336(6089), 1715–1719.
- Foster, P. G. (2004). Modeling compositional heterogeneity. *Systematic Biology*, 53(3), 485–495.
- Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659), 799–805.
- Friedrich, M., & Tautz, D. (1995). Ribosomal DNA phylogeny of the major extant arthropod classes and the evolution of myriapods. *Nature*, 376, 165–167.
- Gabriel, W. N., McNuff, R., Patel, S. K., Gregory, T. R., Jeck, W. R., Jones, C. D., & Goldstein, B. (2007). The tardigrade *Hypsibius dujardini*, a new model for studying the evolution of development. *Developmental Biology*, 312(2), 545–559.
- Gadagkar, S. R., & Kumar, S. (2005). Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous. *Molecular Biology and Evolution*, 22(11), 2139–2141.
- Gelbart, W. M., Bayraktaroglu, L., Bettencourt, B., Campbell, K., Crosby, M., Emmert, D., ... Weil, C. (2003). The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Research*, 31(1), 172–175.
- Gene, Ontology, & Consortium. (2004). The gene ontology (GO) database and informatics resource. *Nucleic Acids Research*, 32, 258–261.
- Ghirardelli, E. (1968). Some aspects of the biology of chaetognaths. *Advances in Marine Biology*, 6, 271–375.
- Ghirardelli, E. (1981). Chaetognatha: Taxonomic position, affinity and evolution of the phylum. *Origine Dei Grande Phyla Dei Metazoi*, 191–233.
- Giribet, G. (2008). Assembling the lophotrochozoan (spiralian) tree of life. *Philosophical Transactions of the Royal Society B*, 363, 1513–1522.

- Giribet, G., Edgecombe, G. D., & Wheeler, W. C. (2001). Arthropod phylogeny based on eight molecular loci and morphology. *Nature*, *413*, 157–161.
- Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(333–351).
- Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., ... Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–52.
- Graur, D., & Martin, W. (2004). Reading the entrails of chickens: Molecular timescales of evolution and the illusion of precision. *Trends in Genetics*, *20*(2), 80–86.
- Guidetti, R., Bertolani, R., & Nelson, D. (1999). Ecological and faunistic studies on tardigrades in leaf litter of beech forests. *Proceedings of the Seventh International Symposium on the Tardigrada*, *238*, 215–223.
- Guidetti, R., & Bertolani, R. (2005). Tardigrade taxonomy: An updated check list of the taxa and a list of characters for their identification. *Zootaxa*, *845*, 1–46.
- Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology*, *59*(3), 307–321.
- Guindon, S., & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, *52*(5), 696–704.
- Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., ... William, T. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*, 1494–1512.
- Haggerty, L. S., Jachiet, P. A., Hanage, W. P., Fitzpatrick, D. A., Lopez, P., O'Connell, M. J., ... McInerney, J. O. (2014). A pluralistic account of homology: Adapting the models to the data. *Molecular Biology and Evolution*, *31*, 501–516.
- Halanych, K. M., Bacheller, J. D., Aguinaldo, A. M., Liva, S. M., Hillis, D. M., & Lake, J. A. (1995). Evidence from 18S ribosomal DNA that the lophophorates are protostome animals. *Science*, *268*(5210), 1641–1643.
- Halanych, K. M. (2004). The new view of animal phylogeny. *Annual Review of Ecology, Evolution, and Systematics*, *35*, 229–256.
- Hashimoto, T., Horikawa, D. D., Saito, Y., Kuwahara, H., Kozuka-Hata, H., Shin-I, T., ... Kunieda, T. (2016). Extremotolerant tardigrade genome and improved

- radiotolerance of human cultured cells by tardigrade-unique protein. *Nature Communications*, 7, 12808.
- Hastings, B. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1), 97–109.
- Hausdorf, B., Helmkampf, M., Nesnidal, M. P., & Bruchhaus, I. (2010). Phylogenetic relationships within the lophophorate lineages (Ectoprocta, Brachiopoda, and Phoronida). *Molecular Phylogenetics and Evolution*, 55(3), 1121–1127.
- Hayward, P. J., & Ryland, J. S. (1995). Crustaceans (phylum Crustacea). *Handbook of the Marine Fauna of North-West Europe, Oxford Uni*, 289–462.
- Hejnal, A., Obst, M., Stamatakis, A., Ott, M., Rouse, G. W., Edgecombe, G. D., ... Love, S. (2009). Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proceedings of the Royal Society B*, 267(1677), 4261–4270.
- Higgs, P. G., & Attwood, T. K. (2013). Bioinformatics and molecular evolution. *John Wiley & Sons*.
- Ho, S. Y. W. (2014). The changing face of the molecular evolutionary clock. *Trends in Ecology & Evolution*, 29(9), 496–503.
- Holt, R. A., & Jones, S. J. M. (2008). The new paradigm of flow cell sequencing. *Genome Research*, 18, 839–846.
- Holt, R. A., Subramanian, G. M., Halpern, A., Sutton, G. G., Charlab, R., Nusskern, D. R., ... Hoffman, S. L. (2002). The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science*, 298(5591), 129–149.
- Horikawa, D. D., Kunieda, T., Abe, W., Watanabe, M., Nakahara, Y., & Yukuhiro, F. (2008). Establishment of arearing system of the extremotolerant tardigrade *Ramazzottius varieornatus*: A new model animal for astrobiology. *Astrobiology*, 8(3), 549–556.
- Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., ... Ban, J. F. (2016). A new view of the tree of life. *Nature Microbiology*, 1, 1–6.
- Hugot, J.-P., Baujard, P., & Morand, S. (2001). Biodiversity in helminths and nematodes as a field of study: An overview. *Nematology*, 3(3), 199–208.
- Hurles, M. (2004). Gene duplication: The genomic trade in spare parts. *PLoS Biology*, 2(7), 1–5.
- Hyman, L. H. (1959). The invertebrates: Smaller coelomate groups. *Mc Graw-Hill, New York, Volume 5*.

- Inagaki, Y., Susko, E., Fast, N. M., & Roger, A. J. (2004). Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaeobacteria in EF-1 phylogenies. *Molecular Biology and Evolution*, *21*(7), 1340–1349.
- Innan, H., & Kondrashov, F. (2010). The evolution of gene duplications: Classifying and distinguishing between models. *Nature Reviews Genetics*, *11*, 97–108.
- Jeffroy, O., Brinkmann, H., Delsuc, F., & Philippe, H. (2006). Phylogenomics: The beginning of incongruence? *Trends in Genetics*, *22*(4), 225–231.
- Jékely, G. (2013). Global view of the evolution and diversity of metazoan neuropeptide signaling. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(21), 8702–8707.
- Jensen, R. a. (2001). Orthologs and paralogs - we need to get it right. *Genome Biology*, *2*(8), 1002.1-1002.2.
- Jensen, S., Gehling, J. G., & Droser, M. L. (1998). Ediacara-type fossils in Cambrian sediments. *Nature*, *393*, 576–569.
- Jeram, A. J., Selden, P. A., & Edwards, D. (1990). Land animals in the Silurian: Arachnids and myriapods from Shropshire, England. *Science*, *250*(4981), 658–661.
- Jönsson, K. I., Rabbow, E., Schill, R. O., Harms-Ringdahl, M., & Rettberg, P. (2008). Tardigrades survive exposure to space in low Earth orbit. *Current Biology*, *18*(17), 729–731.
- Jonsson, K. I., & Rebecchi, L. (2002). Experimentally induced anhydrobiosis in the tardigrade *Richtersius coronifer*: Phenotypic factors affecting survival. *Journal of Experimental Zoology*, *293*, 574–584.
- Ju, J., Kim, D. H., Bi, L., Meng, Q., Bai, X., Li, Z., ... Turro, N. J. (2006). Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(52), 19635–19640.
- Jukes, C. (1969). Evolution of protein molecules. *Mammalian Protein Metabolism*, *111*.
- Kenrick, P., Wellman, C. H., Schneider, H., & Edgecombe, G. D. (2012). A timeline for terrestrialization: Consequences for the carbon cycle in the Palaeozoic. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*, 519–536.
- Kevan, P. G., Clark, E. A., & Thomas, V. G. (1990). Insect pollinators and sustainable agriculture. *American Journal of Alternative Agriculture*, *5*(1), 13–22.

- Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, *16*(2), 111–120.
- Kishino, H., & Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *Journal of Molecular Evolution*, *29*, 170–179.
- Kishino, H., Thorne, J. L., & Bruno, W. J. (2001). Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Molecular Biology and Evolution*, *18*(3), 352–61.
- Knoll, A. H., & Carroll, S. B. (1999). Early animal evolution: Emerging views from comparative biology and geology. *Science*, *284*(5423), 2129–2137.
- Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., & Mardis, E. R. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell*, *155*(1), 27–38.
- Kocot, K. M., Cannon, J. T., Todt, C., Citarella, M. R., Kohn, A. B., Meyer, A., ... Halanych, K. M. (2011). Phylogenomics reveals deep molluscan relationships. *Nature*, *477*(7365), 452–456.
- Kocot, M., Struck, T. H., Merkel, J., Waits, D. S., Todt, C., Brannock, P. M., ... Halanych, K. M. (2016). Phylogenomics of Lophotrochozoa with consideration of systematic error. *Systematic Biology*, *66*(2), 256–282.
- Kolaczkowski, B., & Thornton, J. W. (2004). Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature Letters*, *431*, 980–984.
- Koonin, E. V. (2005). Orthologs, Paralogs, and Evolutionary Genomics. *Annual Review of Genetics*, *39*(1), 309–338.
- Koutsovoulos, G., Kumar, S., Laetsch, D. R., Stevens, L., Daub, J., & Conlon, C. (2016). No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*. *Proceedings of the National Academy of Sciences*, *113*(18), 5053–5058.
- Krause, A., Stoye, J., & Vingron, M. (2005). Large scale hierarchical clustering of protein sequences. *BMC Bioinformatics*, *6*, 15.
- Kück, P., & Longo, G. C. (2014). FASconCAT-G: Extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. *Frontiers in Zoology*, *11*(1), 81.
- Kumar, S. (2005). Molecular clocks: Four decades of evolution. *Nature Reviews Genetics*, *6*, 654–662.



- Labandeiraemail, C. C. (2005). Invasion of the continents: Cyanobacterial crusts to tree-inhabiting arthropods. *Trends in Ecology & Evolution*, *20*, 253–262.
- Lake, J. A. (1990). Origin of the Metazoa. *Proceedings of the National Academy of Sciences of the United States of America*, *87*, 763–766.
- Lander, E. S., Heaford, A., Sheridan, A., Linton, L. M., Birren, B., Subramanian, A., ... Szustakowki, J. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921.
- Lartillot, N., Blanquart, S., & Lepage, T. (2009). PhyloBayes 3.3: A Bayesian software for phylogenetic reconstruction and molecular dating using mixture models. *User Manual*.
- Lartillot, N., Brinkmann, H., & Philippe, H. (2007). Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evolutionary Biology*, *7 Suppl 1*, S4.
- Lartillot, N., & Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, *21*(6), 1095–1099.
- Laumer, C. E., Hejnol, A., & Giribet, G. (2015). Nuclear genomic signals of the “microturbellarian” roots of platyhelminth evolutionary innovation. *E Life*, *4*, 1–31.
- Lee, Y., & Rio, D. (2015). Mechanism and regulation of alternative pre-mRNA splicing. *Annual Review of Biochemistry*, *84*, 291–323.
- Leinonen, R., Sugawara, H., & Shumway, M. (2011). The Sequence Read Archive. *Nucleic Acids Research*, *39*(2010), 2010–2012.
- Lemey, P., Rambaut, A., Drummond, A. J., & Suchard, M. A. (2009). Bayesian phylogeography finds its roots. *PLoS Computational Biology*, *5*(9), 1–16.
- Lepage, T., Bryant, D., Philippe, H., & Lartillot, N. (2007). A general comparison of relaxed molecular clock models. *Molecular Biology and Evolution*, *24*(12), 2669–2680.
- Lewis, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, *50*, 913–925.
- Lio, P. & Goldman, N. (1998) Models of molecular evolution and phylogeny. *Genome Research*, *8*(12), 1233-1244.
- Lipmanl, D. J., & Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science*, *227*(4693), 1435–1441.
- Lipps, J., & Signor, P. (1992). Origin and early evolution of the Metazoa. *Springer*, 1–569.

- Little, C. (1983). The colonisation of land: Origins and adaptations of terrestrial animals. *Cambridge University Press*.
- Liu, J., & Rost, B. (2003). Domains, motifs and clusters in the protein universe. *Current Opinion in Chemical Biology*, 7(1), 5–11.
- Lobo, I. (2008). Basic Local Alignment Search Tool (BLAST). *Nature Education*, 1(1), 215.
- Lozano-Fernandez, J., Carton, R., Tanner, A. R., Puttick, M. N., Blaxter, M., Vinther, J., ... Pisani, D. (2016). A molecular palaeobiological exploration of arthropod terrestrialization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1699), 20150133.
- Luo, Z.-X., Chen, P., Gang, L., & Meng, C. (2007). A new eutriconodont mammal and evolutionary development in early mammals. *Nature*, 446, 288–293.
- Lynch, M. (2000). The evolutionary fate and consequences of puplicate genes. *Science*, 290(5494), 1151–1155.
- Ma, J. (2006). Gene expression and regulation. *Springer NY*, 1–582.
- MacNaughton, R. B., Cole, J. M., Dalrymple, R. W., Braddy, S. J., Briggs, D. E. G., & Lukie, T. D. (2002). First steps on land: Arthropod trackways in Cambrian-Ordovician eolian sandstone, southeastern Ontario, Canada. *Geology*, 30, 391–394.
- Mallatt, J. M., Garey, J. R., & Shultz, J. W. (2004). Ecdysozoan phylogeny and Bayesian inference: First use of nearly complete 28S and 18S rRNA gene sequences to classify the arthropods and their kin. *Molecular Phylogenetics and Evolution*, 31(1), 178–191.
- Maloof, A. C., Porter, S. M., Moore, J. L., Dudás, F. Ö., Bowring, S. A., Higgins, J. A., ... Eddy, M. P. (2010). The earliest Cambrian record of animals and ocean geochemical change. *The Geological Society of America Bulletin*, 122, 1731–1774.
- Marcotte, E. M., Pellegrini, M., Ng, H. L., Rice, D. W., Yeates, T. O., Eisenberg, D., ... Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285(5428), 751–753.
- Marcotte, E. M., Pellegrini, M., Ng, H.-L., Rice, D. W., Yeates, T. O., & Eisenberg, D. (1999). Detecting protein function and protein-protein interactions from genome sequences. *Science*, 30, 751–753.
- Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. a, ... Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057), 376–80.

- Marlétaz, F., Gilles, A., Caubit, X., Perez, Y., Dossat, C., Samain, S., ... Le Parco, Y. (2008). Chaetognath transcriptome reveals ancestral and unique features among bilaterians. *Genome Biology*, 9(6), 94.
- Marlétaz, F., Martin, E., Perez, Y., Papillon, D., Caubit, X., Lowe, C. J., ... Le Parco, Y. (2006). Chaetognath phylogenomics: A protostome with deuterostome-like development. *Current Biology*, 16(15), 577–578.
- Martín-durán, J. M., Passamanek, Y. J., Martindale, M. Q., & Hejnal, A. (2016). The developmental basis for the recurrent evolution of deuterostomy and protostomy. *Nature Ecology and Evolution*, 1, 1–10.
- Martin, T., & Luo, Z.-X. (2005). Homoplasy in the mammalian ear. *Science*, 307(5711), 861–862.
- Matus, D. Q., Copley, R. R., Dunn, C. W., Hejnal, A., Eccleston, H., Halanych, K. M., ... Telford, M. J. (2006). Broad taxon and gene sampling indicate that chaetognaths are protostomes. *Current Biology*, 16(15), 575–576.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *Journal Chemical Physics*, 21(6), 1087–1092.
- Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews Genetics*, 11, 31–46.
- Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6), 315–327.
- Minelli, A., Boxshall, G., & Fusco, G. (2013). Arthropod biology and evolution: Molecules, development, morphology. *Springer*.
- Misof, B. et al. (2014). Phylogenomics resolves the timing and pattern of insect evolution. *Science*, 346, 763–767.
- Miya, M., & Nishida, M. (2000). Use of mitogenomic information in teleostean molecular phylogenetics: A tree-based exploration under the maximum-parsimony optimality criterion. *Molecular Phylogenetics and Evolution*, 17(3), 437–455.
- Moroz, L. L. (2009). On the Independent origins of complex brains and neurons. *Brain Behavior and Evolution*, 74, 177–190.
- Moroz, L. L. (2009). On the independent origins of complex brains and neurons. *Brain, Behavior and Evolution*, 74(3), 177–190.
- Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., & Erlich, H. (1986). Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction. *Cold Spring Harbor Symposia on Quantitative Biology*.

- Munkres, J. R. (1984). Elements of algebraic topology. *Perseus Publishing*.
- Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3), 443–453.
- Nei, M. (1986). Stochastic errors in DNA evolution and molecular phylogeny. *Progress in Clinical and Biological Research*, 218, 133–147.
- Nelson, D. R., & Marley, N. J. (2000). The biology and ecology of lotic Tardigrada. *Freshwater Biology*, 44(1), 93–108.
- Nelson, D. R. (2002). Current status of the Tardigrada: Evolution and ecology. *Integrative and Comparative Biology*, 42, 652–659.
- Nielsen, C. (2001). Animal Evolution: Interrelationships of the living phyla. *Oxford University Press*.
- Niimura, Y., & Nei, M. (2007). Evolutionary dynamics of olfactory receptor genes in *Drosophila* species. *Proceedings of the National Academy of Sciences of the United States of America*, 104(17), 7122–7127.
- Nirenberg, M. W., & Matthaei, J. H. H. (1961). The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. *Proceedings of the National Academy of Sciences of the United States of America*, 47(10), 1588–1602.
- Novozhilov, N. I. (1957). Un nouvel ordre d'arthropodes particuliers: Kazacharthra du Lias des monts Ketmen (Kazakhstan, SE., URSS). *Société Géologique de France*, 7, 171–184.
- Oakley, T. H., Wolfe, J. M., Lindgren, A. R., & Zaharoff, A. K. (2013). Phylotranscriptomics to bring the understudied into the fold: Monophyletic Ostracoda, fossil placement, and pancrustacean phylogeny. *Molecular Biology and Evolution*, 30(1), 215–233.
- Olmstead, R. G. (1995) Species concepts and plesiomorphic species. *Systematic Botany*, 20(4), 623-630
- Omilian, A. R., & Taylor, D. J. (2001). Rate acceleration and long-branch attraction in a conserved gene of cryptic daphniid (Crustacea) species. *Molecular Biology and Evolution*, 18(12), 2201–2212.
- Owen, R. (1843). Lectures on the comparative anatomy and physiology of the invertebrate animals. *Presented at the Royal College of Surgeons*.
- Pagel, M. (1994). Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society B*, 255, 37–45.

- Palazzo, A. F., & Gregory, T. R. (2014). The Case for Junk DNA. *PLoS Genetics*, *10*(5).
- Papillon, D., Perez, Y., Caubit, X., & Le Parco, Y. (2004). Identification of chaetognaths as protostomes is supported by the analysis of their mitochondrial genome. *Molecular Biology and Evolution*, *21*(11), 2122–2129.
- Papillon, D., Perez, Y., Fasano, L., Le Parco, Y., & Caubit, X. (2003). Hox gene survey in the chaetognath *Spadella cephaloptera*: Evolutionary implications. *Development Genes and Evolution*, *213*(3), 142–8.
- Paps, J., Baguñ, J., & Riutort, M. (2009). Bilaterian phylogeny: A broad sampling of 13 nuclear genes provides a new lophotrochozoa phylogeny and supports a paraphyletic basal Acoelomorpha. *Molecular Biology and Evolution*, *26*(10), 2397–2406.
- Paps, J., Baguñà, J., & Riutort, M. (2009). Lophotrochozoa internal phylogeny: New insights from an up-to-date analysis of nuclear ribosomal genes. *Proceedings of the Royal Society B*, *276*, 1245–1254.
- Parsons, A. B., Brost, R. L., Ding, H., Li, Z., Zhang, C., Sheikh, B., ... Boone, C. (2004). Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. *Nature Biotechnology*, *22*(1), 62–69.
- Parsons, T. R. (1988). Comparative oceanic ecology of the plankton communities of the Subarctic Atlantic and Pacific oceans. *Oceanography and Marine Biology*, *26*, 317–359.
- Penn, S., Rank, D., Hanzel, D., & Barker, D. (2000). Mining the human genome using microarrays of open reading frames. *Nature Genetics*, *26*, 316–318.
- Peterson, K. J., & Butterfield, N. J. (2005). Origin of the Eumetazoa: Testing ecological predictions of molecular clocks against the Proterozoic fossil record. *Proceedings of the National Academy of Sciences*, *102*(27), 9547–9552.
- Peterson, K. J., & Eernisse, D. J. (2001). Animal phylogeny and the ancestry of bilaterians: Inferences from morphology and 18S rDNA gene sequences. *Evolution and Development*, *3*(3), 170–205.
- Petrovich, R. (2001). Mechanisms of fossilization of the soft-bodied and lightly armored faunas of the Burgess Shale and of some other classical localities. *American Journal of Science*, *301*(8), 683–726.
- Pevzner, P. A., Tang, H., & Waterman, M. S. (2001). An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(17), 9748–9753.

- Philippe, H., Brinkmann, H., Copley, R. R., Moroz, L. L., Nakano, H., Poustka, A. J., ... Telford, M. J. (2011). Acoelomorph flatworms are deuterostomes related to *Xenoturbella*. *Nature*, *18*(11), 1492–1501.
- Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., & Baurain, D. (2011). Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLoS Biology*, *9*(3), e1000602.
- Philippe, H., & Douady, C. J. (2003). Horizontal gene transfer and phylogenetics. *Current Opinion in Microbiology*, *6*(5), 498–505.
- Philippe, H., Lartillot, N., & Brinkmann, H. (2005). Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Molecular Biology and Evolution*, *22*(5), 1246–1253.
- Philippe, H., & Lopez, P. (2001). On the conservations of protein sequences in evolution. *Trends in Biochemical Science*, *26*(7), 414–416.
- Philippe, H., Vienne, D. M. De, Ranwez, V., Roure, B., Baurain, D., & Delsuc, F. (2017). Pitfalls in supermatrix phylogenomics. *European Journal of Taxonomy*, *283*, 1–25.
- Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N., & Delsuc, F. (2005). Heterotachy and long-branch attraction in phylogenetics. *BMC Evolutionary Biology*, *5*, 50.
- Phillips, M. J., Delsuc, F., & Penny, D. (2004). Genome-scale phylogeny and the detection of systematic biases. *Molecular Biology and Evolution*, *21*(7), 1455–1458.
- Pisani, D., Carton, R., Campbell, L. I., & Akanni, W. A. (2013). An overview of arthropod genomics, mitogenomics and the evolutionary origins of the arthropod proteome. *Arthropod Biology and Evolution: Molecules, Development, Morphology, Chapter 3*, 41–60.
- Pisani, D., Poling, L. L., Lyons-Weiler, M., & Hedges, S. B. (2004). The colonization of land by animals: Molecular phylogeny and divergence times among arthropods. *BMC Biology*, *2*, 1.
- Posada, D., & Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology*, *53*(5), 793–808.
- Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, *33*, 501–504.
- Pugh, P. J. A., & McInnes, S. J. (1998). The origin of Arctic terrestrial and freshwater tardigrades. *Polar Biology*, *19*(3), 177–182.

- Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., ... Res, A. (2005). InterProScan: Protein domains identifier. *Nucleic Acids Research*, *33*, 116–120.
- Rahm, P. G. (1921). Biologische und physiologische Beiträge zur Kenntnis der Moosfauna. *Z. Allgem. Physiol*, *10*, 1–35.
- Raup, D. M., & Stanlet, S. M. (1978). Principles of paleontology. *Macmillan Press*, 1–480.
- Rebers, J. E., & Willis, J. H. (2001). A conserved domain in arthropod cuticular proteins binds chitin. *Insect Biochemistry and Molecular Biology*, *31*(11), 1083–1093.
- Regier, J. C., Shultz, J. W., & Kambic, R. E. (2005). Pancrustacean phylogeny: Hexapods are terrestrial crustaceans and maxillopods are not monophyletic. *Proceedings. Biological Sciences / The Royal Society*, *272*(1561), 395–401.
- Regier, J. C., Shultz, J. W., Hussey, A., Ball, B., Wetzer, R., Martin, J. W., & Cunningham, C. W. (2010). Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature*, *462*, 1079–1083.
- Reumont, M. Von, Blanke, A., Richter, S., Alvarez, F., Bleidorn, C., & Jenner, R. A. (2014). The first venomous crustacean revealed by transcriptomics and functional morphology: Remipede venom glands express a unique toxin cocktail dominated by enzymes and a neurotoxin article fast track. *Molecular Biology and Evolution*, *31*(1), 48–58.
- Reyes, A., Pesole, G., & Saccone, C. (2000). Long-branch attraction phenomenon and the impact of among-site rate variation on rodent phylogeny. *Gene*, *259*(1–2), 177–187.
- Richard A. Gibbs, Weinstock, G. M., Metzker, M. L., Muzny, D. M., Sodergren, E. J., Scherer, S., ... Dunn, D. M. (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, *428*(April), 493–521.
- Riesgo, A., Farrar, N., Windsor, P., Giribet, G., & Leys, S. (2014). Early evolution of molecular complexity in metazoans: An analysis of transcriptomes of sponges from all four Porifera classes. *Molecular Biology and Evolution*, *31*, 1102–1120.
- Ringwald, M., Eppig, J. T., Kadin, J. A., Richardson, J. E., & the Gene Expression Database, G. (2000). GXD: A gene expression database for the laboratory mouse: current status and recent enhancements. *Nucleic Acids Research*, *28*(1), 115–119.
- Roberts, D. (1981). Pycnogonids from Strangford Lough, Northern Ireland. *Irish Naturalists' Journal*, *205*, 189–192.
- Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., & Jackman, S. D. (2010). De novo assembly and analysis of RNA-seq data. *Nature Methods*, *7*, 909–912.

- Robinson, A. S., Franz, G., & Atkinson, P. W. (2004). Insect transgenesis and its potential role in agriculture and human health. *Insect Biochemistry and Molecular Biology*, *34*(2), 113–120.
- Rodriguez-Ezpeleta, N., Brinkmann, H., Roure, B., Lartillot, N., Lang, B. F., & Philippe, H. (2007). Detecting and overcoming systematic errors in genome-scale phylogenies. *Systematic Biology*, *56*(3), 389–399.
- Roeding, F., Hagner-Holler, S., Ruhberg, H., Ebersberger, I., von Haeseler, A., Kube, M., ... Burmester, T. (2007). EST sequencing of Onychophora and phylogenomic analysis of Metazoa. *Molecular Phylogenetics and Evolution*, *45*(3), 942–951.
- Ronquist, F., & Huelsenbeck, J. P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, *19*(12), 1572–1574.
- Ronquist, F., Klopfstein, S., Vilhelmsen, L., Schulmeister, S., Murray, D. L., & Rasnitsyn, A. P. (2012). A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. *Systematic Biology*, *61*(6), 973–999.
- Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., ... Huelsenbeck, J. P. (2012). MrBayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, *61*(3), 539–542.
- Rota-Stabelli, O., Campbell, L., Brinkmann, H., Edgecombe, G. D., Longhorn, S. J., Peterson, K. J., ... Telford, M. J. (2011). A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. *Proceedings of the Royal Society B*, *278*(1703), 298–306.
- Rota-Stabelli, O., Daley, A. C., & Pisani, D. (2013). Molecular timetrees reveal a cambrian colonization of land and a new scenario for ecdysozoan evolution. *Current Biology*, *23*(5), 392–398.
- Rouse, G. W. (1999). Trochophore concepts: Ciliary bands and the evolution of larvae in spiralian Metazoa. *Biological Journal of the Linnean Society*, *66*(4), 411–464.
- Ruggiero, M. A., Gordon, D. P., Orrell, T. M., Bailly, N., Bourgoin, T., Brusca, R. C., ... Kirk, P. M. (2015). A higher level classification of all living organisms. *PLoS ONE*, *10*(4), 1–60.
- Saiki, R., Scharf, S., Faloona, F., Mullis, K., Horn, G., Erlich, H., & Arnheim, N. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*.
- Sanderson, M. J. (1996). A nonparametric approach of rate constancy to estimating divergence times in the absence. *Molecular Biology and Evolution*, (1989), 1218–1231.



- Sanger, F., Donelson, J. E., Coulson, a R., Kössel, H., & Fischer, D. (1973). Use of DNA polymerase I primed by a synthetic oligonucleotide to determine a nucleotide sequence in phage fl DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 70(4), 1209–1213.
- Sanger, F., Nicklen, S., & Coulson, a R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), 5463–7.
- Savard, J., Tautz, D., Richards, S., Weinstock, G. M., Gibbs, R. A., Werren, J. H., ... Lercher, M. J. (2006). Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of holometabolous insects. *Genome Research*, 16, 1334–1338.
- Schill, R. O., Mali, B., Dandekar, T., Schnolzer, M., Reuter, D., & Frohme, M. (2009). Molecular mechanisms of tolerance in tardigrades: New perspectives for preservation and stabilization of biological material. *Biotechnology Advances*, 27(4), 348–352.
- Schram, F. R. (1973). Pseudocoelomates and a Nemertine from the Illinois Pennsylvanian. *Journal of Paleontology*, 47(5), 985–989.
- Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., ... Ragg, T. (2006). The RIN: An RNA integrity number for assigning integrity values to RNA measurements. *BMC Molecular Biology*, 7(1), 3.
- Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nature Methods*, 5(1), 16.
- Scourfield, D. J. (1926). On a new type of crustacean from the old Red Sandstone (Rhynie Chert Bed, Aberdeenshire)-*Lepidocaris rhyniensis*. *Philosophical Transactions of the Royal Society B*, 214, 153–187.
- Sharma, P., Kaluziak, S., Pérez-Porro, A., González, V., Hormiga, G., Wheeler, W., & Giribet, G. (2014). Phylogenomic interrogation of Arachnida reveals systemic conflicts in phylogenetic signal. *Molecular Biology and Evolution*, 31(11), 2963–2984.
- Shear, W. A. (1991). The early development of terrestrial ecosystems. *Nature*, 351, 283–289.
- Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., ... Church, G. M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309(5741), 1728–1732.
- Shimodaira, H., & Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution*, 16(8), 1114–1116.

- Sibley, C. G., & Ahlquist, J. E. (1990). Phylogeny and classification of birds: A study in molecular evolution. *Yale University Press*, 1–196.
- Simmons, M. P., & Freudenstein, J. V. (2002). Artifacts of coding amino acids and other composite characters for phylogenetic analysis. *Cladistics*, *18*(3), 354–365.
- Simon, P., Philippe, H., Baurain, D., Jager, M., Richter, D. J., Di Franco, A., ... Manuel, M. (2017). A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Current Biology*, *In Press*.
- Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., & Jones, S. J. M. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Research*, *19*, 1117–1123.
- Smith, M. R., & Ortega-herna, J. (2014). Hallucigenia's onychophoran-like claws and the case for Tactopoda. *Nature*, *514*, 363–366.
- Sonnhammer, E. L. L., & Koonin, E. V. (2002). Orthology, paralogy and proposed classification for paralog subtypes. *Trends in Genetics*, *18*(12), 619–620.
- Sperling, E. A., Frieder, C. A., Raman, A. V., Girguis, P. R., Levin, L. A., & Knoll, A. H. (2013). Oxygen, ecology, and the Cambrian radiation of animals. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(33), 13446–13451.
- Stanke, M., & Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, *19*, 215–225.
- Stefanović, S., Rice, D. W., & Palmer, J. D. (2004). Long branch attraction, taxon sampling, and the earliest angiosperms: Amborella or monocots? *BMC Evolutionary Biology*, *4*(35), 1–19.
- Stenderup, J. T., Olesen, J., & Glenner, H. (2006). Molecular phylogeny of the Branchiopoda (Crustacea) multiple approaches suggest a “diplostracan” ancestry of the Notostraca. *Molecular Phylogenetics and Evolution*, *41*, 182–194.
- Stevens, P. F. (1984). Homology and phylogeny: Morphology and systematics. *Systematic Botany*, *9*(4), 395–409.
- Stiller, J. W., & Hall, B. D. (1999). Long-branch attraction and the rDNA model of early eukaryotic evolution. *Molecular Biology and Evolution*, *16*(9), 1270–1279.
- Strother, P. K., Battison, L., Brasier, M. D., & Wellman, C. H. (2011). Earth's earliest nonmarine eukaryotes. *Nature*, *473*, 505–509.
- Struck, T. H., & Fisse, F. (2001). Phylogenetic position of Nemertea derived from phylogenomic data. *Molecular Biology and Evolution*, *25*(4), 728–736.

- Struck, T. H., Wey-fabrizius, A. R., Golombek, A., Hering, L., Weigert, A., Bleidorn, C., ... Hankeln, T. (2014). Platyzoan paraphyly based on phylogenomic data supports a noncoelomate ancestry of Spiralia. *Molecular Biology and Evolution*, *31*(7), 1833–1849.
- Swofford, D. L. (2002). Phylogenetic analysis using parsimony (and other methods). Version 4. *Sunderland, MA: Sinauer Associates*.
- Szaniawski, H. (2005). Cambrian chaetognaths recognized in Burgess Shale fossils. *Acta Palaeontologica Polonica*, *50*(1), 1–8.
- Tatusova, T. A., & Madden, T. L. (1999). BLAST 2 SEQUENCES, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiology Letters*, *174*, 247–250.
- Tautz, D., & Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nature Reviews Genetics*, *12*(10), 692–702.
- Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in Life Sciences*, *17*, 57–86.
- Telford, M. J., & Copley, R. R. (2005). Animal phylogeny: Fatal attraction. *Current Biology*, *15*(8), 296–299.
- Telford, M. J., Bourlat, S. J., Economou, A., Papillon, D., & Rota-stabelli, O. (2008). The evolution of the Ecdysozoa. *Philosophical Transactions of the Royal Society B*, *363*, 1529–1537.
- Telford, M. J., & Holland, P. W. H. (1993). The phylogenetic affinities of the chaetognaths: A molecular analysis. *Molecular Biology and Evolution*, *10*(3), 660–676.
- The *C. elegans* Sequencing Consortium. (1998). Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science*, *282*(5396), 2012–2018.
- The UniProt Consortium. (2008). The universal protein resource (UniProt). *Nucleic Acids Research*, *36*, 190–195.
- Thompson, J. D., Higgins, D., & Gibson, T. (1994) ClustalW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Research*, *22*(22), 4673–4680.
- Thorne, J. L., Kishino, H., & Painter, I. S. (1998). Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution*, 1647–1657.
- Todaro, M. A., & Shirley, T. C. (2003). A new meiobenthic priapulid (Priapulida, Tubiluchidae) from a Mediterranean submarine cave. *Italian Journal of Zoology*, *70*(1), 79–87.

- Tokioka, T. (1965). The taxonomical outline of chaetognatha. *Publications of the Seto Marine Biological Laboratory*, 12(5), 335–357.
- Turcatti, G., Romieu, A., Fedurco, M., & Tairi, A. (2008). A new class of cleavable fluorescent nucleotides: Synthesis and optimization as reversible terminators for DNA sequencing by synthesis y. *Nucleic Acids Research*, 36(4), 25.
- Vannier, J., Steiner, M., Renvoisé, E., Hu, S., Casanova, J., Hu, S., & Casanova, J. (2007). Early Cambrian origin of modern food webs: Evidence from predator arrow worms. *Proceedings of the Royal Society B*, 274, 627–633.
- Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley-Interscience, 1st Editio, 1–768.
- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., ... Zhu, X. (2001). The sequence of the human genome. *Science*, 291(5507), 1304–1351.
- Vieira, F. G., & Rozas, J. (2011). Comparative genomics of the odorant-binding and chemosensory protein gene families across the arthropoda: Origin and evolutionary history of the chemosensory system. *Genome Biology and Evolution*, 3(1), 476–490.
- Von Reumont, B. M., Jenner, R. A., Wills, M. A., Dell’Ampio, E., Pass, G., Ebersberger, I., ... Misof, B. (2012). Pancrustacean phylogeny in the light of new phylogenomic data: Support for remipedia as the possible sister group of hexapoda. *Molecular Biology and Evolution*, 29(3), 1031–1045.
- Wägele, J. W. (1996) Identification of apomorphies and the role of groundpatterns in molecular systematics. *Journal of Zoological Systematics and Evolutionary Research*, 34(1) 31-39.
- Wallace, R. B., Shaffer, J., Murphy, R. F., Bonner, J., Hirose, T., & Itakura, K. (1979). Hybridization of synthetic oligodeoxyribonucleotides to ??X 174 DNA: The effect of single base pair mismatch. *Nucleic Acids Research*, 6(11), 3543–3558.
- Walossek, D. (1993). The Upper Cambrian Rehbachiella and the phylogeny of Branchiopoda and Crustacea. *Lethaia*, 26, 318–318.
- Wasmuth, J. D., & Blaxter, M. L. (2004). prot4EST: Translating expressed sequence tags from neglected genomes. *BMC Bioinformatics*, 14, 1–14.
- Waterman, M. S., & Arratia, R. (1984). Recognition in several sequences: Consensus and alignment. *Bulletin of Mathematical Biology*, 46(4), 515–527.
- Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., & Abril, J. F. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915), 520–562.

- Watson, J. D., & Crick, F. H. C. (1953). Molecular structure of nucleic acids: A structure for Deoxyribose Nucleic Acid. *Nature*, *171*, 737–738.
- Wenzel, J. W., & Siddall, M. E. (1999). Noise. *Cladistics*, *15*(1), 51–64.
- Wheat, C. W., & Wahlberg, N. (2013). Phylogenomic insights into the cambrian explosion, the colonization of land and the evolution of flight in Arthropoda. *Systematic Biology*, *62*(1), 93–109.
- Wheeler, B. M., Heimberg, A. M., Moy, V. N., Sperling, E. A., Holstein, T. W., Heber, S., & Peterson, K. J. (2009). The deep evolution of metazoan microRNAs. *Evolution and Development*, *11*(1), 50–68.
- Wilkinson, R. D., Steiper, M. E., Soligo, C., Martin, R., Yang, Z., & Tavaré, S. (2011). Dating primate divergences through an integrated analysis of palaeontological and molecular data. *Systematic Biology*, *60*, 16–31.
- Wilson, H. M., & Anderson, L. I. (2004). Morphology and taxonomy of Paleozoic millipedes (Diplopoda: Chilognatha: Archipolypoda) from Scotland. *Journal of Paleontology*, *78*(1), 169–184.
- Winnebeck, E. C., Millar, C. D., & Warman, G. R. (2010). Why does insect RNA look degraded? *Journal of Insect Science*, *10*(159), 1–7.
- Wu, H., Huang, H., Yeh, L., & Barker, W. (2003). Protein family classification and functional annotation. *Computational Biology and Chemistry*, *27*(1), 37–47.
- Xia, X., Xie, Z., Salemi, M., Chen, L., & Wang, Y. (2003). An index of substitution saturation and its application. *Molecular Phylogenetics and Evolution*, *26*(1), 1–7.
- Yamasaki, H., Fujimoto, S., & Miyazaki, K. (2015). Phylogenetic position of Loricifera inferred from nearly complete 18S and 28S rRNA gene sequences. *Zoological Letters*, *1*(1), 18.
- Yang, Z., & Donoghue, P. C. J. (2016). Dating species divergences using rocks and clocks. *Philosophical Transactions of the Royal Society B*, *371*.
- Yang, Z., & Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: A Markov chain monte carlo method. *Molecular Biology and Evolution*, *14*(7), 717–724.
- Yang, Z., & Rannala, B. (2012). Molecular phylogenetics: Principles and practice. *Nature Reviews Genetics*, *13*, 303–314.
- Yeates, D. K., & Wiegmann, B. M. (1999). Congruence and controversy: Toward a higher-level phylogeny of Diptera. *Annual Review of Entomology*, *44*, 397–428.

- Zapata, F., Goetz, F. E., Smith, S. A., Howison, M., & Siebert, S. (2015). Phylogenomic analyses support traditional relationships within Cnidaria. *PLoS ONE*, *10*(10), 1–13.
- Zelditch, M. L., Fink, W. L., & Swiderski, D. L. (1995). Morphometrics, homology, and phylogenetics: Quantified characters as synapomorphies. *Systematic Biology*, *44*(2), 179–189
- Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829.
- Zhang, G., Li, C., Li, Q., Li, B., Larkin, D. M., Lee, C., ... Ganapathy, G. (2014). Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*, *346*(6215), 1311–1320.
- Zhao, Q.-Y., Wang, Y., Kong, Y.-M., Luo, D., Li, X., & Hao, P. (2011). Optimizing de novo transcriptome assembly from short-read RNA-Seq data: A comparative study. *BMC Bioinformatics*, *12*(Suppl 14), S2.
- Zuckerkandl, E., & Pauling, L. (1965). Molecules as documents of evolutionary history. *Journal of Theoretical Biology*, *8*(2), 357–366.
- Zuckerkandl, E., & Pauling, L. (1962). Molecular disease, evolution, and genic heterogeneity. *Horizons in Biochemistry*, 189–225.
- Zwickl, D. J., & Hillis, D. M. (2002). Increased taxon sampling greatly reduces phylogenetic error. *Systematic Biology*, *51*(4), 588–598.

## Appendices

### 2.2.2 DNA and RNA Extraction Protocols

#### 2.2.2 A DNA Extractions

DNA extraction protocols were provided by Eoin Mulvihill. The DNA extraction protocol was followed directly from the Qiagen Genomic DNA Handbook (<https://www.qiagen.com/ie/resources/resourcedetail?id=97640bc9-e4fe-4c4b-83f6-ac7ca4181597&lang=en>).

#### *Equipment*

- Baked mortar and pestle (autoclaved overnight at 200<sup>0</sup>C)
- Qiagen Genomic-tip 20/G
- Electronic pipette
- Vortex
- Falcon centrifuge tube (15ml)
- Centrifuge (Eppendorf 5417R)
- Nanodrop (Thermo Scientific)
- Agarose gel electrophoresis

#### *Reagents*

- Buffer G2 (RNase A + 0.1ml Qiagen Proteinase K stock solution)
- Buffer QBT (750mM NaCl + 50mM MOPS pH 7.0 + 15% isopropanol v/v + 0.15% Triton X-100 v/v)
- Buffer QC (1.0M NaCl + 50mM MOPS pH 7.0 + 15% isopropanol v/v)
- Buffer QF (1.25M NaCl + 50mM Tris-Cl pH8.5 + 15% isopropanol v/v)

- Tris-EDTA Buffer solution
  - Isopropanol (sigma, I-9516)
  - Ethanol, molecular grade (Sigma, E7023)
  - Agarose gel
  - Ethium bromide
1. The specimen was thawed from the  $-80^{\circ}\text{C}$  freezer and the tissue was weighed. According to Qiagen extraction protocols, spleen and liver tissue are the most bountiful tissue types for DNA & RNA extraction because to their high protein content. Due to the anatomical restrictions of invertebrates this is not possible so most of the tissue was used. This was also necessary due to the size of the specimens we were working on in order to achieve a high enough yield of DNA and RNA from the concentration of tissue at our disposal.
  2. The Buffer G2, QBT, and QC concentrations were prepared.
  3. The tissue was placed in an autoclaved mortar, covered in liquid nitrogen ( $\text{N}_2$ ) and ground into a fine powder.
  4. The powder was then placed in a 15ml falcon tube and mixed well with 2ml of Buffer G2. This buffer lyses the nuclei allowing the release of DNA into the cell, strips it of any bound proteins such as histones, denatures enzymes that breakdown DNA, and finally denatures any RNA in the sample. The solution was incubated at  $50^{\circ}\text{C}$  for 2 hours.
  5. The Qiagen Genomic-tip 20/G was equilibrated with 1ml of Buffer QBT. The Triton X-100 component of the buffer instigates the flow. Once the Genomic-tip drained, the buffer reached the frit, preventing the tip from running dry.



The tissue sample was vortexed for 10 seconds, to prevent clogging, and then applied to the Genomic-tip.

6. Similarly to step 5, the Genomic tip was allowed to drain via gravity and the DNA bound to the resin. To further purify the DNA, the Genomic-tip was washed three times with 1ml Buffer QC. The tip was once again allowed to drain via gravity flow.
7. After these purification steps, the DNA was eluted from the resin and collected into a new falcon tube using two aliquots of 1ml Buffer QF.
8. The eluted DNA was precipitated by adding 1.4ml of isopropanol and mixed thoroughly through 20 inversions. Immediately after this the DNA was split into 1.5ml tubes and centrifuged at 20,000g, 4<sup>0</sup>C, for 10 minutes.
9. The DNA pellets were washed with 1ml of 70% ethanol, inverted, and centrifuged again at the same temperature but at 10,000g for 10 minutes. The supernatant was carefully removed and the pellets were allowed to dry.
10. The DNA was re-suspended in 1ml of Tris-EDTA Buffer to protect it from degradation enzymes such as DNase. Finally, a small aliquot of the extracted DNA was taken for concentration analysis by nanodrop and integrity via agarose gel electrophoresis.

### **2.2.2 B RNA Extractions**

RNA extraction protocols were provided by Eoin Mulvihill. The RNA extraction protocol was followed directly from the Qiagen Handbook.

#### *Equipment*

- Baked mortar and pestle (autoclaved overnight at 200<sup>0</sup>C)
- Sureone filter tips (Fisherbrand, FB78098 / FB78108)
- RNase free LoBind tubes, 1.5ml (Eppendorf, 0030.108.051)
- Electronic pipette
- Bioanalyzer (Agilent)
- Nanodrop (Thermo Scientific)
- Refrigerated centrifuge (Eppendorf 5417R)

#### *Reagents*

- DEPC treated H<sub>2</sub>O (Invitrogen, 750024)
- Trisure (Bioline, BOI-38033)
- Chloroform (Sigma, C-2432)
- Isopropanol (sigma, I-9516)
- Phenol:chloroform:isoamyl alcohol (25:24:1, v/v) (Invitrogen, 15593-031)
- Sodium acetate (3M, pH 5.2) (Sigma, S-7899)
- RNase free DNase (Promega, M6101)
- Ethanol, molecular grade (Sigma, E7023)
- RNase Zap (Sigma, R-2020)

1. Specimen vials were removed from the  $-80^{\circ}\text{C}$  freezers and kept on ice, keeping specimens cold is critical during RNA extraction procedures as RNA begins to degrade at room temperature.  
  
As per the aseptic technique, all equipment was either autoclaved or washed in an ethanol solution before use.
2. The bench was washed down with RNase Zap, gloves were worn throughout the procedure, rinsed with RNase Zap, and replaced regularly.
3. 70%, 75%, and 95% ethanol solutions were made up with molecular grade ethanol and DEPC treated  $\text{H}_2\text{O}$ .
4. The specimen was weighed, then placed in a mortar and pestle and covered in liquid  $\text{N}_2$ , ground into a powder, and rinsed with  $400\mu\text{l}$  Trisure. The solution was left for 30 minutes until it became liquid and transferred to a new Eppendorf. The mortar was rinsed with a further  $400\mu\text{l}$  Trisure and added to the Eppendorf.
5.  $200\mu\text{l}$  of chloroform was added to the Eppendorf and gently shaken for 15 seconds, the solution was left for 5 minutes, and then centrifuged at 20,000 RCF for 15 minutes at  $4^{\circ}\text{C}$ .
6. The top, clear, aqueous layer was removed with a pipette and placed in a new Eppendorf. Care was taken to ensure that none of the pink organic layer was taken also.
7.  $500\mu\text{l}$  isopropanol was added to the new Eppendorf, mixed well, and left at room temperature for 10 minutes. The solution was then spun in the centrifuge (20,000 RCF) for 10 minutes at  $4^{\circ}\text{C}$ .

8. At this point an RNA pellet formed at bottom of the Eppendorf. All isopropanol was removed using a pipette. The pellet was re-suspended in 1ml of 75% ethanol in DEPC H<sub>2</sub>O.
9. The Eppendorf was inverted several times and spun at 7,500RPM, after which all ethanol was removed from the Eppendorf and the pellets were left to dry for 5 minutes. The pellet was then re-suspended in 43μl of DEPC H<sub>2</sub>O.
10. 5μl of DNase buffer (10X) and 2μl of RNase-free DNase were added to the RNA solution and incubated at 37<sup>0</sup>C for 30 minutes, afterwards they were immediately put on ice.
11. A phenol:chloroform solution was used to inactivate the enzymes and further purify the RNA. This solution was made up of 150μl of DEPC treated H<sub>2</sub>O and 200μl of ultra pure phenol:chloroform:isoamyl alcohol. The Eppendorf was vortexed and centrifuged at 20,000 RCF for 5 minutes. The top layer of the solution was then extracted into a new 1.5ml tube.
12. The RNA was made again into pellet form through an ethanol precipitation; 20μl of sodium acetate and 440μl of 95% ethanol. The sample was then left for 30 minutes at -80<sup>0</sup>C and then centrifuged at 20,000 RCF for 20 minutes.
13. The RNA was desalted by adding 500μl of 70% ethanol to the tube and then inverted. The tube was once again centrifuged at 20,000 RCF for 5 minutes and the ethanol was removed using a pipette. The pellet was then left to dry and then re-suspended in 55μl of DEPC treated H<sub>2</sub>O.
14. Finally, 5μl of the purified RNA solution was taken in order to measure its purity and concentration (nanodrop), and integrity (bioanalyzer).

## Publications

Pisani, D., Carton, R., Campbell, L. I., & Akanni, W. A. (2013). An overview of arthropod genomics, mitogenomics and the evolutionary origins of the arthropod proteome. *Arthropod Biology and Evolution: Molecules, Development, Morphology*, Chapter 3, 41–60.

Lozano-Fernandez, J., Carton, R., Tanner, A. R., Puttick, M. N., Blaxter, M., Vinther, J., ... Pisani, D. (2016). A molecular palaeobiological exploration of arthropod terrestrialization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1699)

# An Overview of Arthropod Genomics, Mitogenomics, and the Evolutionary Origins of the Arthropod Proteome

Davide Pisani, Robert Carton, Lahcen I. Campbell, Wasiu A. Akanni, Eoin Mulville and Omar Rota-Stabelli

## Contents

<b>3.1</b>	<b>Introduction</b> .....	41	<b>3.2.4</b>	<b>The Hazards of Using Arthropod Mitochondrial Genomes for Phylogenetics</b> .....	46
<b>3.2</b>	<b>Arthropod Mitogenomes: Useful, but Hazardous Small Genomes</b> .....	42	<b>3.3</b>	<b>Arthropod Comparative Genomics</b> .....	46
3.2.1	Mitogenomic Studies.....	42	3.3.1	Uneven Taxonomic Sampling.....	48
3.2.2	The Structure of the Arthropod Mitochondrial Genome .....	44	3.3.2	Heterogeneity of Genome Sizes and Shortage of microRNA .....	49
3.2.3	Arthropod Mitogenomes: A Composition Nightmare .....	44	<b>3.4</b>	<b>A Genomic Phylostratigraphic Analysis of the Arthropod Proteomes</b> .....	50
			3.4.1	A Robust Phylogenetic Framework for Genomic Studies .....	50
			3.4.2	Expanding Our Understanding of the Arthropod Comparative Genomics .....	51
			3.4.3	The Evolution of Orphan Gene Families in Arthropoda .....	52
			3.4.4	Conserved Rate of Gene Gain with Some Surprises .....	55
			<b>3.5</b>	<b>Conclusions</b> .....	56
			<b>Appendix: Methods for the Analyses Presented in this Chapter</b> .....		56
			A. Generation of the Onychophoran Transcriptome.....		56
			B. Mitogenomic Compositional Analyses.....		57
			C. Phylogenetic Analyses.....		57
			<b>References</b> .....		58

D. Pisani (✉)

School of Biological Sciences and School of Earth Sciences, University of Bristol, Woodland Road, Bristol, BS8 1UG, UK  
e-mail: Davide.Pisani@bristol.ac.uk

R. Carton · L. I. Campbell · W. A. Akanni · E. Mulville

Department of Biology, The National University of Ireland, Callan Building, Maynooth, County Kildare, Ireland  
e-mail: robcarton@gmail.com

L. I. Campbell

e-mail: lahcencampbell@yahoo.ie

W. A. Akanni

e-mail: waakanni13@gmail.com

E. Mulville

e-mail: eoin.d.mulvihill@nuim.ie

O. Rota-Stabelli

IASMA Research and Innovation Centre, Fondazione Edmund Mach, Via E. Mach 1, 38010 San Michele all'Adige (TN), Italy  
e-mail: omar.rota@fmach.it

## 3.1 Introduction

Arthropods represent the largest majority of animal biodiversity and include organisms of economic interest and key model species. It is thus unsurprising that the genome of an arthropod, the fruit fly *Drosophila melanogaster*, was among the very first to be sequenced (Adams et al. 2000) and that to date, about 21 *Drosophila*

genomes as well as a variety of other arthropod genomes have been sequenced. Despite this promising start, current sampling is biased towards economically relevant species, and a suitable close outgroup to the arthropods, which is necessary to polarise genomic studies, is still missing. Among the suitable outgroups to the Arthropoda, the Nematoda represent one of the largest components of the extant animal biomass, and their economic importance is comparable to that of the more biodiverse arthropods. As with the Arthropoda, the importance of the nematodes is reflected in the fact that the very first animal genome to be sequenced was that of the nematode *Caenorhabditis elegans* (The C. elegans genome consortium 1998). Despite the nematodes being phylogenetically close to the arthropods (Aguinaldo et al. 1997; Copley et al. 2004; Dopazo and Dopazo 2005; Philippe et al. 2005; Irimia et al. 2007; Roy and Irimia 2008; Dunn et al. 2008; Belinky et al. 2010; Hejnol et al. 2009; Holton and Pisani 2010), this group is composed of highly derived species, both genetically and morphologically. Accordingly, their genomes are unlikely to be of great utility in understanding arthropod genome evolution. Some genomic data (mostly in the form of transcriptomes) are now available for other smaller ecdysozoan phyla, and some genomes (Priapulida and Tardigrada) are on the horizon. Nonetheless, enough genomic information is now available for the Arthropoda (Table 3.1) to justify an investigation into the evolution of their genome. Such an analysis, however, is intimately dependent on the availability of a robust phylogenetic background, and to a lesser extent, robust divergence times for the nodes in the background phylogeny.

In this chapter, we present an overview of arthropod mitochondrial genomics (Sect. 3.2) and nuclear genomics (Sect. 3.3). We then exploit the available genomic information to investigate the evolutionary origin of novel proteins (orphan gene families) in the arthropod proteome (Sect. 3.4). We notably present the first genomic-scale data set for the Onychophora and include it in our analyses to be able to

consider the closest sister group of the Arthropoda (see Campbell et al. 2011) when identifying orphan gene families. Inclusion of new data for the Onychophora is key to this study as it allows the correct identification of the orphan protein families that arose in the stem arthropod lineage.

---

## 3.2 Arthropod Mitogenomes: Useful, but Hazardous Small Genomes

Each cell contains up to hundreds of mitochondria, and each mitochondrion possesses many copies of their own small, typically circular, genome (mitogenome or mtDNA). Therefore, mitochondrial genes largely outnumber the nuclear ones in terms of their copy number by several orders of magnitude, making mitochondrial genes easy to extract and amplify. Accordingly, there has been an exceptional amount of articles published that attempted (not always successfully) to resolve the phylogenetic relationships within Arthropoda (and more broadly Metazoa) using mtDNA. Other reasons behind the fortunes of mtDNA are as follows: a relatively conserved gene set, the unambiguous orthology of genes, the presence of rare genetic changes, and the availability of universal primers for many lineages. Other characteristics of the mitogenome, however, make it a double-edged sword. These are accelerated mutation rate due to uniparental inheritance, and severe biases in the composition of nucleotides that are often responsible for the dilution of the phylogenetic signal in mtDNA (Bernt et al. 2012). In this section, we review some of these aspects.

### 3.2.1 Mitogenomic Studies

Mitogenomic studies have helped throughout the 1990s and 2000s to elucidate some arthropod affinities. For example, one of the earliest studies providing robust, non-rRNA based, evidence in support of the Pancrustacea used mtDNA gene order comparisons (Boore et al. 1998) and

**Table 3.1** The most important of the available Arthropod genomes

	Species	Genome size (Mb)	GC (%)	Chromosomes	Genes	Transcripts
Chelicerata Acari-Acariformes	<i>Tetranychus urticae</i>	89.6	32.3	N/A	N/A	18,414
Chelicerata Acari-Parasitiformes	<i>Ixodes scapularis</i>	1,896.32	45.5	15	7,112	5,867
Myriapoda Chilopoda	<i>Strigamia maritima</i>	173.61	35.7	N/A	N/A	N/A
Crustacea Branchiopoda	<i>Daphnia pulex</i>	158.62	40.8	N/A	30,613	30,611
Hexapoda Phthiraptera	<i>Pediculus humanus</i>	108.37	27.5	N/A	10,993	10,775
Hexapoda Coleoptera	<i>Tribolium castaneum</i>	210.27	38.4	10	10,132	9,833
Hexapoda Hemiptera	<i>Acyrtosiphon pisum</i>	464	29.6	4	N/A	11,089
Hexapoda Hymenoptera	<i>Apis mellifera</i>	250.29		16	N/A	N/A
Hexapoda Lepidoptera	<i>Bombyx mori</i>	431.75	37.7	28	N/A	N/A
Hexapoda Lepidoptera	<i>Heliconius melpomene</i>	269		21	12,669	N/A
Hexapoda Diptera	<i>Drosophila melanogaster</i>	139.73	42.2	6	15,431	24,113
Hexapoda Diptera	<i>Aedes aegypti</i>	1,310.11	38.3	3	16,684	16,785
Hexapoda Diptera	<i>Anopheles gambiae</i>	265.03	44.5	5	13,240	14,099

N/A not available. All the values in the table were obtained either from the NCBI website or from the original genome paper

mtDNA sequence phylogeny (Hwang et al. 2001). However, in some cases, mitogenomic studies have pointed towards likely incorrect topologies, for example, suggesting a Myriapoda plus Chelicerata grouping (Hwang et al. 2001; Negrisol et al. 2004; Pisani et al. 2004), which has also been uncovered by some analyses of nuclear coding genes (e.g. Pisani et al. 2004; Dunn et al. 2008; Roeding et al. 2009; Hejnol et al. 2009; Meusemann et al. 2010) and that most likely represent a long-branch attraction artefact (Pisani 2004; Rota-Stabelli and Telford 2008; Rota-Stabelli et al. 2010; Campbell et al. 2011; Rota-Stabelli et al. 2011). This topology was most likely the result (in the case of the mtDNA analyses) of a systematic error caused by the use of distant outgroups and compositionally biased taxa (Rota-Stabelli and Telford 2008). Such features of the mitochondrial

genomes may seriously affect phylogenetic reconstruction unless they are taken into account when inferring phylogenies (Rota-Stabelli et al. 2010).

Utility of the mitochondrial genomes is not restricted to phylogeny. The most widely used arthropod barcode is a region of approximately 650 nucleotides of the subunit I of the cytochrome oxidase complex (COX1)—a mitochondrial gene. Other mitochondrial genes (NADH4, for example) are occasionally added to COX1 to improve resolution. A possible risk with mtDNA-based barcoding is the amplification of pseudo-genes numts (nuclear copies of mitochondrial genes), which may disrupt barcoding studies. In addition, single gene barcoding has been shown to fail occasionally and the advent of NGS makes it an obsolete approach (Taylor and Harris 2012). Nevertheless, barcoding remains the



method of choice for biodiversity studies (likely because its simplicity and low cost makes it appealing to founding agencies).

To date, there are more than 300 complete arthropod mitochondrial genomes, and partial sequences are in excess of a million. The taxonomic sampling is, however, extremely biased towards economically relevant species: 47 chelicerates (mostly ticks and mites), 53 crustaceans (mostly malacostracans), 198 insects (mostly beetles, dipterans, and hemipterans), and only 9 myriapods. Still, most major orders and classes are now represented, thus providing an invaluable starting point for comparative analyses.

### 3.2.2 The Structure of the Arthropod Mitochondrial Genome

Arthropod mtDNA varies in size from less than 14,000 bp in the spider *Ornithoctonus huwena* to more than 19,000 bp in *D. melanogaster*. This difference is almost entirely due to non-coding intergenic regions, particularly the major non-coding region commonly called *control region*. Due to its low structural constraints and high tendency to accumulate A and T nucleotides, this region is also called the AT-rich region. The AT-rich region is involved in both replicative and transcriptional processes and typically contains structural elements like hairpin loops and thymidine stretches (Zhang and Hewitt 1997), elements that do not seem to be conserved throughout the arthropods.

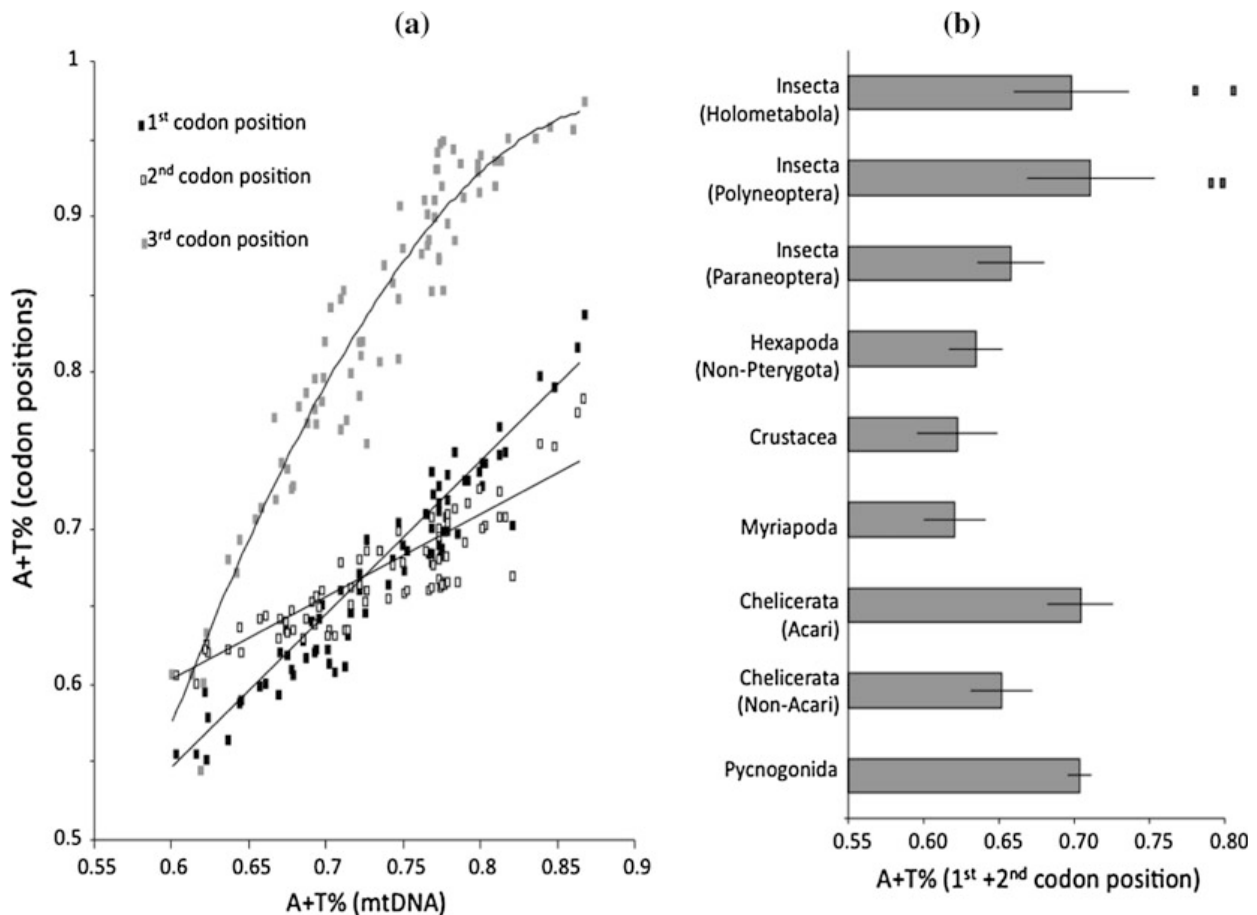
The gene content of the arthropod mtDNA is the same as in most other bilaterians; it typically consists of 13 coding genes, 2 ribosomal RNA subunits, and 20 tRNAs (Boore 1999). This gene set is highly conserved throughout the phylum, although a few exceptions can be found. Examples include a tRNA-Ser duplication in *Thrips imaginis* (Hexapoda: Thysanoptera) (Shao and Barker 2003), a tRNA-His duplication in *Speleonectes tulumensis* ('Crustacea': Remipedia) (Lavrov et al. 2004), and a tRNA-Cys triplication in *Pollicipes polymerus* ('Crustacea': Cirripedia) (Lavrov et al. 2004). Many arthropod mitochondrial coding genes lack a stop codon

(TAA or TAG) and possess a single T or TA at the 3-terminal end. The correct stop codon is then assembled by the polyadenylation of an excised, presumably polycistronic, transcript. Although most arthropod mitogenomes use the invertebrate genetic code, it has been shown that some lineages use a slightly different code (Abascal et al. 2006). Remarkably, this new genetic code is scattered throughout the arthropod tree.

Although the gene content is conserved throughout the arthropods, the gene order may vary significantly (Lavrov et al. 2004). Comparative studies have determined an arthropod ancestral gene order, which is represented (retained) by *Limulus polyphemus*, while the pancrustacean gene order differs from that of all the other arthropods by the position of one of the two leucine tRNAs. tRNAs in general are mostly responsible for variation in gene order as they are hot spots of recombination. Less often, coding genes change their position or swap strand, allowing for variation in gene-specific strand asymmetry, as detailed below.

### 3.2.3 Arthropod Mitogenomes: A Composition Nightmare

The main source of compositional heterogeneity in mtDNA is mutational pressure, which is correlated with a deficiency in the mtDNA repair system and with a consequent inefficiency at replacing erroneous insertions of A nucleotides (Reyes et al. 1998). Compared to other metazoans, arthropod lineages are typically enriched in A and T. In the absence of strong purifying selection, this mutational pressure affects also encoded proteins, which are enriched in amino acids encoded by A+T-rich codons (Foster et al. 1997; Foster and Hickey 1999; Rota-Stabelli et al. 2010). The effect of this mutational pressure depends on structural constraints acting on the genes: more conserved genes such as COX1 accumulate fewer A+T mutations than poorly constrained genes such as ATP8. In addition, not all positions of a gene are affected in a similar way: while the 1st and 2nd codon positions are



**Fig. 3.1** Compositional heterogeneity in arthropod mitogenomes. **a** A+T % content of the three codon positions plotted against that calculated on the whole mtDNA. Second codon position is the most constrained,

more constrained by the genetic code, the 3rd codon positions are more prone to accumulate A+T mutations and experience saturation of replacement events (Fig. 3.1a). Interestingly, 1st codon positions show a different A+T replacement pattern from the 2nd. This advocates the employment of different models of evolution for the 1st and 2nd codon positions and the exclusion of the 3rd codon positions when performing phylogenetic reconstruction from nucleotide sequences. This would, at least partially, compensate for possible artefactual attraction in the case that unrelated species have a similarly increased A+T content.

The A+T content is not homogeneously distributed throughout the arthropods: some groups such as Pycnogonida, Acari, and some insects are more A+T rich than other lineages (Fig. 3.1b). This uneven distribution of nucleotide content may have been responsible for the

while 3rd codon position changes so dramatically that reaches plateau in some species. **b** A+T % calculated on the whole mtDNA in different arthropod lineages. Nucleotide content varies between and within classes

artefactual attraction of, for example, Acari and Pycnogonida in published phylogenetic studies (Podsiadlowski and Braband 2006). In some species such as the bees *Apis mellifera* and *Melipona bicolor* and the hemipterans *Schizaphis graminum* and *Aleurodicus dugesii* (grey dots in Fig. 3.1b), the A+T content reaches extremely high values, the highest ever reported for eukaryotic coding genes.

Strand asymmetry is another type of compositional heterogeneity affecting mtDNA. This bias is related to the origin and direction of mtDNA replication (Reyes et al. 1998) and leads one strand to become enriched in G (and to a lesser extent in T), while the other strand become enriched in C (and less in A). Strand asymmetry is generally expressed in terms of GC-skew. Although all genes in a mitochondrial genome usually have a similar A+T content, homologous genes from different organisms may

have extremely different, sometimes opposite, GC (and AT)-skew: this depends on the strand on which the gene is located, and on its position relative to the origin of replication (Lavrov et al. 2000). Therefore, there is a link between strand asymmetry and gene order.

In arthropods, most mtDNA coding genes are characterised by a negative GC-skew (they have more C than G), while four genes that lie on the opposite strand are characterised by a positive GC-skew. This situation is characteristic, in particular, of species characterised by the arthropod ancestral gene order (as in Fig. 3.2a). In some species, the GC-skew is opposite for all the genes, although the gene order is substantially identical to that of the ancestral arthropods (Fig. 3.2b). In such cases, it is the origin of replication (the control region) that underwent a modification, for example, a duplication or an inversion of strand. In other cases, all genes may have been translocated on the same strand, so that all the genes possess either a positive or a negative GC-skew (Fig. 3.2c).

### 3.2.4 The Hazards of Using Arthropod Mitochondrial Genomes for Phylogenetics

It has been shown that both sources of compositional heterogeneity (A+T mutational pressure and strand asymmetry) may play strong roles in generating artefactual mitogenomic phylogenies (Hassanin et al. 2005; Rota-Stabelli et al. 2010). Compositional problems are worsened by the accelerated rate of evolution of mitogenomic sequences, which is related to the uniparental inheritance characterising mitochondria. An effective approach to deal with these problems is to improve models of mitochondrial sequence evolution both at the nucleotide (Hassanin et al. 2005) and protein level (Abascal et al. 2007; Rota-Stabelli et al. 2009). However, if the biases are too strong to be accounted for using models, one might have to try to highlight potentially incorrect topologies by experimenting with character exclusion strategies targeting more affected genes or codon positions (e.g.

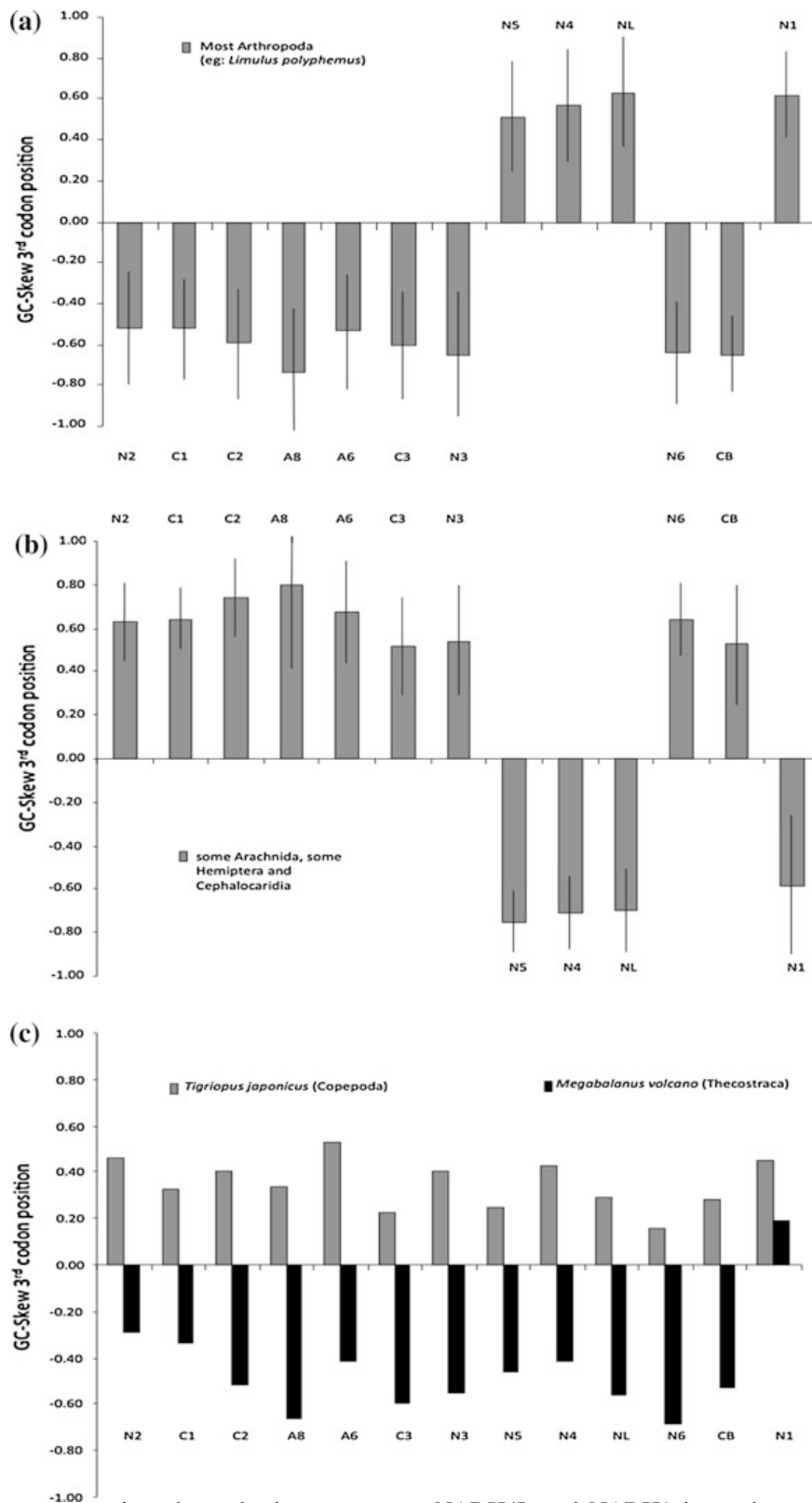
Rota-Stabelli et al. 2010). Sophisticated evolutionary models which account for among site and among branch heterogeneity (Foster 2004; Blanquart and Lartillot 2008) are useful to lessen the effects of these mitochondrial compositional biases. Another obvious approach is to enlarge or modify taxonomic sampling. More taxa may break problematic branches and reduce the number of homoplasies responsible for long-branch (or compositional) attractions. In some conditions, when addition of more taxa does not seem to be breaking long branches, it might be useful to carry out experiments in which taxon sampling is modified (by taxon removal) and the effect of these taxonomic reductions on the analyses is monitored (e.g. Rota-Stabelli et al. 2012; Campbell et al. 2011). More generally, it is advisable to conduct an exploratory compositional analysis of the properties of the mitochondrial genomes under consideration prior to phylogenetic inference. This is particularly true for the arthropods, which include some highly derived lineages, parasites, for example, whose particular lifestyle is responsible for bottleneck events and therefore extreme acceleration of substitution rates or divergent nucleotide compositions.

Compositional biases (and related phylogenetic artefacts) have been primarily studied using mitogenomic data sets (Foster et al. 1997). The advent of the phylogenomic-type (nuclear) data sets has been initially seen as a relief in terms of compositionally related biases. This may, however, not be the case: the community is just noticing that even large genomic data sets are not free from compositional problems that can cause serious phylogenetic artefacts (Nabholz et al. 2012; Rota-Stabelli et al. 2012). Still, the origins of such biases in nuclear genomic data are largely not known.

---

## 3.3 Arthropod Comparative Genomics

The study of arthropod genomics started with the sequencing of the genome of the fruit fly *D. melanogaster* (Adams et al. 2000). Currently,



**Fig. 3.2** Strand asymmetry in arthropod mitogenomes. Each gene in the mtDNA is characterised by a different propensity to accumulate mutations towards G or C. This is because different genes lie on different strands and each strand has his own mutational pressure, described here by the GC-skew statistics. **a** In most arthropods, the majority of genes are on the same strand and possess a negative GC-skew; the ORF of NADH4, NADH5

NADH4L and NADH1 is on the opposite strand; as a consequence, these genes accumulate more G and have a positive GC-skew. **b** Some phylogenetically unrelated arthropods experienced an inversion of the replicative system, which leads to a complete inversion of GC-skew for each of the genes. **c** Some taxa underwent genomic rearrangement, so that all genes are on the same strand

genomic data are available for a relatively large number of arthropods allowing the first attempts at performing comparative genomic analyses of the Arthropoda (Vieira and Rozas 2011). However, the majority of the currently available arthropod genomes are from closely related species (mostly insects), and a coherent set of conclusions about the arthropod nuclear genomes (as presented for the mitochondrial genomes above) is still lacking.

### 3.3.1 Uneven Taxonomic Sampling

The biased taxonomic distribution of the available arthropod genomes is a persistent problem. This is because it does not allow detailed investigations into key questions in arthropod evolution, like the origin of the arthropod subphyla. Initiatives exist that aim at increasing the amount of available genomic information for the Arthropoda. Paramount among these projects are the 1KITE project—1,000 Insects Transcriptome Evolution project (<http://1kite.org/>), and the i5K (<http://www.arthropodgenomes.org/wiki/i5K>) project which plans to sequence the complete genomes of 5,000 insects and related arthropod species. Unfortunately, as commendable as these projects are, they fall short of adequately capturing the breadth of the evolutionary diversity within the Arthropoda. The 1KITE project will not even attempt to generate data for non-hexapod species, while about 87 % of the species currently nominated for sequencing as part of the i5K project are hexapods. Only 0.7 % belongs to Myriapoda and only 2.8 % to Crustacea. This is an important issue with the current initiatives, as this heterogeneous species sampling, even if reflective of species diversity, does not reflect arthropod disparity. As such, it might bias future comparative analyses and might not allow a clear understanding of the genomic factors underlying the great morphological and physiological variation observed in Arthropoda. Disparity (e.g. the morphological diversities observed between a tick and a millipede) is underlined by variation in the genomes of the considered organisms, and the way these

genomes are wired. To understand arthropod disparity, therefore, genomic data as well as protein–protein interaction networks (e.g. Giot et al. 2003) and gene regulatory networks (Davidson and Erwin 2006) would be necessary for representatives of each major lineage within each subphylum. Even though hundreds of insect genomes will be a welcomed resource, it can be expected that, while they will allow to a significant increase in our understanding of adaptations, they will not be particularly useful to explain the origin of arthropod disparity, of the arthropod subphyla and of the main lineages within these subphyla.

An important aspect to which current large-scale genome sequencing projects are not given sufficient attention is that of the arthropod outgroups. To increase the power of comparative analyses, adequate outgroups should also be sequenced, but large-scale sampling initiatives are not considering the outgroups of the Arthropoda. Indeed, to date, the only arthropod outgroups available with at the least one fully sequenced genome are the nematodes. Yet, species belonging to this phylum are too distantly related and too divergent from the Arthropoda (see also above) to be of significant utility in arthropod comparative genomics. Other more closely related genomes (those of the Onychophora and the Tardigrada) should be sequenced and used instead. As part of this chapter, to obviate the lack of genomic-scale data sets for the arthropod outgroups, we shall present a genome-wide transcriptomic data set obtained using next generation sequencing.

The 1KITE and i5K projects have not produced data yet. However, a relative abundance of arthropod genomes has been accumulating in recent years, albeit with a biased taxonomic distribution. The genomes of 21 *Drosophila* species have been sequenced and made publicly available. Transcriptomic, proteomic, and genomic data, as well as abundant functional annotations, for 12 of these species can be found in the specialised database Flybase (<http://flybase.org/>). Other key insects for which genomic information is available include the mosquitoes *Aedes aegypti* (Nene et al. 2007) and *Anopheles gambiae* (Holt

et al. 2002), the honeybee *A. mellifera* (The honeybee genome consortium 2006), the beetle *Tribolium castaneum* (Richards et al. 2008), the body louse *Pediculus humanus* (Kirkness et al. 2010), the pea aphid *Acyrtosiphon pisum* (The pea aphid genome consortium 2010), and the silk moth *Bombyx mori* (The silkworm genome consortium 2008). A variety of other insects, for example, ants and other butterflies, have also been sequenced (Suen et al. 2011; The Heliconius genome consortium 2012). Results from these more recent studies (which generally used next generation sequencing strategies) allowed some truly surprising conclusions to be reached. For example, the *Heliconius* genome consortium was able to demonstrate the repeated exchange of large (~100-kb) adaptive regions among multiple butterfly species in a recent radiation. In this way, they were also able to uncover the pervasiveness and importance of introgressive adaptation and its role in hybrid speciation. For many of these more recently sequenced species, taxon-specific databases exist (e.g. Butterflybase—<http://butterflybase.ice.mpg.de/>). Differently from Flybase, which is a mature database providing, for example, a genome browser, and allowing complex searches (using Gene Ontology—GO terms and developmental stages), most of these species-specific databases are still quite immature. In any case, they represent an important resource and their utility is bound to increase with time.

While hexapod genomes are relatively abundant, the situation changes drastically when moving to other arthropod subphyla. Only one complete crustacean genome (that of the water flea *Daphnia pulex*—Colbourne et al. 2011), and one complete chelicerate genome, that of the two-spotted spider mite *Tetranychus urticae* (Grbic et al. 2011) have been released. Finally, the complete genome of one myriapod, the centipede *Strigamia maritima* (GenBank access id: GCA\_000239455.1), and that of a second chelicerate *Ixodes scapularis* (GenBank access id: GCA\_000208615.1) are now publicly available, although they have not yet been released.

Apart from standard genomic studies, a variety of large-scale transcriptome-wide

sequencing studies have been performed, and EST data are thus available for other taxa. Even though these studies do not provide information about untranslated genomic regions, a large amount of useful data has been provided using these approaches. One of the earliest studies that employed EST generated using next generation sequencing (in that specific case it was 454 sequencing) to gain a complete snapshot of an arthropod genome was the transcriptome sequencing of the emperor scorpion *Pandinus imperator* (Roeding et al. 2009). More recently, Illumina and other sequencing techniques have been applied to other important groups for which genomic data are not available, like the harvestmen (Opiliones; Hedin et al. 2012), and the amphipod crustacean *Parhyale hawaiiensis* (Zeng et al. 2011; Blythe et al. 2012). Similar approaches have started to generate extremely interesting insights into chelicerate venoms, allowing the development of the new science of venomics (Rendon-Anaya et al. 2012) and arthropod developmental biology (Ewen-Campen et al. 2011).

### 3.3.2 Heterogeneity of Genome Sizes and Shortage of microRNA

Important aspects of the key, publicly available, arthropod genomes are reported in Table 3.1. From this table, it is clear that the arthropod genomes are fairly variable. Their lengths in MB vary substantially with one of the chelicerate genomes being the smallest, while the other is the biggest overall. Similarly, GC content is quite variable with *Ixodes* having the highest GC content and the pea aphid the lowest. Also, the number of predicted protein coding genes varies substantially between genomes, with *Daphnia* having 30,613 and *Ixodes* only 7,112. A notable aspect of Table 3.1 is the difference in the number of protein coding genes and known, corresponding transcripts, for *D. melanogaster*. The fruit fly is the only species in Table 3.1 for which the number of known transcripts largely exceeds the number of predicted protein coding genes. The difference between the number of

genes and the number of transcripts is most likely caused by alternative splicing. It is in fact known that approximately 40 % of the protein coding genes in *D. melanogaster* correspond to more than one transcript (Hartmann et al. 2009). The lack of knowledge of alternatively spliced genes for other taxa in Table 3.1 is likely to reflect our ignorance rather than biology. For *D. melanogaster*, deep sequencing of specimens in specific developmental stages, specific tissues, and organs allowed identification of a larger number of transcripts. It is to be expected that as knowledge of the transcriptomes of the other species in Table 3.1 will increase, the number of their known transcripts will also increase. An obvious observation emerging from an analysis of Table 3.1 is that the sequenced chelicerate taxa cannot be particularly good resources for evolutionary biologists. *Ixodes* and *Tetranychus* are highly specialised species unlikely to reflect what the analysis of more standard chelicerate genomes will uncover.

Next generation sequencing approaches have also allowed our understanding of regulatory (non-coding) microRNA to increase substantially. Genome-wide screening performed for taxa belonging to all arthropod subphyla and to the arthropod outgroups (Campbell et al. 2011; Rota-Stabelli et al. 2011) allowed identification of several arthropod-specific microRNA (miR-275 and iab-4), mandibulate-specific ones (miR-965 and miR-282), and chelicerate-specific ones (miR-3931). These studies also showed that arthropods, in contrast to other lineages (such as the mammals or annelids), have significantly less lineage specific microRNAs, suggesting that arthropod genomes, from this point of view, evolve quite differently from those of other animal lineages.

Overall, current genomic-scale information available across the Arthropoda is still too fragmentary to allow the development of a coherent view of arthropod genome evolution. However, in the last section of this chapter, we shall attempt to start obviating this problem, by presenting an evolutionary analysis of the arthropod proteomes that exploits the transcriptomic data we generated for the Onychophora.

### 3.4 A Genomic Phylostratigraphic Analysis of the Arthropod Proteomes

An interesting aspect of the arthropod genome evolution that availability of current metazoan and arthropod genomes allows us to address (given also the data we generated for the Onychophora) is that of the origin of the arthropod-specific protein coding genes (i.e. genes found only within Arthropoda). Studies of this type have been named *genomic phylostratigraphic analyses* by Domazet-Loso et al. (2007). To complete such studies (in addition to genomic information), one needs information about phylogeny and divergence times. The relationships between the arthropods and divergence times used are summarised below.

#### 3.4.1 A Robust Phylogenetic Framework for Genomic Studies

Comparative genomics must be anchored on a phylogenetic tree. Significant progress in our understanding of the ecdysozoan relationships has been made (Dunn et al. 2008; Hejnol et al. 2009; Campbell et al. 2011). Similarly, some agreement on the phylogenetic relationships within the Arthropoda has recently emerged (Regier et al. 2010; Rota-Stabelli et al. 2011), but see Rota-Stabelli et al. (2012). For this study, it is important that the tree used to anchor our analyses is resolved. However, some level of incongruence still exists among the various phylogenetic studies addressing the relationships within Ecdysozoa. With reference to the current study, we shall consider the Lobopodia (Arthropoda plus Onychophora) to be the sister group of the Tardigrada within a monophyletic Panarthropoda. We shall further assume Nematoida (Nematoda plus Nematomorpha) to be the sister group of Panarthropoda, with the Scalidophora (here Priapulida and Kinorhyncha) representing the sister group of Nematoida plus Panarthropoda. That is, we shall assume the ecdysozoan relationships of Campbell et al. (2011) and Rota-Stabelli et al. (2011) to represent our working hypothesis. These

relationships differ from those of Dunn et al. (2008) with reference to the placement of Nematoida that the study of Dunn and co-workers was found to be a member of Cycloneuralia, that is, more closely related to the Scalidophora than to the Arthropoda. However, because Campbell et al. (2011) only performed a Bayesian analysis of their data set and did not present bootstrap support for their results. Given that they did not find particularly strong support (low posterior probabilities) for some key contested nodes (Nematoida + Panarthropoda and Mandibulata—which are not supported in other studies, for example, Dunn et al. 2008), and given that there are few other studies (e.g. Meusemann et al. 2010) whose results contradict those of Campbell et al. (2011) and Rota-Stabelli et al. (2011) with reference to the placement of Tardigrada and the monophyly of Mandibulata, we present here a novel statistical analysis—nonparametric bootstrapping—of the data set used in Campbell et al. (2011). A detailed explanation of the methods used in this analysis is presented in the Appendix to this chapter.

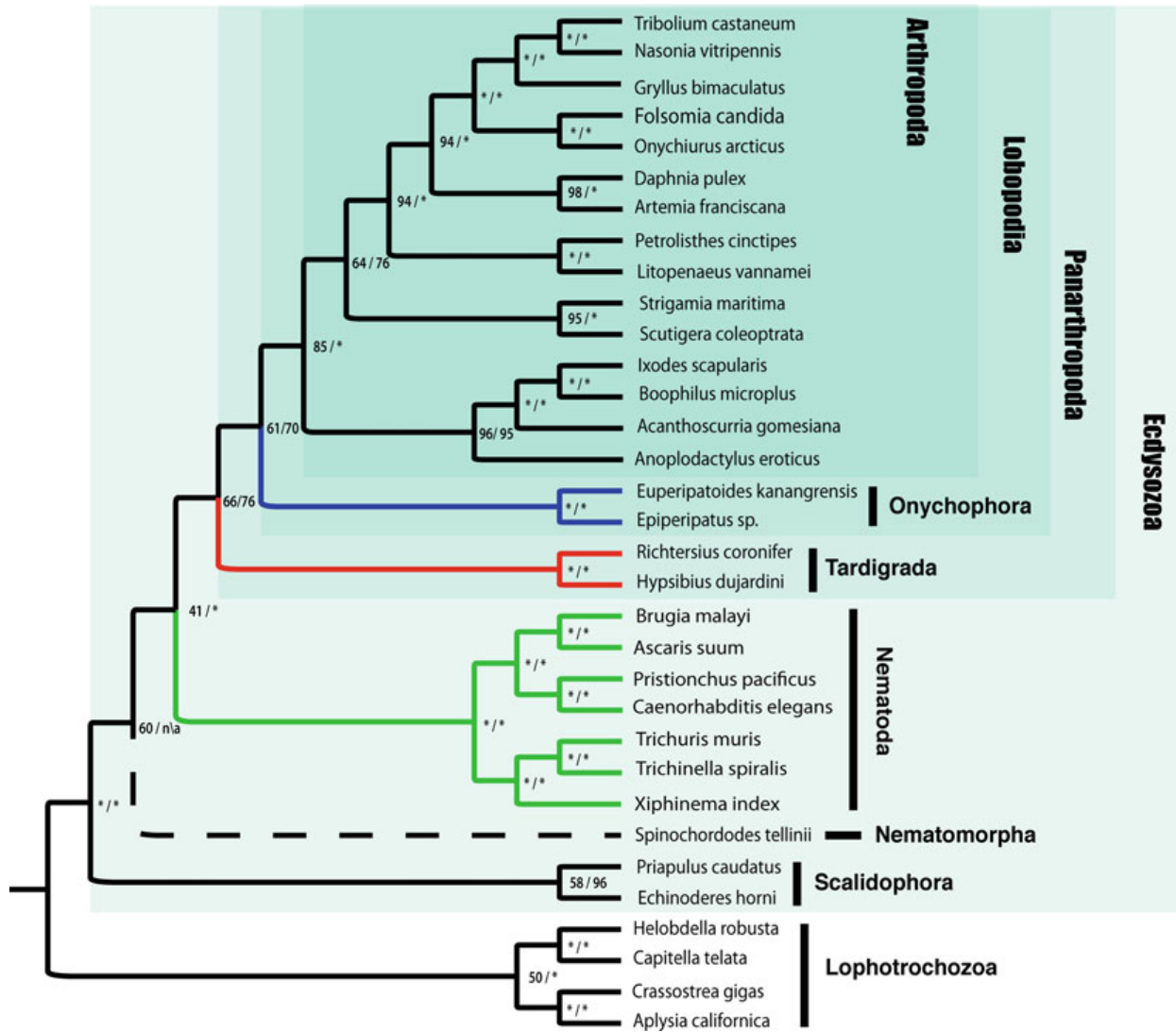
Results of the bootstrap analysis that considers all the taxa in Campbell et al. (2011) are in agreement with the Bayesian analyses in that paper. This analysis shows a lack of support for many important nodes, including Nematoida (which was not recovered), Nematoda plus Panarthropoda (BP = 41), Panarthropoda (BP = 66), Lobopodia (BP = 61), and Mandibulata (BP = 64), see Fig. 3.3. We performed a leaf stability analysis (results not shown—but see Appendix) illustrating that Nematomorpha is the most unstable taxon in the data set. The nematomorph in Campbell et al. (2011) emerged as the sister group of the Nematoda in agreement with Dunn et al. (2008) and Hejnol et al. (2009). Yet, in Fig. 3.3, Nematomorpha is not the sister group of the Nematoda. Instead, it emerges as the sister of a Nematoda + Arthropoda clade. This is an artefact caused by high volume of missing data in the Nematomorpha (which is the most incomplete taxon in Campbell et al. 2011) and that is unstable in bootstrapped data sets.

Upon removal of the unstable Nematomorpha, the bootstrap support for all the other nodes increases significantly. Arthropoda plus Nematoda reaches 100 %, Panarthropoda increases to 76 %, and Lobopodia to 70 %. In conclusion, when accounting for unstable taxa, Arthropoda has a bootstrap support of 100 % and Mandibulata of 76 %. This confirms that there is a good level of support for the clades in Fig. 3.3 and those in Campbell et al. (2011).

### 3.4.2 Expanding Our Understanding of the Arthropod Comparative Genomics

Given our poor understanding of the processes through which the arthropod (nuclear) genomes evolved, we shall here present a genomic phylostratigraphic analysis (Domazet-Loso et al. 2007) of their genome. The aim of this analysis is to gain some information on the evolutionary processes responsible for the origin and evolution of the Arthropoda. Domazet-Loso et al. (2007) performed a similar analysis, but various new genomes have been published since their study, allowing for a much greater precision in the identification of orphan genes along the ecdysozoan and arthropod phylogeny. To better identify proteins that are arthropod specific, we extended our analyses to include a variety of ecdysozoans and non-ecdysozoan genomes. Particularly, we included representatives of the Lophotrochozoa, of the Deuterostomia and two non-bilaterian metazoans (a sponge, *Amphimedon queenslandica*, and a cnidarian, *Hydra magnipapillata*)—see Fig. 3.4. In addition, and most importantly, here we added data for an onychophoran transcriptome, which allowed pinpointing protein families that are specific to the Arthropoda (i.e. that originated after the Onychophora–Arthropoda split). Finally, more reliable molecular clock divergence times (Erwin et al. 2011) are now available and they have been used here to define rates of orphan gene acquisitions through time allowing for





**Fig. 3.3** A phylogeny of the Ecdysozoa. The tree represents a Bayesian bootstrap analysis performed under CAT+G of the data set of Campbell et al. (2011). Values at the nodes represent bootstrap proportions. *Asterisk* = 100 % support. The leftmost value represents the

bootstrap proportion obtained for a data set including all the sequences in Campbell et al. (2011). The rightmost value represents the bootstrap proportion obtained when the most unstable taxon in the data set (the nematomorph *Spiniochordodes*) was excluded

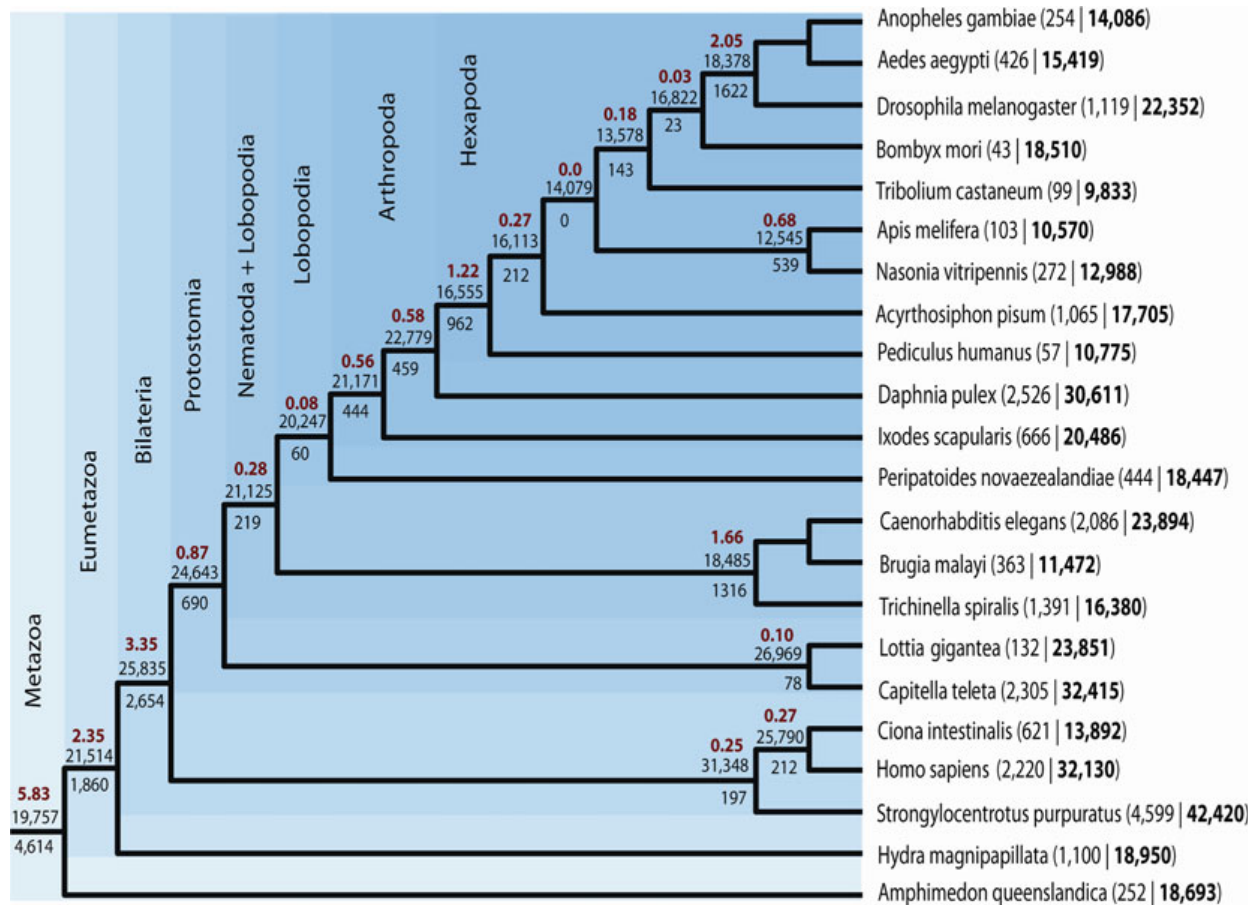
better estimation of rates of new protein family acquisitions in Ecdysozoa and Arthropoda.

### 3.4.3 The Evolution of Orphan Gene Families in Arthropoda

We used the MCL algorithm (Enright et al. 2002) to identify protein families in the set of considered genomes, and identified, for each internal node in Fig. 3.4, all the proteins universally distributed in the taxa descending from each given node. These are orphan families that evolved in the branch underlying the considered

node. The average number of new families acquired across all the internodes of the considered phylogeny is 1,025. When this value is normalised (dividing by the total number of proteins in the considered set of genomes (79,052 protein coding genes)), the 1,025 protein families that are gained as novel orphan genes correspond to  $\sim 1.2\%$ .

Within Arthropoda, and more broadly Panarthropoda, only the origin of the Diptera (with 2.05 % of new protein families being acquired) shows a statistically significantly higher rate of novel gene families acquisition (Figs. 3.4 and 3.5). Genomic data were not available for the



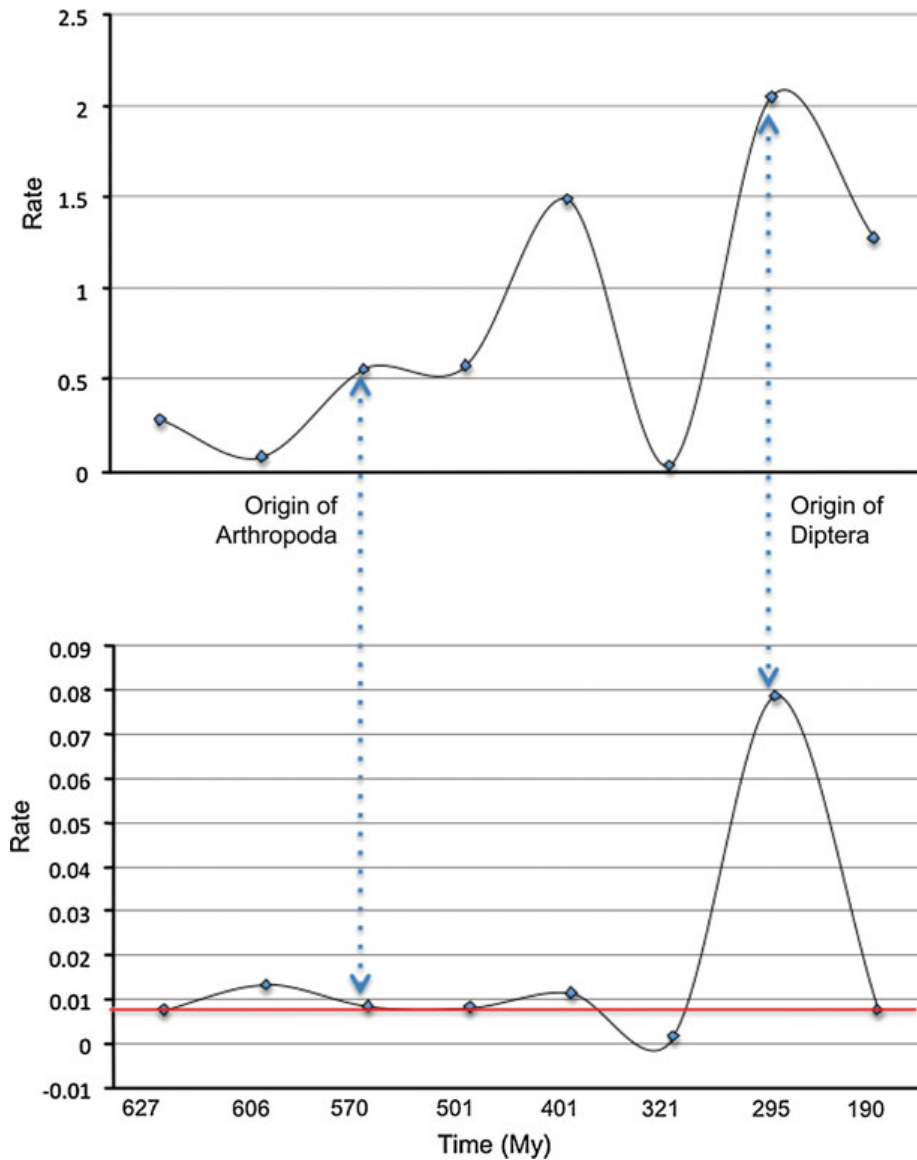
**Fig. 3.4** Orphan protein gains in Arthropoda. The number below each node quantifies the orphan families that evolved along the branch subtending the considered node. The number in black above each node represents the number of protein coding genes inferred to have existed (using squared parsimony) in the common ancestor represented by the considered node. The red value above the node represents the rate of orphan gene acquisition along the branch subtending the considered node. These values are normalised (calculated as the number of orphans divided by the total number of proteins in the collection of considered proteomes). The

numbers reported for each terminal taxon are the number of orphan families that originated along the terminal branch and the number of genes in the genome of the corresponding organism (*in bold*). Note that the numbers of orphans for the terminal taxa are misleading and should not be considered to represent the number of new genes that emerged in the species at the tip of the tree. Instead, they represent the number of orphans in the group the species represent. For example, the number of orphans in *Hydra* represents the orphans that were acquired by the Cnidaria (to which *Hydra* belong and that *Hydra* represents) rather than by *Hydra* itself

Myriapoda when we assembled our data set, but it is clear, given the low level of proteins that originated in the branch separating Arthropoda and Pancrustacea (1.49 %) that also the origin of Mandibulata cannot be marked by a spike in the origin of new protein families (Figs. 3.4 and 3.5).

The most surprising result emerging from this analysis is that the deepest nodes in the Ecdysozoan phylogeny (origin of Nematoda plus Arthropoda, origin of Lobopodia, and origin of Arthropoda) are not characterised by above average acquisitions of new gene families

(Fig. 3.5). When the number of orphan families (N-orph) acquired along a branch is divided by the length (in millions of years) of the branch along which the N-orph accumulated, the pattern in Fig. 3.5a changes quite significantly: even the mild, but somewhat continuous, increment in the rate of N-orph acquisition disappears (Fig. 3.5b). All internodes within Ecdysozoa (on the path leading to Arthropoda and within Arthropoda) roughly exhibit the same rate of new protein acquisition per million of years. Constancy of the rate of protein family



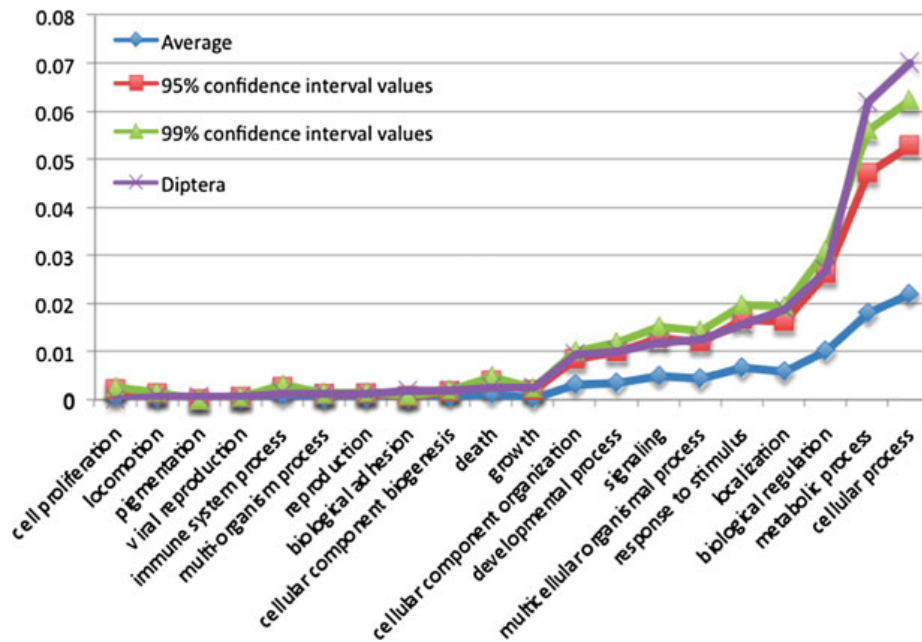
**Fig. 3.5** Protein gains through time. **a** Normalised rates of orphan acquisition (*red values* in Fig. 3.4). This panel illustrates that the normalised rates are quite variable across all the considered nodes. Note that the values were ordered from oldest to youngest to make the figure more readable. **b** Rates of orphan acquisition per millions of years. This chart was derived dividing the values in Fig. 3.5a by the length (in million years) of the branch along which the considered orphans originated. This

figure clearly illustrates how the raw rates and the rates per million years are substantially different, and that normalising for the time of duration of the considered internodes is key to obtain values that are biologically meaningful. The *red line* represents the average rate across the considered lineages (but excluding the Diptera). This was done to estimate the average rate of orphan protein acquisition (i.e. the neutral rate)

acquisition through time (from the Precambrian to the Jurassic—see Fig. 3.5b) suggests that this rate (identified with a red line in Fig. 3.5b) might represent the neutral background rate of new protein family origination in Ecdysozoa. The only internode where this neutral rate is modified is represented by the stem dipteran lineage. Along this lineage (Fig. 3.5b), the rate is significantly increased, suggesting that orphan

gene family acquisition was an important phenomenon in the evolution of this group.

A functional analysis of the orphan proteins that originated along the stem dipteran lineage (see Appendix for methodological details) provides a view of what kind of gene families are acquired along this branch (Fig. 3.6). When comparing the average trend estimated across all the considered stem lineages but the dipteran,



**Fig. 3.6** The function of the newly acquired families. This graph displays the average number of orphans (across all the internodes but the Diptera) for each GO (Gene Ontology) category. We also reported the values representing, respectively, the limits of the 95 and 99 % confidence intervals. Values observed for the dipteran

stem lineage are reported. This figure shows that for two GO categories, the number of orphans acquired in Diptera is higher than the value bounding the 99 % confidence interval over all the other internodes, and that various other GO categories are overrepresented with reference to the other internal branches considered

with the trend observed in the dipteran, two conclusions can be reached. The first is that the trends observed are comparable in shape (i.e. there is a proportionality in the number of new genes acquired on average across the Arthropoda and specifically in Diptera). The second is that when the numbers of genes in each Gene Ontology (GO) category is analysed, it is clear that for two GO terms (metabolic processes and cellular processes), the increase observed in Diptera is significantly higher (greater than the limiting values of a 99 % confidence interval calculated across all the other internodes; Fig. 3.6). A further significantly increased category (exceeding the 95 % confidence interval calculated across all the other non-dipteran internodes) is the localisation proteins category. Finally, other GO categories for which new proteins are accumulated in Diptera to levels that are above average (but not significantly so) are as follows: biological regulation, response to stimulus, multicellular organismal processes, signalling, developmental processes, and cellular component organisation.

### 3.4.4 Conserved Rate of Gene Gain with Some Surprises

It is fairly obvious from the above results that, at least within Ecdysozoa, the origin of new protein families (orphan gene accumulation) did not play a particularly significant role in the evolution of what we recognise as high-level taxonomic groups (phyla and assemblages of phyla). In particular, we have shown here that the origin of the arthropod body plan was not characterised by an unusual rate of new protein families acquisition. One can thus argue that other processes, like the re-wiring of developmental networks (and more generally protein-protein interaction networks), might have been much more important (see also Erwin et al. 2011). Yet, these hypotheses need to be tested and will be tested in the future when more data become available.

On the other hand, the origin of the Diptera is marked by a substantial increase in the origin of orphan families. This is interesting because it suggests that (1) if increases in rate existed somewhere else in the ecdysozoan tree, we

should have been able to identify them (i.e. our results do not seem to represent a methodological artefact), and (2) orphan gene acquisition is not always an unimportant process in animal evolution: hence, the need to investigate it. With reference to the Diptera, it is clear that the strong acceleration in rate of acquisition of new families observed implies that new functionalities emerged in this part of the ecdysozoan tree, and it is clear that these protein families played a role in the origin of this group. Our current GO analyses did not allow us to obtain a detailed description of what the newly acquired dipteran functions are. However, as more precise functional annotations will become available, it will become possible to pinpoint the functions of the orphan genes originating along the dipteran branch much more precisely.

One can only conjecture, given also the unimpressive amount of orphan families being fixed on the stem lineage of the Holometabola, that the origin of key innovations affecting the emergence of novel life cycles or substantially modified morphological features is generally fuelled by re-wiring of the developmental networks and by differential expressions of genes, while origin of novel protein families probably has a greater impact on adaptations to novel environmental challenges.

---

### 3.5 Conclusions

Here, we have tried to summarise mitogenomic and nuclear genomic information currently available for the Arthropoda. There are a large number of mitochondrial genomes available to date, but it is unclear if something that will be of any utility will be gained from the analyses of these genomes. They might have some limited utility in phylogenetics compositional bias studies, and DNA barcoding, but probably not much utility in understanding large-scale evolutionary patterns in Arthropoda.

Arthropod genomics, on the other end, is still in its infancy, very few genomes are available at this stage but within five years, we will probably have thousands of genes available (particularly

thanks to large-scale efforts like the i5k). One wonders what will be gained from having so many genomes. Perhaps a lot, but their biased taxonomic distribution might prove to be a limitation of these data sets. Data analysis will be prohibitively complex, and serious bioinformatic resources will be necessary for these data to be of any utility. In any case, the initial analysis we present in this chapter suggests that, if adequate bioinformatic resources are available, a multitude of arthropod genomes will allow us to gain detailed information on the origin and evolution of this important phylum. Yet, sequencing projects should not forget that arthropod outgroups are necessary and important to increase the power of comparative analyses.

No matter what the future will hold, it is clear that arthropod comparative genomics is still in its infancy. We are just at the dawn of what will be a laborious and complex research task which will involve the continuous effort of many research groups, from all around the world for, probably, several research cycles.

**Acknowledgments** We would like to thank Alessandro Minelli for inviting us to contribute a chapter to this book and for the patience demonstrated during the editing process. DP and RC are supported by a Science Foundation Ireland Research Frontiers Programme (SFI-RFP) grant SFI-RFP 11/RFP/EOB/3106. ORS by a Marie Curie-Trento Province COFUND Fellowship. WAA by an IRCSET PhD studentship.

---

## Appendix: Methods for the Analyses Presented in this Chapter

### A. Generation of the Onychophoran Transcriptome

Total RNA was extracted from three individuals of “*Peripatoides novaezealandiae* complex” (Trewick 1998), which were commercially purchased, using TriZol©. A transcriptome-wide cDNA library was generated and sequenced using two IlluminaHiSeqII lanes at TrinSeq (Trinity College Dublin, Institute of Molecular Medicine, Genome Sequencing Laboratory) to an estimated coverage of <100, using 100-bp paired end reads. Raw data were inspected for

its quality and assembled using Abyss (Simpson et al. 2009) with k-mer of 45. This resulted in ~27,000 assembled transcripts (with lengths variable between ~70 and 1,750 base pairs). Approximately 17,000 of these transcripts had a significant blast hit against an annotated gene, while ~5,000 hit a known gene of unknown function. This set of ~22,000 genes was used to investigate the origin of orphan genes in Arthropoda. However, the 5,000 non-annotated genes were not considered for the Blast2go analysis (see below).

## B. Mitogenomic Compositional Analyses

We downloaded a set of mitochondrial genomes of 90 arthropods in order to represent the whole phylum as homogeneously as possible. Coding genes were extracted and processed with DAMBE (Xia and Xie 2001) to obtain composition for each codon position.

## C. Phylogenetic Analyses

We investigated whether the low posterior probabilities observed for some nodes by Campbell et al. (2011) were caused by the presence of unstable taxa. We estimated leaf stability indices (Thorley and Wilkinson 1999) using P4 (Foster 2004) and performed Bayesian bootstrap analysis under CAT+G—the same model used by Campbell et al. (2011)—using the entire data set of Campbell et al. (2011). To perform the Bayesian bootstrap analyses, 100 bootstrapped data sets were generated starting from the alignment of Campbell et al. (2011). For each bootstrapped data set, a Bayesian analysis (2 independent runs) was performed under CAT+G (using Phylobayes; Lartillot et al. 2009). Results from each Bayesian analysis were summarised to generate a Bayesian majority rule consensus tree, and the resulting 100 trees were then summarised to generate a bootstrap majority rule consensus (results in Fig. 3.3).

## Identification of Novel Gene Families

We downloaded the entire proteomes for the taxa in Fig. 3.4 and used MCL (Enright et al. 2002) to define protein families. A Perl script written by LC was used to partition these gene families with reference to their taxon coverage. This allowed the identification of protein families that are exclusive and universally distributed within each one of the clades in Fig. 3.4. These protein families must have been present in the clade's last common ancestor (LCA) and must have been gained along the stem lineage of the considered clade. Because different genomes have different numbers of protein coding genes, the absolute numbers of newly acquired protein coding families for each internode can be misleading. We thus normalised numbers of orphan families by dividing these numbers by the total number of protein coding genes in the set of considered genomes (sum of the values in bold at the tips of Fig. 3.4). The normalised orphan counts (N-orph) can be interpreted as the fraction of some, abstract, pan-metazoan genome that was acquired at each internode of Fig. 3.4. Finally, we calculated rates of new orphan acquisition per million of years, dividing the N-orph values by the length of the internode along which the N-orph was acquired. As above, this allows the amount of orphan families gained each million year, along each internode in Fig. 3.4, to be expressed as proportions of a reference (abstract) “pan-metazoan” genome. The estimates of divergence times of Erwin et al. (2011) were used to calculate branch durations in million of years. For each internal node in our phylogeny, we also estimated (using squared parsimony—as implemented in Mesquite—<http://mesquiteproject.org>) the expected size of the genome of the corresponding LCA. This was done to allow evaluation of what proportion of each LCA genome was gained via new orphan family acquisition, along the corresponding stem lineage. Because squared parsimony is unlikely to be a particularly robust estimator of ancestral size, we suggest these numbers should be considered with caution, and only to represent a

rough approximation of the true LCA-genomes dimensions.

Once the orphan gene families were identified for every internode of Fig. 3.4, BLAST2Go ([www.blast2go.com](http://www.blast2go.com)) was used to obtain functional information for each of these families. For each protein family, the BLAST2Go analysis was performed for one protein family member only, and we assumed, by homology implication that all the other proteins in the same orphan family had the same (or similar) function.

## References

- Abascal F, Posada D, Knight RD, Zardoya R (2006) Parallel evolution of the genetic code in arthropod mitochondrial genomes. *Plos Biol* 4:e127. doi: [10.1371/journal.pbio.0030127](https://doi.org/10.1371/journal.pbio.0030127)
- Abascal F, Posada D, Zardoya R (2007) MtArt: a new model of amino acid replacement for Arthropoda. *Mol Biol Evol* 24:1–5
- Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, rews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferreira S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C et al (2000) The genome sequence of *Drosophila melanogaster*. *Science* 287:2185–2195
- Aguinaldo AMA, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA (1997) Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* 387:489–493
- Belinky F, Cohen O, Huchon D (2010) Large-scale parsimony analysis of metazoan indels in protein-coding genes. *Mol Biol Evol* 27:441–451
- Bernt M, Braband A, Middendorf M, Misof B, Rota-Stabelli O, Stadler PF (2012) Bioinformatics methods for the comparative analysis of metazoan mitochondrial genome sequences. *Molecular Phylogenetics and Evolution* published on-line ahead of press. doi: [10.1016/j.ympev.2012.09.019](https://doi.org/10.1016/j.ympev.2012.09.019) [to be replaced by volume: page range on publication]
- Blanquart S, Lartillot N (2008) A site- and time-heterogeneous model of amino acid replacement. *Mol Biol Evol* 25:842–858
- Blythe MJ, Malla S, Everall R, Shih YH, Lemay V, Moreton J, Wilson R, Aboobaker AA (2012) High through-put sequencing of the *Parhyale hawaiiensis* mRNAs and microRNAs to aid comparative developmental studies. *Plos One* 7(3):e33784. doi: [10.1371/journal.pone.0033784](https://doi.org/10.1371/journal.pone.0033784)
- Boore JL (1999) Animal mitochondrial genomes. *Nucleic Acids Res* 27:1767–1780
- Boore JL, Lavrov DV, Brown WM (1998) Gene translocation links insects and crustaceans. *Nature* 392:667–668
- Campbell LI, Rota-Stabelli O, Edgecombe GD, Marchioro T, Longhorn SJ, Telford MJ, Philippe H, Rebecchi L, Peterson KJ, Pisani D (2011) MicroRNAs and phylogenomics resolve the relationships of Tardigrada and suggest that velvet worms are the sister group of Arthropoda. *Proc Natl Acad Sci USA* 108:15920–15924
- Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, Tokishita S, Aerts A, Arnold GJ, Basu MK, Bauer DJ, Caceres CE, Carmel L, Casola C, Choi JH, Detter JC, Dong Q, Dusheyko S, Eads BD, Frohlich T, Geiler-Samerotte KA, Gerlach D, Hatcher P, Jogdeo S, Krijgsveld J, Kriventseva EV, Kultz D, Laforsch C, Lindquist E, Lopez J, Manak JR, Muller J, Pangilinan J, Patwardhan RP, Pitluck S, Pritham EJ, Rechtsteiner A, Rho M, Rogozin IB, Sakarya O, Salamov A, Schaack S, Shapiro H, Shiga Y, Skalitzky C, Smith Z, Souvorov A, Sung W, Tang Z, Tsuchiya D, Tu H, Vos H, Wang M, Wolf YI, Yamagata H, Yamada T, Ye Y, Shaw JR, Rews J, Crease TJ, Tang H, Lucas SM, Robertson HM, Bork P, Koonin EV, Zdobnov EM, Grigoriev IV, Lynch M, Boore JL (2011) The ecoresponsive genome of *Daphnia pulex*. *Science* 331:555–561
- Copley RR, Aloy P, Russell RB, Telford MJ (2004) Systematic searches for molecular synapomorphies in model metazoan genomes give some support for Ecdysozoa after accounting for the idiosyncrasies of *Caenorhabditis elegans*. *Evol Dev* 6:164–169
- Davidson EH, Erwin DH (2006) Gene regulatory networks and the evolution of animal body plans. *Science* 311:796–800
- Domazet-Loso T, Brajkovic J, Tautz D (2007) A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet* 23:533–539
- Dopazo H, Dopazo J (2005) Genome-scale evidence of the nematode-arthropod clade. *Genome Biol* 6:R41. doi:[10.1186/gb-2005-6-5-r41](https://doi.org/10.1186/gb-2005-6-5-r41)
- Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe

- GD, Sørensen MV, Haddock SHD, Schmidt-Rhaesa A, Okusu A, Kristensen RM, Wheeler WC, Martindale MQ, Giribet G (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* 452:745–749
- Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* 30:1575–1584
- Erwin DH, Laflamme M, Tweedt SM, Sperling EA, Pisani D, Peterson KJ (2011) The Cambrian conundrum: early divergence and later ecological success in the early history of animals. *Science* 334:1091–1097
- Ewen-Campen B, Shaner N, Panfilio KA, Suzuki Y, Roth S, Extavour CG (2011) The maternal and early embryonic transcriptome of the milkweed bug *Onco-peltus fasciatus*. *BMC Genomics* 12:61. doi: [10.1186/1471-2164-12-61](https://doi.org/10.1186/1471-2164-12-61)
- Foster PG (2004) Modeling compositional heterogeneity. *Syst Biol* 53:485–495
- Foster PG, Hickey DA (1999) Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *J Mol Evol* 48:284–290
- Foster PG, Jermini LS, Hickey DA (1997) Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. *J Mol Evol* 44:282–288
- Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, Vijayadamar G, Pochart P, Machineni H, Welsh M, Kong Y, Zerhusen B, Malcolm R, Varrone Z, Collis A, Minto M, Burgess S, McDaniel L, Stimpson E, Spriggs F, Williams J, Neurath K, Ioime N, Agee M, Voss E, Furtak K, Renzulli R, Aanensen N, Carrolla S, Bickelhaupt E, Lazovatsky Y, DaSilva A, Zhong J, Stanyon CA, Finley RL Jr, White KP, Braverman M, Jarvie T, Gold S, Leach M, Knight J, Shimkets RA, McKenna MP, Chant J, Rothberg JM (2003) A protein interaction map of *Drosophila melanogaster*. *Science* 302:1727–1736
- Grbic M, Van Leeuwen T, Clark RM, Rombauts S, Rouze P, Grbic V, Osborne EJ, Dermauw W, Ngoc PC, Ortego F, Hernandez-Crespo P, Diaz I, Martinez M, Navajas M, Sucena E, Magalhaes S, Nagy L, Pace RM, Djuranovic S, Smagghe G, Iga M, Christiaens O, Veenstra JA, Ewer J, Villalobos RM, Hutter JL, Hudson SD, Velez M, Yi SV, Zeng J, Pires-daSilva A, Roch F, Cazaux M, Navarro M, Zhurov V, Acevedo G, Bjelica A, Fawcett JA, Bonnet E, Martens C, Baele G, Wissler L, Sanchez-Rodriguez A, Tirry L, Blais C, Demeestere K, Henz SR, Gregory TR, Mathieu J, Verdon L, Farinelli L, Schmutz J, Lindquist E, Feyereisen R, Van de Peer Y (2011) The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. *Nature* 479:487–492
- Hartmann B, Castelo R, Blanchette M, Boue S, Rio DC, Valcarcel J (2009) Global analysis of alternative splicing regulation by insulin and wingless signalling in *Drosophila* cells. *Genome Biol* 10:R11. doi: [10.1186/gb-2009-10-1-r11](https://doi.org/10.1186/gb-2009-10-1-r11)
- Hassanin A, Leger N, Deutsch J (2005) Evidence for multiple reversals of asymmetric mutational constraints during the evolution of the mitochondrial genome of metazoa, consequences for phylogenetic inferences. *Syst Biol* 54:277–298
- Hedin M, Starrett J, Akhter S, Schönhofer AL, Shultz JW (2012) Phylogenomic resolution of Paleozoic divergences in harvestmen (Arachnida, Opiliones) via analysis of next-generation transcriptome data. *Plos One* 7(8):e42888. doi: [10.1371/journal.pone.0042888](https://doi.org/10.1371/journal.pone.0042888)
- Hejnal A, Obst M, Stamatakis A, Ott M, Rouse GW, Edgecombe GD, Martinez P, Bagnuà J, Bailly X, Jondelius U, Wiens M, Müller WEG, Seaver E, Wheeler WC, Martindale MQ, Giribet G, Dunn CW (2009) Assessing the root of Bilaterian animals with scalable phylogenomic methods. *Proc R Soc B* 276:4261–4270
- Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, Wincker P, Clark AG, Ribeiro JM, Wides R, Salzberg SL, Loftus B, Yandell M, Majoros WH, Rusch DB, Lai Z, Kraft CL, Abril JF, Anthouard V, Arensburger P, Atkinson PW, Baden H, de Berardinis V, Baldwin D, Benes V, Biedler J, Blass C, Bolanos R, Boscus D, Barnstead M, Cai S, Center A, Chaturverdi K, Christophides GK, Chrystal MA, Clamp M, Cravchik A, Curwen V, Dana A, Delcher A, Dew I, Evans CA, Flanigan M, Grundschober-Freimoser A, Friedli L, Gu Z, Guan P, Guigo R, Hillenmeyer ME, Hladun SL, Hogan JR, Hong YS, Hoover J, Jaillon O, Ke Z, Kodira C, Kokoza E, Koutsos A, Letunic I, Levitsky A, Liang Y, Lin JJ, Lobo NF, Lopez JR, Malek JA, McIntosh TC, Meister S, Miller J, Mobarry C, Mongin E, Murphy SD, O’Brochta DA, Pfannkoch C, Qi R, Regier MA, Remington K, Shao H, Sharakhova MV, Sitter CD, Shetty J, Smith TJ, Strong R, Sun J, Thomasova D, Ton LQ, Topalis P, Tu Z, Unger MF, Walenz B, Wang A, Wang J, Wang M, Wang X, Woodford KJ, Wortman JR, Wu M, Yao A, Zdobnov EM, Zhang H, Zhao Q et al (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* 298:129–149
- Holton TA, Pisani D (2010) Deep genomic-scale analyses of the metazoa reject Coelomata: evidence from single and multigene families analyzed under a supertree and supermatrix paradigm. *Genome Biol Evol* 2:310–324
- Hwang UW, Friedrich M, Tautz D, Park CJ, Kim W (2001) Mitochondrial protein phylogeny joins myriapods with chelicerates. *Nature* 413:154–157
- Irimia M, Maeso I, Penny D, Garcia-Fernandez J, Roy SW (2007) Rare coding sequence changes are consistent with Ecdysozoa, not Coelomata. *Mol Biol Evol* 24:1604–1607
- Kirkness EF, Haas BJ, Sun W, Braig HR, Perotti MA, Clark JM, Lee SH, Robertson HM, Kennedy RC, Elhaik E, Gerlach D, Kriventseva EV, Elsik CG, Graur D, Hill CA, Veenstra JA, Walenz B, Tubio JM, Ribeiro JM, Rozas J, Johnston JS, Reese JT, Popadic



- A, Tojo M, Raoult D, Reed DL, Tomoyasu Y, Kraus E, Mittapalli O, Margam M, Li HM, Meyer JM, Johnson RM, Romero-Severson J, Vanzee JP, Alvarez-Ponce D, Vieira FG, Aguade M, Guirao-Rico S, Anzola JM, Yoon KS, Strycharz JP, Unger MF, Christley S, Lobo NF, Seufferheld MJ, Wang N, Dasch GA, Struchiner CJ, Madey G, Hannick LI, Bidwell S, Joardar V, Caler E, Shao R, Barker SC, Cameron S, Bruggner RV, Regier A, Johnson J, Viswanathan L, Utterback TR, Sutton GG, Lawson D, Waterhouse RM, Venter JC, Strausberg RL, Berenbaum MR, Collins FH, Zdobnov EM, Pittendrigh BR (2010) Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proc Natl Acad Sci USA* 107:12168–12173
- Lartillot N, Lepage T, Blanquart S (2009) PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25:2286–2288
- Lavrov DV, Boore JL, Brown WM (2000) The complete mitochondrial DNA sequence of the horseshoe crab *Limulus polyphemus*. *Mol Biol Evol* 17:813–824
- Lavrov DV, Brown WM, Boore JL (2004) Phylogenetic position of the Pentastomida and (pan)crustacean relationships. *Proc Biol Sci* 271:537–544
- Meusemann K, von Reumont BM, Simon S, Roeding F, Strauss S, Kuck P, Ebersberger I, Walz M, Pass G, Breuers S, Achter V, von Haeseler A, Burmester T, Hadrys H, Wägele JW, Misof B (2010) A phylogenomic approach to resolve the arthropod tree of life. *Mol Biol Evol* 27:2451–2464
- Nabholz B, Ellegren H, Wolf JB (2012) High levels of gene expression explain the strong evolutionary constraint of mitochondrial protein-coding genes. *Molecular Biology and Evolution* published online ahead of press. doi: [10.1093/molbev/mss238](https://doi.org/10.1093/molbev/mss238) [to be replaced by volume: page range on publication]
- Negrisololo E, Minelli A, Valle G (2004) The mitochondrial genome of the house centipede *Scutigera* and the monophyly versus paraphyly of myriapods. *Mol Biol Evol* 21:770–780
- Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu ZJ, Loftus B, Xi Z, Megy K, Grabherr M, Ren Q, Zdobnov EM, Lobo NF, Campbell KS, Brown SE, Bonaldo MF, Zhu J, Sinkins SP, Hogenkamp DG, Amedeo P, Arensburger P, Atkinson PW, Bidwell S, Biedler J, Birney E, Bruggner RV, Costas J, Coy MR, Crabtree J, Crawford M, Debruyne B, Decaprio D, Eglmeier K, Eisenstadt E, El-Dorry H, Gelbart WM, Gomes SL, Hammond M, Hannick LI, Hogan JR, Holmes MH, Jaffe D, Johnston JS, Kennedy RC, Koo H, Kravitz S, Kriventseva EV, Kulp D, Labutti K, Lee E, Li S, Lovin DD, Mao C, Mauceli E, Menck CF, Miller JR, Montgomery P, Mori A, Nascimento AL, Naveira HF, Nusbaum C, O’Leary S, Orvis J, Pertea M, Quesneville H, Reidenbach KR, Rogers YH, Roth CW, Schneider JR, Schatz M, Shumway M, Stanke M, Stinson EO, Tubio JM, Vanzee JP, Verjovski-Almeida S, Werner D, White O, Wyder S, Zeng Q, Zhao Q, Zhao Y, Hill CA, Raikhel AS, Soares MB, Knudson DL, Lee NH, Galagan J, Salzberg SL, Paulsen IT, Dimopoulos G, Collins FH, Birren B, Fraser-Liggett CM, Severson DW (2007) Genome sequence of *Aedes aegypti*, a major arbovirus vector. *Science* 316:1718–1723
- Philippe H, Lartillot N, Brinkmann H (2005) Multi gene analyses of Bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Mol Biol Evol* 22:1246–1253
- Pisani D (2004) Identifying and removing fast-evolving sites using compatibility analysis: an example from the Arthropoda. *Syst Biol* 53:978–989
- Pisani D, Poling LL, Lyons-Weiler M, Hedges SB (2004) The colonization of land by animals: molecular phylogeny and divergence times among arthropods. *BMC Biol* 2:1. doi:[10.1186/1741-7007-2-1](https://doi.org/10.1186/1741-7007-2-1)
- Podsiadlowski L, Braband A (2006) The complete mitochondrial genome of the sea spider *Nymphon gracile* (Arthropoda: Pycnogonida). *BMC Genomics* 7:284
- Regier JC, Shultz JW, Zwick A, Hussey A, Ball B, Wetzer R, Martin JW, Cunningham CW (2010) Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature* 463:1079–1083
- Rendon-Anaya M, Delaye L, Possani LD, Herrera-Estrella A (2012) Global transcriptome analysis of the scorpion *Centruroides noxius*: new toxin families and evolutionary insights from an ancestral scorpion species. *Plos One* 7:e43331. doi:[10.1371/journal.pone.0043331](https://doi.org/10.1371/journal.pone.0043331)
- Reyes A, Gissi C, Pesole G, Saccone C (1998) Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Mol Biol Evol* 15:957–966
- Richards S, Gibbs RA, Weinstock GM, Brown SJ, Denell R, Beaman RW, Gibbs R, Bucher G, Friedrich M, Gimmelikhuijzen CJ, Klingler M, Lorenzen M, Roth S, Schroder R, Tautz D, Zdobnov EM, Muzny D, Attaway T, Bell S, Buhay CJ, Chandrabose MN, Chavez D, Clerk-Blankenburg KP, Cree A, Dao M, Davis C, Chacko J, Dinh H, Dugan-Rocha S, Fowler G, Garner TT, Garnes J, Gnirke A, Hawes A, Hernandez J, Hines S, Holder M, Hume J, Jhangiani SN, Joshi V, Khan ZM, Jackson L, Kovar C, Kowis A, Lee S, Lewis LR, Margolis J, Morgan M, Nazareth LV, Nguyen N, Okwuonu G, Parker D, Ruiz SJ, Santibanez J, Savard J, Scherer SE, Schneider B, Sodergren E, Vattahil S, Villasana D, White CS, Wright R, Park Y, Lord J, Oppert B, Brown S, Wang L, Weinstock G, Liu Y, Worley K, Elsik CG, Reese JT, Elhaik E, Landan G, Graur D, Arensburger P, Atkinson P, Beidler J, Demuth JP, Drury DW, Du YZ, Fujiwara H, Maselli V, Osanai M, Robertson HM, Tu Z, Wang JJ, Wang S, Song H, Zhang L, Werner D, Stanke M, Morgenstern B, Solovyev V, Kosarev P, Brown G, Chen HC, Ermolaeva O, Hlavina W, Kapustin Y et al (2008) The genome of the model

- beetle and pest *Tribolium castaneum*. *Nature* 452:949–955
- Roeding F, Borner J, Kube M, Klages S, Reinhardt R, Burmester T (2009) A 454 sequencing approach for large scale phylogenomic analysis of the common emperor scorpion (*Pandinus imperator*). *Mol Phylogenet Evol* 53:826–834
- Rota-Stabelli O, Telford MJ (2008) A multi criterion approach for the selection of optimal outgroups in phylogeny: recovering some support for Mandibulata over Myriochelata using mitogenomics. *Mol Phylogenet Evol* 48:103–111
- Rota-Stabelli O, Yang Z, Telford MJ (2009) MtZoa: a general mitochondrial amino acid substitutions model for animal evolutionary studies. *Mol Phylogenet Evol* 52:268–272
- Rota-Stabelli O, Kayal E, Gleeson D, Daub J, Boore JL, Telford MJ, Pisani D, Blaxter M, Lavrov DV (2010) Ecdysozoan mitogenomics: evidence for a common origin of the legged invertebrates, the Panarthropoda. *Genome Biol Evol* 2:425–440
- Rota-Stabelli O, Campbell L, Brinkmann H, Edgecombe GD, Longhorn SJ, Peterson KJ, Pisani D, Philippe H, Telford MJ (2011) A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. *Proc R Soc B* 278:298–306
- Rota-Stabelli O, Lartillot N, Philippe H, Pisani D (2012) Serine codon usage bias in deep phylogenomics: pancrustacean relationships as a case study. *Systematic Biology*. doi: [10.1093/sysbio/sys077](https://doi.org/10.1093/sysbio/sys077) [to be replaced by volume: page range on publication]
- Roy SW, Irimia M (2008) Rare genomic characters do not support Coelomata: intron loss/gain. *Mol Biol Evol* 25:620–623
- Shao R, Barker SC (2003) The highly rearranged mitochondrial genome of the plague thrips, *Thrips imaginis* (Insecta: Thysanoptera): convergence of two novel gene boundaries and an extraordinary arrangement of rRNA genes. *Mol Biol Evol* 20:362–370
- Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res* 19:1117–1123
- Suen G, Teiling C, Li L, Holt C, Abouheif E, Bornberg-Bauer E, Bouffard P, Caldera EJ, Cash E, Cavanaugh A, Denas O, Elhaik E, Fave MJ, Gadau J, Gibson JD, Graur D, Grubbs KJ, Hagen DE, Harkins TT, Helmkamp M, Hu H, Johnson BR, Kim J, Moeller JA, Munoz-Torres MC, Murphy MC, Naughton MC, Nigam S, Overson R, Rajakumar R, Reese JT, Scott JJ, Smith CR, Tao S, Tsutsui ND, Viljakainen L, Wissler L, Yandell MD, Zimmer F, Taylor J, Slater SC, Clifton SW, Warren WC, Elsik CG, Smith CD, Weinstock GM, Gerardo NM, Currie CR (2011) The genome sequence of the leaf-cutter ant *Atta cephalotes* reveals insights into its obligate symbiotic lifestyle. *Plos Genet* 7:e1002007. doi:[10.1371/journal.pgen.1002007](https://doi.org/10.1371/journal.pgen.1002007)
- Taylor HR, Harris WE (2012) An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. *Mol Ecol Resour* 2:377–388
- The *C elegans* genome consortium (1998) Genome sequence of the nematode *C elegans*: a platform for investigating biology. *Science* 282:2012–2018
- The Heliconius genome consortium (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature* 487:94–98
- The honeybee genome consortium (2006) Insights into social insects from the genome of the honeybee *Apis mellifera*. *Nature* 443:931–949
- The pea aphid genome consortium (2010) Genome sequence of the pea aphid *Acyrtosiphon pisum*. *Plos Biol* 8:e1000313. doi:[10.1371/journal.pbio.1000313](https://doi.org/10.1371/journal.pbio.1000313)
- The silkworm genome consortium (2008) The genome of a lepidopteran model insect, the silkworm *Bombyx mori*. *Insect Biochem Mol Biol* 38:1036–1045
- Thorley JL, Wilkinson M (1999) Testing the phylogenetic stability of early tetrapods. *J Theor Biol* 200:343–344
- Trewick SA (1998) Sympatric cryptic species in New Zealand Onychophora. *Biol J Linn Soc* 63:307–329
- Vieira FG, Rozas J (2011) Comparative genomics of the odorant-binding and chemosensory protein gene families across the Arthropoda: origin and evolutionary history of the chemosensory system. *Genome Biol Evol* 3:476–490
- Xia X, Xie Z (2001) DAMBE: software package for data analysis in molecular biology and evolution. *J Hered* 92:371–373
- Zeng V, Villanueva KE, Ewen-Campen BS, Alwes F, Browne WE, Extavour CG (2011) De novo assembly and characterization of a maternal and developmental transcriptome for the emerging model crustacean *Parhyale hawaiiensis*. *BMC Genomics* 12:581. doi:[10.1186/1471-2164-12-581](https://doi.org/10.1186/1471-2164-12-581)
- Zhang DX, Hewitt GM (1997) Insect mitochondrial control region: a review of its structure, evolution and usefulness in evolutionary studies. *Biochem Syst Ecol* 25:99–120



**Cite this article:** Lozano-Fernandez J *et al.* 2016 A molecular palaeobiological exploration of arthropod terrestrialization. *Phil.*

*Trans. R. Soc. B* **371**: 20150133.

<http://dx.doi.org/10.1098/rstb.2015.0133>

Accepted: 29 February 2016

One contribution of 15 to a discussion meeting issue 'Dating species divergences using rocks and clocks'.

#### Subject Areas:

molecular biology, palaeontology, evolution, taxonomy and systematics

#### Keywords:

terrestrialization, molecular palaeobiology, arthropod evolution, molecular clock, phylogeny

#### Author for correspondence:

Davide Pisani

e-mail: [davide.pisani@bristol.ac.uk](mailto:davide.pisani@bristol.ac.uk)

Electronic supplementary material is available at <http://dx.doi.org/10.1098/rstb.2015.0133> or via <http://rstb.royalsocietypublishing.org>.

# A molecular palaeobiological exploration of arthropod terrestrialization

Jesus Lozano-Fernandez<sup>1,2</sup>, Robert Carton<sup>3</sup>, Alastair R. Tanner<sup>2</sup>, Mark N. Puttick<sup>1</sup>, Mark Blaxter<sup>4</sup>, Jakob Vinther<sup>1,2</sup>, Jørgen Olesen<sup>5</sup>, Gonzalo Giribet<sup>6</sup>, Gregory D. Edgecombe<sup>7</sup> and Davide Pisani<sup>1,2</sup>

<sup>1</sup>School of Earth Sciences, and <sup>2</sup>School of Biological Sciences, University of Bristol, Life Sciences Building, 24 Tyndall Avenue, Bristol BS8 1TQ, UK

<sup>3</sup>Department of Biology, The National University of Ireland Maynooth, Maynooth, Kildare, Ireland

<sup>4</sup>Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3TF, UK

<sup>5</sup>Natural History Museum of Denmark, University of Copenhagen, Universitetsparken 15, 2100 Copenhagen, Denmark

<sup>6</sup>Museum of Comparative Zoology, Department of Organismic and Evolutionary Biology, Harvard University, 26 Oxford Street, Cambridge, MA 02138, USA

<sup>7</sup>Department of Earth Sciences, The Natural History Museum, Cromwell Road, London SW7 5BD, UK

**id** JL-F, 0000-0003-3597-1221; ART, 0000-0001-8045-2856; MNP, 0000-0002-1011-3442; MB, 0000-0003-2861-949X; JV, 0000-0002-3584-9616; JO, 0000-0001-9582-7083; GG, 0000-0002-5467-8429; DP, 0000-0003-0949-6682

Understanding animal terrestrialization, the process through which animals colonized the land, is crucial to clarify extant biodiversity and biological adaptation. Arthropoda (insects, spiders, centipedes and their allies) represent the largest majority of terrestrial biodiversity. Here we implemented a molecular palaeobiological approach, merging molecular and fossil evidence, to elucidate the deepest history of the terrestrial arthropods. We focused on the three independent, Palaeozoic arthropod terrestrialization events (those of Myriapoda, Hexapoda and Arachnida) and showed that a marine route to the colonization of land is the most likely scenario. Molecular clock analyses confirmed an origin for the three terrestrial lineages bracketed between the Cambrian and the Silurian. While molecular divergence times for Arachnida are consistent with the fossil record, Myriapoda are inferred to have colonized land earlier, substantially predating trace or body fossil evidence. An estimated origin of myriapods by the Early Cambrian precedes the appearance of embryophytes and perhaps even terrestrial fungi, raising the possibility that terrestrialization had independent origins in crown-group myriapod lineages, consistent with morphological arguments for convergence in tracheal systems.

This article is part of the themed issue 'Dating species divergences using rocks and clocks'.

## 1. The long road to terrestrial life

Animals and life more broadly have marine origins, and the colonization of land started early in life's history. Possible evidence for subaerial prokaryotic life dates back to the Archaean [1,2], and terrestrial communities (either freshwater or subaerial) with a eukaryotic component are known from the Torridonian of Scotland approximately 1.2–1.0 billion years ago (Gya) [3]. These deposits include multicellular structures, cysts and thalli that can have a diameter of almost 1 mm [3]. While there is no evidence for land plants, animals and fungi, these deposits indicate that at approximately 1 Ga relatively complex terrestrial ecosystems already existed [4]. Definitive evidence for the existence of land plants is much more recent. The oldest embryophyte body fossils are from the Late Silurian [5]. The oldest spores of indisputable embryophyte origin (trilete spores) extend the history of plants only a little deeper, into the Ordovician (449 million years

ago—Ma) [4,5], and the oldest embryophyte-like spores (which do not necessarily indicate the existence of embryophytes) barely reach the Late Cambrian [4]. Similarly, the fossil record of the terrestrial Fungi does not extend beyond the Ordovician, with the oldest known fungal fossils dating to approximately 460 Ma [6]. However, terrestrial rock sequences from the Cambrian and the Ediacaran are rare, and the late appearance of land plants and Fungi in the fossil record might represent preservational artefacts of the rock record [4].

Only few animal phyla include lineages that can complete every phase of their life cycle outside of water-saturated environments (from moisture films to the oceans) and are thus fully terrestrial. The most diverse and biologically important of the phyla with lineages that attained full terrestriality are the Vertebrata (with the reptiles, birds and mammals, i.e. Amniota); the Mollusca (with the land snails and the slugs); and the Arthropoda (e.g. insects, spiders, scorpions, centipedes) [7]. While the terrestrial vertebrates colonized the land only once even if some members (such as the cetaceans) secondarily reverted to life in water, molluscs and arthropods colonized the land multiple times independently and at different times in Earth history, constituting better model systems to study terrestrial adaptations at the genomic, physiological and morphological levels. In Arthropoda, there have been a minimum of three ancient (Palaeozoic) terrestrialization events: that of the Hexapoda, that of the Myriapoda and that of the Arachnida [8]. In addition, there have been multiple, more recent, land colonization events within malacostracans. These events correspond to the origin of terrestrial isopods (i.e. the woodlice) and amphipods (e.g. the landhoppers), and of a variety of semi-terrestrial species such as the coconut crab (*Birgus latro*), a decapod that lives its adult life on land but still retains marine larvae (see also [9]).

Previous studies [7,10–13] discussed at length the problems faced by animals crossing the water-to-land barrier, with [11] addressing them specifically in the case of the Arthropoda. These problems mostly relate to the different physical properties of air and water, and affect reproduction, sensory reception, locomotion, gas exchange, osmoregulation and protection from an increased exposure to ultraviolet radiation. A classic example of adaptation to terrestriality at the genomic level is observed, in both vertebrates and arthropods, when comparing the olfactory receptors of marine and terrestrial forms. Terrestrialization is associated with massive, independent, parallel changes in the olfactory receptor gene repertoires of both lineages probably because water-soluble and airborne odorants differ and cannot be efficiently bound by the same receptors [14–16].

Multiple independent terrestrialization events within the same lineage permit rigorous comparison of alternative solutions adopted by different (but genomically and morpho-physiologically comparable) groups to the same adaptive challenge, and represent a powerful tool for understanding evolution in a comparative framework [17]. To carry out meaningful comparative studies of animal terrestrialization, however, it is necessary to (i) clarify how many independent terrestrialization events happened in the lineage under scrutiny, (ii) estimate when these terrestrialization events happened and how long they took, and (iii) robustly identify the aquatic sister group of each terrestrial lineage. This information is, in turn, necessary to enable comparative analyses and to estimate the rate at which terrestrial adaptations emerged.

Here we explore the three deepest (Palaeozoic) arthropod terrestrialization events (those of the Hexapoda, Myriapoda

and Arachnida), and summarize and expand current evidence about processes that led to their terrestrialization. We particularly focus on Hexapoda, because hexapod terrestrialization, an event that led to the origin of the majority of terrestrial animal biodiversity [18], is particularly poorly understood.

## 2. The phylogenetic perspective

Phylogenetic relationships among the major arthropod lineages have long been debated [19]. However, some consensus has emerged. Myriapoda, the first of the three major terrestrial arthropod groups we shall consider, is now generally accepted to represent the sister group of Pancrustacea (Hexapoda plus all the crustacean lineages). The Myriapoda–Pancrustacea clade is generally referred to as Mandibulata [20–23]. Alternative hypotheses of myriapod relationships have been previously proposed. Among these are the Atelocerata or Tracheata hypothesis, which suggested myriapods as the sister of hexapods, and the Myriochelata hypothesis, which saw the myriapods as the sister group of chelicerates. Atelocerata was based on morphological considerations (e.g. both myriapods and hexapods use tracheae to carry out gas exchange) and continues to have a few adherents among morphologists [24]. However, Atelocerata has only been recovered once in analyses combining molecular, morphological and fossil data [25]. The Myriochelata hypothesis was derived entirely from molecular analyses [26–30], and is now generally considered to have been the result of a long-branch attraction artefact caused by the faster-evolving pancrustaceans attracting to the outgroup and pushing Myriapoda and Chelicerata into an artefactual clade [20]. Both Myriochelata and Atelocerata are disfavoured by current available analyses, with strong molecular and morphological support favouring a placement of hexapods within ‘Crustacea’ (the Pancrustacea or Tetraconata concept—e.g. [20,23,26,31–35]), and a placement of Myriapoda as the sister group of Pancrustacea within Mandibulata (see references above and [19] for a recent review). Accordingly, there is now general agreement that the sister group of the terrestrial Myriapoda is the (primitively) marine Pancrustacea.

The sister group relationships of the Arachnida are quite well understood. This group includes all the terrestrial chelicerates and has two extant successively more distant marine sister taxa: Xiphosura (horseshoe crabs) and Pycnogonida (sea spiders) [23,36,37]. In contrast, the exact relationships of the Hexapoda within Pancrustacea are still unclear, and it is not obvious whether their sister taxon was a marine-, brackish- or freshwater-adapted organism.

Early analyses of eight molecular loci combined with morphological data provided some support for Hexapoda as the sister group of a monophyletic Crustacea, barring a long-branch clade [38], with Branchiopoda as the sister group of Remipedia plus Cephalocarida (the latter two taxa constituting Xenocarida *sensu* [23]). Subsequently, a taxonomically well-sampled molecular phylogeny of three protein coding genes [34] found support for Branchiopoda as the sister group of Hexapoda, and Remipedia as the sister group of those two taxa. While mitogenomic data have also been used in an attempt to resolve hexapod relationships, this type of data is notoriously difficult to analyse [39,40] and has frequently recovered misleading results (contrast [41,42]). With reference to the relationships of Pancrustacea, mitogenomic data were found to be unable to resolve

hexapod relationships with confidence [43] and we shall not consider them further.

Based on a large dataset of 62 protein coding genes analysed as nucleotide sequences, support for a sister group relationship between Xenocarida (Remipedia + Cephalocarida—see also above) and Hexapoda was found [23,35]. This clade was called Miracrustacea [23]. In the same analysis, Branchiopoda grouped with Malacostraca, Copepoda and Thecostraca in a novel clade named Vericrustacea [23] rather than allying with Hexapoda. However, these findings were shown to be affected by an artefact of serine codon bias [37]. The close association between Remipedia and Hexapoda (to the exclusion of Cephalocarida) was the only high-level pancrustacean relationship proposed by [23] that was confirmed by [37], which reinstated Branchiopoda as a close relative of Hexapoda, finding Remipedia, Hexapoda, Branchiopoda and Copepoda to constitute an unresolved clade that was referred to as ‘clade A’ in [37]. Other recent studies found similar results, suggesting a Branchiopoda + Hexapoda + Remipedia [21,22,44] (and perhaps Cephalocarida [45]) clade, but with different internal resolutions. In particular, [21,44,45] found Remipedia as the closest relative of Hexapoda (as in [34]), whereas [22] found Branchiopoda as the sister taxon of Hexapoda. Oakley *et al.* [45] was the only one, among the studies mentioned above, that included Cephalocarida, and found Remipedia as the sister group of Hexapoda and Branchiopoda as the sister group of Cephalocarida. Overall, from the perspective of molecular phylogenetics, a strong case can be made that Hexapoda, Branchiopoda and Remipedia belong to the same clade. In addition, evidence exists that Cephalocarida might also be a member of this group of hexapod relatives, which was named Allotriocarida [45]. Yet, to date, molecular phylogenetics has not robustly resolved internal allotriocarid relationships.

A close association between Remipedia and Hexapoda had been suggested based on the presence of a duplication of the haemocyanin gene (haemocyanin being the respiratory pigment used by most arthropods) that is uniquely shared between Remipedia and Hexapoda [46]. This duplication could represent a rare genomic event indicative of a possible sister group relationship between Remipedia and Hexapoda. However, Branchiopoda use haemoglobin as a respiratory pigment rather than haemocyanin. Because haemoglobin is an autapomorphy of Branchiopoda, the presence of two haemocyanin genes in Remipedia and Hexapoda and one in Cephalocarida [46] would conclusively resolve the sister group relationship between these taxa only if the relationships between Cephalocarida and Branchiopoda delineated by [45] were correct. This is because if Cephalocarida (which has only one haemocyanin) is not closely related to Remipedia, Branchiopoda and Hexapoda, then the haemocyanin duplication could have happened in the stem lineage subtending Remipedia, Branchiopoda and Hexapoda, with Branchiopoda having lost both paralogues as it shifted to using haemoglobin as a respiratory pigment. To validate the haemocyanin evidence, it is thus of paramount importance that further studies be carried out to either reject or confirm the results of [45], as bootstrap support values for the monophyly of Allotriocarida and the deepest relationships within this clade were variable and never higher than 85% [45]. Similarities between Remipedia and Hexapoda were also previously suggested based on neurological characters [47,48]. However, more recent studies showed that

while neuroanatomical similarities between Hexapoda and Remipedia exist, brain morphology suggests a closer association between Remipedia and Malacostraca [49]. Given that hexapods are generally not found to be close relatives to Malacostraca by other lines of evidence (see above for molecular analyses), similarities in the nervous systems of these three lineages might be subject to evolutionary convergence.

Knowledge of the sister group of each terrestrial arthropod lineage is important not only to increase the power of comparative studies to test adaptive strategies to life on land (see above), but also to understand the route to terrestrialization taken by different lineages. While the sister groups of Myriapoda and Arachnida were undoubtedly marine, most branchiopods inhabit freshwater, and a freshwater route to hexapod terrestrialization was proposed based on this [50]. In contrast, Remipedia is exclusively found in coastal anchialine settings generally with some connection to the sea. Accordingly, a sister group relationship between Remipedia and Hexapoda would better support a direct, marine [10] route to terrestrialization [44].

### 3. The timescale of arthropod terrestrialization

The oldest arthropod fossils are undoubtedly marine. They include trilobites, the oldest representatives of which date back to the Early Cambrian (*ca* 521 Ma [51]); Trilobita is variably interpreted as either stem mandibulates [20] or as stem chelicerates [52]. Other Cambrian, marine fossils include chelicerates (pycnogonids [53]), and crustaceans; both cuticular fragments from Branchiopoda, and possibly also Ostracoda and Copepoda [54] and complete body fossils such as the allotriocarid (most likely stem branchiopod) *Rehbachella kinnekullensis* [55].

The oldest subaerial arthropod traces (ichnofossils) are from the Mid- to Late Cambrian–Early Ordovician age. Examples include trackways impressed on eolian dune sands by an amphibious myriapod-like arthropod, perhaps a euthycarinoid [56]. Other Cambrian (Mid-Cambrian to Furongian) locomotory traces have been documented from subaerially exposed tidal flats in Wisconsin and Quebec [57]. A euthycarcinoid tracemaker has been confidently associated with these traces, further cementing the view that arthropod subaerial activities (if not terrestrial arthropods) were common on Cambrian shorelines. The oldest terrestrial myriapod body fossil (which is also the oldest undisputably terrestrial animal) is the *ca* 426 Ma millipede *Pneumodesmus newmani*, from the Silurian of Scotland [58]. The subaerial ecology of *P. newmani* is indisputable, because spiracles (segmental openings that allow air to enter the tracheal system) are present on the lateral part of its sternites. The Siluro-Devonian fossil record of Myriapoda consists only of taxa that can be assigned with confidence to the crown groups of extant classes (Diplopoda and Chilopoda), as well as the apparent diplopod-allied Kampecarida, and to date no well corroborated candidates for stem-group Myriapoda have been identified [59]. Critical reviews of the diagnostic/apomorphic characters of myriapods have outlined a search image for a stem-group myriapod that could potentially be recognized in Early Palaeozoic marine strata [60]. Arachnid fossils are just a little younger than those of the oldest Myriapoda, the earliest unequivocally terrestrial examples (trigonotarbid) being present in Silurian deposits dated at

approximately 422 Ma [61]. Early Silurian arachnids are represented by the oldest scorpions, which have long been considered to be aquatic because of their associated biota and sediments, but phylogenomic evidence for Scorpiones being nested within terrestrial clades of Arachnida [36] is more compatible with terrestrial habits [62]. The stem group of Arachnida has an aquatic fossil record as far back as the Late Cambrian, the earliest fossils being resting traces of chasmataspidids [63], resolved as sister group to a eurypterid–arachnid clade [64]. Evidence for complex terrestrial ecosystems with land plants, fungi and a variety of arthropods is known from the Upper Silurian onward [65] and is confirmed in the beautifully preserved, and widely celebrated, Lower Devonian (approx. 411 Ma), Rhynie chert Konservat-Lagerstätte [66]. The latter includes the oldest examples of Hexapoda in the fossil record, including Collembola and Insecta.

Recent molecular clock analyses of the arthropod radiation (or of parts of it) generally corroborate the palaeontological evidence and suggest times of origin for Arachnida that are broadly consistent with the fossil evidence [8,21,67–70]. However, molecular divergence times for the origin of crown-group Hexapoda and Myriapoda substantially predate fossils, and this discrepancy is more pronounced in the case of Myriapoda, for which divergence estimates firmly place the modern representatives of this phylum deep in the Cambrian, despite the oldest known crown myriapod fossil being only 426 Ma [58]. This is problematic, because all crown myriapods are terrestrial, and all use tracheae for gas exchange. If tracheae have a single origin in Myriapoda, then current molecular clock results suggest a Cambrian terrestrialization for this lineage, which is not documented in the fossil record. Ephemeral, terrestrial ecosystems existed since approximately 1 Ga [3], and the fossil record of embryophyte-like spores suggests that some form of vegetation existed on land in the Cambrian [2,4,5]. Such limited terrestrial environments, as well as coastal environments [56,57], could have already been conducive to myriapod life on land in the Cambrian [2].

One recent molecular clock study of the arthropod radiation [71], despite being in agreement with other studies with reference to arthropod terrestrialization, is in disagreement with both the fossil record and other molecular clock studies with reference to the deepest divergences in the arthropod tree. However, this study was based on the gene set of [23], that was shown to be affected by strong codon-usage biases [37]. In the absence of correction, this dataset recovered a large number of otherwise unsupported pancrustacean clades (e.g. Vericrustacea and Miracrustacea, see [71]) and consequent erroneous estimation of branch lengths and divergence times. Indeed, subsequent analysis of the same data that attempted to correct for such biases [37] yielded results generally comparable to those obtained in other molecular clock studies.

## 4. A freshwater route to life on land?

An interesting question in the study of terrestrialization is whether land was invaded directly from the sea (the marine route [10,44]), or whether animals first colonized freshwater environments and only subsequently moved to the land (the freshwater route [50]). To address this question, we can look at the fossil record of stem terrestrial lineages when available, and to the sister group of these terrestrial

lineages. A freshwater route would imply that the last common ancestor of the considered terrestrial taxa and its sister aquatic lineage separated in a freshwater habitat [50], whereas a marine route would imply that they separated either in a marine or brackish (estuarine) environment [44]. Myriapods and arachnids have marine sister groups. In the case of the Hexapoda, a freshwater route was suggested based on presumed sister-group relationships between Branchiopoda and Hexapoda [50]. While the freshwater origins hypothesis is challenged by the proposal that Remipedia are the sister group of Hexapoda [44], this is far from well established (see above), leaving space for the possibility that hexapod ancestors might have first colonized fresh water and only after that the land. Here we investigate whether hexapods took a marine or a freshwater route to the colonization of land.

## 5. Material and methods

### (a) Dataset assembly

We expanded a published dataset [72] to include new arthropod taxa (see electronic supplementary material, table S1) mostly obtained from NCBI. Transcriptomes of the sea spider *Pycnogonus* sp. and of the horseshoe crab *Limulus polyphemus* were obtained as part of this study and sequenced, respectively, at Edinburgh Genomics and at the Geogenomic Center in Copenhagen. We also added other bilaterian taxa to increase the number of calibration points available for molecular clock analyses (electronic supplementary material, table S1 and figures S1–S5). The core dataset included 57 taxa and 246 genes. This dataset was then pruned of all non-panarthropod species, to avoid systematic biases that might have been induced by the presence of distant outgroups, and create a smaller dataset (including 30 species and 246 genes) used for phylogenetic analyses only. We developed a series of PERL scripts (available at [github.com/jairly/MoSuMa\\_tools](http://github.com/jairly/MoSuMa_tools)) to add species to the existing dataset. BLASTp [73] was used, with an *E*-value cut-off of less than  $10^{-20}$  to identify potential orthologues. The new potential orthologues were aligned with the existing orthologue set using MUSCLE [74], and a maximum-likelihood (ML) tree was generated using PhyML [75] under the LG + G model. Tree distances (branch length distances) were used to distinguish orthologues from paralogues using a few simple rules. (1) If only one putative orthologue existed and its average tree distance from all previously identified orthologues in the dataset was within 3 standard deviations of the average of the tree distances calculated across all previously identified orthologues, then the putative orthologue was retained. (2) If there was only one putative orthologue and its distance to other previously identified orthologues exceeded 3 standard deviations from the average of the tree distances calculated across all previously identified orthologues, then the tree and the alignment were visually inspected. (2a) If the sequence was misaligned, then the alignment was corrected and the procedure repeated. (2b) If the sequence was correctly aligned and the sequence clustered in a phylogenetically unexpected position (e.g. a new *Daphnia* sequence that clustered with a human sequence), then the sequence was deemed a possible paralog and not retained. Note that here ‘phylogenetically unexpected’ simply means obviously incorrect. A myriapod sequence clustering with a chelicerate, for example, was considered to cluster in an expected position, in contrast to *Daphnia* clustering with a human. (2c) If the sequence was correctly aligned and the sequence clustered in a phylogenetically plausible position (e.g. a new *Drosophila* sequence that clustered within insects) the sequence was retained but flagged to allow for directed exclusion (if necessary) in subsequent analyses. (3) If more than one putative

orthologue was present in the dataset, then the tree was first visually inspected to evaluate whether all putative orthologues formed a monophyletic group (i.e. to make sure they constituted a set of in-paralogs). (3a) If they did and their average tree distance from other sequences was less than 3 standard deviations from the average distance across all previously identified orthologues, then the putative orthologue of minimal branch length was retained. (3b) If the putative orthologues did not cluster together and all but one had significant distance (in excess of 3 standard deviations) from the average distance across all previously identified orthologues, the putative orthologue of acceptable distance was retained if it also clustered in a phylogenetically plausible position. (3c) If all putative orthologues had excessively long branches (more than 3 standard deviations from the average), then they were all rejected. Each set of orthologues was realigned using MUSCLE [74] and trimmed using Gblocks [76] to exclude ambiguously aligned sections. Gblocks settings were: minimum number of sequences for a conserved position = 50% of the sequences in the protein family; minimum number of sequences for a flank position = 75% of the sequences in the protein family; minimum length of a block = 5; allowed gap positions = half. The final dataset of curated sequences was concatenated using FASCONCAT v. 1.0 [77]. It included 58 taxa across all Protostomia and Deuterostomia and 40 657 amino acid positions. Taxa were deleted from this dataset to generate the taxonomically reduced alignment used for phylogenetic reconstruction (see above). The latter included 30 panarthropod species and 40 657 amino acid positions.

### (b) Phylogenetic reconstruction

Phylogenetic trees were inferred using PHYLOBAYES MPI v. 1.5 [78] under the site-heterogeneous CAT – GTR + G model of amino acid substitution [79]. Convergence was assessed by running two independent Markov chains and using the bpcomp and tracecomp tools from PHYLOBAYES to monitor the maximum discrepancy in clade support (maxdiff), the effective sample size (effsize) and the relative difference in posterior mean estimates (rel\_diff) for several key parameters and summary statistics of the model. The appropriate number of samples to discard as ‘burn in’ was determined first by visual inspection of parameter trace plots, and then by optimizing convergence criteria.

### (c) Molecular clock analyses

Divergence time estimation was performed using PHYLOBAYES 3.3f (serial version) [80] on a fixed topology (see electronic supplementary material, figures S1–S5). We used two alternative relaxed molecular clock models: the autocorrelated CIR model [81] and the uncorrelated gamma multipliers model (UGAMMA) [82], as in [83]. The tree was rooted on the Deuterostomia–Protostomia split. A set of 24 calibrations (see electronic supplementary material, table S2) was used, with a root prior defined using a Gamma distribution of mean 636 Ma and standard deviation of 30 Ma. However, previously we had also tested the effect of a much more relaxed root prior that used an exponential distribution of average 636 Ma (see electronic supplementary material, table S2 for justifications). The substitution model used to estimate branch lengths was the CAT – GTR + G model, as in the phylogenetic analysis. All analyses were conducted using soft bounds with 5% of the probability mass outside the calibration interval. A birth–death model was used to define prior node ages. Analyses were run under the priors to evaluate the effective joint priors induced by our choice of priors. Convergence was tested running the tracecomp tool as specified above.

### (d) Ancestral environment reconstructions

Maximum-likelihood-based ancestral character state reconstruction was carried in R ([www.R-project.org](http://www.R-project.org) [84]) using

maximum-likelihood estimation under the Mk model [85,86] to infer whether the last common ancestor of Branchiopoda was a freshwater-, marine- or brackish-adapted animal. The branchiopod phylogeny of [87] was modified to include key fossils from [88]: *Rehbachella*, *Lepidocaris*, *Castracollis* and *Almatium*. *Rehbachella kinnekullensis* (from the Upper Cambrian) is particularly important as it was initially described as a marine stem-group anostracan [55], and subsequently reassigned to a stem-group branchiopod [89]. This systematic placement has not been universally accepted, with some analyses instead allying *Rehbachella* closer to cephalocarids than to branchiopods [45,90]. Whereas *Rehbachella* is found in association with marine taxa [55], and the geological context of the bituminous limestones in which the fossils are preserved indicates dysoxic marine sediments, most extant branchiopods are found in fresh water or in continental brackish waters (vernal pools, saline lakes, etc.). *Lepidocaris rhyniensis* [91] and *Castracollis wilsonae* [92] are freshwater branchiopod fossils from the Early Devonian Rhynie chert. Kazacharthra (represented herein by *Almatium gusevi* [93]), are Triassic–Jurassic relatives of Notostraca limited to non-marine (lacustrine) deposits from Kazakhstan, Mongolia and China. A matrix representing ecological preferences for all considered taxa was assembled from the literature (see electronic supplementary material, table S3). The time-calibrated tree was generated by adding the fossils from [88] to the tree in [87] using 10 calibrations from [94] and setting tip taxa to their occurrence times. The time-calibrated topology was generated using the R package paleotree [95]. We calculated marginal likelihood under Mk for internal nodes in this time-calibrated tree and present the scaled marginal likelihoods of the three possible root states for total-group Branchiopoda.

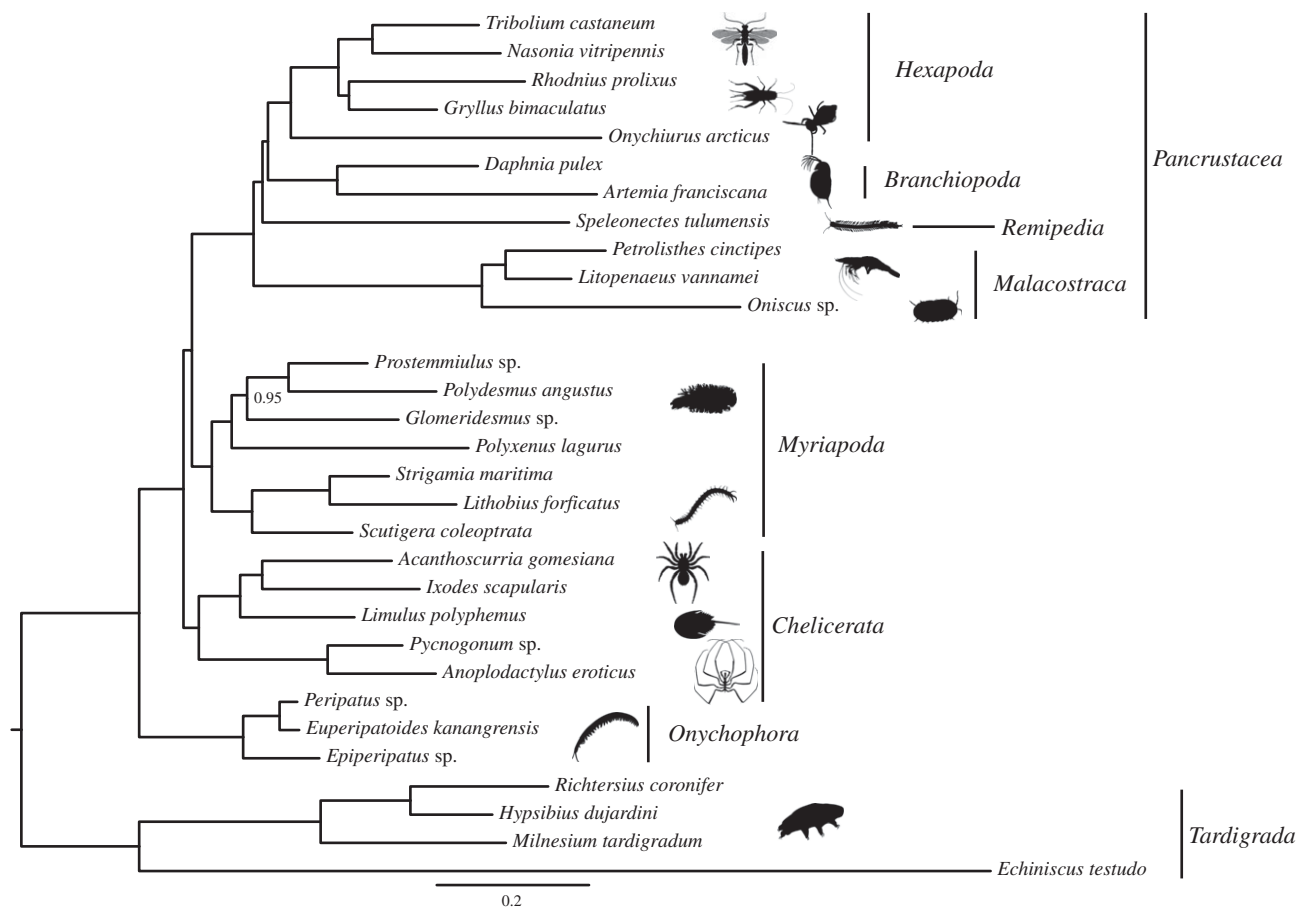
## 6. Results

### (a) Phylogeny

Our phylogenetic analyses are presented in figure 1. They clearly support monophyly of Arthropoda and of the three main arthropod lineages (Chelicerata, Myriapoda and Pancrustacea). While a few studies have suggested that Tardigrada, rather than Onychophora, might be the closest sister group of Arthropoda [96], evidence for this phylogenetic arrangement is limited to only a few morphological characters. Our choice of Tardigrada as outgroup is thus guided by results of previous phylogenomic studies [72,97,98]. The relationships among the arthropod lineages are resolved according to current convention and depict a Mandibulata clade (PP = 1) as the sister group of Chelicerata (PP = 1). Within Chelicerata, the sea spiders are recovered as the sister group of the other chelicerates, Euchelicerata (PP = 1), with xiphosurans as sister group to arachnids. Myriapods are likewise well resolved, dividing into Chilopoda and Diplopoda, and each group follows the currently well-accepted relationships [69,99]. Within Pancrustacea, we recovered an arrangement of taxa that is consistent with the monophyly of Allotriocarida. Of particular relevance to terrestrialization is the partial allotriocarid clade, including Branchiopoda, Remipedia and Hexapoda. Within this clade, we found Branchiopoda to be the sister group of Hexapoda (PP = 1), in agreement with [22,37] but contrasting with other studies (as summarized above [21,44,45]).

### (b) Molecular divergence times

Molecular divergence times among arthropod major clades are presented in figure 2 and table 1 and in electronic



**Figure 1.** Bayesian phylogeny of Panarthropoda. This tree was obtained under the CAT + GTR + G model. All nodes but one had a posterior probability of 1. bpcomp maxdiff = 0; minimum effective size = 55; maximum rel\_diff = 0.2. Most silhouettes from organisms are from Phylopic ([phylopic.org/](http://phylopic.org/)).

supplementary material, figures S1–S5. Results obtained using the UGAMMA model are shown in figure 2a, the autocorrelated CIR model in figure 2b. Results obtained using the UGAMMA model but with a more permissive exponential root prior are reported in figure 2a. Using UGAMMA, 95% credibility intervals surrounding the average divergence times were significantly larger than when the autocorrelated CIR model was used. However, it was evident that for the three nodes of interest (those representing Palaeozoic terrestrialization events) the values in the 95% credibility interval obtained under CIR always represented subsets of the values in the 95% credibility interval obtained using UGAMMA. While the two sets of results are thus statistically indistinguishable, they differ in their congruence with the fossil record. While the more permissive UGAMMA analyses did not reject a Late Cambrian to Silurian origin of the three terrestrial arthropod lineages (the upper limit consistent with the fossil evidence), the CIR model rejected an Ordovician origin for the Myriapoda, suggesting a Precambrian origin instead. Under UGAMMA, arachnid terrestrialization happened in the Silurian, whereas CIR suggests an Ordovician colonization of land. In the case of the Hexapoda, UGAMMA analysis suggested an Ordovician origin, whereas CIR suggested a Cambrian origin and statistically rejected an Early Ordovician origin for this group. Thus, in general, CIR results suggest deeper divergence times. The use of the exponential root, while affecting divergence times of the deepest nodes in our tree (e.g. the age of the Deuterostomia–Protostomia split which is not presented in figure 2, but see electronic supplementary material, figures S1–S5), did not

have any effect on the divergence times of the nodes of interest (figure 2 and electronic supplementary material, figure S2).

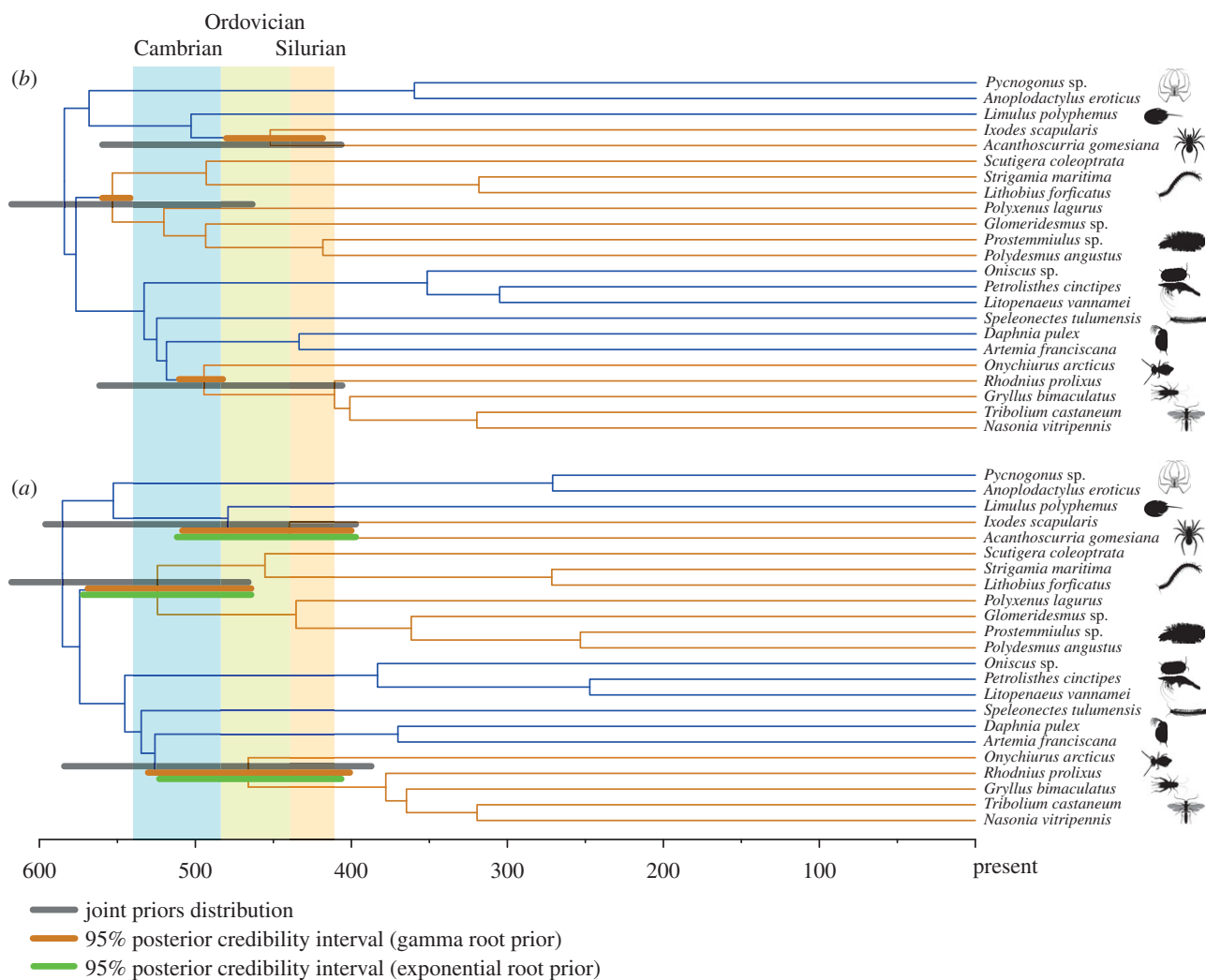
### (c) Ancestral environmental reconstruction

Our ancestral environmental reconstructions (figure 3) aimed to clarify whether the hexapods colonized the land through a freshwater route if their sister group is Branchiopoda rather than Remipedia (figure 1). We found that the last common ancestor of the stem-group Branchiopoda most likely inhabited a marine environment ( $p = 0.84$ ; figure 3). A lower, but not negligible, probability is found for an ancestral freshwater habitat ( $p = 0.15$ ), whereas a brackish ancestry for the total-group Branchiopoda can be confidently rejected ( $p = 0.002$ ; figure 3). Note that these results used a topology where the marine *Rehbachella* was considered the sister group of the extant branchiopods. As pointed out above, some studies suggested this fossil might instead be allied to cephalocarids [45,90]. If that were the case, given the sister group relationship between cephalocarids and branchiopods suggested in these studies, then a marine origin of Branchiopoda would be inevitable, thus not changing the results of our analyses.

## 7. Discussion

Terrestrialization is the process through which aquatic organisms adapt to a subaerial lifestyle [7], and abundant literature has addressed this process at the physiological level [9,10,12]. However, most of these studies were performed on isolated





**Figure 2.** Results of molecular clock analyses. (a) Divergence times obtained under the CIR autocorrelated, relaxed, molecular clock model. (b) Divergence times obtained using the Uncorrelated Gamma Multipliers model. In both cases, nodes in the tree represent average divergence times estimated using the root prior with 636 Ma mean and 30 Ma SD. Brown bars represent 95% credibility intervals from the considered analysis. Grey bars represent the joint priors (for the considered nodes and analyses). Green bars in figure 2b indicate 95% credibility intervals obtained using the exponential prior of average 636 Ma. Blue branches indicate marine lineages. Brown branches terrestrial lineages. In the timescale, numbers represent Myr before the present.

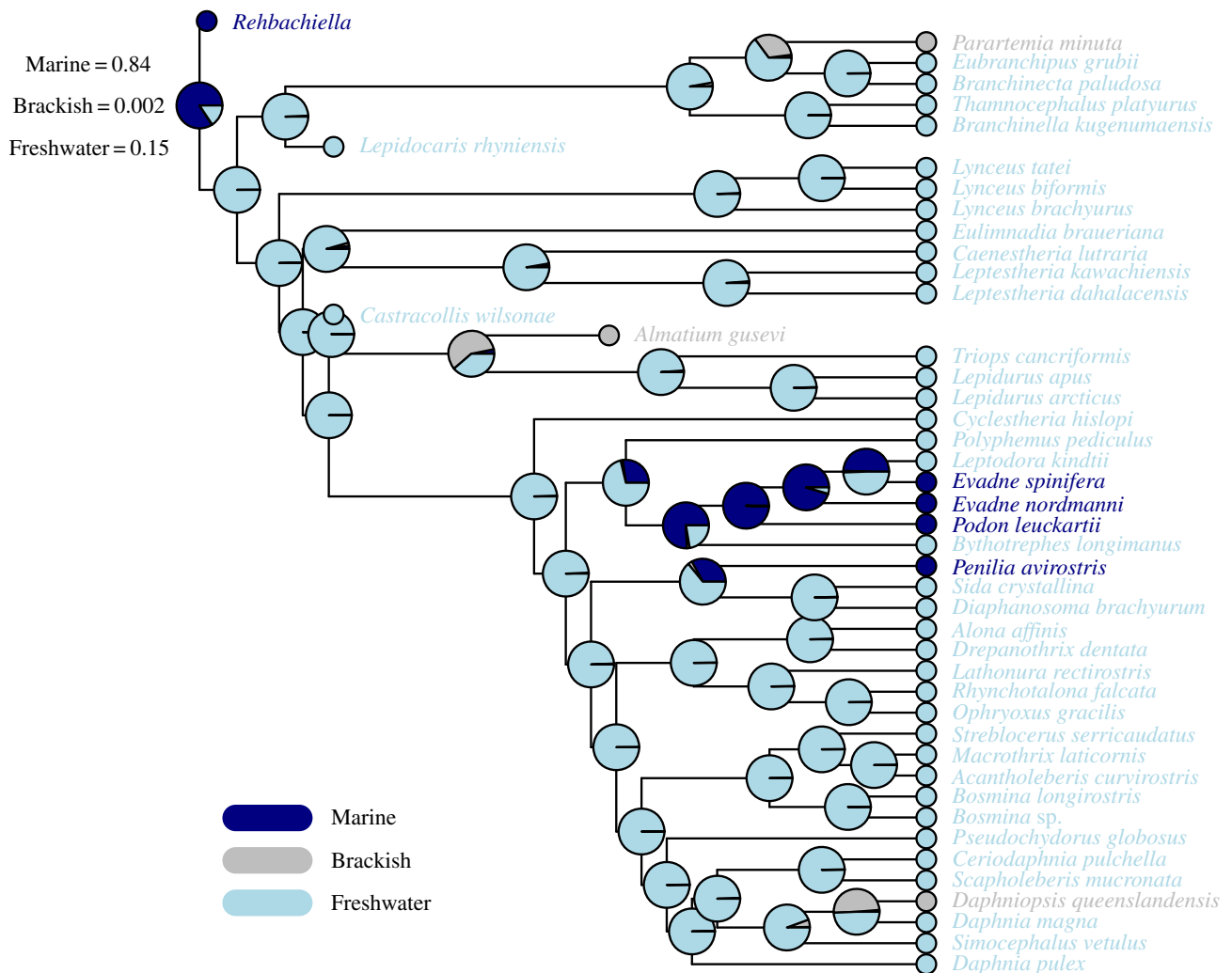
**Table 1.** Molecular divergence times for key terrestrial arthropod lineages.

taxon	molecular clock model		mean age (Ma)	95% credibility interval
	UGAMMA	CIR		
Myriapoda	528	568–463	558	572–544
Chilopoda	457	526–408	490	511–452
Diplopoda	439	537–317	519	541–486
Hexapoda	468	512–407	499	431–394
Arachnida	440	518–397	460	493–413

lineages and did not take full advantage of the comparative approach [17], in part because the application of modern comparative methods [100] needs detailed phylogenetic information and divergence times for terrestrial lineages and their close relatives. Such information has only recently started to be available in sufficient detail.

Our phylogenetic analyses used an expanded multigene dataset of wide systematic scope. While our results are

consistent with the monophyly of Allotriocarida, in contrast to [45] and other studies [21,23,35,44], we did not find support for a sister group relationship between Remipedia and Hexapoda. We instead recovered Branchiopoda as the sister group of Hexapoda, as has been proposed previously [22]. Our results cannot be taken as definitive, most importantly because, as with all previous relevant analyses we were able to include only one remipede species, and similar to all



**Figure 3.** Results of the ancestral environment reconstruction analysis indicating that the last common total-group branchiopod ancestor was most likely a marine organism. The pie charts show the scaled marginal likelihoods of ancestral states for all nodes, with the scaled likelihoods of the total-group ancestor also shown in the text. Branch lengths are proportional to time.

previous studies except that of [45], we did not include cephalocarids. With reference to molecular divergence times, whereas [28] obtained the first set of estimates specifically aiming at clarifying terrestrialization in Arthropoda, their study used a dataset composed of only few genes and taxa and molecular clock methods and calibrations that are now obsolete [101]. The most relevant previous molecular clock study specifically addressing arthropod terrestrialization is that of [8], although divergence times among terrestrial lineages can be found in a variety of other studies [21,67–70,102]. Summarizing results from these previous studies indicates that crown (terrestrial) Myriapoda emerged at 554 Ma, crown (terrestrial) Arachnida emerged at 495 Ma, and crown terrestrial Hexapoda emerged at 495 Ma. These divergence times are broadly in line with the results of our analyses (figure 2 and table 1 and electronic supplementary material, figures S1–S5). In the case of Arachnida, this is broadly compatible with the fossil evidence, whereas in the cases of Hexapoda and particularly Myriapoda the molecular divergences are significantly older. Interpretation of the amphibious euthycarcinoids, which first appear in the Cambrian, as stem-group hexapods [103], goes some way to reconciling early estimates for the origin of Hexapoda and the substantially later appearance of crown-group fossils in the Early Devonian.

A recent fossil-independent attempt at dating the metazoan radiation [104] suggested that divergence times

that are substantially in line with the fossil record, like all those reported above except [71], represent artefacts caused by over-constrained calibrations, and that the history of animals is much more in line with previous, outdated, findings that suggested the existence of metazoans approximately 1.5 Ga [105]. Indeed, Battistuzzi *et al.* [104] also suggested that the analyses of Wheat & Walberg [71], despite being in strong disagreement with the arthropod fossil record and with other molecular clock studies of the arthropod radiation, may be accurate. As discussed above, however, the results of [71] are based on a dataset affected by strong compositional biases, and used a pancrustacean topology that has now mostly been contradicted. In addition, it has now been shown that there is not enough information left in genomic datasets to correctly estimate rates of evolution in the deepest part of the animal tree without reference to fossils [102], as advocated by Battistuzzi *et al.* [104]. Tellingly, an analysis of the relative rates of substitution per branch inferred by Battistuzzi *et al.* [104] shows them to be identical (and set to the median rate across their entire tree) in 64.5% of the internal branches in their chronogram (electronic supplementary material, figure S6). Furthermore, these constant strict-clock rates are asymmetrically clustered in the root-ward part of their tree. In other words, the relative divergence time approach used in [104] did not relax the clock in the deepest part of their chronogram, and inferred that more than half of

opisthokont history (the outgroup in their chronogram is Fungi) was strictly clocklike. The existence of a deep clock for Metazoa and Opisthokonta is clearly unrealistic and is rejected by the data [102], confirming Pisani & Liu's [101] suggestion that relative divergence times cannot meaningfully be applied in deep time. Given the results of [102], and the rate distribution in electronic supplementary material, figure S6, it is not unsurprising that [104] found results comparable to those found in outdated strict-clock studies [105] from two decades ago. From the point of view of arthropod evolution, the convergence of the results of [104] and [71] further suggests that deep divergence times for the origin of Arthropoda are likely to be artefactual.

Considering hexapod terrestrialization, both the freshwater [50] and the marine [44] routes should be considered valid alternatives. Key to distinguishing between the two is understanding whether the last common ancestor of the Hexapoda and either Remipedia or Branchiopoda inhabited a marine, brackish or freshwater habitat. If the last common ancestor of Hexapoda and its sister clade was a freshwater organism, then the colonization of land could have started from a freshwater habitat. If Remipedia (or Remipedia plus Cephalocarida—if Xenocarida were confirmed in future studies) is confirmed as the sister group of Hexapoda, then a marine route would be strongly favoured as there is no evidence that the anchialine–water dwelling remipedes might have ever been living away from the coasts, whereas cephalocarids are marine. If Branchiopoda is confirmed as the sister group of the hexapods, then the situation would be more ambiguous, as modern branchiopods are mostly found in continental waters, leaving the question of the environmental preferences of the last common branchiopod ancestor unresolved. To address this problem, we used ancestral character reconstruction which suggests that, when both extant and fossil taxa are considered, the last common ancestor of Branchiopoda and Hexapoda was most likely a marine organism. Thus, current evidence, when considering phylogenetic uncertainty of hexapod relationships and fossil evidence, seems to favour a marine route to land also for the Hexapoda. Future discoveries of additional Cambrian stem-group branchiopods could better clarify this problem.

## 8. Conclusion

Ephemeral, terrestrial habitats have long existed on the Earth, at the very least since approximately 1 Ga. However, animal terrestrialization was a much more recent process. This was first of all because animals originated in the Cryogenian and radiated close to the base of the Cambrian, in disagreement with [104], and in agreement with [83,102]. Our molecular

clock results cannot reject fossil-based divergence times for Arachnida and Hexapoda, and we thus conclude that the most likely scenario, given the current evidence, is that these lineages colonized the land in the Ordovician or the Silurian (Arachnida) and the Ordovician (Hexapoda). Estimates that Myriapoda may have colonized land earlier are in disagreement with the myriapod fossil record, even allowing that terrestrial ecosystems already existed in the Cambrian. A mid-late Cambrian diversification of Diplopoda has, however, been predicted based on geographic distributions of extant millipedes and palaeogeography [106]. We do, however, note that our results for the origins of Chilopoda and Diplopoda are consistent with current fossil evidence (figure 2 and electronic supplementary material, figures S1–S5). One possible scenario that would partly resolve this clash between fossils and molecules would be that these two lineages independently colonized the land; but for that to be the case, tracheae should have evolved independently. This possibility has been suggested previously based on differences in structure of the tracheae and position of the spiracles [107] and should be subjected to critical testing. Irrespective of the precise time at which different arthropods colonized land, it seems currently more likely that the process of animal terrestrialization did not begin before the Late Cambrian and proceeded from the coastline towards the centre of the continents.

**Data accessibility.** Supplementary Information are available with the paper <http://dx.doi.org/10.1098/rstb.2015.0133>. The multiple-sequence alignment used for the analyses is available for download at [https://bitbucket.org/bzxdp/terrestrialisation\\_arthropoda](https://bitbucket.org/bzxdp/terrestrialisation_arthropoda).

**Authors' contributions.** J.L.-F. carried out analyses and wrote the manuscript. R.C. assembled the dataset and helped writing the manuscript. A.T. provided scripts for dataset assembly and helped writing the manuscript. J.O., M.B. and J.V. provided data and helped writing the manuscript. G.G. and G.E. contributed to write the manuscript. M.P. carried out ancestral character state reconstruction and helped writing the manuscript. DP contributed to all steps of the analyses and to writing the manuscript.

**Competing interests.** We confirm that we have no competing interests.

**Funding.** This work was supported by a Marie Skłodowska-Curie Fellowship to J.L.-F. R.C. was supported by a Science Foundation Ireland grant to D.P. (11/RFP/EOB/3106), A.R.T. was supported by a University of Bristol (STAR) PhD studentship. M.P. was supported by a NERC PhD studentship. Edinburgh Genomics is partially supported by core grants from (NERC R8/H10/56), MRC (MR/K001744/1) and BBSRC (BB/J004243/1). J.O. was supported by a grant from the Danish Agency for Science, Technology and Innovation (0601-12345B).

**Acknowledgements.** We thank the editors for having invited us to contribute to this issue of Philosophical Transactions of the Royal Society and two anonymous reviewers for providing constructive criticism. Thanks to staff at Edinburgh Genomics and the Geogenetics center in Copenhagen for help with sequencing the pycnogonid and *Limulus* transcriptomes.

## References

- Labandeira CC. 2005 Invasion of the continents: cyanobacterial crusts to tree-inhabiting arthropods. *Trends Ecol. Evol.* **20**, 253–262. (doi:10.1016/j.tree.2005.03.002)
- Shear WA. 1991 The early development of terrestrial ecosystems. *Nature* **351**, 283–289. (doi:10.1038/351283a0)
- Strother PK, Battison L, Brasier MD, Wellman CH. 2011 Earth's earliest non-marine eukaryotes. *Nature* **473**, 505–509. (doi:10.1038/nature09943)
- Clarke JT, Warnock R, Donoghue PCJ. 2011 Establishing a time-scale for plant evolution. *New Phytol.* **192**, 266–301. (doi:10.1111/j.1469-8137.2011.03794.x)
- Kenrick P, Wellman CH, Schneider H, Edgecombe GD. 2012 A timeline for terrestrialization: consequences for the carbon cycle in the Palaeozoic. *Phil. Trans. R. Soc. B* **367**, 519–536. (doi:10.1098/rstb.2011.0271)
- Redecker D, Kodner R, Graham LE. 2000 Glomalean fungi from the Ordovician. *Science*

- 289, 1920–1921. (doi:10.1126/science.289.5486.1920)
7. Little C. 1983 *The colonisation of land: origins and adaptations of terrestrial animals*, 300 p. Cambridge, UK: Cambridge University Press.
  8. Rota-Stabelli O, Daley AC, Pisani D. 2013 Molecular timetrees reveal a Cambrian colonization of land and a new scenario for ecdysozoan evolution. *Curr. Biol.* **23**, 392–398. (doi:10.1016/j.cub.2013.01.026)
  9. Richardson A, Araujo PB. 2015 Lifestyles of terrestrial crustaceans. In M Thiel, L Walting (eds), *The natural history of the Crustacea. Lifestyles and feeding biology*, pp. 299–336. New York, NY: Oxford University Press.
  10. Little C. 1990 *The terrestrial invasion: an ecophysiological approach to the origins of land animals*, 304 p. Cambridge, UK: Cambridge University Press.
  11. Dunlop JA, Scholtz G, Selden PA. 2013 Water-to-Land Transitions. In *Arthropod Biology and Evolution*, pp. 417–439. Berlin, Germany: Springer Berlin Heidelberg.
  12. Gordon MS, Olson EC. 1995 *Invasions of the land: the transitions of organisms from aquatic to terrestrial life*. New York, NY: Columbia University Press.
  13. Selden PA. 2001 Terrestrialization (Precambrian–Devonian). In *eLS*. Hoboken, NJ: John Wiley & Sons, Ltd. (doi:10.1038/npg.els.0001641)
  14. Niimura Y. 2009 On the origin and evolution of vertebrate olfactory receptor genes: comparative genome analysis among 23 chordate species. *Genome Biol. Evol.* **1**, 34–44. (doi:10.1093/gbe/evp003)
  15. Niimura Y, Nei M. 2005 Evolutionary dynamics of olfactory receptor genes in fishes and tetrapods. *Proc. Natl Acad. Sci. USA* **102**, 6039–6044. (doi:10.1073/pnas.0501922102)
  16. Vieira FG, Rozas J. 2011 Comparative genomics of the odorant-binding and chemosensory protein gene families across the Arthropoda: origin and evolutionary history of the chemosensory system. *Genome Biol. Evol.* **3**, 476–490. (doi:10.1093/gbe/evr033)
  17. Felsenstein J. 1985 Phylogenies and the comparative method. *Am. Nat.* **125**, 1–15. (doi:10.1086/284325)
  18. Stork NE, McBroom J, Gely C, Hamilton AJ. 2015 New approaches narrow global species estimates for beetles, insects, and terrestrial arthropods. *Proc. Natl Acad. Sci. USA* **112**, 7519–7523. (doi:10.1073/pnas.1502408112)
  19. Giribet G, Edgecombe GD. 2012 Reevaluating the arthropod tree of life. *Annu. Rev. Entomol.* **57**, 167–186. (doi:10.1146/annurev-ento-120710-100659)
  20. Rota-Stabelli O, Campbell L, Brinkmann H, Edgecombe GD, Longhorn SJ, Peterson KJ, Pisani D, Philippe H, Telford MJ. 2011 A congruent solution to arthropod phylogeny: phylogenomics, microRNAs and morphology support monophyletic Mandibulata. *Proc. R. Soc. B* **278**, 298–306. (doi:10.1098/rspb.2010.0590)
  21. Misof B *et al.* 2014 Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**, 763–767. (doi:10.1126/science.1257570)
  22. Borner J, Rehm P, Schill RO, Ebersberger I, Burmester T. 2014 A transcriptome approach to ecdysozoan phylogeny. *Mol. Phylogenet. Evol.* **80**, 79–87. (doi:10.1016/j.ympev.2014.08.001)
  23. Regier JC, Shultz JW, Zwick A, Hussey A, Ball B, Wetzer R, Martin JW, Cunningham CW. 2010 Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature* **463**, 1079–1083. (doi:10.1038/nature08742)
  24. Wägele JW, Kück P. 2013 Arthropod phylogeny and the origin of Tracheata (=Atelocerata) from Remipedia-like ancestors. In *Deep metazoan phylogeny: the backbone of the tree of life*, pp. 285–341. Berlin, Germany: De Gruyter.
  25. Wheeler WC, Giribet G, Edgecombe GD. 2004 Arthropod systematics. The comparative study of genomic, anatomical, and paleontological information. In *Assembling the tree of life*, pp. 281–295. New York, NY: Oxford University Press.
  26. Friedrich M, Tautz D. 1995 Ribosomal DNA phylogeny of the major extant arthropod classes and the evolution of myriapods. *Nature* **376**, 165–167. (doi:10.1038/376165a0)
  27. Cook CE, Smith ML, Telford MJ, Bastianello A, Akam M. 2001 *Hox* genes and the phylogeny of the arthropods. *Curr. Biol.* **11**, 759–763. (doi:10.1016/S0960-9822(01)00222-6)
  28. Pisani D, Poling LL, Lyons-Weiler M, Hedges SB. 2004 The colonization of land by animals: molecular phylogeny and divergence times among arthropods. *BMC Biol.* **2**, 1. (doi:10.1186/1741-7007-2-1)
  29. Mallatt JM, Garey JR, Shultz JW. 2004 Ecdysozoan phylogeny and Bayesian inference: first use of nearly complete 28S and 18S rRNA gene sequences to classify the arthropods and their kin. *Mol. Phylogenet. Evol.* **31**, 178–191. (doi:10.1016/j.ympev.2003.07.013)
  30. Meusemann K *et al.* 2010 A phylogenomic approach to resolve the arthropod tree of life. *Mol. Biol. Evol.* **27**, 2451–2464. (doi:10.1093/molbev/msq130)
  31. Boore JL, Lavrov DV, Brown WM. 1998 Gene translocation links insects and crustaceans. *Nature* **392**, 667–668. (doi:10.1038/33577)
  32. Zrzavý J, Štys P. 1997 The basic body plan of arthropods: insights from evolutionary morphology and developmental biology. *J. Evol. Biol.* **10**, 353–367. (doi:10.1046/j.1420-9101.1997.10030353.x)
  33. Richter S. 2002 The Tetraconata concept: hexapod-crustacean relationships and the phylogeny of Crustacea. *Org. Divers. Evol.* **2**, 217–237. (doi:10.1078/1439-6092-00048)
  34. Regier JC, Shultz JW, Kambic RE. 2005 Pancrustacean phylogeny: hexapods are terrestrial crustaceans and maxillopods are not monophyletic. *Proc. R. Soc. B* **272**, 395–401. (doi:10.1098/rspb.2004.2917)
  35. Regier JC *et al.* 2008 Resolving arthropod phylogeny: exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence. *Syst. Biol.* **57**, 920–938. (doi:10.1080/106351432#50802570791)
  36. Sharma PP, Kaluziak ST, Pérez-Porro AR, González VL, Hormiga G, Wheeler WC, Giribet G. 2014 Phylogenomic interrogation of Arachnida reveals systemic conflicts in phylogenetic signal. *Mol. Biol. Evol.* **31**, 2963–2984. (doi:10.1093/molbev/msu235)
  37. Rota-Stabelli O, Lartillot N, Philippe H, Pisani D. 2013 Serine codon-usage bias in deep phylogenomics: pancrustacean relationships as a case study. *Syst. Biol.* **62**, 121–133. (doi:10.1093/sysbio/sys077)
  38. Giribet G, Edgecombe GD, Wheeler WC. 2001 Arthropod phylogeny based on eight molecular loci and morphology. *Nature* **413**, 157–161. (doi:10.1038/35093097)
  39. Bernt M, Braband A, Middendorf M, Misof B, Rota-Stabelli O, Stadler PF. 2013 Bioinformatics methods for the comparative analysis of metazoan mitochondrial genome sequences. *Mol. Phylogenet. Evol.* **69**, 320–327. (doi:10.1016/j.ympev.2012.09.019)
  40. Rota-Stabelli O, Kayal E, Gleeson D, Daub J, Boore JL, Telford MJ, Pisani D, Blaxter M, Lavrov DV. 2010 Ecdysozoan mitogenomics: evidence for a common origin of the legged invertebrates, the Panarthropoda. *Genome Biol. Evol.* **2**, 425–440. (doi:10.1093/gbe/evq030)
  41. Nardi F, Spinsanti G, Boore JL, Carapelli A, Dallai R, Frati F. 2003 Hexapod origins: monophyletic or paraphyletic? *Science* **299**, 1887–1889. (doi:10.1126/science.1078607)
  42. Delsuc F, Phillips MJ, Penny D. 2003 Comment on ‘Hexapod origins: monophyletic or paraphyletic?’ *Science* **301**, 1482; author reply 1482. (doi:10.1126/science.1086558)
  43. Hassanin A. 2006 Phylogeny of Arthropoda inferred from mitochondrial sequences: strategies for limiting the misleading effects of multiple changes in pattern and rates of substitution. *Mol. Phylogenet. Evol.* **38**, 100–116. (doi:10.1016/j.ympev.2005.09.012)
  44. von Reumont BM *et al.* 2012 Pancrustacean phylogeny in the light of new phylogenomic data: support for Remipedia as the possible sister group of Hexapoda. *Mol. Biol. Evol.* **29**, 1031–1045. (doi:10.1093/molbev/msr270)
  45. Oakley TH, Wolfe JM, Lindgren AR, Zaharoff AK. 2013 Phylotranscriptomics to bring the understudied into the fold: monophyletic Ostracoda, fossil placement, and pancrustacean phylogeny. *Mol. Biol. Evol.* **30**, 215–233. (doi:10.1093/molbev/mss216)
  46. Ertas B, von Reumont BM, Wägele J-W, Misof B, Burmester T. 2009 Hemocyanin suggests a close relationship of Remipedia and Hexapoda. *Mol. Biol. Evol.* **26**, 2711–2718. (doi:10.1093/molbev/msp186)
  47. Fanenbruck M, Harzsch S, Wägele JW. 2004 The brain of the Remipedia (Crustacea) and an alternative hypothesis on their phylogenetic relationships. *Proc. Natl Acad. Sci. USA* **101**, 3868–3873. (doi:10.1073/pnas.0306212101)
  48. Fanenbruck M, Harzsch S. 2005 A brain atlas of *Godzillignomus frondosus* Yager, 1989 (Remipedia, Godzillidae) and comparison with the brain of *Speleonectes tulumensis* Yager, 1987 (Remipedia, Speleonectidae): implications for arthropod

- relationships. *Arthropod Struct. Dev.* **34**, 343–378. (doi:10.1016/j.asd.2005.01.007)
49. Stemme T, Iliffe TM, Bicker G, Harzsch S, Koenemann S. 2012 Serotonin immunoreactive interneurons in the brain of the Remipedia: new insights into the phylogenetic affinities of an enigmatic crustacean taxon. *BMC Evol. Biol.* **12**, 168. (doi:10.1186/1471-2148-12-168)
50. Glenner H, Thomsen PF, Hebsgaard MB, Sorensen MV, Willerslev E. 2006 Evolution: the origin of insects. *Science* **314**, 1883–1884. (doi:10.1126/science.1129844)
51. Maloof AC, Porter SM, Moore JL, Dudás Fő, Bowring SA, Higgins JA, Fike DA, Eddy MP. 2010 The earliest Cambrian record of animals and ocean geochemical change. *GSA Bull.* **122**, 1731–1774. (doi:10.1130/B30346.1)
52. Legg DA. 2014 *Sanctacaris uncata*: the oldest chelicerate (Arthropoda). *Naturwissenschaften* **101**, 1065–1073. (doi:10.1007/s00114-014-1245-4)
53. Waloszek D, Dunlop JA. 2002 A larval sea spider (Arthropoda: Pycnogonida) from the Upper Cambrian 'Orsten' of Sweden, and the phylogenetic position of pycnogonids. *Palaeontology* **45**, 421–446. (doi:10.1111/1475-4983.00244)
54. Harvey THP, Vélaz MI, Butterfield NJ. 2012 Exceptionally preserved crustaceans from western Canada reveal a cryptic Cambrian radiation. *Proc. Natl Acad. Sci. USA* **109**, 1589–1594. (doi:10.1073/pnas.1115244109)
55. Walossek D. 1993 *The Upper Cambrian Rebbachiella and the phylogeny of Branchiopoda and Crustacea*. Fossils and Strata no. 32, 202 p. Oslo, Norway: Scandinavian University Press.
56. MacNaughton RB, Cole JM, Dalrymple RW, Braddy SJ, Briggs DEG, Lukie TD. 2002 First steps on land: arthropod trackways in Cambrian–Ordovician eolian sandstone, southeastern Ontario, Canada. *Geology* **30**, 391–394. (doi:10.1130/0091-7613(2002)030<0391:FSOLAT>2.0.CO;2)
57. Collette JH, Gass KC, Hagadorn JW. 2012 *Protichnites eremita* unshelled? experimental model-based neoichnology and new evidence for a euthycarcinoid affinity for this ichnospecies. *J. Paleontol.* **86**, 442–454. (doi:10.1666/11-056.1)
58. Wilson HM, Anderson LI. 2004 Morphology and taxonomy of Paleozoic millipedes (Diplopoda: Chilognatha: Archipolypoda) from Scotland. *J. Paleontol.* **78**, 169–184. (doi:10.1666/0022-3360(2004)078<0169:MATOPM>2.0.CO;2)
59. Shear WA, Edgecombe GD. 2010 The geological record and phylogeny of the Myriapoda. *Arthropod Struct. Dev.* **39**, 174–190. (doi:10.1016/j.asd.2009.11.002)
60. Edgecombe GD. 2004 Morphological data, extant Myriapoda, and the myriapod stem-group. *Contrib. Zool.* **73**, 207–252.
61. Jeram AJ, Selden PA, Edwards D. 1990 Land animals in the Silurian: arachnids and myriapods from Shropshire, England. *Science* **250**, 658–661. (doi:10.1126/science.250.4981.658)
62. Scholtz G, Kamenz C. 2006 The book lungs of Scorpiones and Tetrapulmonata (Chelicerata, Arachnida): evidence for homology and a single terrestrialisation event of a common arachnid ancestor. *Zoology* **109**, 2–13. (doi:10.1016/j.zool.2005.06.003)
63. Dunlop JA, Anderson LI, Braddy SJ. 2003 A redescription of *Chasmataspis laurencii* Caster and Brooks, 1956 (Chelicerata: Chasmataspidida) from the Middle Ordovician of Tennessee, USA, with remarks on chasmataspid phylogeny. *Trans. R. Soc. Edinb. Earth Sci.* **94**, 207–225. (doi:10.1017/S0263593303000130)
64. Lamsdell JC. 2013 Revised systematics of Palaeozoic 'horseshoe crabs' and the myth of monophyletic Xiphosura. *Zool. J. Linn. Soc.* **167**, 1–27. (doi:10.1111/j.1096-3642.2012.00874.x)
65. Edwards D, Selden PA, Richardson JB, Axe L. 1995 Coprolites as evidence for plant–animal interaction in Siluro–Devonian terrestrial ecosystems. *Nature* **377**, 329–331. (doi:10.1038/377329a0)
66. Parry SF, Noble SR, Crowley QG, Wellman CH. 2011 A high-precision U–Pb age constraint on the Rhynie Chert Konservat-Lagerstätte: time scale and other implications. *J. Geol. Soc. Lond.* **168**, 863–872. (doi:10.1144/0016-76492010-043)
67. Rehm P, Borner J, Meusemann K, von Reumont BM, Simon S, Hadry H, Misof B, Burmester T. 2011 Dating the arthropod tree based on large-scale transcriptome data. *Mol. Phylogenet. Evol.* **61**, 880–887. (doi:10.1016/j.ympev.2011.09.003)
68. Rehm P, Meusemann K, Borner J, Misof B, Burmester T. 2014 Phylogenetic position of Myriapoda revealed by 454 transcriptome sequencing. *Mol. Phylogenet. Evol.* **77**, 25–33. (doi:10.1016/j.ympev.2014.04.007)
69. Brewer MS, Bond JE. 2013 Ordinal-level phylogenomics of the arthropod class Diplopoda (millipedes) based on an analysis of 221 nuclear protein-coding loci generated using next-generation sequence analyses. *PLoS ONE* **8**, e79935. (doi:10.1371/journal.pone.0079935)
70. Tong KJ, Duchêne S, Ho SYW, Lo N. 2015 Insect phylogenomics. Comment on 'Phylogenomics resolves the timing and pattern of insect evolution'. *Science* **349**, 487. (doi:10.1126/science.aaa5460)
71. Wheat CW, Wahlberg N. 2013 Phylogenomic insights into the Cambrian explosion, the colonization of land and the evolution of flight in Arthropoda. *Syst. Biol.* **62**, 93–109. (doi:10.1093/sysbio/sys074)
72. Campbell LI *et al.* 2011 MicroRNAs and phylogenomics resolve the relationships of Tardigrada and suggest that velvet worms are the sister group of Arthropoda. *Proc. Natl Acad. Sci. USA* **108**, 15 920–15 924. (doi:10.1073/pnas.1105499108)
73. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402. (doi:10.1093/nar/25.17.3389)
74. Edgar RC. 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797. (doi:10.1093/nar/gkh340)
75. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010 New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321. (doi:10.1093/sysbio/syq010)
76. Castresana J. 2000 Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540–552. (doi:10.1093/oxfordjournals.molbev.a026334)
77. Kück P, Meusemann K. 2010 FASconCAT: convenient handling of data matrices. *Mol. Phylogenet. Evol.* **56**, 1115–1118. (doi:10.1016/j.ympev.2010.04.024)
78. Lartillot N, Rodrigue N, Stubbs D, Richer J. 2013 PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst. Biol.* **62**, 611–615. (doi:10.1093/sysbio/syt022)
79. Lartillot N, Philippe H. 2004 A Bayesian mixture model for across-site heterogeneities in the amino acid replacement process. *Mol. Biol. Evol.* **21**, 1095–1109. (doi:10.1093/molbev/msh112)
80. Lartillot N, Lepage T, Blanquart S. 2009 PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286–2288. (doi:10.1093/bioinformatics/btp368)
81. Lepage T, Bryant D, Philippe H, Lartillot N. 2007 A general comparison of relaxed molecular clock models. *Mol. Biol. Evol.* **24**, 2669–2680. (doi:10.1093/molbev/msm193)
82. Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006 Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88. (doi:10.1371/journal.pbio.0040088)
83. Erwin DH, Laflamme M, Tweedt SM, Sperling EA, Pisani D, Peterson KJ. 2011 The Cambrian conundrum: early divergence and later ecological success in the early history of animals. *Science* **334**, 1091–1097. (doi:10.1126/science.1206375)
84. Paradis E, Claude J, Strimmer K. 2004 APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290. (doi:10.1093/bioinformatics/btg412)
85. Pagel M. 1994 Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc. R. Soc. B* **255**, 37–45. (doi:10.1098/rspb.1994.0006)
86. Lewis PO. 2001 A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* **50**, 913–925. (doi:10.1080/106351501753462876)
87. Stenderup JT, Olesen J, Glenner H. 2006 Molecular phylogeny of the Branchiopoda (Crustacea) – Multiple approaches suggest a 'diplostracan' ancestry of the Notostraca. *Mol. Phylogenet. Evol.* **41**, 182–194. (doi:10.1016/j.ympev.2006.06.006)
88. Olesen J. 2007 Monophyly and phylogeny of Branchiopoda, with focus on morphology and homologies of branchiopod phyllopodous limbs.

- J. Crustacean Biol.* **27**, 165–183. (doi:10.1651/S-2727.1)
89. Olesen J. 2009 Phylogeny of Branchiopoda (Crustacea)—character evolution and contribution of uniquely preserved fossils. *Arthropod Syst. Phylogeny* **67**, 3–39.
90. Wolfe JM, Hegna TA. 2014 Testing the phylogenetic position of Cambrian pancrustacean larval fossils by coding ontogenetic stages. *Cladistics* **30**, 366–390. (doi:10.1111/cla.12051)
91. Scourfield DJ. 1926 On a new type of crustacean from the Old Red Sandstone (Rhynie Chert Bed, Aberdeenshire)—*Lepidocaris rhyniensis*, gen. et sp. nov. *Phil. Trans. R. Soc. Lond. B* **214**, 153–187. (doi:10.1098/rstb.1926.0005)
92. Fayers SR, Trewin NH. 2002 A new crustacean from the Early Devonian Rhynie chert, Aberdeenshire, Scotland. *Earth Environ. Sci. Trans. R. Soc. Edinb.* **93**, 355–382. (doi:10.1017/S0263593302000196)
93. Novozhilov NI. 1957 Un nouvel ordre d'arthropodes particuliers: Kazacharthra du Lias des monts Ketmen (Kazakhstan, SE., URSS). *Bull. Soc. Géol. Fr.* **7**, 171–184.
94. Mathers TC, Hammond RL, Jenner RA, Hänfling B, Gómez A. 2013 Multiple global radiations in tadpole shrimps challenge the concept of 'living fossils'. *PeerJ* **1**, e62. (doi:10.7717/peerj.62)
95. Bapst DW. 2012 paleotree: an R package for paleontological and phylogenetic analyses of evolution. *Methods Ecol. Evol.* **3**, 803–807. (doi:10.1111/j.2041-210X.2012.00223.x)
96. Smith MR, Ortega-Hernández J. 2014 *Hallucigenia's* onychophoran-like claws and the case for Tactopoda. *Nature* **514**, 363–366. (doi:10.1038/nature13576)
97. Dunn CW *et al.* 2008 Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature* **452**, 745–749. (doi:10.1038/nature06614)
98. Laumer CE *et al.* 2015 Spiralian phylogeny informs the evolution of microscopic lineages. *Curr. Biol.* **25**, 2000–2006. (doi:10.1016/j.cub.2015.06.068)
99. Fernández R, Laumer CE, Vahtera V, Libro S, Kaluziak S, Sharma PP, Pérez-Porro AR, Edgecombe GD, Giribet G. 2014 Evaluating topological conflict in centipede phylogeny using transcriptomic data sets. *Mol. Biol. Evol.* **31**, 1500–1513. (doi:10.1093/molbev/msu108)
100. Paradis E. 2012 *Analysis of phylogenetics and evolution with R*, 2nd edn. 386 p. New York, NY: Springer.
101. Pisani D, Liu AG. 2015 Animal evolution: only rocks can set the clock. *Curr. Biol.* **25**, 1079–1081. (doi:10.1016/j.cub.2015.10.015)
102. dos Reis M, Thawornwattana Y, Angelis K, Telford MJ, Donoghue PCJ, Yang Z. 2015 Uncertainty in the timing of origin of animals and the limits of precision in molecular timescales. *Curr. Biol.* **25**, 2939–2950. (doi:10.1016/j.cub.2015.09.066)
103. Legg DA, Sutton MD, Edgecombe GD. 2013 Arthropod fossil data increase congruence of morphological and molecular phylogenies. *Nat. Commun.* **4**, 2485. (doi:10.1038/ncomms3485)
104. Battistuzzi FU, Billing-Ross P, Murillo O, Filipski A, Kumar S. 2015 A protocol for diagnosing the effect of calibration priors on posterior time estimates: a case study for the Cambrian explosion of animal phyla. *Mol. Biol. Evol.* **32**, 1907–1912. (doi:10.1093/molbev/msv075)
105. Wray GA, Levinton JS, Shapiro LH. 1996 Molecular evidence for deep Precambrian divergences among metazoan phyla. *Science* **274**, 568–573. (doi:10.1126/science.274.5287.568)
106. Shelley RM, Golavatch SI. 2011 Atlas of myriapod biogeography. I. Indigenous ordinal and supra-ordinal distributions in the Diplopoda: Perspectives on taxon origins and ages, and a hypothesis on the origin and early evolution of the class. *Insecta Mundi* **158**, 1–134.
107. Dohle W. 1998 Myriapod–insect relationships as opposed to an insect–crustacean sister group relationship. In *Arthropod relationships*, pp. 305–315. London, UK: Chapman & Hall.