THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Horizontal gene transfer in the sponge *Amphimedon queenslandica*

Simone Summer Higgie
BEnvSc (Honours)

A thesis submitted for the degree of Doctor of Philosophy at The University of Queensland in 2018

School of Biological Sciences

Abstract

Horizontal gene transfer (HGT) is the nonsexual transfer of genetic sequence across species boundaries. Historically, HGT has been assumed largely irrelevant to animal evolution, though widely recognised as an important evolutionary force in bacteria. From the recent boom in whole genome sequencing, many cases have emerged strongly supporting the occurrence of HGT in a wide range of animals. However, the extent, nature and mechanisms of HGT in animals remain poorly understood. Here, I explore these uncertainties using 576 HGTs previously reported in the genome of the demosponge *Amphimedon queenslandica*.

The HGTs derive from bacterial, plant and fungal sources, contain a broad range of domain types, and many are differentially expressed throughout development. Some domains are highly enriched; phylogenetic analyses of the two largest groups, the Aspzincin_M35 and the PNP_UDP_1 domain groups, suggest that each results from one or few transfer events followed by post-transfer duplication. Their differential expression through development, and the conservation of domains and duplicates, together suggest that many of the HGT-derived genes are functioning in *A. queenslandica*. The largest group consists of aspzincins, a metallopeptidase found in bacteria and fungi, but not typically in animals. I detected aspzincins in representatives of all four of the sponge classes, suggesting that the original sponge aspzincin was transferred after sponges diverged from their last common ancestor with the Eumetazoa, but before the contemporary sponge classes emerged. In *A. queenslandica*, the aspzincins may have been co-opted for multiple functions, since 54 of the 90 fit into one of four ontogenetic expression profiles, each of which is putatively co-expressed with different suites of native genes. Based on secretion signals and the conservation of key catalytic residues, I propose that proteolytic activity is maintained in at least one of the aspzincin roles in *A. queenslandica*.

Mobile elements are capable of genomic excision, movement and integration, they enable HGT in bacteria, and some animal HGTs are genomically close to eukaryotic transposable elements (TEs); as such, mobile elements are speculated as possible players in the mechanisms of interkingdom and interdomain HGT to animals. In *A. queenslandica*, the overall repeats densities around predicted unduplicated HGTs are not different to those around predicted unduplicated native genes. However,

the surrounding repeats content of unduplicated HGTs has slightly increased proportions of the *helitron* DNA transposon and of simple repeats. Further, 29% of the HGT-derivatives are TEs, half of which are unknown in class, a quarter are *copia* long terminal repeats retrotransposons and the other quarter are *helitrons*. These data suggest that repeats and TEs may have putative roles in the HGT process in animals, such as simple repeats possibly conferring increased chances of genomic integration through recombination and *helitrons* perhaps increasing chances of HGT functionality via their reservoir of regulatory elements. Another seven per cent of the HGT-derivatives have high sequence similarities to proteins present on bacterial plasmids, suggestive that plasmids were involved in some of the transfers. Together, these results offer novel insight on HGT in *A. queenslandica* and demonstrate that HGT has a varied nature in this animal with an extensive impact, partially due to post-transfer duplications. Further, this work offers insights on possible mechanisms that led to the HGT derived genes of this sponge.

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, financial support and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my higher degree by research candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis and have sought permission from co-authors for any jointly authored works included in the thesis.

Publications during candidature

Conference Abstracts

Higgie, S. S. and S. M. Degnan, 2016. Trans-domain horizontal gene transfer from bacterium to animal: the post-transfer evolution of the aspzincins in the sponge *Amphimedon queenslandica*. Society for Molecular Biology and Evolution (SMBE), Gold Coast, QLD, Australia, 3-7 July 2016.

Publications included in this thesis

No publications included.

Contributions by others to the thesis

Sandie M. Degnan contributed to the conception and design of this research, advised on methods and analysis, and provided critical comments on the thesis.

Selene L. Fernandez-Valverde produced several of the datasets analysed in this thesis (as specifically acknowledged in the relevant chapters).

William L. Hatleberg produced the genome-wide BLAST2Go and Pfam annotations for Aqu2.1 gene models, as specifically acknowledged in Chapter 3. William L. Hatleberg also modified a published script that was used to test for Pfam domain enrichments (as specifically acknowledged in Chapter 3).

Federico Gaiti developed the script used in Chapter 3 for the co-expression analysis (as specifically acknowledged in Chapter 3).

Statement of parts of the thesis submitted to qualify for the award of another degree

None.

Research involving human or animal subjects

No animal or human participants were involved in this research.

Acknowledgements

For the last seven years, I have been a member of the Degnan Laboratories in a few roles, so it is with great warmth (and disbelief!) that I write these acknowledgements.

First, to my advisor Sandie Degnan. Thank you for sharing your expertise and for all your help, guidance, and enthusiasm during my project. You have given me a lot of patience, freedom, and time to find my way and I appreciate that. I have always respected your seemingly endless supply of curiosity and passion for science and feel privileged to have you as a role model in both science and life, thank you for your continued support.

Thank you to my co-advisor Sassan Asgari, I appreciate your time and comments. I am also very appreciative to my readers Mark Ragan and Matt Sweet for always being at important milestones throughout my project, for their critique, suggestions, and encouragement. In addition, thank you to my committee chair Jan Engelstädter for his time and for providing valuable feedback.

I am grateful to the Australian Government for an Australian Postgraduate Award and thus the financial opportunity to undertake this quest for greater understanding. I also appreciate a travel award from the School of Biological Sciences, which enabled me to attend the Society for Molecular Biology and Evolution (SMBE) 2016 conference. Thanks also to Gail Walter, the efficient and patient postgraduate administrator for the School of Biological Sciences.

To both joint lab Heads, Sandie and Bernie Degnan, thank you for developing my research questions, skills and career over the last seven years. I am grateful for the world-class research facilities, environment and opportunities you have given me, including two weeks in the field on the beautiful Heron Island of the Great Barrier Reef and a trip to Japan to attend the 12th ISDCI congress. Related, a huge thanks to all past and present Deglabsters for your large part in making the lab a vibrant and diverse community, both scientifically and socially. Laura, Federico, Andrew, Carmel, William, Tahsha, Kerry, Selene, Jabin, Kevin, Bec, Ben, Aude, Maely, Katia, Daniel, Shun, Jo, Felipe, Jaret, Nobuo, Gemma, Nick, Markus,

Claire, Mel, Yasu, Emily, Arun, Eunice, Xueyan, Romy, and Lisa – your comradeship, assistance and advice with all things science, computers, and life in general are much appreciated.

More specifically, I am hugely grateful to Selene Fernandez-Valverde for her invaluable help with the mysterious "black screen", that is, for her bioinformatics and programming expertise and help! Special thanks also to Sandie and Selene for our collaborative HGTracker project, the results of which formed the starting point of my thesis. To Laura, Carmel, William, Andrew, and Fede, thank you for your general computing and analysis help. Laura, you have always been so helpful with analysis ideas, interpretations, and problem-solving: thank you for your time and help. Also, thanks Laura and Ben for reading drafts and taking the time to give thoughtful and helpful feedback – I really appreciate it. William and Federico, thanks for your scripts. Yet another thank you to William, for your help with anything and everything, including your help with InDesign and formatting. Andrew – thanks for your early help with analyses for transposons and the "aspzicvndlknregokm"! Kerry, Ben, Tahsha and Carmel, thank you for your assistance during my foray in the wet lab. Further, a heartfelt thanks to Kerry for being such an amazing, efficient and kind lab manger, as well as a wonderful role model. Your integrity, kindness and hard-working nature is truly inspiring. Thank you too for starting to take the aspzincin story even further into exciting territory with *in situ* hybridisation.

Over my last ten years (!) at the University of Queensland I have had the good fortune of associating with incredible researchers, some of whom I now consider mentors. Gimme Walter, Claudia Vickers and Daniel Ortiz-Barrientos, you have all inspired and encouraged me, and have developed my thinking as a scientist and person. I have immense respect for each of you and will always be grateful to you all.

To my family and friends in general: thank you for being interested in my research, for your support and for your understanding when I was mentally and/or physically absent with no excuse but "my PhD". To my aunt Dell – a sincere thank you for your incredible support during my whole time at university and also for setting an inspiring example in your life and career. To Daniel, a deep thank you for being so supportive when I decided to start my PhD, as well as for all your tremendous support throughout my undergraduate and honours. Jabin – thanks for your friendship and for showing me how to toughen up and grow thicker skin. A huge shout out to my climbing friends in Brisbane, especially Adnan, Shessy,

Robbie, Davide, Sandra, Karina, Bernie, Amanda, Andy, Jai, Ross, Ella W. and Frank! You guys were a constant and refreshing source of positivity, fun and support. Alongside you, I faced fears, pushed hard and grew a lot – all of that has permeated every other aspect of my life.

The last five years have presented the biggest personal upheavals and challenges of my life. I cannot fully express my gratitude for the wonderful people who I consider "my Brisbane family" and who kept me going through some tough times. Ellia, William, Tobias, Lauren, Cody, Christina, Ariel and Skye – thank you so very much. You are all incredible people, scientists, and friends; your help, support, unwavering loyalty, and wisdom mean the world to me. Importantly, the last five years have also presented my biggest highs to date, often thanks to you guys and none of which would have been as great if I couldn't have shared it with you.

Ivan, as I write this at 10:37pm on a Saturday night, you are next door sculpting. Your passion, dedication, and hard-working nature has often fuelled my energy and inspired me to keep on at it. Thank you for enduring the long distance part of our relationship and for all your support and patience.

Finally, to my parents and my grandparents: merci buckets! Maman and Dad, or Nicole and Clive Higgie, all my life you have been fine examples of the importance in being interested and passionate. Thank you for your support, love, and friendship. To my Grandpère, Merwyn Norrish, I am so lucky to have you as my grandfather and friend. You, your ethics and your incredible career have often motivated me: thank you for all your support. I am grateful to my other grandparents too; they supported my studies and always encouraged me to work hard and to be my best self. Both my grandmothers were passionate about higher education for women and I always had them sitting on my shoulders as I worked. I dedicate my thesis to my Grandmère, Françoise Céline Norrish, née Honoré.

Financial support

This research was supported by an Australian Government Research Training Program Scholarship.

Keywords

horizontal gene transfer, animal, aspzincin, evolution, mobile elements, sponge, metallopeptidase, duplication, domain architecture, transposable elements

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 060409, Molecular Evolution, 40%

ANZSRC code: 060405, Gene Expression, 30%

ANZSRC code: 060309, Phylogeny and Comparative Analysis, 30%

Fields of Research (FoR) Classification

FoR code: 0603, Evolutionary Biology, 45%

FoR code: 0604, Genetics, 45%

FoR code: 0608, Zoology, 10%

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Definition
aa	Amino acid
AAA	Adenosine 5'-triphosphatase associated with diverse cullular activites
ACLAME	A classification of mobile genetic elements
AI	Alien index
AIG1	AvrRpt2 induced gene 1 domain
Ank	Ankyrin repeat
ApoL	Apolipoprotein L domain
Aq	*Amphimedon queenslandica*
AqAspz	*A. queenslandica* aspzincin
AqAspzsSc	AqAspz that show co-localisation with silicatein
AqHGT	HGTracker-identified HGT in *A. queenslandica*
AqHGT_Conaco	*A. queenslandica* HGT as identified by Conaco et al. (2016)
AqHGT_NM	*A. queenslandica* HGT containing a nonmetazoan domain
AqHGT_PNP	*A. queenslandica* horizontally transferred PNP_UDP_1 domain containing gene
AqPNP	*A. queenslandica* PNP domain-containing genes
AqSilicatein	Silicatein genes of *A. queenslandica*
Aqu1	First generation gene model predictions of *A. queenslandica*
Aqu2.1	Second generation gene model predictions of *A. queenslandica*
AS	Aspartic proteinase
AsaP1	Aspzincin peptidase of *Aeromonas salmonicida* subsp. *achromogenes*
Aspz	Aspzincin domain
Aspzincin_M35	Aspzincin catalytic domain
ATPase	Adenosine 5'-triphosphatase
BIC	Bayesian Information Criterion
BLAST	Basic Local Alignment Search Tool
BLASTn	Nucleotide Basic Local Alignment Search Tool
BLASTp	Protein Basic Local Alignment Search Tool
BLIND	Basic linear index determination of transcriptomes
bp	Base pair
CA	Cysteine clan A
CArG	DNA promoter sequence
CATH	Class Architecture Topology Homologous superfamily database
CD	Cysteine clan D
CDD	Conserved Domain Database
CDF	Cation diffusion facilitator
cDNA	Complementary DNA
CDS	Coding DNA sequence

CEL-Seq	Cell Expression by Linear amplification and Sequencing
CMS-EnSpm	DNA transposon family
COR	C-terminal of Roc domain
CR1	Chicken repeat 1 non-LTR retrotransposon
DAPI	4',6-diamididino-2-phenylindole fluorescent stain
db	Database
DDE	Superfamily endonuclease
DDE	Endonuclease domain
DEAD/DEAH	helicase domain
DNA	Deoxyribonucleic acid
DSB	Double-strand break
DUF	Domain of unknown function
e	Expect
E1_DerP2_DerF2	MD-2-related lipid-recognition domain found in Epididymal secretory protein E1 and Der p 2 and Der f 2 proteins
ECM	Extracellular matrix
EGT	Endosymbiotic gene transfer
ENCODE	Encyclopedia of DNA elements
env	Envelop-like gene of retroviruses
EST	Expressed sequence tag
FHV	Flock house virus
gag_pre-intergs	GAG-pre-integrase domain
GH5	Cellulose GH5 domains
GI	Genomic islands
GlyzipOmp	Glycine zipper domain
GlyzipOmpA	Glycine zipper domain
GlyzipYMGG	Glycine zipper domain
GO	Gene ontology
hAT-Ac	DNA transposon family
Helitron_like_N	Helitron helicase-like domain at N terminus
HGT	Horizontal gene transfer
HhMAN1	Mannanase gene
HIV	Human immunodeficiency virus
HMM	Hidden Markov model
hpe	Hours post emergence
HTH	Helix-turn-helix domain
HTT	Horizontal transfer of transposable elements
INT	Integrase
IRT	Inverted terminal repeat
IS	Insertion sequence
JGI	Joint Genome Institute

kb	Kilobase
kbp	Kilo base pair
kDa	Kilo dalton
LCA	Last common ancestor
LF	Lobe-finned fishes
LGT	Lateral gene transfer
LINE	Long interspersed nuclear elements
LRAT	Lecithin retinol acyltransferase domain
LRR	Leucine rich repeats
LTR	Long terminal repeat
LUD	Likely unduplicated
M	Metallopeptidase family
MA	Metallopeptidase clan A
MAFFT	Multiple Alignment using Fast Fourier Transform
ME	Mobile element
MH	Metallopeptidase clan H
ML	Maximum Likelihood
MMP	Matrix metalloproteinase family
MP	Metallopeptidase
mRNA	Messenger RNA
mya	Million years ago
n	Number
NAD	Nicotinamide adenine dinucleotide
NCBI	National Centre for Biotechnology Information
NHEJ	Nonhomologous end joining
NLS	Nuclear localisation signal
nonHGT	A gene that is not a predicted HGT (but could be contamination, ambiguous or native in a genome)
nr	Nonredundant
NTPase	Nucleoside-triphosphatase
ORF	Open reading frame
p	Probability
P-type ATPase	Primary transporters superfamily
PA	Mixed endopeptidase clan PA
PCR	Polymerase chain reaction
Pfam	Protein family
PIF1	Helicase domain
PNP_UDP_1	Phosphorylase domain
Poly-A	Polyadenylated adenosine
PPOD	Putative peroxidase protein
RGD	Arginine-glycine-aspartate motif

RH	Rnase H
Rhs	Recombination hotspot
RHS	RGD helper sequence
RNA	Ribonucleic acid
RNA-Seq	RNA sequencing
RND	Resistance-nodulation-cell division superfamily
RT	Reverse transcriptase
RT-qPCR	Real time quatitative polymerase chain reaction
rve	integrase core domain
RVT	Reverse transcriptase domain
SINE	Short interspersed nuclear elements
SP	Signal peptide
SR	Simple repeat
SRCR	Scavenger receptor cysteine-rich
SseC	Secretion system effector C
SUD	Strictly unduplicated
t	Test
T-DNA	Transferring DNA
T4SS	Type IV secretion system
T6SS	Type VI secretion system
TATA	DNA promoter sequence
taxid	Taxonomic identity
TCp3.2	Member of the Tc1/mariner-like transposable element superfamily
TE	Transposable element
TED	Transposon Ellen Dempsey
Ti	Tumor inducing
TIG	Transcription factor immunoglobulin-like fold domain
TIMP	Tissue inhibitor of metallopeptidases
Tox-HDC	Toxin with an H, D/N and C signature
TPR	Tetratricopeptide repeat
Ty1	retrotransposon
UvrD_C_2	UvrD-like helicase C-terminal domain
v	Version
vir	Bacterial virulence genes
WD	Short structural motif often ending a tryptophan-aspartic acid (W-D) dipeptide
WLM	Wss1p-like metalloprotease domain
WMISH	Whole mount in situ hybridisation
yHGT	Putatively younger horizontal gene transfer
ZU5	Domain found in zona occludens 1 proteins, unc5-like netrin receptors and ankyrins

Chapter 1 - Introduction

1.1 The acquisition of genetic variation and novelty

Understanding the acquisition of genetic variation is a central interest to biologists, since variation is the raw fuel for evolution and thus broadly underlies much of evolutionary biology (Carroll 2008; Brakefield 2011; Moczek et al. 2015). Further, reconciling the astonishing morphological variation in life with the underlying molecular diversity is an intriguing challenge since at the molecular level, life forms on Earth are strikingly similar (King and Wilson 1975; Carroll 2008; Brakefield 2011; Cleland 2013; Moczek et al. 2015). Various biological disciplines are showing that variation is created in a multitude of processes, far beyond our classical neo-Darwinism understanding of variation being generated from nucleotide substitutions, insertions, and deletions (Koonin 2009a; Brakefield 2011; Wijayawardena et al. 2013).

Genetic variation can arise from the modification, co-option or duplication of existing coding sequence (Haldane 1933; Fisher 1935; Muller 1936; Ohno 1970; Andersson et al. 2015). Transposable elements (TEs) can be reservoirs of genetic variation, since they have potentially useful enzymatic machinery with abilities for binding, cutting and interacting with DNA and proteins, and have been reported as recruited and co-opted in a diverse range of eukaryotes (Miller et al. 1999; Volff 2006; Sinzelle et al. 2009; Piskurek and Jackson 2012). The variation in sexually reproducing populations is influenced by recombination and random segregation during meiosis, gene flow, hybridisation, and introgression (Barrett and Schluter 2008; Masel and Trotter 2010; Hoffmann and Sgrò 2011). Further, eukaryotes obtain genetic diversity through post-transcriptional modifications such as alternative splicing, which results in differential exon combinations (Matlin et al. 2005; Blencowe 2006). While thought rare, novel genes themselves can be born *de novo* from noncoding sequences (Ding et al. 2012; Andersson et al. 2015; Dunn and Ryan 2015). Variation in phenotype is subject to natural selection and often large changes in morphological form result from differential expression of the same basic genetic tool kit via modification of regulatory regions (Carroll 2008; Brakefield 2011; Wenger and Galliot 2013; Dunn and Ryan 2015). Bacteria are unicellular organisms that do not undergo sexual reproduction; rather,

they produce clonal replicates by binary fission and thus do not generate genetic variation via meiotic recombination (Gogarten et al. 2002). In addition to variation generated from mutation, bacteria also gain variation by acquiring DNA from other individuals and the environment – such horizontal gene transfer (HGT) is the nonsexual transfer and integration of genetic sequence between species and can directly transmit novel genes of immediate adaptive significance (Beiko et al. 2005; Polz et al. 2013; Soucy et al. 2015).

In sum, these are some of the many ways in which variation is gained; yet, many of these processes described involve either slow and small changes, or the modification of existing functions. Evolution was once thought to be a slow and gradual process; however, historic records such as fossils have influenced more recent theories that suggest in fact, there are long periods of stability punctuated with rapid bursts of change – termed "punctuated equilibrium" (Eldredge and Gould 1972; Eldredge and Eldredge 1977; Zeh et al. 2009). How then can these slow and small variation-generating processes lead to the diverse morphology that we observe in life and enable rapid evolutionary changes? Certainly, as mentioned, evolutionary developmental research is showing that large changes can result from modifications in gene regulation. In addition, comparative genomic data across broad taxa are also showing that life is less orderly and more dynamic than previously thought, and that genomes in all kingdoms are not isolated entities but are interconnected (Koonin 2009a; Schaack et al. 2010; Wijayawardena et al. 2013). HGT is a mechanism accepted to be driving such movement and connectedness in bacteria; however, until recently, it was assumed that both eukaryotic cells and the multicellular condition posed too many barriers for foreign DNA to become incorporated into eukaryotic heritable genomes (Gladyshev and Arkhipova 2009; Koonin 2009a; Whitaker et al. 2009; Raoult 2010; Marcet-Houben and Gabaldón 2010; Wybouw et al. 2012; Wijayawardena et al. 2013; Boto 2015). These assumptions are being challenged now, as reports of HGTs in eukaryotes are fast growing, though we do not yet understand how this occurs or the extent of its impact on eukaryotic evolution. These gaps are a piece in the longstanding and fascinating puzzle of understanding how genomic variation is created.

1.2 Research context

Comparisons of genomes across broad taxa are revealing levels of genomic connectedness from HGT amongst diverse species that are not compatible with the tree of life concept, which suits a strict model of vertical descent of genes in species through time (Boto 2015; Soucy et al. 2015). Rather, a web of life is proposed as more appropriately representing the apparent fluidity of genes across species boundaries (Koonin 2009b; Raoult 2010; Wijayawardena et al. 2013; Soucy et al. 2015). HGT is a recognised driver of such connectedness in bacteria, and increasing levels of DNA sequencing over the last decade has uncovered data best explained as HGT in animals (see cases in Table 1.1; Gladyshev et al. 2008; Dunning Hotopp 2011; Wijayawardena et al. 2013). These cases have demanded a change in the historic assumptions made about the irrelevance of HGT to animals; however, both the nature and extent of HGT in animals are not well understood.

For my thesis, I explore genes computationally identified as likely HGTs in a morphologically simple animal, the demosponge *Amphimedon queenslandica,* with a particular focus on a large group of bacterial-like metalloendopeptidase aspzincin genes. A better understanding of the putative HGTs of *A. queenslandica* and their impact on their host is both important and fascinating since the notion of HGT in animals has a contentious past and is still not well understood. In addition, my work further develops biological understanding of the emerging model species *A. queenslandica*, which is a representative of one of the oldest extant lineages of animals (Srivastava et al. 2010; Ryan et al. 2013), and thus holds a key phylogenetic position for comparative genomics. Last, more broadly, my PhD project contributes to a more robust understanding of HGT in animals.

1.3 Horizontal gene transfer

The horizontal transfer of genetic material across species boundaries increases genetic variation, can directly transmit novel functions of adaptive significance, and provides sequence from which new genes and regulatory elements can arise (Dunning Hotopp et al. 2007; Baños et al. 2009; Koonin 2009a; Boto 2010; Schaack et al. 2010; Anderson and Seifert 2011; Gophna and Ofran 2011; Matveeva et al. 2012; Wheeler et al. 2013). The terms horizontal and lateral gene transfer (HGT and LGT, HGT hereafter) are used interchangeably in the literature and reflect the movement of sequence between individuals “horizontally”, as opposed to vertically from parent to offspring (Keeling and Palmer 2008;

Park and Zhang 2012). The process involves DNA from an exogenous source entering a new host cell and becoming incorporated into the genome of the new host (Ochman et al. 2000; Cooper 2014). If the cell is eukaryotic, the DNA must also pass through the nuclear envelop (Danchin et al. 2010). To have any evolutionary impact, the new DNA must be heritable; therefore, if the new host reproduces sexually, the receiving host cell must be in the germline (Keeling 2009; Danchin et al. 2010; Pauchet and Heckel 2013; Boto 2014; Crisp et al. 2015). Upon integration into the host genome, HGTs may be expressed and selected for, thereby becoming part of the functional genome (Park and Zhang 2012; Baltrus 2013; Soucy et al. 2015). Alternatively, expressed HGTs may have detrimental fitness costs and may experience negative selection (Park and Zhang 2012; Baltrus 2013; Soucy et al. 2015). Finally, HGTs may not be expressed, though still offer sequence and/or structural variation to the host genome (Gophna and Ofran 2011; Baltrus 2013; Soucy et al. 2015).

1.3.1 HGT in bacteria

HGT was first described by Griffith (1928) in a study that revealed the direct transfer of virulence from virulent to non-virulent bacteria by an unknown factor. The transferring factor was later identified as DNA (Avery et al. 1944) and before long, a detailed examination of HGT between bacteria was presented (Tatum and Lederberg 1947). By 1960, HGT was recognised as a significant evolutionary factor in multidrug resistance in bacteria (Akiba et al. 1960; Davies and Davies 2010). Now, HGT is accepted as a major bacterial evolutionary force that partially drives the amazing diversity in bacterial metabolic properties, lifestyles and cellular structures (Ochman et al. 2000; Dunning Hotopp 2011; Baltrus 2013). HGT greatly impacts the adaptability of bacteria particularly since their clonal reproduction via binary fission generates little genetic variation (Gogarten et al. 2002; Keeling 2009; Soucy et al. 2015). The benefits of HGT to bacteria span from the refinement of enzymatic functions to major shifts in ecological lifestyles, with a large diversity of gene types reported as HGTs across a wide array of bacterial species, including genes involved in metabolism, biosynthesis, transcription and translation (Gogarten et al. 2002; Boucher et al. 2003; Baltrus 2013). As well, HGT has been implicated in the evolution of major bacterial physiological processes, including aerobic respiration, quorum sensing, nitrogen fixation, photosynthesis, and sulphate reduction (Boucher et al. 2003).

There are three well-established mechanisms by which HGT occurs in bacteria – the process can occur by recombination following transformation, conjugation or transduction (Ochman et al. 2000; Soucy et al. 2015). Transformation is the uptake of naked exogenous DNA from the environment (Ochman et al. 2000; Soucy et al. 2015). This process has been reported in both bacteria and archaea, and enables transferal of DNA from very distantly related organisms (Soucy et al. 2015). Conjugation requires direct physical contact of the donor and recipient cells by a conjugation pilus or bridge (Ochman et al. 2000; Soucy et al. 2015). The transferring sequence, typically a plasmid, travels through the pilus (Ochman et al. 2000; Soucy et al. 2015). Some chromosomal sequences can also be transferred, by either plasmids that can integrate with the recipient chromosome or by conjugative transposons, which contain genes for their excision, for the pilus formation and for their transposition (Ochman et al. 2000). Conjugation allows bacterium to bacterium transfers, but in addition, it can occur naturally from *Agrobacterium* species to plants (Buchanan-Wollaston et al. 1987) and in laboratory experiments, from bacterium to yeast (Heinemann and Sprague 1989). The third well-known mechanism is transduction, an effect of phage predation (Ochman et al. 2000; Soucy et al. 2015). Sequence is transferred to the receiving cell when the phage has previously replicated within a donor cell and in the process, has also packaged donor DNA fragments into its own genome (Ochman et al. 2000). This mechanism is documented for bacteria and archaea and the transmission possibilities depend on the phage receptors (Ochman et al. 2000; Soucy et al. 2015).

1.3.2 HGT in animals

Until recently, the chances of HGT in eukaryotes were presumed too low because of the greater separation of foreign DNA from the genome by the additional barriers of the nuclear membranes, and for many eukaryotes and animals, the sequestration of the germline (Gladyshev et al. 2008; Koonin 2009a; Koonin 2009b; Whitaker et al. 2009; Raoult 2010; Marcet-Houben and Gabaldón 2010; Wybouw et al. 2012; Wijayawardena et al. 2013). Further, the post-transfer barrier of vastly different genomic environments between putative transfer partners was viewed as too great (Gladyshev et al. 2008; Koonin 2009b; Whitaker et al. 2009; Raoult 2010; Marcet-Houben and Gabaldón 2010; Wybouw et al. 2012; Wijayawardena et al. 2013). Consequently, foreign sequence in animal genome data was often presumed to be contamination and was discarded, thus leaving less signs of HGT in animal genomes, thereby further supporting the initial assumptions (Dunning Hotopp 2011).

Table 1.1 Details of some relevant putative HGT cases in animals

Animal recipient	Donor organism	Sequence transferred	Detection method	Reference
Rotifers *Adineta vaga* and *A. ricciae*	Bacteria, fungi, plants	Thousands of genes	Sequence similarity, PCR, transcriptomics	Gladyshev et al. 2008; Boschetti et al. 2012; Flot et al. 2013
Hydra magnipapillata	Bacteria	Many, including PPOD (protein component of the cuticle)	Sequence similarity based genome survey, phylogeny, correlation of gene and phenotype	Chapman et al. 2010; Böttger et al. 2012
Silkmoth *Bombyx mori*	Bacteria, plant, fungi	Ten enzymes	Sequence similarity	Zhu et al. 2011; Wheeler et al. 2013
Gall midges including *Asteromyia carbonifera*	Fungal	Carotenoid biosynthesis gene homologues: one lycopene cyclase/phytoene synthase and two phytoene desaturase homologues	Transcriptomics, PCR, phylogeny	Cobbs et al. 2013
Spider mite *Tetranychus uriticae*	Fungal	Carotenoid biosynthesis gene homologues cyanase	Phylogeny	Grbić et al. 2011; Wybouw et al. 2012; Wybouw et al. 2014
Coffee borer beetle *Hypothenemus hampei*	Bacillus clade of bacteria	Mannanase (HhMAN1 gene) that allows the beetle to metabolise galactomannan, the major storage polysaccharide coffee berries	Functional support, eukaryotic gene traits, PCR, correlation of phenotype with HGT	Acuña et al., 2012
Tunicate *Ciona intestinalis*	Bacteria	Cellulose synthase gene	Sequence similarity, fluorescence in situ hybridisation	Nakashima et al. 2004
Mustard leaf beetle *Phaedon cochleariae*	Bacteria	Two xylanase genes	Function support, eukaryotic gene traits	Pauchet and Heckel 2013
Bean beetle *Callosobruchus chinensis*	Bacteria	Large parts of *Wolbachia* genomes	PCR, southern blot, antibiotic treatments, sex-linked inheritance	Kondo et al. 2002; Nikoh et al. 2008
Longicorn beetle *Monochamus alternatus*	Bacteria	Large parts of *Wolbachia* genomes	PCR surveys across populations, fluorescence in situ hybridisation	Aikawa et al. 2009
Fruit fly *D. ananassae* (and other Drosophila species)	Bacteria	Large parts of *Wolbachia* genomes	Sequence similarities, PCR, fluorescence in situ hybridisation	Dunning Hotopp et al. 2007
12 Fruit fly *Drosophila* species	Protist and bacteria	On average, each species has 173 HGTs	HGT index, phylogenetic analyses and genomic linkage with native genes	Crisp et al. 2015
Tardigrade *Hypsibius dujardini*	Bacteria, plants, fungi and archaea	Up to ~ 6663 genes	BLAST/ sequence similarity, HGT index, phylogeny, PCR, codon use and intron characteristics	Boothby et al. 2015; Bemm et al. 2016; Koutsovoulos et al. 2016; Richards and Monier 2016
Demosponge *Amphimedon queenslandica*	Bacteria, plants, fungi	Hundreds of genes	Alien index, BLAST/sequence similarity, phylogeny, sequence composition characteristics; HGTracker	Conaco et al. 2016; Fernandez-Valverde et al. in preparation
Parasitoid wasp *Nasonia vitripennis*	Bacteria	Large parts of *Wolbachia* genomes	Sequence similarities, phylogeny, genomic characteristics, developmental expression, phylogeny	Werren et al. 2010
Many species in the *Nematoda* phylum, including *Pristionchus pacificus, Brugia malayi, Meloidogyne hapla, M. incognita, Caenorhabditis japonica, C. elegans, C. brenneri, C. briggsae*	Bacteria, including the endosymbiont *Wolbachia*	Many genes, including genes related to parasitism and plant cell-wall degradation	Sequence similarities, phylogeny, HGT index, genome linkage, intron features	Dunning Hotopp et al. 2007; Danchin et al. 2010; McNulty et al. 2010; Schuster and Sommer 2012; Crisp et al. 2015
Trichoplax adhaerens (Placozoan)	Bacteria	12 genes	Genomic location, homology	Driscoll et al. 2013
Stick insects (Phasmatodea)	Bacteria	Pectinase genes	RNA sequencing, phylogenetic analyses	Shelomi et al. 2016
Pea aphid *Acyrthosiphon pisum*	Fungi	Carotenoid biosynthesis genes	Phylogeny, function support	Moran and Jarvik 2010
Sponge *Astrosclera willeyana*	Bacteria	Biomineralisation associated gene	WMISH showing localised gene expression, sequence characteristics, genomic location	Jackson et al. 2011
Starlet sea anemone *Nematostella vectensis*	Bacteria	Glyoxylate and shikimic acid pathway genes	Sequence similarities, phylogeny, sequence composition	Kondrashov et al. 2006; Starcevic et al. 2008
Octocorals	Bacteria or virus	Mismatch repair gene	Phylogeny, expression	Bilewitch and Degnan 2011
10 primate species	Protist and bacteria	On average, each species has 109 HGTs	HGT index, phylogenetic analyses and genomic linkage with native genes	Crisp et al. 2015

Identifying HGTs in animals and providing convincing support for their case is particularly complicated due to the close associations animals have with other life forms (Becq et al. 2010; Dunning Hotopp 2011). Distinguishing between the sequence of an animal and that of its symbionts, environmental contaminants and even experimental contaminants can be difficult, but is crucial (Dunning Hotopp 2011). Further, high quality support is especially important because the increasing number of reported animal HGTs has included some controversial and/or unsubstantiated claims. These cases include the reported transfer of glyoxylate cycle genes to non-placental vertebrates (Kondrashov et al. 2006), and the famous photosynthesising sea slug *Elysia chlorotica* (Rumpho et al. 2008; Rumpho et al. 2010; Pierce et al. 2012; Bhattacharya et al. 2013; Schneider and Thomas 2014; Schwartz et al. 2014). Recently, the degree of HGT in tardigrades is in dispute, with some claiming extensive HGT in *Hypsibius dujardini* (Boothby et al. 2015), while others conclude those claims are erroneous and that contamination is a better explanation (Bemm et al. 2016; Koutsovoulos et al. 2016). The tardigrade case attracted attention and has received commentary from others as well, with Richards and Monier (2016) echoing previously made cautions for the reporting of animal HGT and advising on appropriate detection methods (Dunning Hotopp 2011; Boto 2014).

Detecting HGTs in animals requires finding foreign-like sequence that is integrated into the animal genome and is not the result of contamination. Foreign-like sequence can be discovered through phylogenetic incongruence or by using sequence compositional parameters such as GC content bias, since there is variation in such characters among species, but the genes within one species are relatively similar (Whitaker et al. 2009; Becq et al. 2010; Boschetti et al. 2012). Ideally, a mix of both parametric and phylogenetic approaches is used for more robust computational prediction of HGTs (Becq et al. 2010; Schonknecht et al. 2013; Conaco et al. 2016). Demonstrating the integration of the foreign-like sequence into the host genome is crucial (Moran and Jarvik 2010; Boschetti et al. 2012). Gold star methods include showing functionality of the HGT by temporal and/or spatial specific transcription, correlating HGT expression with a phenotype, and sequence verification of the HGT with flanking native sequences (Dunning Hotopp 2011; Richards and Monier 2016). Table 1.2 critically assesses published approaches used to detect HGTs.

Table 1.2 A summary of published approaches used to detect HGT in animals

(Part 1 of 2)

COMPUTATIONAL PREDICTION				
Broad approach	**Method**	**Strengths**	**Weakness**	**References**
Compositional methods	Parametric approaches (e.g. GC content bias, codon usage bias, oligonucleotide bias, di- tri- or tetra-nucleotide composition, *k*-mer frequencies)	- Only information needed is the genome of interest - Many methods of different levels of specificity and sensitivity. - Fast and suitable for whole genome analyses	- No detection of ancient HGT events due to their adaption to the genomic environment of their host - No indication of sequence origin - Many false positives and negatives - Methods developed for bacteria and are not necessarily relevant to eukaryotic genomes e.g. due to differences in gene regulation and greater variation in GC composition in eukaryotic genomes - No distinction between HGT and contamination, or between HGT and native over-expressed genes, or native genes with biased aa composition, or native repetitive genes. - Different methods target different types of HGT	Chor et al. 2009; Whitaker et al. 2009; Becq et al. 2010; Mallet et al. 2010; Azad and Lawrence 2011; Schönknecht et al. 2013; Wybouw et al. 2016
Sequence similarities	Phylogenetic incongruence	- Powerful - More accurate than only considering best blast hit - Can detect ancient transfer events	- Susceptible to false positives if poor sequence representation of phylogenetic neighbours - No distinction between HGT and contamination - Computationally expensive for large datasets - Complex and difficult to automate - Outcome dependent on correct tree rooting - Transfers between neighbours on the reference tree are not detected - Requires a reference tree	Whitaker et al. 2009; Becq et al. 2010; Mallet et al. 2010; Moran and Jarvik 2010; Jackson et al. 2011; Wybouw et al. 2012; Schönknecht et al. 2013; Crisp et al. 2015; Wybouw et al. 2016
	Alien Index	Simple and intuitive ratio	- Susceptible to database errors such as contamination and to false positives if poor sequence representation of phylogenetic neighbours - AI thresholds change due to changing database size impacting e-values - Relies on the best blast hit	Gladyshev et al. 2008; Chapman et al. 2010; Boschetti et al. 2012
	Blast	Fast and suitable for large datasets	- Top hit can be misleading - Susceptible to database biases, errors such as contamination, and to false positives if poor sequence representation of phylogenetic neighbours - Inappropriate for transfers between phylogenetically close species	Whitaker et al. 2009; Porter and Golding 2011; Driscoll et al. 2013
	HGT Index	- Simple and intuitive ratio - Uses bitscores instead of e-values so changes in database size have no impact	- Susceptible to database errors such as contamination and to false positives if poor sequence representation of phylogenetic neighbours -Relies on the best blast hit	Boschetti et al. 2012; Crisp et al. 2015

Table 1.2 A summary of published approaches used to detect HGT in animals

(Part 2 of 2)

Broad approach	Method	Strengths	Weakness	References
		COMPUTATIONAL SUPPORT		
Genomic location	Symbiont genome lacks the transferred gene(s)			Ros and Hurst 2009; Dunning Hotopp 2011
	Does transferred gene(s) sit within native sequences?	Uses existing data.	Relies on quality of the genome assembly.	Moran and Jarvik 2010; Acuña et al. 2012; Driscoll et al. 2013
	Does transferred gene(s) sit within mobile genetic elements such as TEs?	Offers a mode of transfer.	These areas are notoriously difficult to assembly so could be due to misassembly.	Dunning Hotopp et al. 2007; Gladyshev et al. 2008; Moran and Jarvik 2010; Acuña et al. 2012; Boschetti et al. 2012; Crisp et al. 2015
Sequence characteristics	Phylogenetic hallmarks in an otherwise foreign-like gene, e.g.: - Larger introns - Larger intergenic spacers - Lack of Shine-Delgarno (5'AGGAGG3') - Presence of a poly-A signal and poly-A tail		Will not detect recent transfers that have not yet adapted.	Jackson et al. 2011; Acuña et al. 2012; Moran et al. 2012
Sequence verification	Deep sequencing coverage	Increases confidence in assembly		Dunning Hotopp et al. 2007
		WET LAB VERIFICATION		
Broad approach	**Method**	**Strengths**	**Weakness**	**References**
Function	Temporal or spatial specific transcription (RT-qPCR or fluorescence in situ hybridisation of mRNA)	Suggests the HGT is functional.	Does not distinguish between host or symbiont expression	Gladyshev et al. 2008; Dunning Hotopp 2011; Jackson et al. 2011
	Correlate transferred gene(s) with phenotype/ecological role		Often difficult to demonstrate the role of HGTs and is not feasible for large datasets	Moran and Jarvik 2010; Dunning Hotopp 2011; Acuña et al. 2012; Wybouw et al. 2012
Inheritance	Breeding experiments to show how the HGT is inherited, e.g. if symbionts are maternally inherited, yet the HGT is paternally inherited then it cannot be the sequence of a symbiont.	Not relevant to many systems.	Distinguishes symbiont/host sequences and shows integration of HGT into the host genome.	Dunning Hotopp et al. 2007
Sequence verification	PCR and sequencing of putative HGT and flanking native sequence	Supports HGT integration and presence in genome		Dunning Hotopp et al. 2007; Wheeler et al. 2013

There are several examples of HGT in animals that are exceptionally well supported. For example, Moran and Jarvik (2010) discovered multiple fungal carotenoid biosynthesis enzymes incorporated into the genome of the red phenotype of pea aphid *Acyrthosiphon pisum*. Support for this case include phylogenetic analyses, Polymerase Chain Reaction (PCR) confirmation of the transferred genes in red but not green individuals, and the loss of red body colour when the carotenoid desaturase enzyme is mutated (Moran and Jarvik 2010). Further, expressed sequence tags (ESTs) from independent samples and labs show that the HGTs are on large scaffolds that have insect-like genes, and the transferred sequence has larger introns and intergenic regions than expected of a fungus (Moran and Jarvik 2010). Another well supported case is the transfer of a bacterial mannanase gene to the coffee berry borer beetle *Hypothenemus hampei* (Acuña et al. 2012). Phylogenetic analyses show a *Bacillus* origin of the beetle gene that hydrolyses galactomannan, the major storage polysaccharide of coffee berries, thus enabling the beetle to reproduce and feed entirely inside coffee berries. The HGT sits between eukaryotic TEs and shows eukaryotic traits, such as a polyadenylation signal and no prokaryotic Shine-Delgarno sequence (Acuña et al. 2012). Further, the HGT was found in individuals with the coffee berry lifestyle across a broad geographic range, but not in relatives who do not colonise coffee berries – thus it appears this HGT event enabled the beetle to adapt to a novel ecological niche (Acuña et al. 2012).

Other documented cases of HGTs in animals involve a diverse range of species throughout the animal kingdom, from basal sponges to primates, and with sequences transferred from bacteria, fungi, protists and plant, as detailed in Table 1.1. Accompanying the ever-growing list of such HGT cases, some of which are better supported than others, are proposals that HGT has a greater impact on the evolution of many animals than previously understood (e.g., Dunning Hotopp 2011; Crisp et al. 2015; Danchin 2016) and alternatively, that many animal HGT claims are not valid, but result from misinterpretation of genome sequencing contamination and/or differential gene loss (Ku and Martin 2016; Salzberg 2017). Therefore, the evolutionary impact of HGT on animals awaits resolution.

In addition to not knowing the full impact of HGT on animals, the mechanisms by which foreign DNA is transferred to animal genomes are not understood; however, two mechanistic themes have emerged. First, many of the reported HGTs in animals are from bacterial endosymbiont donors, such as transfers from the endosymbiont *Wolbachia* to many arthropod hosts, including several nematodes

(Dunning Hotopp et al. 2007; McNulty et al. 2010), the beetles *Monochamus alternatus* (Aikawa et al. 2009) and *Callosobruchus chinensis* (Kondo et al. 2002; Nikoh et al. 2008), the fruit fly *Drosophila ananassae* (Dunning Hotopp et al. 2007), and the parasitoid wasp *Nasonia vitripennis* (Werren et al. 2010). The presence of endosymbionts in germ cells simplifies the animal HGT process since foreign sequence is closer to the nucleus and the germ cell sequestrated heritable genome. While this enables increased potential contact and transfer opportunity, it does not offer solutions on how the transferring sequence passes through the nuclear membranes, becomes integrated into the new host genome and becomes functionally assimilated. Indeed, close proximity is not enough for successful HGT, since sequences of endosymbiotic origin are either not found at all or are not enriched in some eukaryotes known to host endosymbionts (Chapman et al. 2010; Nikoh et al. 2010). Further, many cases of HGT in animals are not from endosymbiotic donors (see Table 1.1). Therefore, while endosymbiosis may cause greater chances of HGT, it appears that barriers are differentially overcome and the question remains: how does foreign sequence evade the host defence systems and become integrated into the functional host genome?

The second trend emerging from the literature on HGT in animals suggests that mobile elements (MEs) could be involved in the mechanisms by which foreign DNA integrates into animal genomes. Some studies have reported the genomic proximity of animal HGTs to TEs (e.g., Dunning Hotopp et al. 2007; Gladyshev et al. 2008; Acuña et al. 2012), and it has been proposed that HGT may be mediated by TEs and their inherent ability for genomic mobility (Kidwell 1993; Syvanen and Kado 2002; Gladyshev et al. 2008; Acuña et al. 2012). This proposal is further supported by findings on a related biological process – the horizontal transfer of TEs between different animal species. There is recurrent and prevalent horizontal transfer of TEs among animals including mammals and reptiles (Houck et al. 1991; de Almeida and Carareto 2005; Pace et al. 2008; Gilbert et al. 2012; Gilbert et al. 2013; Ivancevic et al. 2013; Walsh et al. 2013). So ubiquitous, this horizontal transfer of TEs (HTT) is suggested a significant factor in animal genome evolution, contributing to sequence and structural variation, regulatory element modifications, and novelty (Schaack et al. 2010; Ivancevic et al. 2013; Walsh et al. 2013; Boto 2014). Research is documenting not only the huge extent of this horizontal movement within animals, but is also revealing the virus and arthropod vector mechanisms by which TEs jump from one animal to another (Ivancevic et al. 2013). Viruses have narrow host ranges and

bacterial viruses/phages typically can only infect one or limited numbers of bacterial strains – a phage that infects 20 strains is considered to have a broad host range (Mihara et al. 2016). Therefore, a viral vector of HGTs from bacteria to animals is unlikely, since to my knowledge, no virus is currently known to infect both animals and bacteria (Glansdorff et al. 2009; Bilewitch and Degnan 2011; Nilsson 2014; Mihara et al. 2016). However, of significance to animal HGT is that these studies show that TEs can act as vectors too, and can also transfer hitchhiking non-LTR retrotransposons that are not transfer prone (Ivancevic et al. 2013). While from a different circumstance, these findings show that both MEs and genomes are dynamic and that perhaps MEs are involved in mechanisms of animal HGT – certain they have appropriate machinery.

1.4 THE STUDY SYSTEM: THE DEMOSPONGE *AMPHIMEDON QUEENSLANDICA*

A. queenslandica is a marine demosponge that was first formally described from individuals from the southern Great Barrier Reef of Australia (Hooper and Van Soest 2006; Degnan et al. 2008). Sponges are one of the oldest extant phyletic lineages of animals (Figure 1.1; Srivastava et al. 2010; Ryan et al. 2013) and the complete genome of *A. queenslandica* has been publically available since 2010 (Srivastava et al. 2010). Genomic comparisons of *A. queenslandica* with other animals and holozoan sister species can help inform hypotheses and inferences on animal evolution. As such, research effort has been spent developing further resources for *A. queenslandica* and it is an emerging model species (Degnan et al. 2008; Degnan et al. 2015) in a key phylogenetic position for comparative genomics.

Extensive genomic resources exist for *A. queenslandica*; in addition to the complete genome, the microbiome has been characterised (Fieth et al. 2016; Gauthier et al. 2016). Further, there are many transcriptomic datasets; for instance, the expression of *A. queenslandica* genes has been measured at 82 time-points through 17 ontogenetic stages across the developmental lifecycle (Anavy et al. 2014; Levin et al. 2016). These 17 stages include embryogenesis, which occurs in the brooding chamber of the mother, the free-swimming larval planktonic phase, and the benthic developmental stages, which include settlement, metamorphosis, juvenile development and adulthood (Figure 1.2; Degnan et al. 2008; Degnan et al. 2015).

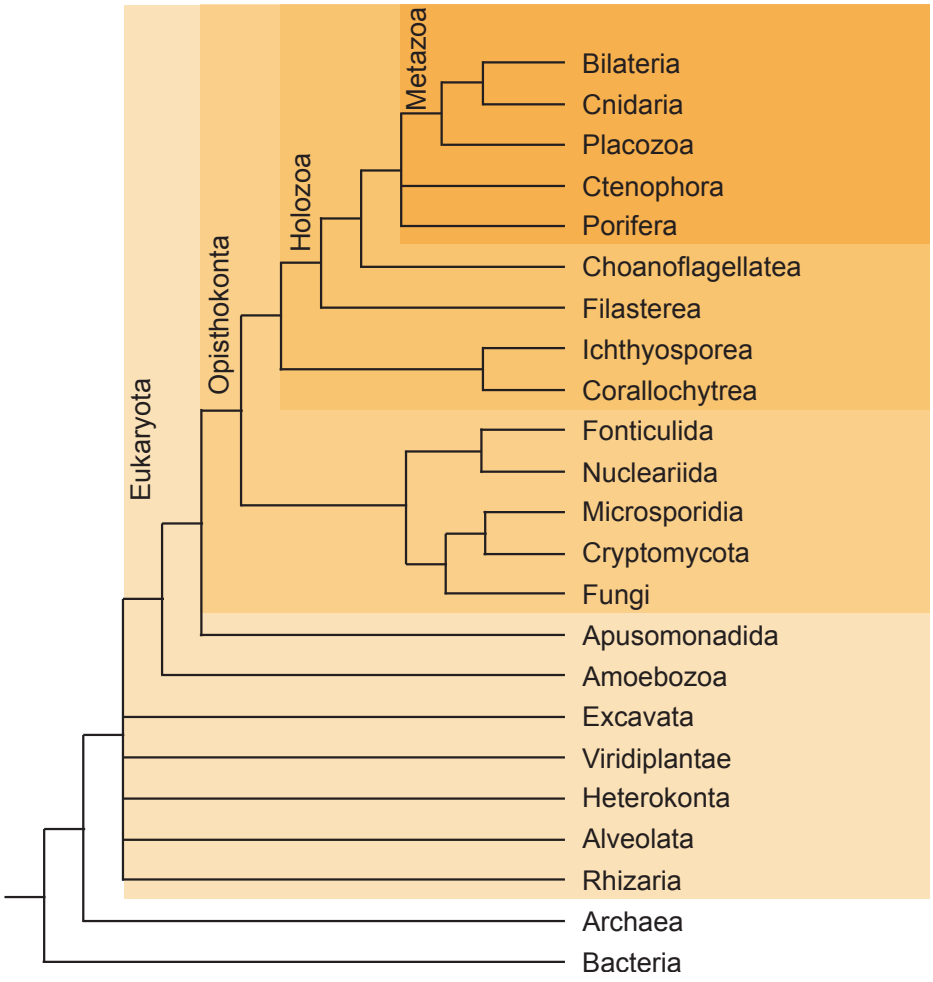


Figure 1.1 Cladogram displaying the phylogenetic position of sponges (phylum Porifera) Groups with uncertain phylogenetic positioning are presented as polytomies. The redacted phylogeny presented is summarised from Baldauf (2003), Niklas and Newman (2013), Torruella et al. (2015), Simion et al. (2017), and Whelan et al. (2017). Not to scale.

The genome of *A. queenslandica* is fully sequenced and annotated, and has high coverage and depth (Srivastava et al. 2010). This is particularly important here, since many of the analyses and outcomes of this thesis derive from this genomic data. The *A. queenslandica* genome was sequenced using 9-fold whole-genome Sanger shotgun coverage and was assembled with a custom method specific for polymorphic genomes (Srivastava et al. 2010). This involved read-to-read pairwise alignments and the designation of overlapping reads into read-clusters, which were assembled into contigs using the phrap algorithm (Srivastava et al. 2010). The contigs were further organised into scaffolds using phrap (Srivastava et al. 2010). The assembly is approximately 167 megabase-pairs, half of which is contained in 2652 contigs larger than 11.2kb or in 310 scaffolds larger than 120kb (Srivastava et al. 2010). The assembly was validated through comparisons with ESTs, itself and with finished fosmids.

The assembled scaffolds cover at least 93% of the coding bases of 66375 *A. queenslandica* ESTs and the scaffolds are estimated to at least partially represent approximately 99% of the genes captured by the ESTs (Srivastava et al. 2010). The scaffolds were aligned to themselves to assess their redundancy; from scaffolds of at least 10kb, only eight per cent of the total bases hit another scaffold with at least 96% identity (Srivastava et al. 2010). Finally, the genome project was further validated with comparisons of the assembly to finished fosmids, revealing high coverage depths of the fosmids by the genome assembly and a high degree of genome completeness (Srivastava et al. 2010).

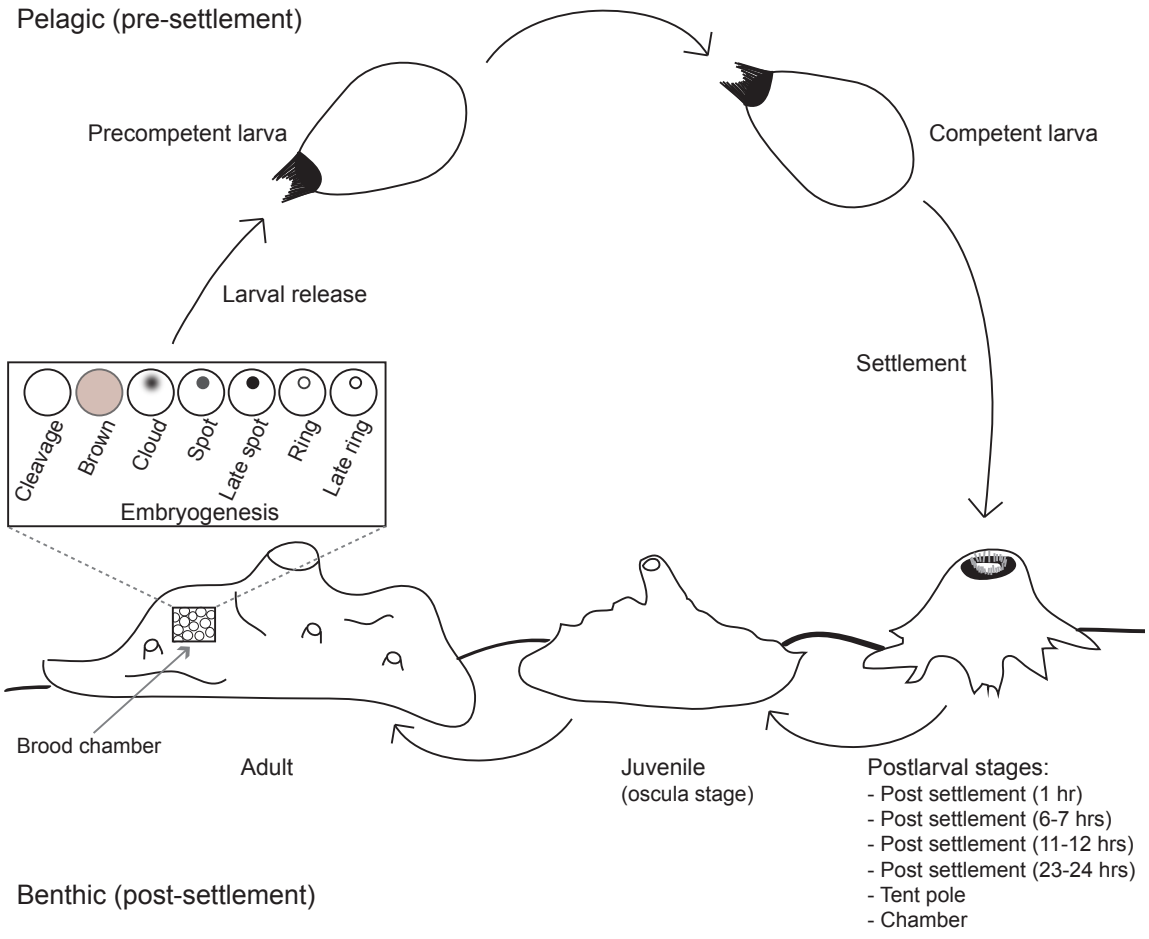


Figure 1.2 Life cycle of *A. queenslandica*
A. queenslandica has a biphasic life cycle with free-swimming pelagic larvae and benthic adults. Embryogenesis is complete after the late ring stage and free-swimming larvae are released from the maternal brooding chamber, but are not yet competent for settlement and metamorphosis. Larvae are typically competent 6-12 hours post emergence (hpe) from the brooding chamber. Following this are six time points of postlarval stages through which individuals settle on a substrate and undergo metamorphosis. The first osculum appears during the oscula stage and allows individuals to filter feed, from which point individuals are considered juveniles, which grow and mature into adults.

Throughout its lifecycle, *A. queenslandica* interacts with bacteria in different capacities; notably, the primary food source of this animal is bacteria and consequently, there is naked bacterial DNA within sponge cells resulting from phagocytosis (Fieth et al. 2016). Further, *A. queenslandica* is a spermcast spawning hermaphrodite – that is, adults release sperm into water teeming with bacteria, but retain eggs, which are fertilised within the sponge and embryos receive maternal provisioning, including nurse cells containing yolk and symbiotic bacteria (Degnan et al. 2008; Srivastava et al. 2010; Degnan et al. 2015; Fieth et al. 2016). Further, *A. queenslandica* has two characterised prevalent symbionts (Gauthier et al. 2016). As such, in the sequencing of the complete genome of *A. queenslandica,* efforts were made to eliminate contamination. This included extensive washing of the sequencing material prior to lysis, the removal of a presumed contaminating fosmid, removal of laboratory sourced contaminant contigs, and the removal of ecological contamination identified by taxonomic analysis of unassembled reads by MEGAN 2.0 (Srivastava et al. 2010). However, despite this approach, bacterial-like sequence has been noted in the genome (Hentschel et al. 2012; Conaco et al. 2016). Further to genomic contamination from the environment and/or the sequencing process, these foreign sequences could result from differential gene loss across multiple lineages creating unusual taxonomic distributions of foreign-seeming genes. Also, convergent evolution may cause genes to appear more closely related than they are, so in fact an unusual taxonomic distribution is not real. Alternatively, differential substitution rates may cause genes to appear more distantly related than they actually are. In addition, poor sequencing representation of some taxonomic groups can falsely create an unusual phylogenetic distribution. Finally, some of the foreign sequences may result from HGT.

The above described close associations of *A. queenslandica* with bacteria not only pose the problem of contamination to a genome sequencing project, they also offer chances of contact with foreign DNA, including that of digested food bacteria, and thus HGT opportunity. In comparison to many animals, *A. queenslandica* has less potential barriers to HGT because sponges have germ cells that are continuously segregated from adult stem cells (Juliano and Wessel 2010); so the chances of transfer events becoming heritable are greater in sponges than in animals that only segregate germ cells early in embryogenesis (Zhaxybayeva and Doolittle 2011; Jensen et al. 2016). Indeed, in an analysis of HGT in the genome of *A. queenslandica,* Conaco et al. (2016) found 227 putative HGTs of bacterial origin using methods based on either sequence similarity and phylogeny or sequence composition characteristics.

An independent analysis of HGT in the genome of *A. queenslandica* using the program HGTracker detected 576 putative HGTs due to their nonmetazoan sequence similarity yet their incorporation into native gene containing scaffolds (Fernandez-Valverde et al. in preparation). The putative HGTs identified by Fernandez-Valverde et al. (in preparation) are predicted to be from various taxonomic sources including bacteria, plants and fungi.

The majority of the predicted HGTs in *A. queenslandica* have assimilated to the sponge genome and thus are probably old HGTs, suggestive that HGT in *A. queenslandica* is not ongoing (Conaco et al. 2016; Fernandez-Valverde et al. in preparation). Therefore, parametric HGT detection methods analysing *A. queenslandica* sequence composition characteristics do not detect the majority of the predicted HGTs in this species since they have adapted to the host genome. At present, sponges and their phylogenetic neighbours are poorly represented in sequence databases; therefore, sequence similarity based HGT detection methods may yield false positives that only later with increased sequencing will be correctly identified as not the result of HGT. However, reducing these false positives by not using sequence similarity methods, but instead requiring predicted HGTs to have atypical sequence composition characteristics, will limit an investigation of HGT in *A. queenslandica* since this approach will overlook the majority of predicted HGTs, the older HGTs. Despite biased sequencing representation, sequence similarity based methods have been used to identify HGTs in at least six species that are phylogenetic neighbours of *A. queenslandica*, specifically these species include cnidarians (Kondrashov et al. 2006; Chapman et al. 2010; Bilewitch and Degnan 2011), a placozoan (Driscoll et al. 2013), another poriferan species (Jackson et al. 2011), and a choanoflagellate (Yue et al. 2013; see Figure 1.1 for phylogenetic positions). Therefore, the risk of potential false positives arising from poor sequencing representation is considered a weakness that will be lessened in time with increased sequencing.

1.5 THESIS AIMS AND CHAPTERS SYNOPSIS

The overall goal of this study is to further explore and better understand HGT in *A. queenslandica.* At a broad level, I describe the *A. queenlandica* HGTs using sequence similarity comparative approaches as well as transcriptomic data. I then focus more closely on a case study of the HGT-derived aspzincin genes of *A. queenslandica* for a deeper level of understanding. Last, I test the hypothesis that MEs are involved in the mechanisms of HGT in animals. More specifically, I pursue the following three aims.

Aim 1. To asses the extent and nature of HGT in A. queenslandica

In Chapter 2 I compare the outcomes of the two independent analyses of HGT in the same *A. queenslandica* data. Further, I survey the HGTs for protein domains to explore the types of genes that have been transferred and to find if there are enrichments of any genes. I characterise the discovered enrichment groups in terms of their broad taxonomic sources, their domain architectures, and their ontogenetic expression profiles. To consider the extent of HGT in *A. queenslandica,* I explore the phylogeny of the two largest gene groups to test whether the genes within each enriched group have been independently transferred, or if less transfers have occurred, but were followed by gene duplications. Last, I explore whether the transferred genes include genes not typically found in the Metazoa and that have possibly lead to greater amounts of innovation.

Aim 2. To characterise the HGT-derived aspzincin gene family in A. queenslandica

Metallopeptidase aspzincins are not typically found in the animal kingdom, yet have undergone large amounts of duplication in *A. queenslandica.* In Chapter 3, I consider the approximate timing of the original transfer event of an aspzincin gene to *A. queenslandica*, or ancestor, by searching for aspzincins in 26 other sponge species. In attempts to decipher the possible function(s) of the aspzincins in *A. queenslandica,* I examine if *A. queenslandica* has a gap in its metallopeptidase repertoire, relative to other sponges and animals, that has been filled by the aspzincins. In addition, I characterise the *A. queenslandica* aspzincins in terms of known aspzincin sequence features and investigate the types of genes with which aspzincins are possibly co-expressed through *A. queenslandica* developmental.

Aim 3. To explore the possible role of mobile elements in the transfer mechanisms of HGT in A. queenslandica

The mechanisms resulting in animal HGT from nonanimal sources are not known and this lack of knowledge has contributed to the historical assumptions that HGT in animals does not occur. Because of their mobility, MEs are hypothesised to be involved in the process of animal HGT. Further, a number of studies have reported animal HGTs that are flanked by TEs. However, to date, these associations have not been systematically explored with an approach that accounts for putatively confounding factors and thus, their significance remains unknown. In Chapter 4, I endeavour to establish if there is a relationship between MEs and HGTs in *A. queenslandica.*

Chapter 2 - Characterisation of horizontally transferred genes present in the animal *Amphimedon queenslandica*

2.1 Abstract

Horizontal gene transfer (HGT) is increasingly accepted as an evolutionary process not only in bacteria, but in animals too. However, the nature and extent of HGT in animal evolution is not well understood. Here, HGT in the demosponge *Amphimedon queenslandica* is explored by manual examination of the previously reported 576 computationally detected HGTs through an analysis of the domain content, broad taxonomic sources and the ontogenetic expression profiles of these genes. Three hundred and fifty domain types are conserved within 519 of the HGTs, 329 of which are expressed - thus a broad range of HGTs are inferred to be functional in their animal host. In addition, I found domain enrichments, with 116 domain types found in two or more HGTs. Phylogenetic analyses of the two largest bacterial-like groups, the metalloendopeptidase Aspzincin_M35 domain group (n=90) and the phosphorylase PNP_UDP_1 domain group (n=49), show that these two enrichments each result from a few transfer events followed by extensive duplication. The aspzincin analysis also revealed another putative independent horizontal transfer of the aspzincin domain to animals along the Vertebrata stem, with aspzincins discovered in 16 vertebrates. Further, my investigation revealed other post-transfer evolutionary trajectories in addition to duplication, including gene chimeras/fusions of sequences creating novel domain combinations in *A. queenslandica*. Last, I found that at least 10% of the *A. queenslandica* HGTs contain domains not typically found in the Metazoa and thus may have enabled greater levels of innovation than the other HGTs. Together, these data show that the evolutionary trajectory of *A. queenslandica* has included considerable adoption, co-option and duplication of HGTs. Further, these findings show an example of how HGT can be overestimated, since a large proportion of originally identified HGTs are in fact HGT-derived genes born from post-transfer duplication.

2.2 INTRODUCTION

Horizontal gene transfer (HGT) is the nonsexual genomic gain of exogenous genetic material – a process that crosses species boundaries and that has long been recognised as important in bacterial evolution (Andersson 2005; Dunning Hotopp et al. 2007; Koonin 2009a; Brakefield 2011; Dunning Hotopp 2011). This gene movement can involve genes of immediate selective advantage to the recipient, such as antibiotic or pesticide resistance; therefore, HGT can greatly impact adaptability (Gogarten et al. 2002; Keeling 2009; Soucy et al. 2015). The advent of whole genome sequencing is now highlighting that HGT in animals from nonanimal sources is more common and has greater evolutionary impact than previously assumed (Dunning Hotopp 2011; Boto 2015; Crisp et al. 2015).

Reported examples of HGT in animals involve a diverse range of recipient species throughout the animal kingdom; these cases show that there is not just one set of animal biological circumstances that can overcome barriers of HGT, which include the sequestration of the heritable animal genome in the nucleus and, for many animals, the germline (Gladyshev et al. 2008; Koonin 2009b; Whitaker et al. 2009; Raoult 2010; Marcet-Houben and Gabaldón 2010; Wybouw et al. 2012; Wijayawardena et al. 2013). Animal HGT recipients include the sponge *Astrosclera willeyana*, which continually develops germ cells from pluripotent somatic cells, asexual bdelloid rotifers, parasitic nematodes, arthropods, and vertebrates, including humans and other primates (e.g., Gladyshev et al. 2008; McNulty et al. 2010; Jackson et al. 2011; Boschetti et al. 2012; Flot et al. 2013; Wheeler et al. 2013; Crisp et al. 2015). Transfers to animal genomes have come from varied origins, including bacteria (both symbionts and nonsymbionts), fungi, plants, and protists (e.g., Dunning Hotopp et al. 2007; Gladyshev et al. 2008; Boschetti et al. 2012).

Historically, animal HGT research has been controversial, with contention over various reported cases (see Chapter 1 for further details), partially because of assumptions on the biological plausibility of HGT to animals and also because it is not straightforward identifying HGTs. More specifically, it is not simple distinguishing between foreign sequence that is truly incorporated into an animal genome and foreign sequence arising from the many organisms living closely with that animal. Therefore, it is encouraging that some exceptionally well-supported animal HGTs have emerged; for example, fungal

carotenoid biosynthesis genes in the pea aphid *Acyrthosiphon pisum* (Moran and Jarvik 2010) and a bacterial mannanase gene in in the coffee berry borer beetle *Hypothenemus hampei* (Acuña et al. 2012).

The demosponge *A. queenslandica* is a basal animal with populations across the Great Barrier Reef of Australia (Degnan et al. 2008). Like all sponges (phylum Porifera), *A. queenslandica* has HGT opportunity through multiple close interactions with bacteria in the environment, as well as from bacterial symbionts and significantly, from bacteria present as a sponge food source (Hentschel et al. 2012; Gauthier et al. 2016). Further, as a sponge, *A. queenslandica* reproduces sexually, but does not sequester germ cells early in development; rather germ cells arise recurrently from pluripotent somatic cells (Juliano and Wessel 2010). Such dedifferentiation of somatic cells to germ cells increases chances of HGTs becoming heritable (Redrejo-Rodriguez et al. 2012; Degnan 2014; Jensen et al. 2016). The complete genome of *A. queenslandica* was published in 2010 (Srivastava et al. 2010) and genomic resources generated since then include genome wide expression analyses and sequencing and characterisation of the *A. queenslandica* microbiome (Fernandez-Valverde et al. 2015; Anavy et al 2014; Levin et al. 2016; Fieth et al. 2016; Gauthier et al. 2016). To eliminate contamination in the assembled genome, the sequencing material was washed prior to lysis and a contaminating fosmid, laboratory sourced contaminant contigs and ecological contamination identified by taxonomic analysis of the unassembled reads by MEGAN 2.0 were all removed from the sequencing project (Srivastava et al. 2010). However, bacterial-like genes were still later noticed in the genome of this animal (Hentschel et al. 2012). Indeed, a genome-wide analysis for putative HGTs of bacterial origin detected 227 such genes in the first generation gene model predictions of *A. queenslandica* (Conaco et al. 2016).

Conaco et al. (2016) detected 227 putative HGTs in the genome of *A. queenslandica* using four methods based on either sequence similarities or sequence composition characteristics. Many of the detected putative HGTs have a large range of expression levels in larval and adult tissues (Conaco et al. 2016). Because some of the HGTs were also found in the transcriptomes of another marine demosponge *Haliclona amboinensis* and/or the freshwater demosponge *Ephydatia muelleri,* Conaco et al. (2016) suggest that some of the transfer events were deeper in the demosponge lineage. Enrichment for catalytic activity was found in the 227 HGTs; therefore, HGT may have contributed to the diversity of biochemical compounds of *A. queenslandica* (Conaco et al. 2016).

An independent and alternative approach to address the bacterial-like sequence in the *A. queenslandica* genome identified 576 putative HGTs using the HGT detection program HGTracker (Fernandez-Valverde et al. in preparation). HGTracker systematically identifies foreign sequence in the genome assemblies of animals using sequence similarity methods, and then distinguishes between putative contaminants and HGTs based on proximity to native genes in the genome assembly (Figure 2.1; Fernandez-Valverde et al. in preparation).

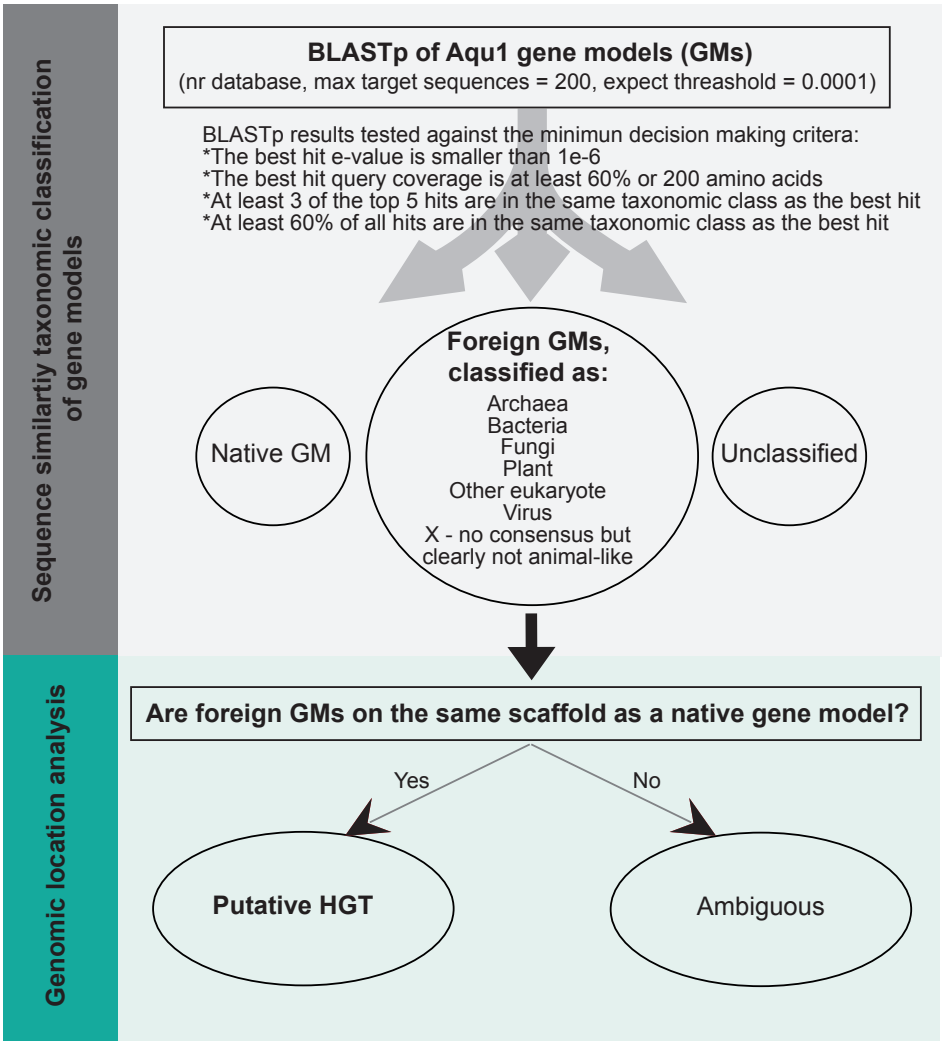


Figure 2.1 Summary of the main principles of the program HGTracker
The pipeline has two main steps; step one taxonomically classifies gene models based on sequence similarity identified by BLASTp, and step two considers the genomic location of any foreign gene models. Those gene models that show best sequence similarity to nonmetazoan organisms, but are found on the same scaffold as native genes are considered putative HGTs. Those that show best sequence similarity to nonmetazoan organisms, but do not have any native genes on the same scaffold are deemed ambiguous and perhaps are more likely to be contaminating sequence. The HGTracker outputs for the Aqu1 gene models of *A. queenslandica* used in my study are (1) the broad level taxonomic classifications based on sequence similarity (in grey) and (2) the putative HGT gene status (in green). Note that the taxonomic class X describes when a sequence is clearly foreign but does not fall plainly into a single taxonomic class.

HGTracker classified 14672 *A. queenslandica* first generation gene models as native and 1467 as alien sequences (Fernandez-Valverde et al. in preparation). A further 12614 gene models were unable to be classified because their sequence similarity results were too ambiguous for the decision making rules of HGTracker, which were created to not classify unclear situations that require a more detailed investigation (Fernandez-Valverde et al. in preparation). 576 of the alien gene models were classified as putative HGTs due to their similarity to nonmetazoan sequences, yet their incorporation into native gene containing scaffolds. The putative HGTs are predicted to be from a variety of taxonomic sources, though the majority were classified as bacterial-like (n=288). 120 other HGTs were definitely nonmetazoan and thus alien, but were not able to be clearly classified to one broad taxonomic group (HGTracker category X). The others were taxonomically classified as fungal-like (n=82), plant-like (n=76), other eukaryotes-like (n=9), and archaeal-like (n=1). HGTracker also predicted 545 putative contaminant genes and 346 ambiguous alien genes, the latter are on scaffolds that only also contain taxonomically unclassified genes, which could be native or alien genes. The GC content of the native and putative HGT gene models is lower than that of the putative contaminant and ambiguous gene models (Fernandez-Valverde et al. in preparation). The HGTs have also gained predicted introns (Fernandez-Valverde et al. in preparation). These results suggest two main points, first that these HGTs have adapted to their new genomic environment in at least two sequence characteristics. Second, while there may be some HGTs lost in the ambiguous and putative contamination groups of genes, it is likely to be a small number.

Here I further characterise the extent and nature of HGT in *A. queenslandica* by building upon the work of Conaco et al. (2016) and Fernandez-Valverde et al. (in preparation). To do this, first I identify the similarities and differences in the results of both these independent HGT-detection analyses of the same data. Next, and similar to the characterisation undertaken by Conaco et al. (2016), I characterise the larger list of putative HGTs of any origin, as identified by HGTracker (Fernandez-Valverde et al. in preparation), by surveying for protein domains to assess what type of genes have been transferred and if there is enrichment of any genes. Based on the domain content results, I consider the broad taxonomic sources, domain architectures, and ontogenetic expression profiles of the fifteen largest groups of HGTs. Further, I consider the phylogenies of the two largest domain groups. Last, I determine which of the *A. queenslandica* HGTs contain domains not typically found in the Metazoa and thus may have more greatly contributed to novelty in this animal genome.

2.3 METHODS

2.3.1 Data sources

Two versions of gene model predictions for *A. queenslandica* were used in this chapter. Generally, the first generation Aqu1 gene model predictions from the genome assembly (described in Chapter 1.4) were used and, along with the genome-project CDS sequences also used in this chapter, are publically available (Srivastava et al. 2010; available from Ensembl Genomes http://metazoa.ensembl.org/Amphimedon_queenslandica/Info/Index and the JGI Genome Portal http://genome.jgi.doe.gov/AmpqueaRenierasp/AmpqueaRenierasp.download.html). The Aqu2.1 gene models were more recently predicted using both the genome and transcriptome; these were used for transcription analyses (Fernandez-Valverde et al. 2015; available from Ensembl Genomes http://metazoa.ensembl.org/Amphimedon_queenslandica/Info/Index and from QCloud http://amphimedon.qcloud.qcif.edu.au/downloads.html).

Details, including Aqu1 gene model accession references, of the *A. queenslandica* HGTs identified by Conaco et al. (2016) were obtained from the publically available supplementary materials (http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0151092#sec023). Also from these public data of Conaco et al. (2016), a list of putatively younger HGTs was extracted. Conaco et al. (2016) scored each identified HGT from one to five with an increasing number reflecting increasing likelihood of older age, using the following five measures of HGT age, each worth one point if met: host typical codon usage; host typical GC per cent; two or more exons; transcript(s) of gene found in sponge *Haliclona amboinensis*; and transcript(s) of gene found in sponge *Ephydatia muelleri.* Any gene scored by Conaco et al. (2016) as only one or two, that is bacterial-like in sequence, and that exists on a scaffold containing at least one native gene was considered a putatively younger HGT in this chapter.

SL. Fernandez-Valverde interrogated the first generation Aqu1 gene models against HGTracker (default settings). The results were shared via personal communication.

The transcription of *A. queenslandica* HGTs throughout development was explored using a genome-wide ontogenetic transcript dataset generated for other projects (Anavy et al. 2014; Levin et al. 2016; Gene Expression Omnibus accession codes GSE54364 and GPL18214). Gene expression was measured using the RNA-Seq method CEL-Seq (Cell Expression by Linear amplification and Sequencing;

Hashimshony et al. 2012) for all gene models for 82 samples through 17 ontogenetic stages across the embryonic, larval, juvenile and adult developmental periods (Anavy et al. 2014; Levin et al. 2016). The 82 time-points were ordered based on increasing transcriptional entropy not morphology, using the BLIND clustering method (Anavy et al. 2014), though the larval stages were manually re-ordered based on their precisely known developmental time points (Anavy et al. 2014; Levin et al. 2016).

For indicated HGTs of potentially novel domain architectures, I considered the support for gene model predictions from the coverage of transcript reads mapping to the models, using already published transcriptomic data (Fernandez-Valverde et al. 2015; available from QCloud http://amphimedon.qcloud.qcif.edu.au/downloads.html).

2.3.2 Comparing classification results

To compare the overlap in the results of Conaco et al. (2016) and Fernandez-Valverde et al. (in preparation), a Venn diagram was generated using the online tool Venny v2.1 (http://bioinfogp.cnb.csic.es/tools/venny/). To manually verify taxonomic classifications of certain Aqu1 gene models made by both studies, gene models of interest were submitted to the nonredundant (nr) database of the National Centre for Biotechnology Information (NCBI) through the protein Basic Local Alignment Search Tool (BLASTp; default settings with organism exclude: *Amphimedon queenslandica* (taxid:400682)).

2.3.3 Protein domain surveys

For predicted domain architectures, the Aqu1 peptide sequences of all HGTracker-identified HGTs (AqHGTs) were submitted to the Pfam v28.0 Batch Search tool (cut-off –E 1.0; http://pfam.xfam.org). Results with an e-value smaller than 1.0 were automatically accepted, unless there were multiple domain types predicted for the same sequence, in which case only the best domain hit was retained. DoMosasics was used to visually present these results (Moore et al. 2014). Using the Pfam database information fields of Clan, Clan Description, HMM Name, HMM Description and Gene Ontology (GO) terms, the list of predicted domain types found in the AqHGTs were classified as enzymes, informational genes, and mobile element (ME) related genes. Informational genes were classified using the definition of Rivera et al. (1998), as outlined in Appendix 2.1.

The hmmsearch program of HMMER version 3.0 (default settings including cut-off –E 10.0; Finn et al. 2011) was used to interrogate all Aqu1 gene models with the Pfam hidden Markov models (HMMs; Eddy 1996) for the metallopeptidase Aspzincin_M35 domain (accession PF14521; aspzincin hereafter) and the phosphorylase PNP_UDP_1 domain (accession PF01048; PNP hereafter). The Aqu2.1 gene models were also searched with the Pfam aspzincin HMM. All hit alignments were manually checked for length and for conservation of key amino acids, as identified by the HMMs. To gain the complete domain compositions and to crosscheck for better-suited domain assignments, all the accepted gene models were submitted to Pfam version 27.0 (cut-off –E 1.0; http://pfam.xfam.org).

To thoroughly search for *A. queenslandica* aspzincins that may have diverged and thus may not be detected by Pfam's HMM, 52 *A. queenslandica* aspzincin gene models of similar sequence structure throughout the whole gene model were aligned in Geneious Pro 5.1.7 (www.geneious.com; Kearse et al. 2012). This alignment was used to build a custom HMM with hmmbuild in HMMER version 3.0 (default settings; Finn et al. 2011). The Aqu1s and Aqu2.1s were interrogated with this custom HMM using hmmsearch (default settings including cut-off –E 10.0; Finn et al. 2011). All hit alignments were manually checked for length and for conservation of key amino acids, as identified by the HMM.

2.3.4 The expression of A. queenslandica HGTs of interest

The ontogenetic expression of AqHGTs was examined using the existing genome-wide ontogenetic transcript dataset described in 2.3.1 (Anavy et al. 2014; Levin et al. 2016). From these data, the 82 normalised count values for each putative HGT of interest were extracted. Unless otherwise specified, the 82 time-point values were averaged into their 17 developmental stages. For all expression data presented, the expression profile of genes without at least one normalised read count of five at any one developmental stage was not considered, since low counts in all stages probably reflect non-meaningful expression and/or data noise. HGTracker interrogated the first generation Aqu1 gene models; however, I converted these to their second generation Aqu2.1 equivalents and used the corresponding expression data. On the occasion that the Aqu1 model differed significantly from the Aqu2.1 model, or that there was no Aqu2.1 prediction covering the sequence of an Aqu1, the Aqu1 expression data was used. I compared a quartile summary of the read counts totalled from each developmental stage for each

AqHGT with that of all the Aqu2.1 genes. Further, these quartile summaries were compared with that of the putative younger HGTs

2.3.5 Multiple sequence alignments and phylogenetic analyses

Phylogenetic analyses of the discovered aspzincin and PNP gene groups in the AqHGTs were made to test for phylogenetic incongruence, as predicted by HGTracker, and to explore the number of transfer events responsible for the large number of these genes and their possible donor taxa. To find sequences representative of the domain sequence diversity across the *A. queenslandica* PNP domains, a multiple alignment and neighbour-joining tree with 500 bootstrap replicates for the PNP domains was made using Geneious Pro 5.1.7 (www.geneious.com; Kearse et al. 2012). On this tree, a vertical line that intersects the four major and well-supported branches (>67% bootstrapping support) was drawn and representative Aqu1s were selected from the clade of each branch intersected by the drawn line. To find similar sequences, those representatives were then submitted against the NCBI 2013 nr database using BLASTp (default settings with organism exclude: *Amphimedon queenslandica* (taxid:400682)). Six of the seven native PNP domains were blasted with the same settings; the last native was dismissed because it was a short partial domain.

Because of the greater sequence diversity within the aspzincin domains than within the PNP domains, for each aspzincin domain the top five sequences were collected from the BLASTp results of searches against the NCBI 2017 nr database (default settings with organism exclude: *Amphimedon queenslandica* (taxid:400682)). For increased statistical power, the total number of bacterial and fungal sequences analysed was lowered by the following steps aiming to reduce sequence redundancy from the retrieved BLASTp results, without reducing representation/diversity. First, exact and near exact hit results were detected using CD-hit (0.95 cut-off; Li and Godzik 2006; Huang et al. 2010) and removed from further analyses. Second, the remaining sequences were aligned and a neighbour-joining tree was made using Geneious Pro 5.1.7 (www.geneious.com; Kearse et al. 2012). A vertical line intercepting 30 branches was drawn down this tree and representative sequences were selected from the clade of each branch intersecting the drawn line. To thoroughly search for any putative animal sequences, a second round of BLASTp searches was conducted, with the search settings restricted to metazoan sequences only.

Aspzincin and PNP domain multiple alignments were performed through the Geneious Pro 5.1.7 Multiple Alignment using Fast Fourier Transform (MAFFT) 6.814b plug-in (Katoh et al. 2002), and were manually trimmed and refined in Geneious Pro 5.1.7, with guidance from the HMM search alignments and by eye. ProtTest 2.4 implemented the Bayesian Information Criterion (BIC) to find the appropriate models of evolution for alignments (Darriba et al. 2011). Maximum Likelihood (ML) trees were constructed using the PhyML 2.0.12 plug-in in Geneious Pro 5.1.7 (Guindon and Gascuel 2003). At least 250 bootstrap replicates provided statistical support for the ML analyses for both of the aspzincin and PNP trees. Trees were drawn with FigTree version 1.4.0. BioEdit v7.0.5 (www.mbio.ncsu.edu/BioEdit/page2.html; Hall 1999) was used to create a sequence identity matrix from the each of the multiple alignments.

2.3.6 Identification of allelic variants in the aspzincin and PNP genes of A. queenslandica

In part, the phylogenetic analyses described above for the two largest domain groups found within the AqHGTs aimed to distinguish between those groups resulting from many independent HGT events or from few HGT events that were followed by duplication. To distinguish between diversification due to polymorphisms causing allelic variants and diversification due to duplication and divergence causing multiple real genes, the cDNA of each relevant Aqu1 gene model was searched against all the Aqu1 CDS sequences using BLASTn in Geneious Pro 5.1.7 (www.geneious.com; Kearse et al. 2012). Genes sitting alone on a scaffold are a flag for allelic variants since they could result from reads for a real gene that already exists on a real scaffold, but those reads are divergent enough that they do not map correctly to the real gene. Hence a new gene and scaffold are mistakenly created solely for those divergent reads, which are actually reads for an allelic variant of the real gene. The expected level of divergence in allelic variants is difficult to predict; therefore, the coverage and identities of hits from the blast results were manually perused on a case-by-case basis.

2.3.7 Characterisation of the HGTs containing domains typically absent in the Metazoa

The Pfam version 29.0 taxonomic distribution of each domain identified in the AqHGTs was consulted to identify those domains not reported in animals. All AqHGTs predicted to contain any of the putatively nonmetazoan domains (AqHGT_NM) were extracted and their domain hits were manually assessed for model coverage and e-values, to verify the nonmetazoan domain hits from the automated Pfam Batch

Search results. To identify combinations suggestive of novel genes evolving through the fusion of native and transferred sequences, while assessing the quality of domain hits, the taxonomic distribution of the other domains within the AqHGT_NMs was recorded.

2.4 RESULTS

2.4.1 Different results from independent analyses of the same data

A greater number of HGTs were detected by Fernandez-Valverde et al. (in preparation; n=576; AqHGTs) than identified by Conaco et al. (2016; n=227; AqHGT_Conacos), in part since the former analysis considered HGTs of any taxonomic origin whereas the latter only considered those of predicted bacterial origin. Using HGTracker, Fernandez-Valverde et al. (in preparation) classified 288 of those 576 AqHGTs as bacterial-like, seemingly quite similar to the result of Conaco et al. (2016); however, while these two gene lists overlap somewhat, there are large differences (Figure 2.2).

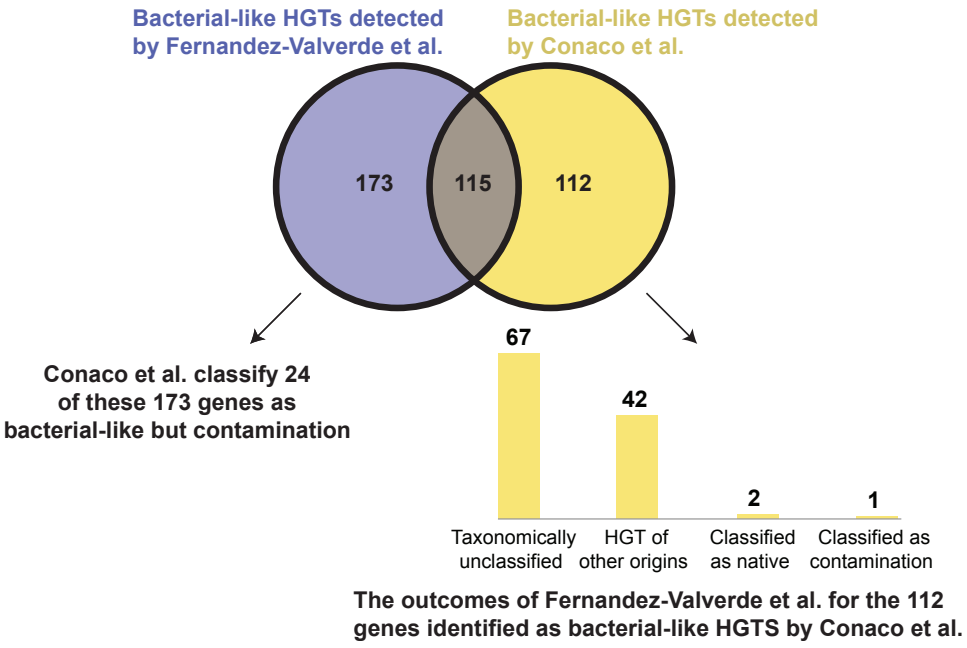


Figure 2.2 Comparison of results from two independent studies of HGT in *A. queenslandica*
Conaco et al. (2016) and Fernandez-Valverde et al. (in preparation) classified many genes differently, with only 115 genes overlapping in both studies.

Many of the differences in the results of the two approaches arise from differences in the taxonomic classifications of the same genes. For instance, Conaco et al. (2016) detected 112 bacterial-like HGTs that were not detected as bacterial-like by HGTracker; however, HGTracker classified 42 of those genes as

HGTs of a nonbacterial origin. Two are deemed fungal-like, one "other eukaryote-like", two plant-like, and 37 are classified as definitely not animal-like but otherwise unclear. Further, HGTracker could not taxonomically classify another 67 of those 112 bacterial-like AqHGT_Conacos. They received BLASTp hits that HGTracker considered; however, those hits either did not meet the minimum requirement for the decision making rules, in terms of the e-values, query coverage values, and/or the number of species in the hits, or, those standards were met, but there was not a clear taxonomic consensus (decision rules described in Figure 2.1). Of the 112 bacterial-like AqHGT_Conacos, HGTracker classifies two as metazoan-like and hence as native genes. Submission of one of these (accession Aqu1.214130|Aqu2.1.21829_001) to the NCBI nr database returns high quality hits from a diverse range of animal species (all of 116 returned hits are from animals; top hit is "PREDICTED: arginyl-tRNA-protein transferase 1-like isoform X1 [*Crassostrea gigas*]", query coverage 85%, e-value 2e-66, identity 31%). The other gene (accession Aqu1.222568|Aqu2.1.34221_001) also appears to be a native gene - when blasted, all 146 of the returned hits are from animal species, of high quality (top hit is "PREDICTED: ankyrin repeat domain-containing protein 45-like [*Acropora digitifera*]", query coverage 93%, e-value 3e-36, identity 33%), and the hits cover the full sequence, not just the ankyrin repeat domain.

Another factor causing differences in the HGT detection results is how filtering for contamination was managed by the two approaches. HGTracker detected 173 genes as bacterial-like HGTs that are not in the list of 227 AqHGT_Conacos. Twenty-four of these were detected as foreign by Conaco et al. (2016), but were filtered as more likely to be contamination due to their GC content, expression or presence on a smaller scaffold (Conaco et al. 2016). HGTracker did not consider the GC content or expression of genes for decision-making, and dealt with scaffolds differently – scaffolds were classified based on if they contain at least one native gene. HGTracker accepted scaffolds with at least one native gene as part of the *A. queenslandica* genome.

Finally, despite both methods being underpinned by similar principles of foreign gene detection based on sequence similarities, differences in specific settings have caused some of the differences in results. The second detection method of Conaco et al. (2016) is based on the top ten hits in BLASTp searches and is therefore very similar to HGTracker determining if a gene is foreign based on the top five unique

species in the BLASTp results. For some genes, considering the top ten hits is a more conservative approach and likely explains some of why Conaco et al. (2016) identified 159 less genes. However, for other genes, considering the top five unique species returned in the results is the more conservative approach, since the top ten hits of a BLASTp results could be different isoforms from the same species, or perhaps unrepresentative sampling/database species biases may result in few bacterial sequences for a given gene existing in the NCBI nr database.

In sum, these comparisons do not reveal a gross error in approach from either study, but show that the described differences in results were created by subtle methodological decisions affecting how BLAST results were parsed and how contamination was filtered out. The decision-making criteria of HGTracker appear more thorough and conservative, yet also less constrained. Specifically, sequence similarity decisions made by HGTracker are more logical because they consider the top five unique species in BLAST results and thereby are not influenced by multiple isoforms or duplicates in one species, which is a problem when considering the top 10 hits as in the approach of Conaco et al. (2016). Further, HGTracker distinguishes between putative HGTs and contaminants based on the presence/absence of at least one native gene on foreign gene-containing scaffolds; this is less arbitrary than the rules used by Conaco et al. (2016), which are based on cut-off points for GC content, expression levels in one dataset, and scaffold size. For these reasons, and also because my thesis is interested in HGTs arising from any taxonomic source and not just from bacteria, unless otherwise specified, I elect to focus hereafter on the AqHGTs predicted by HGTracker.

2.4.2 A wide variety of protein domains exist in the HGTs of A. queenslandica

Of the 576 putative AqHGTs identified by HGTracker, 519 contain at least one predicted recognised domain. In all, 350 different Pfam domains are predicted, the majority of which (234) are found in just one AqHGT (Appendix 2.2), while 116 are present in more than one AqHGT (Table 2.1). Based on the information for each domain type in the Pfam database, 173 of the 350 domain types are predicted enzymatic, nine are predicted to be from informational genes, and 14 are related to MEs (Appendix 2.1). The domains include enzyme members such as metal dependent amidohydrolases, glycosyl hydrolases, cupin enzymes, and ATP-dependent and NAD-dependent DNA ligase enzymes (Table 2.1; Appendix 2.2). There are a variety of peptidases including members of peptidase clans CA, CD, MA,

Table 2.1 The 116 Pfam domains present in more than one of the 576 putative AqHGTs and the proportion of each domain group that are expressed

(Part 1 of 2)

Pfam Accession	Pfam Name	No. of AqHGTs containing at least one	% expressed
PF15466	DUF4635	13	100%
PF14310	Fn3-like	6	100%
PF13385	Laminin_G_3	5	100%
PF00350	Dynamin_N	4	100%
PF04851	ResIII	4	100%
PF00933	Glyco_hydro_3	4	100%
PF02837	Glyco_hydro_2_N	4	100%
PF01915	Glyco_hydro_3_C	4	100%
PF00703	Glyco_hydro_2	4	100%
PF14538	Raptor_N	3	100%
PF16656	Pur_ac_phosph_N	3	100%
PF00149	Metallophos	3	100%
PF01419	Jacalin	3	100%
PF16355	DUF4982	3	100%
PF03283	PAE	3	100%
PF00230	MIP	3	100%
PF07974	EGF_2	2	100%
PF13646	HEAT_2	2	100%
PF13855	LRR_8	2	100%
PF01926	MMR_HSR1	2	100%
PF13166	AAA_13	2	100%
PF13520	AA_permease_2	2	100%
PF08241	Methyltransf_11	2	100%
PF00107	ADH_zinc_N	2	100%
PF01380	SIS	2	100%
PF00884	Sulfatase	2	100%
PF16347	DUF4976	2	100%
PF01263	Aldose_epim	2	100%
PF01501	Glyco_transf_8	2	100%
PF02110	HK	2	100%
PF05708	Peptidase_C92	2	100%
PF14295	PAN_4	2	100%
PF01753	zf-MYND	2	100%
PF13415	Kelch_3	2	100%
PF14592	Chondroitinas_B	2	100%
PF13229	Beta_helix	2	100%
PF08240	ADH_N	2	100%
PF01231	IDO	2	100%
PF13918	PLDc_3	2	100%
PF00618	RasGEF_N	2	100%
PF03747	ADP_ribosyl_GH	2	100%
PF00909	Ammonium_transp	2	100%
PF15879	MWFE	2	100%
PF00080	Sod_Cu	2	100%
PF04306	DUF456	2	100%
PF00581	Rhodanese	2	100%
PF14008	Metallophos_C	2	100%
PF07687	M20_dimer	2	100%
PF05593	RHS_repeat	2	100%
PF00617	RasGEF	2	100%
PF02661	Fic	2	100%
PF06206	CpeT	2	100%
PF01451	LMWPc	2	100%
PF01546	Peptidase_M20	3	100%
PF02836	Glyco_hydro_2_C	3	100%
PF14521	Aspzincin_M35	59	92%
PF10998	DUF2838	6	83%
PF00531	Death	9	78%
PF01048	PNP_UDP_1	25	76%
PF00400	WD40	4	75%
PF02894	GFO_IDH_MocA_C	4	75%
PF13499	EF-hand_7	4	75%

Table 2.1 The 116 Pfam domains present in more than one of the 576 putative AqHGTs and the proportion of each domain group that are expressed

(Part 2 of 2)

Pfam Accession	Pfam Name	No. of AqHGTs containing at least one	% expressed
PF02221	E1_DerP2_DerF2	10	70%
PF01408	GFO_IDH_MocA	6	67%
PF00083	Sugar_tr	3	66%
PF04548	AIG1	3	66%
PF00106	adh_short	3	66%
PF01344	Kelch_1	3	66%
PF07646	Kelch_2	3	66%
PF13964	Kelch_6	3	66%
PF00036	EF-hand_1	3	66%
PF16095	COR	5	60%
PF13847	Methyltransf_31	7	57%
PF03160	Calx-beta	7	57%
PF00112	Peptidase_C1	4	50%
PF07714	Pkinase_Tyr	2	50%
PF13671	AAA_33	2	50%
PF00078	RVT_1	2	50%
PF02784	Orn_Arg_deC_N	2	50%
PF00749	tRNA-synt_1c	2	50%
PF01055	Glyco_hydro_31	2	50%
PF01120	Alpha_L_fucos	2	50%
PF00891	Methyltransf_2	2	50%
PF16282	SANT_DAMP1_like	2	50%
PF13242	Hydrolase_like	2	50%
PF03553	Na_H_antiporter	2	50%
PF00491	Arginase	2	50%
PF04505	CD225	2	50%
PF00278	Orn_DAP_Arg_deC	2	50%
PF16334	DUF4964	2	50%
PF16335	DUF4965	2	50%
PF08760	DUF1793	2	50%
PF00069	Pkinase	9	44%
PF04970	LRAT	7	43%
PF03600	CitMHS	5	40%
PF01594	UPF0118	3	33%
PF00271	Helicase_C	13	8%
PF07727	RVT_2	25	0%
PF00665	rve	15	0%
PF13358	DDE_3	12	0%
PF14214	Helitron_like_N	11	0%
PF00270	DEAD	8	0%
PF05970	PIF1	5	0%
PF00226	DnaJ	5	0%
PF13976	gag_pre-integrs	5	0%
PF13538	UvrD_C_2	3	0%
PF13489	Methyltransf_23	3	0%
PF13592	HTH_33	3	0%
PF13508	Acetyltransf_7	3	0%
PF11252	DUF3051	3	0%
PF07464	ApoLp-III	3	0%
PF02689	Herpes_Helicase	2	0%
PF14227	UBN2_2	2	0%
PF13960	DUF4218	2	0%
PF04021	Class_III signal	2	0%
PF06017	Myosin_TH1	2	0%

Expression data is from a genome-wide ontogenetic transcript dataset (Anavy et al. 2014; Levin et al. 2016). A gene is considered expressed if it has at least one developmental stage with a normalised read count of at least five. See Appendix 2.2 for those 234 domain types found in only one AqHGT. Table shading reflects expression proportions for each domain group, with increasingly darker shading for lower proportions.

MH, and PA. Further, there are domains related to transcription factors, membrane transporters, and protein kinases (Table 2.1; Appendix 2.2). There is a striking enrichment of the metalloendopeptidase aspzincin domain, with 59 gene models predicted to have at least one aspzincin domain (Table 2.1). The two next most common domains are the phosphorylase PNP UDP 1 and the reverse transcriptase RVT 2 domains, which are each detected in 25 AqHGTs. A further 21 domains are found in 5 to 15 AqHGTs, and 92 domains are found in two to four AqHGTs (Table 2.1).

2.4.3 The ontogenetic expression of the HGTs of A. queenslandica

a. Expression profiles throughout development

Forty per cent of the AqHGTs (n=226) do not have at least one normalised transcript count of five at any one stage in the analysed developmental transcriptome, and are thus considered not expressed in this dataset. The remaining 60% are expressed in at least one developmental stage (n=350), with variable profiles. Some genes are more highly expressed early in development in embryos and free-swimming larvae (n=122), some are more highly expressed in postlarval and juvenile stages (n=160), and others are more highly expressed in adult sponges (n=68; Figure 2.3). When the mean normalised read count from each developmental stage is totalled across all stages for each gene, there is variation in the levels to which the genes are expressed (Figure 2.4a). The expressed AqHGTs and all the expressed Aqu2.1s that are not classified as HGTs are similar in the quartile distributions of these total expression sums (Figure 2.4b), with no significant difference between the two groups (Student's *t*-test, $\log_{10}$ transformed, *p*-value=0.34). The three most highly expressed gene models in the entire dataset of expressed Aqu2.1 gene models are native, but the fourth most highly expressed is an AqHGT that contains an aspzincin domain (accession Aqu1.220837|Aqu2.1.31788_001; total read counts = 177 433).

b. Comparing the domains present in expressed HGTs with those in unexpressed HGTs

There is a positive association in the AqHGTs between being expressed and containing detectable domains (Pearson's Chi-squared test, *p*-value=9.8e-05; Appendix 2.3). Further, there is a positive association between being expressed and domain diversity (measured as the number of different Pfam domains contained, two-sample test for equality of proportions with continuity correction, *p*-value=6.3e-09; Appendix 2.3).

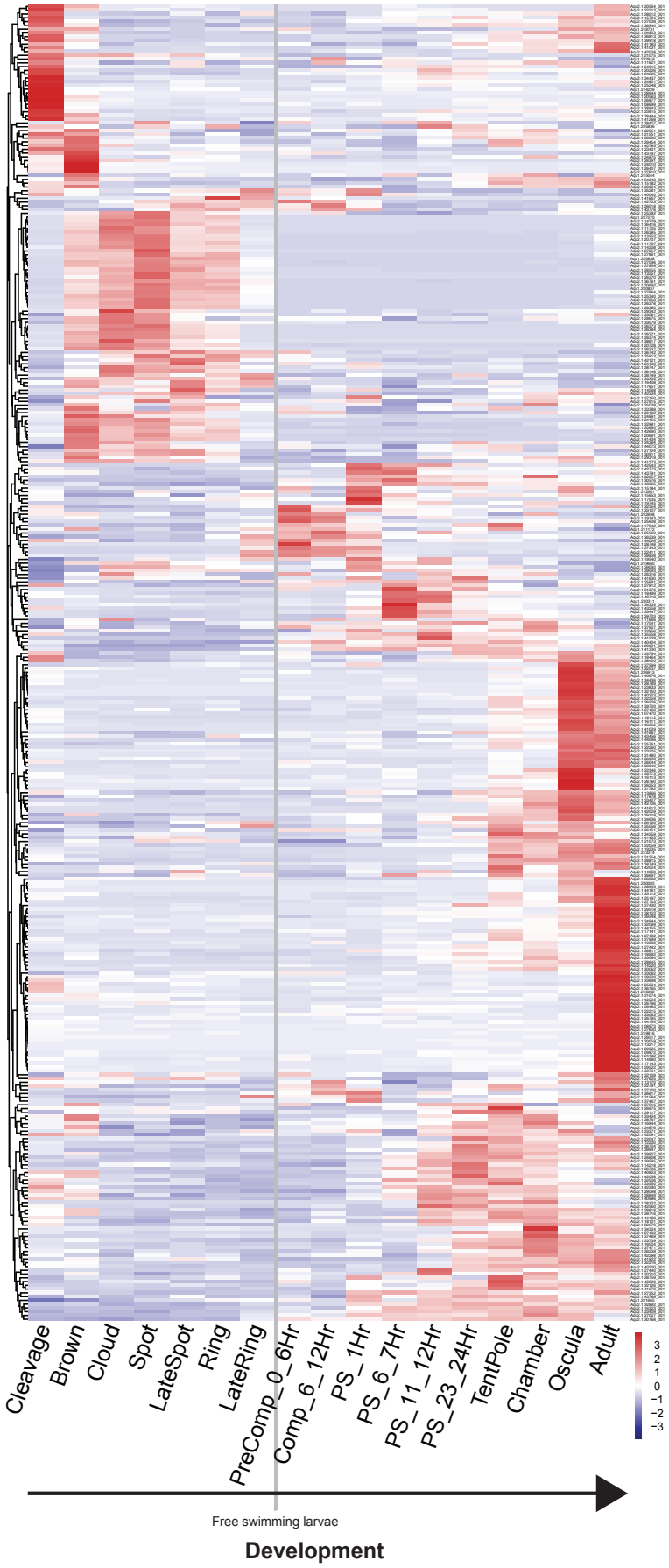


Figure 2.3 Ontogenetic expression profiles of AqHGTs
350 AqHGTs have at least one developmental stage with a normalised CEL-Seq read count of at least five. Here the expression patterns of these genes is visualised by a heat map scaled by row with genes organised by Euclidean distances (R Studio, Pretty Heatmaps package v 0.7.7). Figure 1.2 displays the lifecycle of *A. queenslandica*. Embryogenesis is completed after the Late Ring stage, 0-6 hours post emergence (hpe) from the maternal brooding chamber the larvae are free swimming but not yet competent for settlement and metamorphosis (PreComp_0_6Hr). Larvae are typically competent 6-12 hpe (Comp_6_12Hr). Following this are six time points of post-larval stages through which individuals settle on a substrate and undergo metamorphosis (stages abbreviated: 1 hour post settlement as PS_1Hr, 6-7 hours post settlement as PS_6_7Hr, 11-12 hours post settlement as PS_11_12Hr, 23-24 hours post settlement as PS_23_24Hr). The first osculum appears during the oscula stage, allowing individuals to filter feed, from which point individuals are considered juveniles, before developing into adults.

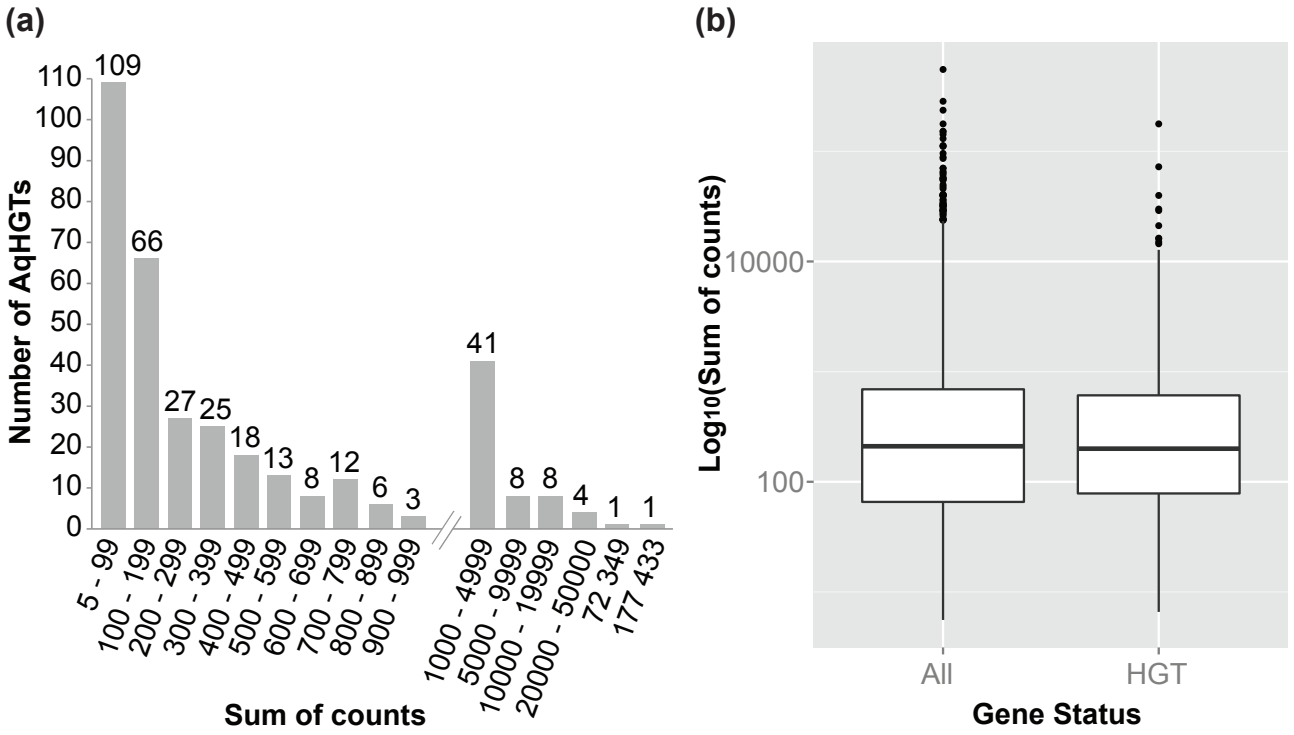


Figure 2.4 Quantification of AqHGT expression
(a) The CEL-Seq normalised count intervals on the x-axis represent the sum for each gene model of all 17 count means throughout developmental time. Note that for ease of visualisation, the intervals span 99 counts only up to 1000, thereafter the intervals vary because of the vast ranges in the sum values. (b) The expression quantities and range seen in the expressed AqHGTs are typical, when compared to the expression levels and range exhibited by the expressed genes of *A. queenslandica* (n=16 010 after removal of the AqHGTs and the 28 366 genes that all had less than five reads at each of the 17 developmental stages). The three genes with the highest sum of all 17 developmental stage means are native genes (read counts = 553 737, 284 670, and 236 279) followed by the highest AqHGT (read counts = 177 433).

Sixty-two domain groups consist of genes that are mostly (75-100%) expressed, 30 domain groups have a lower percentage (51-74%) of expressed genes, and five domain groups have a low proportion of expressed genes (8-44%). Nineteen domain groups only consist of unexpressed genes.

c. Putatively younger HGTs are expressed less

A group of putatively younger HGTs were identified by Conaco et al. (2016) and are used here as a subset of putatively younger HGTs from within the larger pool of AqHGTs. All 227 of the *A. queenslandica* HGTs identified by Conaco et al. (2016) have host-like codon usage and GC per cent. However, 18 of those genes have only one exon and no identified transcripts in two other sponge species, thus receiving an aging score of two (Conaco et al. 2016). All 18 genes are both bacterial-like in sequence and on scaffolds that contain at least one native gene. One of these putatively younger HGTs has not

one developmental stage with a normalised count of at least five; however the other 17 do and are here considered meaningfully expressed. These 17 younger HGTs have a lower quartile distribution of expression sums to both the expressed AqHGTs and the expressed 16024 non-HGT Aqu2.1s (Figure 2.5). These expression sums are significantly lower in the younger HGTs than in the other groups (one-sided Student's *t*-tests, $\log_{10}$ transformed, *p*-values=1.05e-05 and 2.7e-05 respectively).

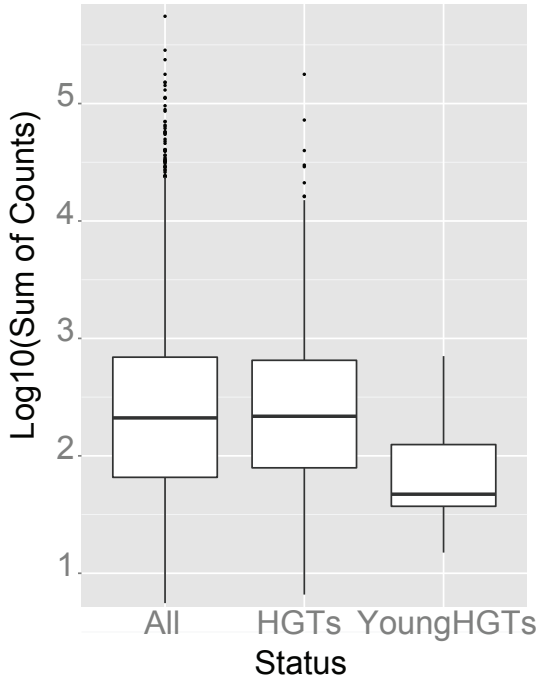


Figure 2.5 Younger AqHGT expression
Quartile summaries of the total sum of the count means from the 17 developmental stages for each gene for three subsets of *A. queenslandica* genes: all Aqu1s not classified as HGTs or contamination (n=16024); all AqHGTs (n=339) and putatively younger HGTs (n=17; these genes are excluded from the preceding subsets). Genes with less than five reads at each of the 17 developmental stages have been removed from each subset (n=28366, 225 and 1 respectively).

2.4.4 Characterisation of the 15 most common domain-based HGT groups in A. queenslandica

To gain deeper insights on the nature and extent of HGT in *A. queenslandica,* in this section I focus on the 15 most common domain-based groups of the AqHGTs.

a. Expression patterns within each domain group

Of the 15 most common domain groups, nine contain genes that are expressed (this excludes the DUF4635 group since all these genes are also in the aspzincin group, see below). With the exception of the sole Helicase C expressed gene, all these nine groups exhibit a range of ontogenetic expression profiles, but also contain genes sharing similar profiles that may reflect co-regulation (Figure 2.6). Excluding the aspzincins, all groups generally have higher expression in the postlarval, juvenile and adult stages (stages described in Nakanishi et al. 2014; Figure 2.6). The aspzincins have 11 genes most highly expressed later in development after larvae settle for metamorphosis; however, 43 are expressed more during embryogenesis between cleavage and spot developmental stages (stages described in Leys and Degnan 2002; Richards 2010), with three expression profiles (Figure 2.6).

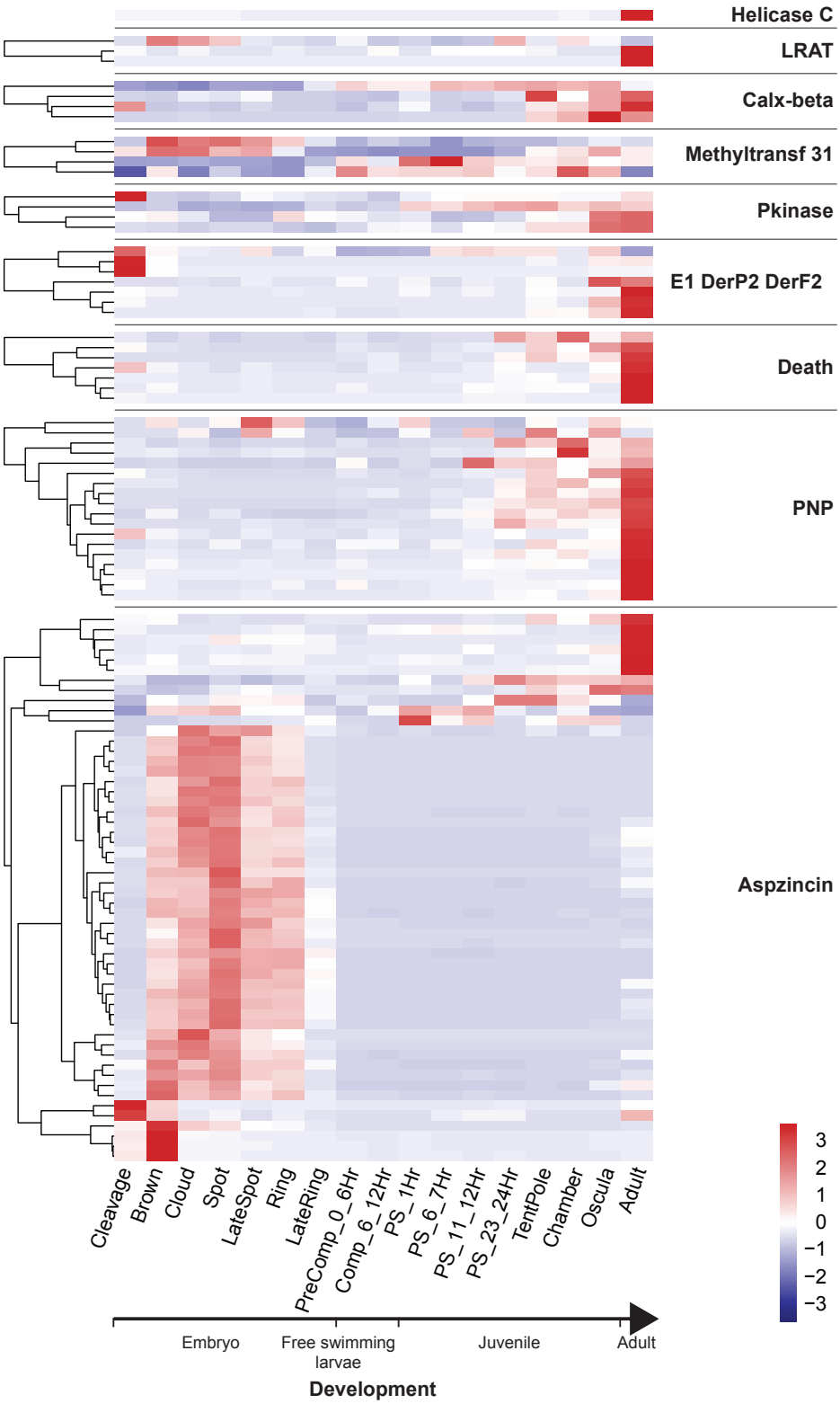


Figure 2.6 Within-group ontogenetic expression patterns of those AqHGTs containing at least one of the 15 most common domains of the AqHGTs

Each AqHGT presented has at least one developmental stage with a normalised CEL-Seq read count of at least five. Here the expression patterns of these genes is visualised by a heat map scaled by row with genes organised first by domain group and then by Euclidean distances (R Studio, Pretty Heatmaps package v 0.7.7). See Figure 1.2 for a summary of the lifecycle of *A. queenslandica*.

The PNP and aspzincin groups are not only the largest groups, but are also by far the largest of the expressed groups (Table 2.1). All but two of the PNP genes are not expressed until the postlarval and juvenile stages, and are particularly high in transcription levels in the adult stage; the other two are expressed mostly mid-development (Figure 2.6). The aspzincin group has at least six expression profiles, each quite specific to embryo stages, postlarval and juvenile stages, or to juvenile and adult stages (Figure 2.6).

Four of the five domain types found only in genes that are not expressed are related to MEs (RVT_2 reverse transcriptase, rve integrase core domain, Helitron helicase-like domain, and DDE superfamily endonuclease). The remaining domain type found only in unexpressed genes is the DEAD/DEAH box helicase domain.

b. The taxonomic origins of the AqHGTS in the top 15 domain groups

Most of the genes containing the most common domains are classified as bacterial-like (n=113), followed by plant-like (n=46) and fungal-like (n=19; Figure 2.7). Twenty-three other genes are classified as clearly not animal-like, but received BLAST hits from an ambiguous mix of nonmetazoan taxa (HGTracker classification X). All of the genes within most domain groups have been classified as either one taxonomic classification or a mix of one specific classification and as ambiguous (Figure 2.7).

The Species Distribution database of Pfam identifies most of these domains as found throughout the web of life (Figure 2.8); therefore the HGTracker classifications are mostly not challenged by these distributions. However, there are two conflicts between the taxonomic classifications of HGTracker and Pfam v29.0, specifically for the DUF4635 and Death domains, which are reported in Pfam solely in animals. The only domain not reported in animals by Pfam is the aspzincin domain (Figure 2.8).

c. Domain architecture, the promiscuity of the top 15 domains, and patterns of ontogenetic expression profiles with domain architectures

Promiscuous domains found in many proteins and/or combined with many different domains may be more readily adaptable (Copley 2003; Fraser et al. 2006; Khersonsky et al. 2006; Conant and Wolfe 2008; Bornberg-Bauer et al. 2010). To test if the AqHGTs more commonly contain putatively

promiscuous domains, I mined the Pfam database for both the number of sequences reported to contain each domain of interest (ranges from 34 to 1251283; Appendix 2.4). Also, the number of different domain architectures involving those domains of interest was retrieved (ranges from one to 4424; Appendix 2.4). I found no relationship between the number of *A. queenslandica* genes in each AqHGT domain group with the number of sequences containing each domain in the Pfam database or with the number of different domain combinations reported in Pfam for each domain (*p*-values 0.45 and 0.57 respectively). Within each domain group, there is no apparent pattern between domain architecture and expression profiles (Appendix 2.5).

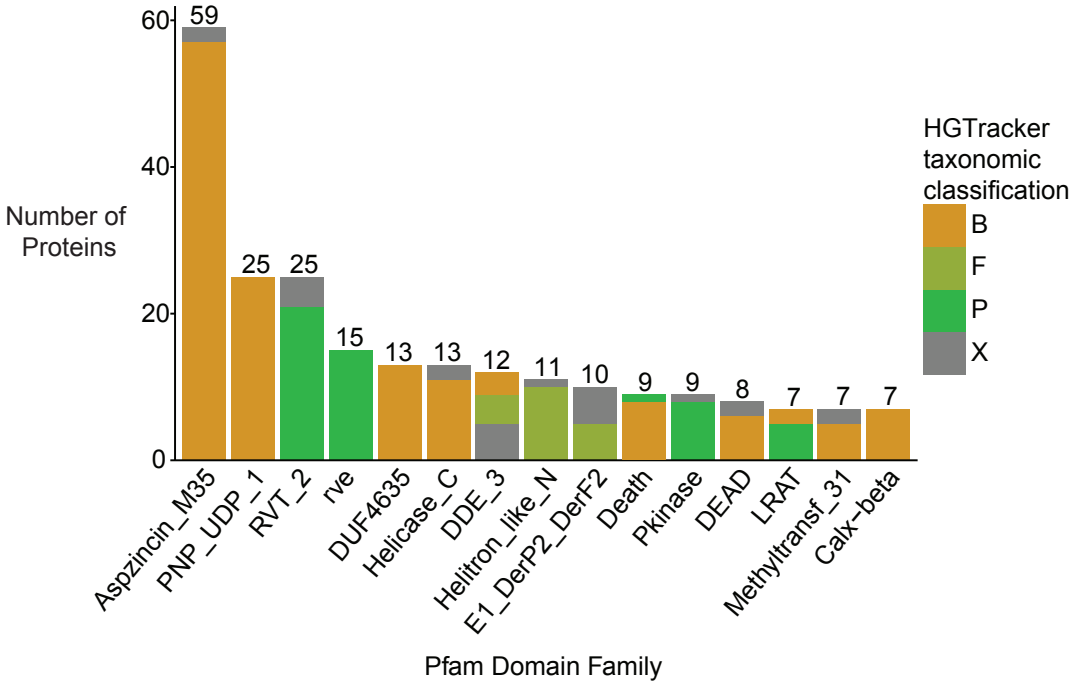


Figure 2.7 Taxonomic sources of the 15 most common AqHGT groups
Groups were determined based on domain content. These classifications are made by HGTracker, which makes broad taxonomic classifications based on sequence similarities obtained from BLASTp results. The taxonomic groups are B (bacterial-like), F (fungal-like), P (plant-like), and X (mixed). The taxonomic class X describes when a sequence is clearly foreign, but does not fall cleanly into a single taxonomic class.

d. The aspzincin and PNP case studies

The aspzincin and PNP gene groups are two of the three largest transferred gene groups; they are just half a per cent of the domain types yet are found in a large proportion (15%) of the AqHGTs. They also have the highest proportions of expressed genes (since the DUF4635s are found within the aspzincins and the Deaths in the PNPs). Thus both groups were explored in more detail.

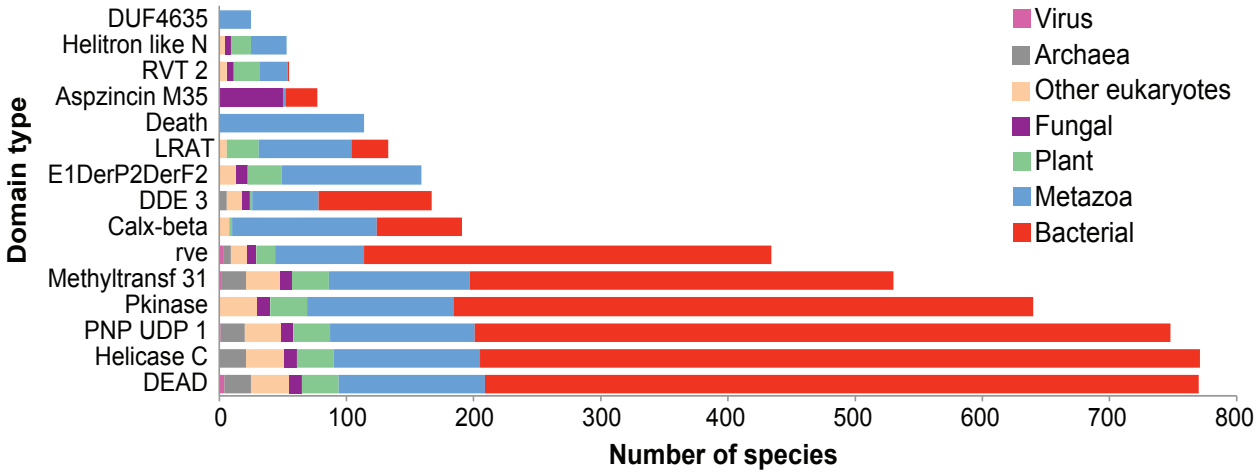


Figure 2.8 Taxonomic distribution of Pfam domains in UniProtKB reference proteomes
As of April 2016, Pfam version 29.0

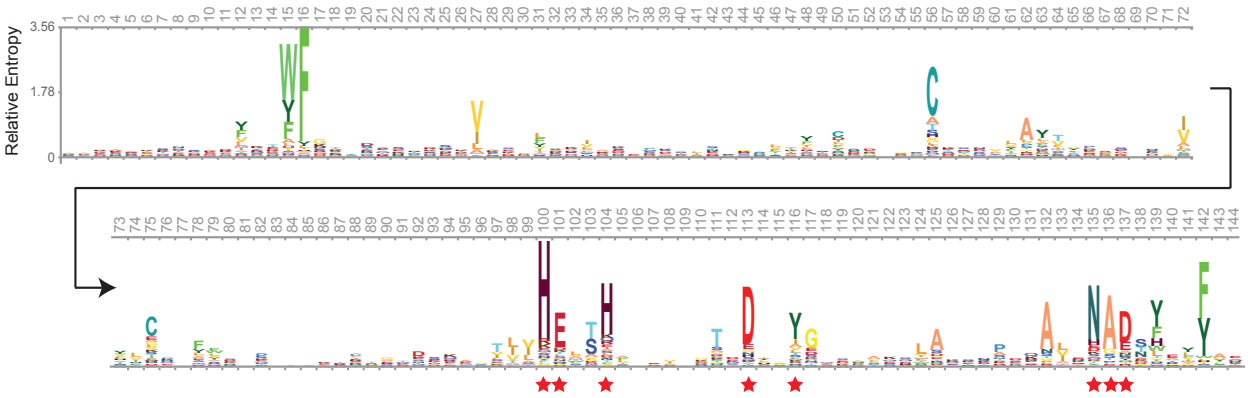
Aspzincins

Searching the *A. queenslandica* predicted proteome with Pfam's aspzincin HMM

Interrogation of all Aqu1 gene models with the Pfam HMM for the aspzincin domain (PF14521; visualised in Figure 2.9a) revealed an additional 17 *A. queenslandica* aspzincin domain-containing genes (AqAspzs) to those identified by HGTracker. Most of the aspzincin hit alignments show excellent conservation of the important amino acids as deemed by the HMM. Some hits have short query coverage values, yet still have high conservation of key amino acids - I have retained these as AqAspzs since their excellent partial domain hits are often for the end of the domain, but at the start of the gene model, or vice versa. Therefore, these partial domains are likely the result of incorrect gene model prediction.

Using BLASTp, I manually examined the 11 AqAspzs that are unclassified by HGTracker. Six receive an ambiguous mix of bacterial and fungal BLASTp hits. Another three of the unclassified genes contain highly repetitive regions - once those regions are excluded from the BLASTp search, the genes receive bacterial hits. Another unclassified gene has bacterial BLASTp results, but could not be classified by HGTracker because the best hit of each result has a query coverage just below the required length of 60%. The last of these unclassified genes (Aqu1.218127|Aqu2.1.27860_001; further described below) receives bacterial and fungal hits to the first half of the gene, but animal hits to the rest of the gene. HGTracker classifies five of the newly identified AqAspzs as likely contaminants. These show

(a) Sequence logo of Pfam's "Aspzincin_M35" HMM



(b) Sequence logo of HMM built from 52 *A. queenslandica* aspzincin gene models

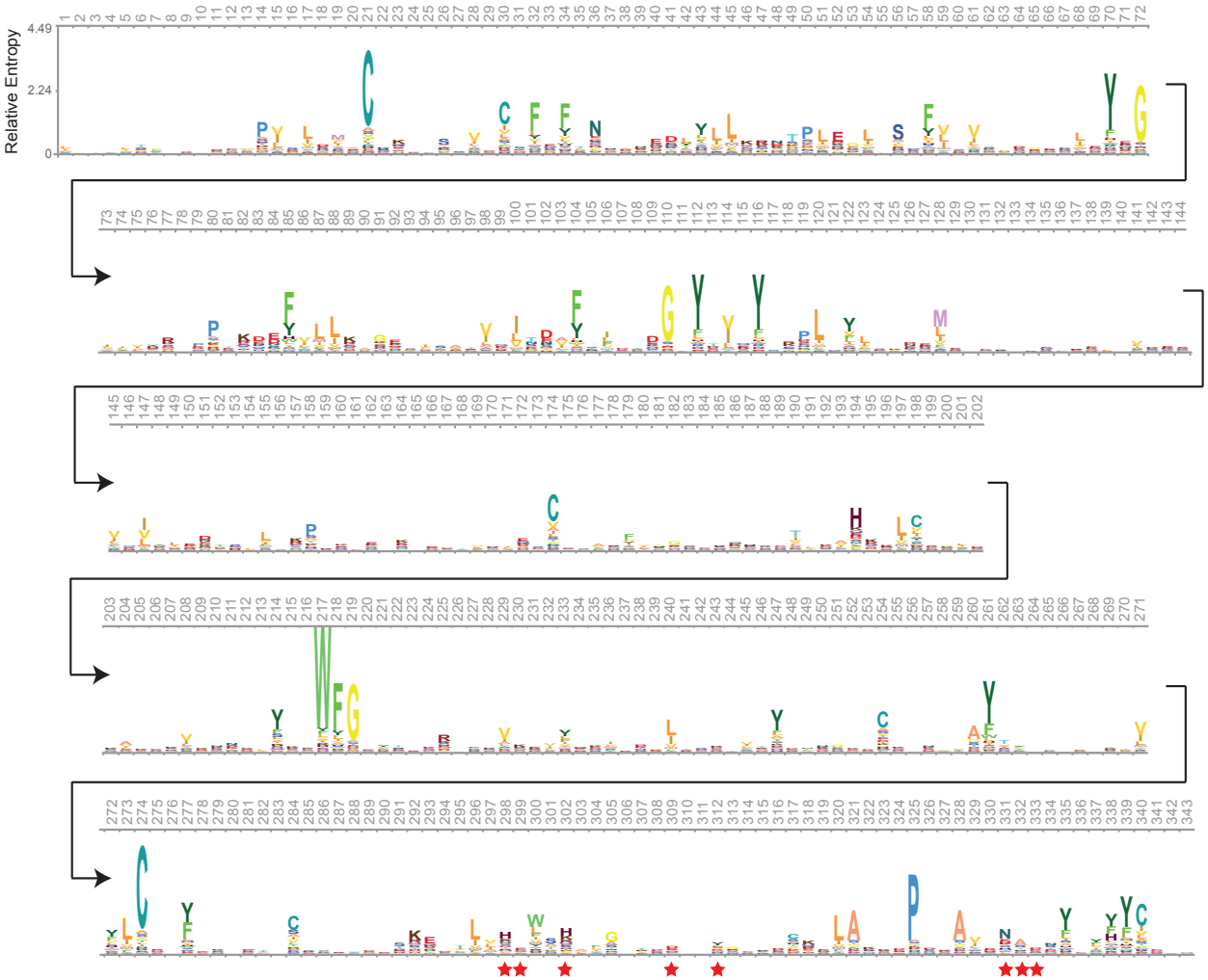


Figure 2.9 Sequence logos of aspzincin HMMs
Letter size depicts the degree of amino acid conservation, with larger letters more frequently conserved than smaller letters and thus inferred functionally more important and also diagnostic. (a) Sequence logo of Pfam's Aspzincin_M35 domain HMM, with the key catalytic and diagnostic amino acids highlighted with red stars (144 aa in length, built from 71 sequences). (b) Sequence logo of a HMM constructed from *(continued over page)*

Figure 2.9 Sequence logos of aspzincin HMMs

(continued from previous page)

52 *A. queenslandica* aspzincin proteins. For ease of comparisons, the sequence corresponding to the Pfam model in (a) starts on a new line at position 203 and the same key amino acids are highlight by red stars (343 aa in length, built from 52 sequences). Both logos were made with Skylign (www.skylign.org/; options: Create HMM – remove mostly-empty columns, Alignment sequences are full length, Information content – all) and adapted in Adobe Illustrator CS3 for ease of viewing. Colours are unique to each amino acid.

sequence similarity to bacterial sequences and exist either alone on a small scaffold or on a scaffold that contains only other foreign gene models.

Searching the *A. queenslandica* predicted proteome with a custom aspzincin HMM

The AqAspzs may have had quite a different evolutionary trajectory to that of those bacterial and fungal aspzincins incorporated in the Pfam aspzincin HMM. Therefore, to search for divergent AqAspzs, I built an HMM from 52 already detected AqAspz gene models, some of which are from the originally detected AqAspzs and some are those identified from the above HMM searching (Figure 2.9b). I used the entire AqAspz gene models because they are similar to each other in length and sequence, but only the latter half is the catalytic domain modelled in the Pfam HMM. Further, the approximate first half of the AqAspz gene models is specific to the AqAspzs within the genome of *A. queenslandica* - a pilot test search using a self-built HMM from only that beginning noncatalytic part of the AqAspzs only matched to aspzincin domain-containing genes, and no other *A. queenslandica* genes.

Interrogation of the Aqu1 models with the full length AqAspz-specific HMM revealed another eight putative AqAspzs. After inspection of their hits to the HMM, HGTracker results, scaffold location and manual inspection of their BLASTp search results, I have included six in the AqAspz group. Five are bacterial-like and on scaffolds with native genes, hence are classified as HGTs, and one is bacterial-like, but alone on a small scaffold, hence is classified as putative contamination. The two discarded cases include a hit to an aspzincin domain better predicted in a different gene model already detected, and a hit to the beginning of a convincing aspzincin, but this is followed a long gap of 193 nucleotides and there is not enough useable aspzincin sequence.

Phylogeny of *A. queenslandica* aspzincins

To test the phylogenetic incongruence of the AqAspzs and to explore the status of the genes classified by HGTracker as possible contamination, native and so on, I analysed their phylogeny in relation with collected bacterial, fungal, and animal aspzincins. 129 bacterial and fungal sequences were compiled from BLASTp searching of the AqAspzs. The majority of these are either from the Proteobacteria phylum (60%) or Agaricomyce members of fungal division Basidiomycota (26%), along with members of phyla Ascomycota (5%), Actinobacteria (8%), Cyanobacteria (0.5%), and Bacteroidetes (0.5%) (Appendix 2.6). The majority of the Proteobacteria species belong to the Gammproteobacteria (62%). The 129 full peptide sequences were submitted to CD-hit to remove near exact sequences (cut-off 0.95; Li and Godzik 2006; Huang et al. 2010), which reduced the number of sequences to 90. The aspzincin catalytic domain as determined by Pfam's HMM was extracted from these sequences using the GetfastaBed tool from the BEDTools package (Quinlan and Hall 2010), and those extracted domains were submitted to CD-hit to further reduce near duplicate domain sequences, which refined the sequences to 79 (cut-off 0.90; Li and Godzik 2006; Huang et al. 2010). In phylogenetic analyses, the probability of recovering the true phylogenetic tree partially depends on the length of the alignment, and the required length is influenced by the number of sequences involved – shorter alignments give less phylogenetic signal and are thus less powerful with increasing numbers of sequences (Moret et al. 2002). Therefore, to further reduce the number of sequences from the phylogenetic analysis, but not reduce sequence diversity, a neighbour-joining tree was made from an alignment of those sequences, and 23 bacterial and seven fungal sequences were selected as representatives from throughout the tree as described in section 2.3.5 (Appendix 2.7).

BLASTp searches of all the AqAspzs against the NCBI nr metazoan database revealed 35 aspzincin domain-containing animal sequences (excluding isoforms identified by the sequence names) from 16 vertebrate, one arthropod, and one cnidarian species. This includes the three sequences from the only other animal along with *A. queenslandica* in the Pfam aspzincin domain database, *Chelonia mydas* (accessed February 2017). Their submission to Pfam confirmed the aspzincin domain assignment, with good conservation of key amino acids, as determined by Pfam's HMM. Table 2.2 presents various details of these animal sequences, but of note, the vertebrate sequences are from six bird, three turtle,

two alligator, two fish, two frog and one lizard species, and all but two of the sequences either have the typical bacterial and fungal domain architecture of just one aspzincin domain (n=18) or an aspzincin and at least one Apolipoprotein L domain (ApoL; accession PF05461; n=15). The hits for all but three of the 35 animal sequences are for the last two thirds of the domain model only (Table 2.2). A key motif in the aspzincin domain is HEXXH…D, where 'X' is any amino acid; within the vertebrates there are five variants of this motif and within taxonomic groups there is sometimes the same conserved motif (e.g., all six bird sequences have HEASH…D), though other groups contain variation (e.g., the four lizard sequences have the three variants HEVSH…D, HELSH…D, and HEVAH…D). The

Table 2.2 Details of all identified non-sponge animal aspzincins

	Species	NCBI Accession	Pfam domain architecture	Aspzincin conservation (HMM start-end; HEXXH…D)
Coral	*Acropora digitifera*	XP_015748838	Aspz	11-144; HELSH..D
Arthropod	*Hyalella azteca*	XP_018028613	Aspz + HTH_11	4-144; HEMMH…D
Alligator	*Alligator sinensis*	XP_006035449	Aspz + ApoL + ApoL	56-117; HEASH..D
	A. mississippiensis	KYO24514	Aspz	56-117; HEASH..D
	A. mississippiensis	KYO25245	Aspz	56-117; HKASH..D
	A. mississippiensis	XP_019333874	Aspz + ApoL + ApoL	56-117; HEASH..D
	A. mississippiensis	XP_019346388	Aspz	57-120; HEASH..D
Birds	*Gallus gallus*	XP_422045	Aspz + ApoL	57-120; HEASH..D
	Meleagris gallopavo	XP_010711948	Aspz + ApoL	58-119; HEASH..D
	Struthio camelus australis	XP_009674420	Aspz + ApoL	49-119; HEASH..D
	Apteryx australis mantelli	XP_013816269	Aspz + ApoL + ApoL	46-119; HEASH..D
	Coturnix japonica	XP_015724402	Aspz + ApoL	53-119; HEASH..D
	Anser cygnoides domesticus	XP_013026629	Aspz	58-119; HEASH..D
Fish	*Astyanax mexicanus*	XP_007228219	Aspz + BTK	41-119; HEVSH..D
	Pygocentrus nattereri	XP_017555497	Aspz	45-118; HEVSH..D
Lizards	*Anolis carolinensis*	XP_008110094	Aspz	57-119; HEVSH..D
	A. carolinensis	XP_008110710	Aspz	57-121; HELSH..D
	A. carolinensis	XP_008123625	Aspz	7-119; HEVSH..D
	A. carolinensis	XP_008110100	Aspz	57-120; HEVAH..D
Turtles	*Chelonia mydas*	EMP29494	Aspz + ApoL + ApoL	56-136; HEASH..D
	C. mydas	EMP41822	Aspz + ApoL	54-119; HEASH..D
	C. mydas	EMP41823	Aspz + ApoL	54-119; HEVSH..D
	C. mydas	XP_007066530	Aspz + ApoL + ApoL	56-136; HEVSH..D
	Chrysemys picta bellii	XP_005293640	Aspz	57-118; HEVSH..D
	C. p. bellii	XP_005314731	Aspz + ApoL	58-113; HEASH..D
	C. p bellii	XP_005314765	Aspz + ApoL + ApoL	57-120; HEASH..D
	C. p bellii	XP_005315205	Aspz + ApoL	58-111; HEVSH…
	C. p bellii	XP_008174002	Aspz	55-119; HEVSH..D
	Pelodiscus sinensis	XP_006110231	Aspz + ApoL	54-141; HEVSH..D
Frogs	*Xenopus laevis*	OCA63368	Aspz	47-113; HEVSH..D
	X. laevis	OCA63371	Aspz	45-108; HEVSH…
	X. tropicalis	OCA23989	Aspz	54-120; HEVSH..D
	X. tropicalis	OCA23991	Aspz	53-135; HEVSH..D
	X. tropicalis	OCA23992	Aspz	57-108; HEVSH…
	X. tropicalis	XP_012826418	Aspz	48-116; HEVSH…

The conservation assessment considers where in the Pfam HMM the animal hits begin and end, and the conservation of the key catalytic residues of HEXXH…D, where X is any amino acid.

two invertebrate sequences are quite different to vertebrate sequences - their hits cover the whole domain model and their key conserved motifs are either unique from those of the other animals (in the arthropod: HEMMH…D) or rare in the other animals (in the coral: HELSH…D). Twenty of these 35 animal sequences were selected from representative branches of a neighbour-joining tree and used in the aspzincin phylogenetic analysis.

The phylogenetic relationships of the described sequences were explored based on a multiple alignment using a ML analysis. Six AqAspzs from the then-known 84 were excluded because, while they show high conservation to some of the catalytic domain, they are only short partial domains and probably reflect gene model prediction issues. The final trimmed alignment is 222 amino acids in length. The best model of sequence evolution is predicted to be WAG + G using BIC in ProtTest 2.4 (Darriba et al. 2011).

The ML analysis reliably groups the AqAspzs with the aspzincins of bacteria and fungi, to the exclusion of all but two of the other animal aspzincins (Figure 2.10). Regardless of their HGTracker classification, the phylogenetic analysis predicts that all the AqAspzs are more closely related to each other, as they group together in one clade to the exclusion of all the other sequences (Figure 2.10). The arthropod and coral aspzincins stand out as anomalies in the animal distribution of the aspzincin domain, since all the other detected animal aspzincins are sequences from either *A. queenslandica* or a small number of vertebrates. This taxonomic anomaly is reflected in the phylogenetic hypothesis - these two sequences are placed within the bacterial and fungal group of sequences (Figure 2.10). While this clade is not well supported, all the other sequences analysed are well supported in branches excluding these two animal sequences and all the bacterial and fungal sequences (Figure 2.10). Appendix 2.8A shows the sequence identity matrix from the same multiple alignment. Quartile distribution summaries of these sequence identities show higher sequence identities between the aspzincins of *A. queenslandica* and bacterial and fungal species than between the aspzincins of *A. queenslandica* and the vertebrates (Appendix 2.8B).

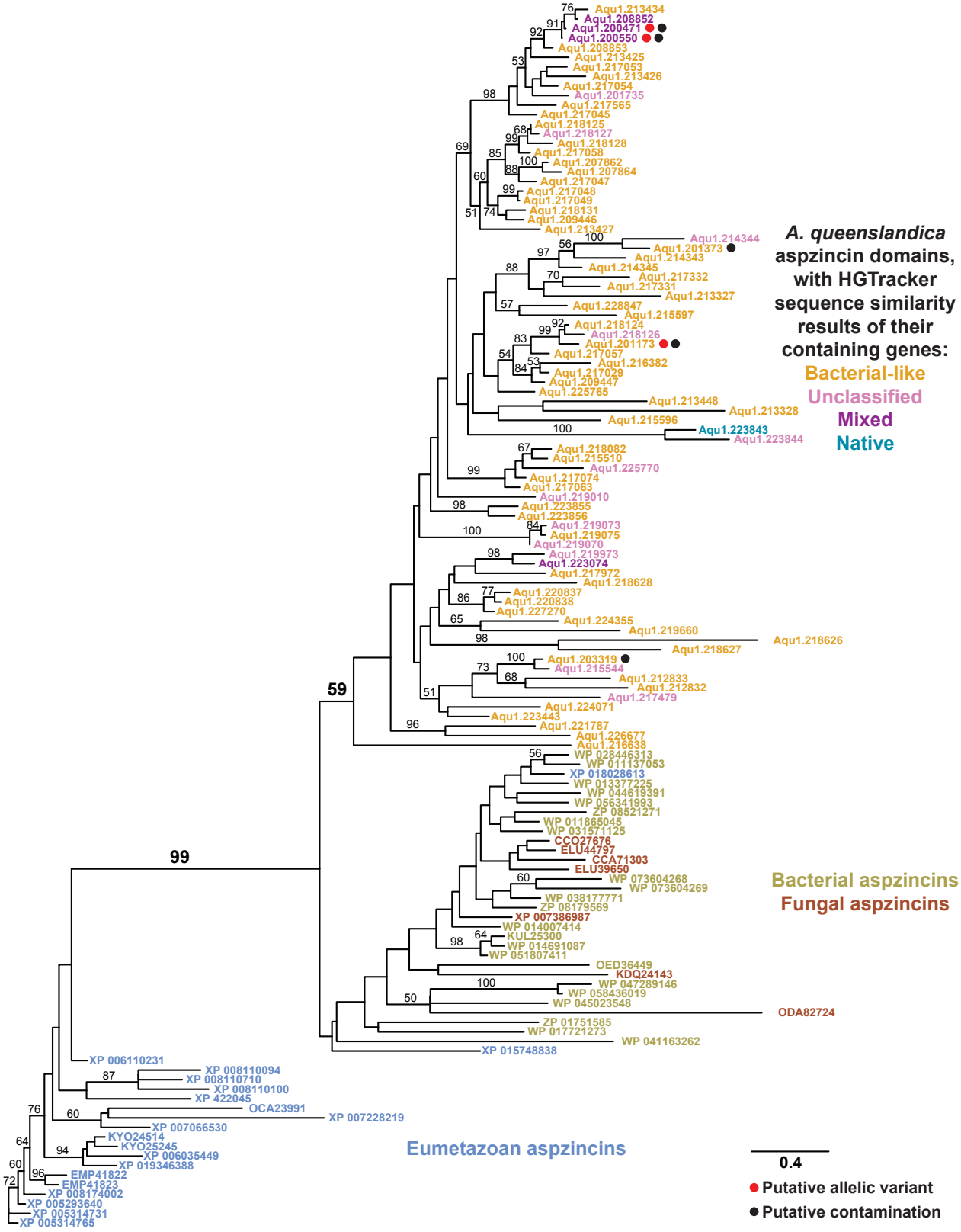


Figure 2.10 Phylogeny of the amino acid sequence for aspzincin domains from animals, fungi, and bacteria

Includes domains from 78 AqAspzs. Phylogeny inferred by ML. Topology support obtained from 250 bootstrap replicates; only support values greater than 50 and on major branches are shown. *(continued over page)*

Figure 2.10 Phylogeny of the amino acid sequences for aspzincin domains from animals, fungi and bacteria

(previous page)

Unrooted tree. Text colour reflects the taxonomy as reported by the NCBI database for the non-sponge domains. For the sponge domains, the text colour reflects the taxonomic classification assigned by HGTracker based on sequence similarity. The native *A. queenslandica* aspzincin sitting with the bacterial-like *A. queenslandica* aspzincins results from a bacterial-like aspzincin domain at the end of a large native gene. HGTracker classifies gene models based on BLASTp results for the whole gene model with at least 60% query coverage (Fernandez-Valverde et al. in preparation); hence this bacterial-like domain has been labelled as native. Red dots demark gene models likely to be allelic variants and are hereafter excluded from analyses. Black dots show the three domains found in gene models classified as more likely to be contamination by HGTracker because they are bacterial-like in sequence, but have no native genes on their scaffold - in these cases because they are alone on a small scaffold. The pink domains are unclassified because their BLASTp results did not reach HGTracker decision-making thresholds. The taxonomic class Mixed reflects when a sequence is clearly foreign, but does not clearly fall into a single taxonomic class (category X).

Three of the six AqAspz that sit alone on a scaffold have high identity for the full CDS with that of the sequences closest to them in the tree of Figure 2.10. These are marked on the tree as allelic variants and hereafter, are considered not to be unique genes, but allelic variants.

Revised domain architecture of *A. queenslandica* aspzincins

Closer inspection of the Pfam domain hits for all the AqAspzs, including those newly found genes, shows that most of the hits to other domains can be clearly dismissed because of poor e-values and query coverage percentages. In the case of three genes, the hits are good, but transcript data show that the gene models are incorrect predictions joining separate genes (Appendix 2.9). In summary, all but one of the AqAspzs are predicted to contain only a single aspzincin domain (Figure 2.11). The exception is one aspzincin gene that also contains two predicted hemopexin domains (Figure 2.11).

The hits to the domain of unknown function DUF4635 in 13 AqAspzs are short, poor and result from enrichment of leucines and tyrosines in both the actual DUF4635 domain and in these parts of the AqAspsz, since the domain and AqAspzs are otherwise quite different. In Figure 2.11 these proteins are demarked with a DUF4635–like domain, to show that they are all within the same clade and that they have a common stretch of sequence. Note that these 13 genes make up one of the other most common domain groups of the HGTs and thus are hereafter not considered as a separate domain group since as aspzincins, they have already been analysed.

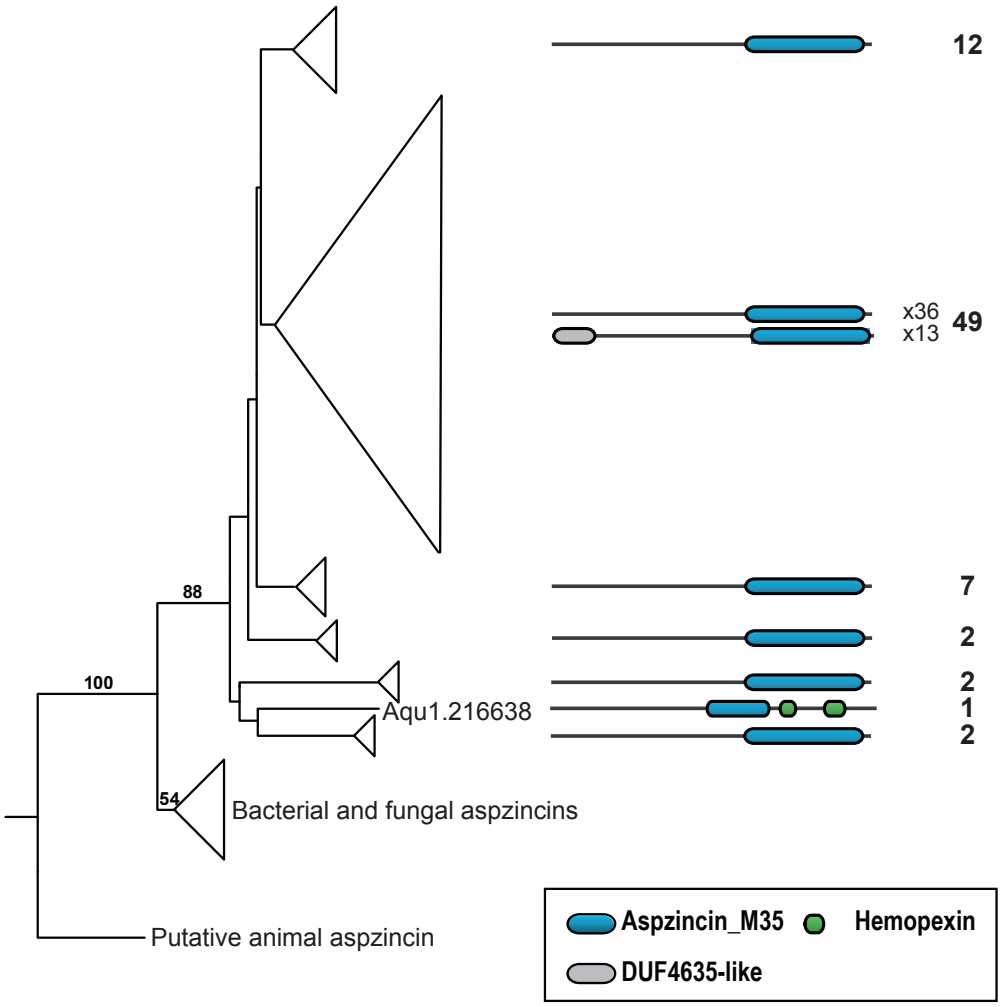


Figure 2.11 Phylogenetic distribution of the AqAspz domain architectures
Domain architecture was determined by searching the Pfam database v27.0 and then manually assessing the hits, including assessment of their length and amino acid conservation. The numbers of the far right side reflect the number of gene models in the relevant clade. Branch values indicate the typology support from 500 bootstrap replications of a ML analysis of the aspzincins (data not shown), WAG+G model of evolution. The figure is not to scale, though for each protein cartoon, the domain and protein length are proportional. While this phylogeny was constructed with 78 AqAspzs, the largest clade contains 3 likely allelic variants that are not considered here, thus the largest clade has 49 domain architectures presented and not 52.

The gene model Aqu1.216638|Aqu2.1.25761_001 contains an aspzincin followed by two hemopexin domains. Most of the hemopexin domain-containing species in the Pfam database are animals, with the exception of five plants, 15 fungi, and 39 bacteria (v31.0, accessed February 2017; http://pfam.xfam.org/family/PF00045#tabview=tab7). Aqu1.216638|Aqu2.1.25761_001 appears to be a correctly predicted gene model. There are no CEL-Seq data that suggest otherwise and four different gene model prediction methods have all generated the same gene hypothesis. Most convincingly, there are transcripts that cover the full gene model. Therefore, this gene model is of particular interest since it may be an example of novelty created by a mostly animal domain fusing with a domain of bacterial origin.

PNPs

Searching the *A. queenslandica* predicted proteome with Pfam's PNP_UDP_1 HMM

Interrogation of all Aqu1s with the Pfam PNP domain HMM (PF01048; visualised in Figure 2.12) revealed an additional 31 *A. queenslandica* PNP domain-containing genes (AqPNPs), bringing the total number to 56. Of the newly found, HGTracker classified seven as native, 15 as putative contamination and nine as unclassified. The PNP domains of *A. queenslandica* have high levels of conservation to each other and with the residues in the HMM, particularly those weighted as more important in the model; therefore, there was no need to create an *A. queenslandica* specific HMM. Submission of the AqPNPs to Pfam confirms the PNP domain assignment for all these sequences, as opposed to some other similar domain.

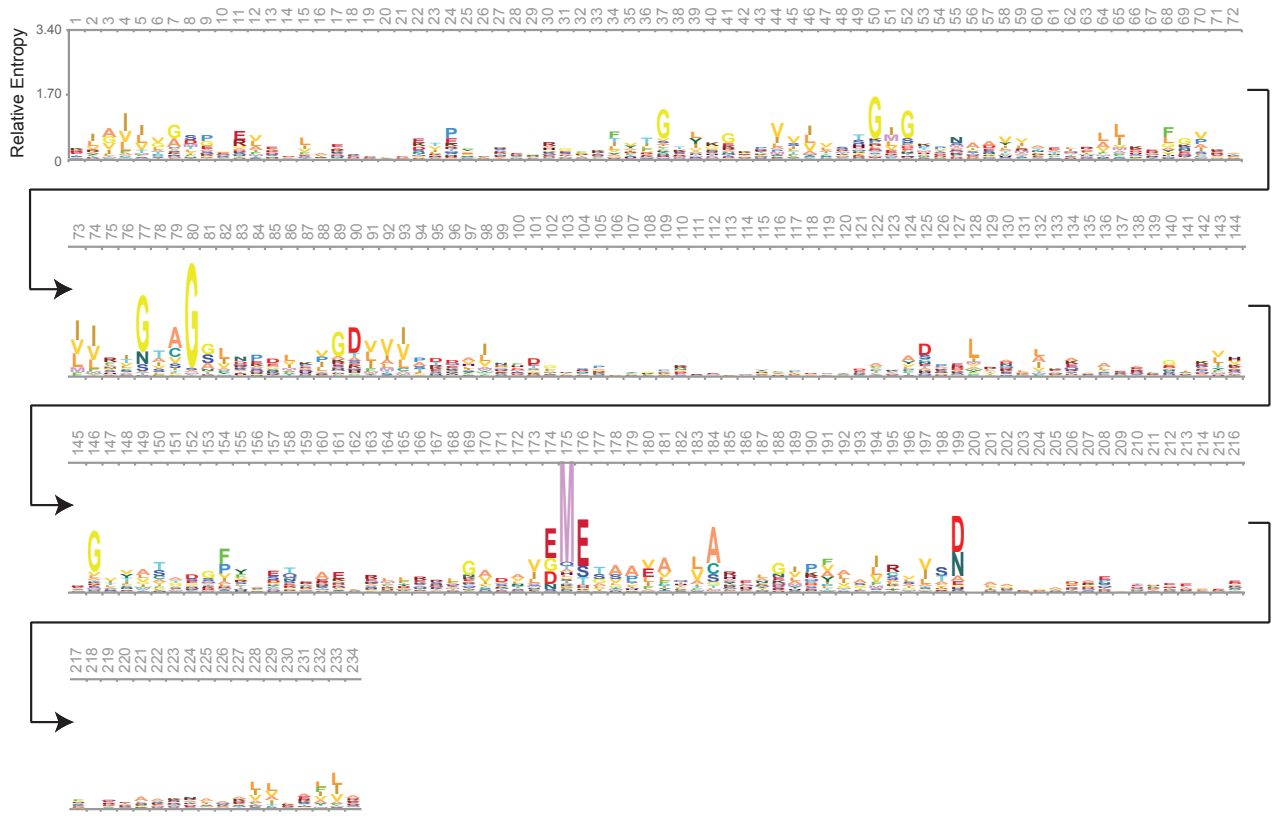


Figure 2.12 Sequence logos of Pfam's PNP_UDP_1 HMM
Model is 234 amino acids in length and built from 114 sequences. The letter size depicts the degree of amino acid conservation, with larger letters more frequently conserved than smaller letters and thus inferred functionally more important and also diagnostic. Logo was made with Skylign (www.skylign.org/; options: Create HMM – remove mostly-empty columns, Alignment sequences are full length, Information content – all) and adapted in Adobe Illustrator CS3 for ease of viewing. Colours are unique to each amino acid.

Six of the unclassified AqPNPs are bacterial-like, but are on scaffolds with other unclassified genes only, and thus could be HGTs or contamination. The other three unclassified genes receive animal BLASTp hits, but also have separate bacterial hits to the PNP domain at the other end of the proteins and are further examined below as potential HGT-native fusion genes. All the AqPNPs classified as putative contamination are bacterial-like, but each gene is alone on a small scaffold; hence, there is no positive support for their linkage with native sequence in the genome assembly.

The Pfam database reports PNP domains in 226 animal, 57 plant, 281 fungal, and 3057 bacterial species (v31.0, accessed August 2017; http://pfam.xfam.org/family/PF01048#tabview=tab7); therefore, HGTracker has probably correctly classified the seven native AqPNPs, though this will be tested below.

Phylogeny of *A. queenslandica* PNPs

To test the HGTracker classifications of the AqPNPs and to assess their phylogenetic incongruence, I collected the best BLASTp hits of representative AqPNPs classified by HGTracker as native, HGT and contamination. Phylogenetic relationships of the resulting 53 *A. queenslandica,* four bacterial and six animal PNP domains were explored based on a multiple alignment using a ML analysis. Including gaps, the final trimmed alignment is 189 amino acids in length. The best model of sequence evolution is predicted to be WAG + I + G using BIC in ProtTest 2.4 (Darriba et al. 2011). In the resulting tree, all the bacterial-like AqPNPs and the bacterial PNPs clade together to the exclusion of the animal and native AqPNPs. However, two groups of bacterial-like AqPNPs are predicted more closely related to the bacterial PNPs than the other two groups. In case more closely related bacteria to these two groups exist in the NCBI nr database yet were somehow missed, I searched the NCBI nr database with the domains from three more sequences from these two groups and reanalysed the sequences. The inclusion of these sequences did not change the length of the alignment, the best model of sequence evolution nor the topology of the tree: the bacterial-like AqPNPs and the bacterial PNPs reliably group together to the exclusion of the animal and native AqPNPs (Figure 2.13). All the bacterial PNPs form one clade, despite being species from four different phyla (Appendix 2.10). There are four main clades of bacterial-like PNPs; two sit within a clade with the bacterial PNPs to the exclusion of the two other groups. Each bacterial-like AqPNP clade of the tree has some members that are classified by HGTracker

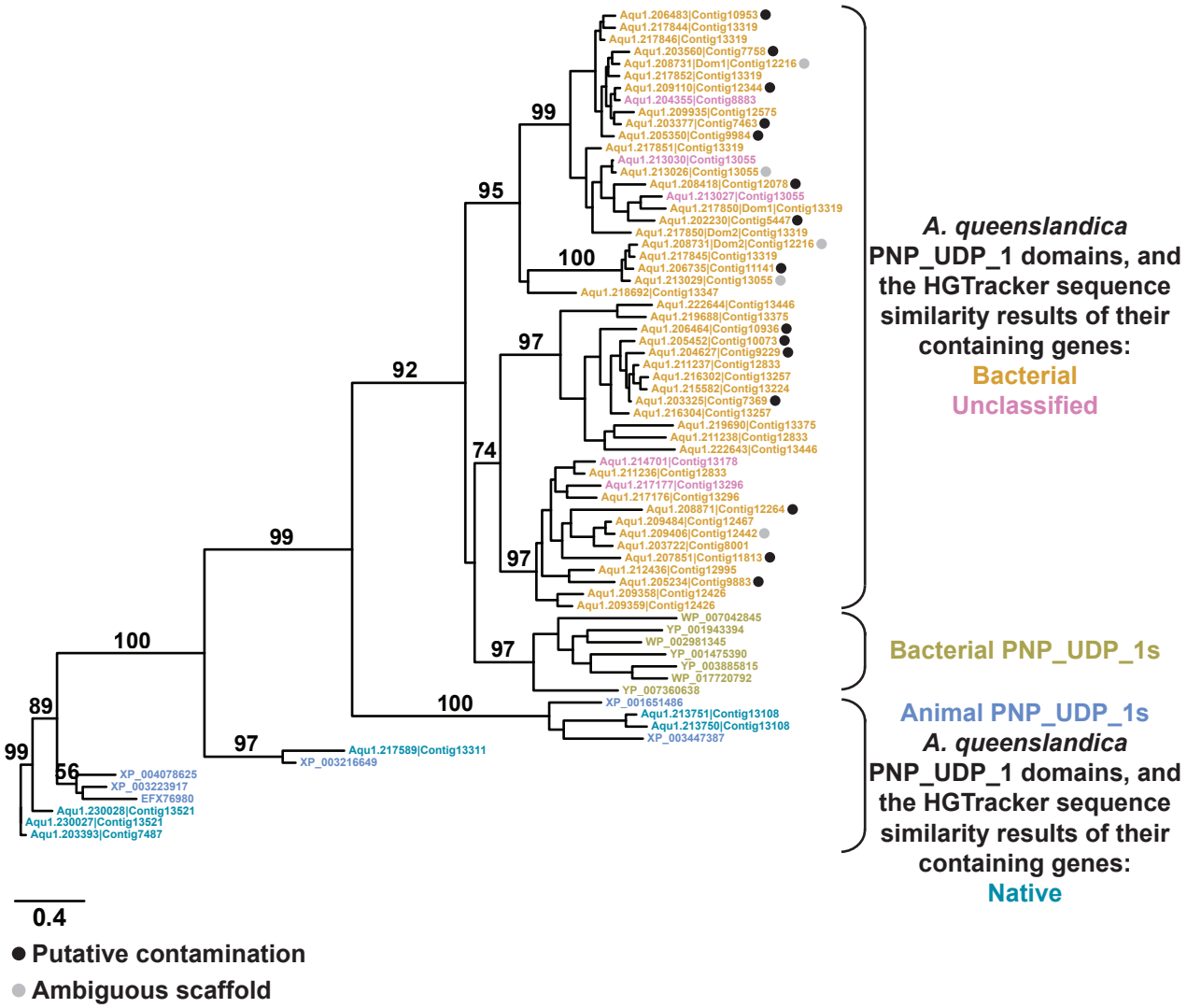


Figure 2.13 Phylogeny of the amino acid sequence for 69 PNP_UDP_1 domains from animals, bacteria, and *A. queenslandica*

Phylogeny inferred by ML. Topology support was obtained from 500 bootstrap replicates; only support values greater than 50 and on major branches are shown. Unrooted tree. Text colour reflects the taxonomy as reported by the NCBI database for the non-sponge domains. For the sponge domains, the text colour reflects the taxonomic classification assigned by HGTracker based on sequence similarities.

as HGT and others that are classified as contamination (Figure 2.13). Quartile distribution summaries of the sequence identity matrix from the multiple alignment show greater sequence identities between the PNPs of *A. queenslandica* HGTs and of bacteria than between PNPs of *A. queenslandica* HGTs and other animals. There are also greater sequence identities between the PNPs of *A. queenslandica* HGTs and of bacteria than between the PNPs of *A. queenslandica* HGT and native genes (Appendix 2.11). None of the AqPNPs that sit alone on a scaffold have high identity for the full CDS of any other *A. queenslandica* sequence. Thus, they are all considered as unique loci and not allelic variants.

Revised domain architecture of *A. queenslandica* PNPs

Manual inspection of the Pfam Batch Search results for all the AqPNPs resulted in the dismissal of some domain hits due to poor e-values and low query coverage percentages. In summary, 33 of 56 gene models contain only a single PNP domain, including all seven native AqPNPs. The domains assigned to the other 23 AqPNPs include the COR, Pkinase, ZU5, Ank_2 and Death domains (Figure 2.14). There is not a phylogenetic pattern in the distribution of domain architectures (Figure 2.14), except that the two Pkinase domain containing gene models are in the same clade together.

The sequence similarity, phylogeny and domain architecture results raise the possibility that PNP domains of bacterial origin have become linked to native sequence and formed novel protein coding genes of partial native and partial bacterial sequence. Twelve genes contain a bacterial-like PNP domain as well as at least one copy of the typically animal-only Death domain. Another gene model contains a bacterial-like PNP domain and a typically animal-only ZU5 domain. Three other genes begin with sequence that matches animal sequences, but the genes become bacterial-like for the PNP-encoding sequence. I investigated any relevant genome assembly data, alternative gene model predictions and transcriptomic data, but could not strengthen or weaken any of the putative native-HGT fusion cases and thus they may be incorrect gene model predictions (Appendix 2.12).

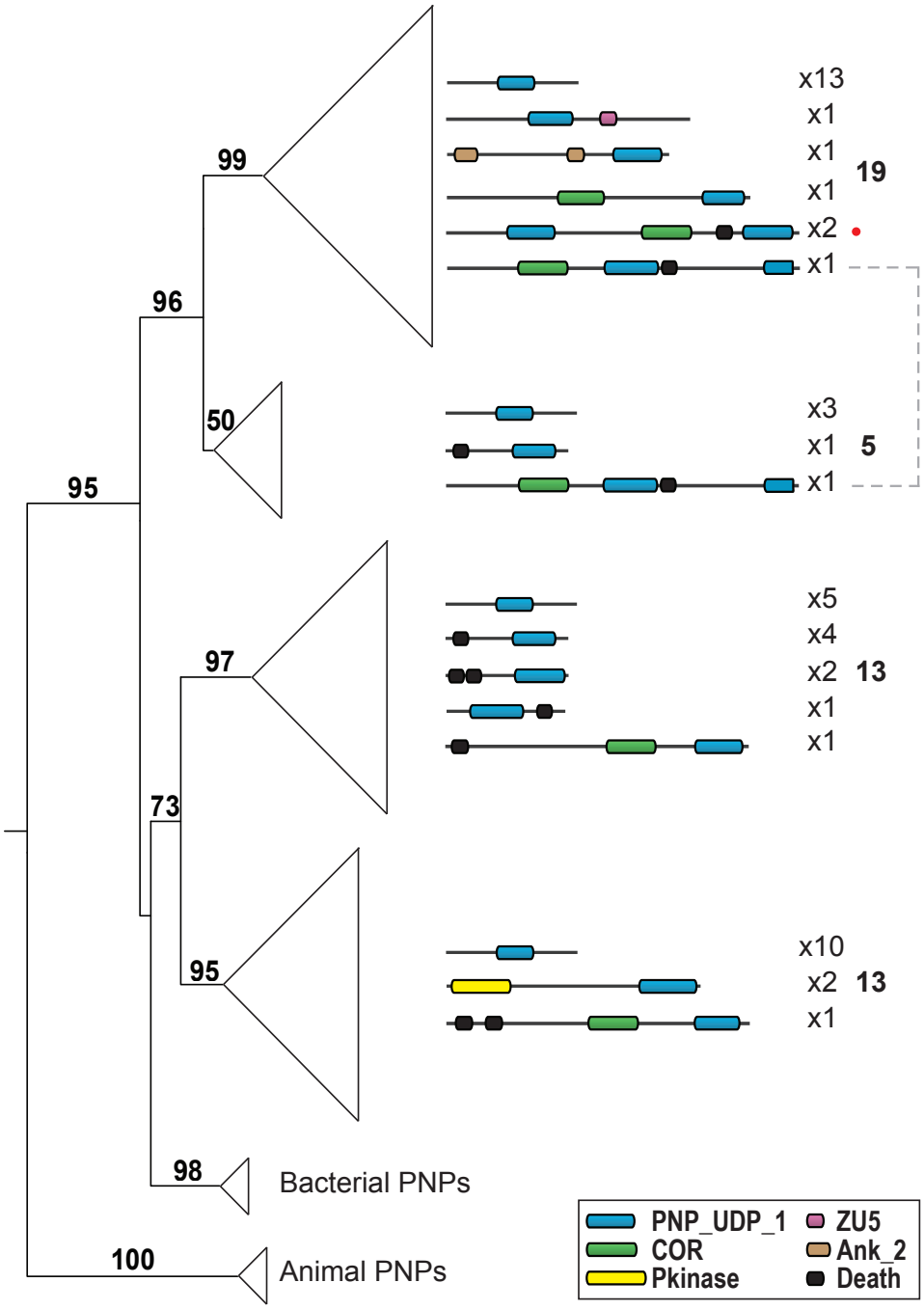


Figure 2.14 Phylogenetic distribution of the AqPNP domain architectures
Domain architecture was determined by searching the Pfam database v27.0 and manually assessing the hit lengths and the degree of key amino acid conservation. The numbers of the far right side reflect number of gene models. Branch values indicate the typology support from 500 bootstrap replications in the analysis of Figure 2.13. Not to scale, though within each protein cartoon, the domain and protein length are proportional. The red dot signifies that two of the PNP domains in the relevant clade are within the same one protein. The grey broken line shows where each domain sits in the only other protein that contains two PNP domains.

2.4.5 A. queenslandica HGTs containing domains typically absent in the Metazoa

Of the most common 15 domains in the AqHGTs, the aspzincin domain is the only one predicted by Pfam as not typically present in animals (v30.0, accessed February 2017; Finn et al. 2016; http://pfam.xfam.org/family/PF14521#tabview=tab7). To find other putative such cases, the taxonomic distribution was retrieved from the Pfam v29.0 database for all 350 domain types found in the AqHGTs. From this, 81 domain types were highlighted as not found in the UniProt reference proteomes of animals. To assess if these specific genes are functional, I examined their domain architectures and developmental expression profiles, excluding the already examined aspzincins.

a. Domain architecture

Manual assessment of the Pfam domain hits in the AqHGTs predicted to contain a nonmetazoan domain (AqHGT_NM) revealed that the hits to 36 of the domain types have poor e-values (e-values>0.1), or query coverage percentages less than 15%, or have better hits to the same region from a different domain model. Consequently, these domain assignments were discarded, reducing the total number of nonmetazoan domains in the AqHGTs from 81 to 57 and the number of implicated AqHGTs from 82 to 60 (domain architectures presented in Appendices 2.13 and 2.14).

I assessed the taxonomic distribution of the domains within the AqHGT_NM specifically looking for domain combinations suggestive of novel genes evolving from native and transferred sequences. Six cases were found and using BLASTp, I submitted the separate parts of each of those genes to the NCBI nr database (Appendix 2.15). Only one of these, Aqu1.216334|Aqu2.1.25248_001 has results suggesting the gene may consist of sequence from different taxonomic sources, with the other five genes receiving either low numbers of hits to only one of the gene parts and/or an ambiguous mix of taxonomies in the hit results (Appendix 2.15). None of the six genes are interrupted by genome assembly gaps, nor do any have different gene model predictions (Appendix 2.15). There is no expression data for three of these genes, including for Aqu1.216344|Aqu2.1.25248_001. For the other three, there are transcripts across the gene models linking the domains of interest (Appendix 2.15), though it remains unclear whether these genes actually comprise of a mix of native and transferred sequence. In sum, the original hypothesis of six putative fusion genes is not strongly supported, though remains a possibility.

The putative fusion gene Aqu1.216334|Aqu2.1.25248_001 contains a eukaryotic P-loop guanosine triphosphatase AIG1 domain and a bacterial glycine zipper motifs-containing DUF456 domain. The AIG1 domain is well conserved and contains the Walker A motif (GXXXXGK(T/S); where X is any amino acid), that is commonly found in most P-loop NTPase domains (Aravind et al. 2004). The other common feature of P-loop NTPases is the Walker B motif, (hhhhD; where h is any hydrophobic amino acid), which is downstream from the Walker A motif (Aravind et al. 2004). This is not so clear to find, but three such possible motifs exist in the gene model: two copies of ALVSD (sum of hydropathy indices of first 4 aa = 9), and QFLLD (sum of hydropathy indices of first 4 aa = 6.9). The DUF456 hit clearly shows glycine zipper motifs, which are (G,A,S)XXXGXXX(G,S,T) (Kim et al. 2005). This gene sits on a scaffold amongst native genes and other AIG1-glycine zipper containing genes, all of which are unclassified in sequence similarity by HGTracker. Therefore, because their unusual domain combination creates an inherently ambiguous situation with which HGTracker cannot cope, I searched all Aqu1 models for glycine zipper motifs and AIG1 domains using Pfam HMMs. These searches found 53 other proteins also containing the two domains, and all are unclassified by HGTracker. Because the Pfam glycine zipper HMMs (PF13436 GlyzipOmpA; PF13488 GlyzipOmp; PF13441 GlyzipYMGG) with which I searched are reported as bacterial only by Pfam (as of October 2013), yet glycine zippers in general are throughout the three domains of life (Kim et al. 2005), and because this AIG1 and glycine zipper domain combination is not found in those domain architectures listed by Pfam for both these domains (as of October 2013), I searched for this pattern in other species using two approaches. First, using the custom BLAST tool in Geneious Pro 5.1.7 (Kearse et al. 2012), I interrogated the Aqu1 glycine zipper sequences against the 810 AIG1 sequences present in the Pfam database v26.0. 194 of these received hits to a glycine rich area in the AIG1 protein. Second, using the hmmsearch program of HMMER version 3.0 (default settings including cut-off –E 10.0; Finn et al. 2011), I searched the human OrthoMCL peptides (Chen et al. 2006) with the Pfam AIG1 and Glycine zipper HMMs specified above. Thirteen human proteins are predicted to contain at least one AIG1 domain, but none of these contain a predicted glycine zipper. Thus, the domain combination of AIG1 and the glycine zipper DUF456 appears unique, but AIG1s are associated with glycine rich sequences.

b. Taxonomic sources

Based on sequence similarities, HGTracker classifies 26 of the 60 AqHGT_NMs as bacterial-like, 13 as fungal-like, three as plant-like, two as eukaryotic (but not animal, plant or fungal), and 16 as nonmetazoan, but otherwise ambiguous in taxonomic results (HGTracker classification X). Comparing the HGTracker classification of each gene in relation to the Pfam distribution of the contained nonmetazoan domain, 22 genes have inconsistent taxonomic classifications. Twenty-two others were consistently classified. Only 44 comparisons were possible because of the 16 genes not specifically classified by HGTracker, beyond the nonmetazoan X classification.

c. Expression profiles throughout developmental time

Eighteen of the 60 AqHGT_NMs do not have a single developmental stage with a normalised count of five or more (Appendix 2.13). These genes include the two putative fusion genes that each has a Class_III signal domain and an E1_DerP2_DerF2 domain. The other 42 genes have inferred meaningful expression. These genes have a range of developmental expression profiles; some are more highly expressed early in development in embryos and larvae (n=8), some are more highly expressed in postlarval and juvenile stages (n=24), and others are more highly expressed in adult sponges (n=10; Appendix 2.14). No apparent pattern exists between these different expression profiles and domain content (Appendix 2.14).

The majority of the genes that have a consensus taxonomic classification from HGTracker and the Pfam domain content are expressed (19 of 22). Similarly, 13 of those 16 genes that were not classified by HGTracker further than the nonmetazoan X are expressed. However, not even half of the genes with conflicting taxonomic classifications are expressed (10 of 22).

2.4.6 A note on the complete record of A. queenslandica aspzincins

After the analyses of this chapter and Chapter 3, the latest gene model predictions, the Aqu2.1s, were searched for aspzincin domains. Four new AqAspzs were found from searches with the Pfam aspzincin HMM and five others were uncovered with the HMM built from 52 *A. queenslandica* aspzincins (as described in section 2.3.3). These sequences are not predicted in any of the Aqu1s. Since these nine gene models are likely not to significantly change the outcomes of this work, for feasibility purposes,

I have not incorporated these genes into already completed work. General properties of all AqAspzs known to date are detailed in Appendix 2.16.

2.5 DISCUSSION

2.5.1 Many A. queenslandica HGTs are predicted to be functioning in diverse roles and some are putatively co-regulated

The majority of the AqHGTs contain at least one recognisable putative domain and collectively, these genes contain a wide range of different predicted domain types. The conservation of domains in HGTs suggests they were possibly quite recently transferred and/or are being used by their new host and thus are functionally constrained (Ober 2010; Podlaha and Zhang 2010). Most of the bacterial-like HGTs identified in *A. queenslandica* by Conaco et al. (2016) likely result from old transfer events, since all have taken on host GC content and codon usage characteristics, and many are found in at least one or two other sponge species. Similarly, Fernandez-Valverde et al. (in preparation) found that *A. queenslandica* native genes and HGTs of any source have comparable GC content and thus also suggest that the HGTs have had time to adapt to their new genomic environment. Therefore, the conserved domains identified in the AqHGTs likely do not result from a lack of decay time because of recent transferal events, but probably result from functional constraint. The notion of many AqHGTs experiencing functional constraint is further supported by their differential expression through development. Transcription of non-functional genes does occur; for instance, the ENCODE pilot project reported that most of the human genome is transcribed (Birney et al. 2007). However, the ontogenetic expression patterns and the signs of possible co-expression found for many of the AqHGTs suggest their expression is not just background transcription. It appears likely that the AqHGT_PNPs are under at least two different suites of regulatory control, with most transcribed in late development, but some mid-development. The AqAspzs have at least six expression profiles and thus likely at least six co-regulated groups, each quite specific to embryo stages, postlarval and juvenile stages, or to juvenile and adult stages.

The inferred active state of many AqHGTs suggests that HGT has considerably impacted the functional genome of *A. queenslandica,* particularly because these genes encode a vast array of capabilities. The 350 different types of domains detected belong to 104 different Pfam clans, though 140 of the domain types detected do not belong to any clan. A Pfam clan contains members related by sequence

or structure: if a domain family has members that cannot be detected by one HMM, Pfam will split the domain family into however many members necessary for complete detection, each member with its own HMM and all members grouped in one clan (Finn et al. 2006). Some of the domains contained in the AqHGTs are members of the large P-loop NTPase clan; these often have chaperone roles assisting in the assembly, disassembly or operation of protein complexes (Neuwald et al. 1999). The AqHGTs also contain tetratricopeptide repeats (TPRs) that are present in many diverse proteins with roles in protein folding, neurogenesis, transcriptional control and cell cycle regulation (Goebl and Yanagida 1991). More specifically, TPR domain-containing proteins have been predicted important in the cross-talk between *A. queenslandica* and its symbionts (Fieth et al. 2016); thus genes possibly involved in host-symbiont co-evolution have been horizontally transferred to the host, putatively enabling communication between the partners (Fieth et al. 2016). The AqHGTs have domains from clans involving a variety of peptidases, transcription factors, membrane transporters, and mobile elements such as retrotransposons, retroviruses, and polymerases from a variety of RNA viruses. Such a broad range in the capacities of the functional HGTs of *A. queenslandica* indicates flexibility in the genome and further supports claims that HGT fuels biochemical diversification in a wide range of animals (Gladyshev et al. 2008; Werren et al. 2010; Boschetti et al. 2012; Crisp et al. 2015; Conaco et al. 2016; Sieber et al. 2017).

A focus of HGT research is on the type of genes more likely to be retained in a new host genome, and thus more likely to be functional and to have evolutionary impact (Jain et al. 1999; Gophna and Ofran 2011; Park and Zhang 2012). Based on empirical studies of the types of genes detected as HGTs in bacteria, the complexity hypothesis predicts that proteins involved in fewer interactions are more likely to become functional and useful to their new host, and thus are more likely retained (Rivera et al. 1998; Jain et al. 1999; Moran et al. 2012). In continuation of this hypothesis, the protein function rule emerged; this rule expects that operational genes, like those involved in cellular processes, are more likely to be successful functional HGTs because their function relies on fewer interactions (Rivera et al. 1998; Jain et al. 1999; Park and Zhang 2012). The protein function rule also predicts that informational genes such as those involved in DNA replication, transcription and translation are less likely to be functional HGTs because they often produce small components of large protein complexes such as the ribosome; therefore, they are predicted to be less adaptable for a host (Rivera et al. 1998; Jain et al.

1999; Park and Zhang 2012). Boschetti et al. (2012) report that 83% of the alien genes in the bdelloid rotifer *Adineta ricciae* are enzymes and Crisp et al. (2015) found a similar trend of enzyme enrichment in the HGTs of primate and nematode genomes. After enzymes, Crisp et al. (2015) found the second largest group of alien genes are membrane-bound proteins, which are also operational genes. In line with these studies, the complexity hypothesis, and the protein function rule, many of the AqHGTs are operational genes, most particularly enzymes. While the AqHGTs also contain informational genes, they are a small proportion. However, trends exactly opposite to these findings and theories are also reported. For example, using computational methods, Gophna and Ofran (2011) found horizontally transferred genes have far more interaction sites in comparison to native genes. Their results suggest that proteins with greater chances of making new interactions in their new environment are more adaptable and more likely to become useful, thus are more likely to be conserved and detected as HGTs (Gophna and Ofran 2011). The lack of consensus in reported findings probably reflects that there are many factors influencing the trajectory of transferred genes in new genomic environments.

HGT success may not only be influenced by the adaptability of HGTs, but also by different strengths of negative or positive selection on HGTs arising from their immediate benefit or harm to the host (Moran et al. 2012). Genes for pore-forming toxins in the aerolysin family have been independently transferred to species of different kingdoms and are conserved because they confer immediate defence and predatory advantages to their new hosts (Moran et al. 2012). Further, independent transfers from bacteria to *Saccharomyces cerevisiae* of genes involved in the biotin synthesis pathway were also apparently immediately advantageous on arrival to their host (Hall and Dietrich 2007). While such immediately beneficial HGTs exist, they may not represent the most common trajectory of HGTs, possibly because such instant beneficial HGT-host compatibility may be rare. Park and Zhang (2012) considered bacterial and archaeal HGTs and found that the level of gene expression is a more important factor than the type of gene, in respect to determining the chances of successful transferal. Strong selection against highly expressed HGTs may occur because of the fitness costs of that expression to the host (Park and Zhang 2012). HGT transcription and translation cost energy and may reduce cellular efficiencies, change optimal concentrations, and cause harmful protein interactions and misfolding (Park and Zhang 2012). Successful HGTs may more commonly become fixed not because they offer benefits to the host, but because their initial low expression levels reduce selection against them, thus

giving them time in the host (Park and Zhang 2012). This is consistent with previous findings that many HGTs in bacteria are pre-adaptations – initially neutral or almost neutral to the new host, but with an environmental change in time, they become beneficial (Gogarten and Townsend 2005).

The lower expression of the putatively younger AqHGTs further supports the hypothesis that higher expression hampers HGT success. An important caveat is that these younger AqHGTs are classified as such because they do not have any introns and they are not present in the transcriptomes of two other sponge species. Spliceosomal introns are absent in bacteria (Rogozin et al. 2012), so the presence of such an intron in a bacterial-like HGT in an animal probably reflects time spent in the host. However, since many animal genes contain no introns, absence of introns does not necessarily mean a gene is not animal-like in gene structure. For example, in the first generation gene models of *A. queenslandica,* the median number of exons is five, but the mode is one (Srivastava et al. 2010). Further, the lack of transcripts in two other sponges may not reflect gene absence, but simply no expression of those genes for those particular samples that were sequenced. In addition, ancient HGTs may have been lost in certain sponge species, but retained in others. Nevertheless, the predicted 17 most recently transferred AqHGTs are expressed less in comparison to both native genes and inferred older HGTs. Their low expression corresponds to lower fitness costs to their new host, thus lowering negative selection against the new arrivals and so they persist in the genome. Just as many bacterial HGTs are/were pre-adaptations (Gogarten and Townsend 2005), the inferred younger HGTs of *A. queenslandica* suggest that also in animals, many HGTs were initially pre-adaptations that in time became advantageous.

The large number of different domain types collectively contained by the AqHGTs implies that the number of apparent HGTs in *A. queenslandica* does not result from only a few transfer events followed by duplication. Rather, it seems many different genes have been transferred independently. This adds to the growing support for the notion that HGT in animals is more common than previously assumed (e.g., Gladyshev et al. 2008; Danchin et al. 2010; Moran and Jarvik 2010; Acuña et al. 2012; Boto 2014; Boothby et al. 2015; Crisp et al. 2015; Danchin et al. 2016; Martinson et al. 2016). Further, Fernandez-Valverde et al. (in preparation) report that the predicted sources of the AqHGTs include archaea, bacteria, fungi, plants and other eukaryotes. The large differences in both the biology of these life forms and in the ways in which they interact with *A. queenslandica* indicate not only that

HGT in animals is more common than previously thought, but also perhaps the vastly different donor taxonomies reflect more than one transferal mechanism. For instance, one common theme of many animal HGTs is their probable origin in symbionts of the host, for instance transfers from the germline cell endosymbiont *Wolbachia* to numerous arthropod hosts (Kondo et al. 2002; Dunning Hotopp et al. 2007; Nikoh et al. 2008; Aikawa et al. 2009; McNulty et al. 2010; Werren et al. 2010). The close physical relationship of the partners results in higher chances of genetic material being transferred, especially if the symbiont is endocellular, already within the cells of the host animal and with one less physical barrier for transfers (Dunning Hotopp et al. 2007; Boto 2014). Further, germline cell endosymbionts have higher chances of transfers becoming heritable, since transfers to somatic cells are not heritable (Dunning Hotopp et al. 2007; Boto 2014; Crisp et al. 2015). All animals interact with bacteria and increasingly the importance of symbiotic bacteria to animal biology is being shown across the Metazoa (Webster et al. 2010; Hentschel et al. 2012; McFall-Ngai 2013; Alegado and King 2014; Degnan 2014; Levin et al. 2014; McFall-Ngai et al. 2015). *A. queenslandica* is not the exception; beyond being in a typical animal environment teeming with bacteria, multiple types of bacterial relationships have been characterised for *A. queenslandica* and for sponges in general (Hentschel et al. 2012; Degnan 2014; Fieth et al. 2016; Gauthier et al. 2016).

The aspzincin and PNP domain phylogenetic analyses presented here have not identified a specific bacterial or fungal taxonomic origin of these genes; however, general comparisons between the taxa in the BLAST results of these analyses and the taxa present in the *A. queenslandica* microbiome show some similarities. The microbiome of *A. queenslandica* includes at least six putative vertically-inherited bacterial symbionts, which are transferred from mother to embryo in nurse cells that also contain bacteria as a food source (Fieth et al. 2016). The majority of both the microbiome and the BLAST hits of the AqAspzs belong to the Gammaproteobacteria class, with the three orders of four *A. queenslandica* symbionts making up 19% of the BLAST results (Chromatiales, Oceanospirillales and Alteromonadales). The AqHGT_PNP BLAST results also contain shared taxa with some of the main symbionts (from the Gammaproteobacteria and the Deltaproteobacteria). Conaco et al. (2016) report of similar sequences to some *A. queenslandica* HGTs, including some AqAspzs and AqHGT_PNPs, in two other sponge species; therefore, it is not surprising that a more precise taxonomic source of these HGTs cannot be identified in the databases, since the sequences available likely belong to species

only descended from the original HGT sources. However, given the similarities at the phylum, class and even order levels, it remains plausible that relatives of the contemporary symbionts, and perhaps historical *A. queenslandica* symbionts, may be one source of some of the AqHGTs.

The broad range of species returned in the BLAST searches of the AqAspzs and AqPNPs do not exclude the possibility of other non-symbiont HGT sources too. Sponges, including *A. queenslandica,* are unlike most animals because they do not segregate germ cells early in embryogenesis; rather, sponges continuously segregate germ cells from stem cells and thus have increased chances of transfer events becoming heritable (Juliano and Wessel 2010; Degnan 2014). Environmental fungal and bacterial DNA released from phagocytosis by *A. queenslandica* are potential reservoirs of DNA ever-present in the sponge ecosystem. The mechanisms behind such foreign DNA becoming incorporated into a new host genome are intriguing and remain unknown (Dunning Hotopp 2011; Wijayawardena et al. 2013).

2.5.2 Enrichment of some domains in the HGTs of A. queenslandica

While there are large amounts of diversity in the domain content of the AqHGTs, there are also enrichments of certain domains. In fact, one third of the domains are found in more than one AqHGT. Most of these domains are found in only a few AqHGTs, but fifteen are found in seven or more genes. The numbers of the reported enriched domains presented in this chapter, for example 25 RVT_2 domain-containing AqHGTs, are probably underestimates, since targeted searches of the aspzincins and PNPs revealed approximately 20 more of each.

Many of the genes within each enriched AqHGT group are expressed. The proper regulation of HGTs in their new genomic environmental is an obscure factor of HGT in animals. Possibly existing regulatory systems in the host are co-opted or regulatory elements are also horizontally transferred, as found by Shin et al. (2016), who show that DNA methylation patterns can also be horizontally transferred along with genes in bacteria. The enriched AqHGTs have a pattern that within each domain group, most genes are expressed or most are not expressed. Within the expressed groups, only two groups show signs of possible co-regulation (the aspzincins and the PNPs); therefore, the mostly on or mostly off pattern does not reflect global regulation of each group. Half of the domain groups that contain no expressed genes are related to transposable elements (TEs) (the following ten out of the total 19

groups: RVT_2, rve, DDE_3. Helitron, PIF1, gag_pre-intergs, UvrD_C_2, HTH_33, DUF3051, Herpes_Helicase). The transposition rate of active TEs probably varies through evolutionary time (Huang et al. 2012); for instance, it may increase under certain environmental conditions including stress (McClintock 1984; Strand and McDonald 1985; Slotkin and Martienssen 2007; Zeh et al. 2009). In *D. melanogaster,* transposition rates are estimated as one per thousand to a million generations, per element copy (Nuzhdin and Mackay 1995; Domínguez and Albornoz 1996). In *S. cerevisiae* the retrotransposon Ty1 has an estimated transposition rate ranging from once every few months to every few years (Paquin and Williamson 1984). Therefore, the lack of transcription of the HGT-derived TEs in *A. queenslandica* may signify that they are no longer active, that they have not undergone molecular domestication and thus are not functionally expressed, or simply that the analysed transcription database is a single snapshot in time and has not captured their activity. Because of the predicted ancient nature of most of the detected HGTs of *A. queenslandica* (Conaco et al. 2016; Fernandez-Valverde et al. in preparation), there may have been time for the evolution of host silencing of these once foreign TEs, possibly via DNA methylation and RNA interference as demonstrated in other species (Miura et al. 2001; Yang and Kazazian 2006; McCue et al. 2012; Saito 2013; Wheeler 2013).

2.5.3 HGT or HGT-derivative? Overestimation of HGT due to post-transfer duplications

The aspzincin and PNP domains are just half a per cent (2 of 350) of all the domain types found in the AqHGTs yet are found in a large proportion of the AqHGTs (146 of 576; 25%). These enrichments prompted the question of if they result from many independent transfer events or alternatively, if they are the outcome of HGT followed by duplication. For both domain groups, phylogenetic analyses support the latter hypothesis.

Regardless of their HGTracker classification (likely HGT, likely contamination or ambiguous), all the AqAspzs clade together in the phylogenetic prediction, thus even those classified as contamination or ambiguous are inferred as truly incorporated in the *A. queenslandica* genome. This conclusion is further supported since all the AqAspzs classified as likely contamination are so classified because they sit alone on a scaffold, and not because they sit in an alien dense scaffold. If some of the AqAspzs were genomic contamination, they would be expected to be more closely related to the bacterial aspzincins and would thus be in the bacterial and fungal clade.

The grouping of the AqAspzs together to the exclusion of all other sequences also suggests that there was just a single horizontal gene transfer event, followed by extensive duplication and divergence. A completely different tree would support the alternative hypothesis of many independent transfer events – the AqAspzs would not all clade together, rather they would be expected to lie amongst the bacterial aspzincins. Based on the clades and branch lengths of the ML topology, it appears that the AqAspzs have had both ancient and more recent duplications; however, that inference depends on the rates of divergence after duplication. An alternative scenario is that the predicted ancient duplications are in fact up to six independent transfer events. Distinguishing between one to six transfer events is not possible to date; however it is a significant finding that these genes exist in *A. queenslandica* not due to 90 transfer events, but because of one or few transfer events followed by huge expansion due to duplication and divergence.

The distribution of the aspzincin domain in animals, along with the phylogenetic analysis and domain architecture results presented here suggest that the aspzincins of *A. queenslandica,* the vertebrates and the invertebrates have separate evolutionary trajectories. The non-sponge invertebrate aspzincins are either HGTs or bacterial contamination, based on their predicted phylogenetic position within the bacterial and fungal group. The vertebrate sequences are different to all the other aspzincins analysed, because of both their clear sequence signal revealed in the phylogenetic analysis and also because approximately half of them have a vertebrate-unique domain architecture of at least one Apolipoprotein L (ApoL) domain in addition to their aspzincin domain. Because they are phylogenetically distinct from the bacterial and fungal aspzincins, they are unlikely to be contamination, unless the contaminating source is very different to taxa already sampled and deposited in the NCBI nr database. In addition, the aspzincin assignment given to these genes is well supported, thus they are unlikely to be a metallopeptidase relative, since the aspzincin model is the best match to these sequences and they have conserved the key catalytic residues typical for aspzincins. Therefore, since they are both true aspzincins and not the result of contamination, there are three possible hypotheses for their presence in these vertebrates. First, they could be the result of vertical inheritance, yet massive amounts of independent gene loss in most animal lineages has resulted in the rare distribution of aspzincins in animals. Gene loss is an important evolutionary force (Salzberg et al. 2001; Iyer et al. 2004; Wolf and Koonin 2013; Ku and Martin 2016; Salzberg 2017); however, given both the rare and phylogenetically widely separated

animal distribution of aspzincins (i.e. some sponges and these select vertebrates), a huge amount of independent gene loss is an unlikely explanation. A second possible scenario is that the aspzincin genes may have evolved *de novo* in vertebrates, were horizontally transferred to bacteria or fungi, and then were transferred from bacteria or fungi to the sponge lineage. While this hypothesis is extraordinary because most contemporary organisms encoding aspzincin genes are bacteria or fungi, it does explain the phylogenetic prediction presented here. Finally, a third hypothesis is that the vertebrate aspzincins are descendants of an independent HGT, quite separate to the sponge aspzincins (Figure 2.15). Under this hypothesis, the donor of the vertebrate aspzincins could be bacteria or fungi, but post-transfer divergence has rendered the vertebrate aspzincins distinct to those of contemporary bacteria, fungi and *A. queenslandica.* Alternatively, the vertebrate aspzincins could be distinct to others in current sequence databases because their donor was a plant or protist that is now extinct, no longer encodes aspzincin genes, or has not yet been sequenced.

The AqAspzs are well supported as more closely related to the bacterial and fungal aspzincins than to the vertebrate aspzincins, although this may be influenced by variation in molecular evolution rates among the lineages analysed. Therefore, the phylogeny of the aspzincin domain can be interpreted as the result of either vertical inheritance followed by extensive gene loss in most lineages in the web of life or of three independent HGT events, followed by far lower amounts of gene loss in only three lineages (Figure 2.15). The vertical inheritance and gene loss hypothesis requires aspzincins to have been conserved in the common ancestors of contemporary eukaryotic species for the long time period from the origin of eukaryotes until the divergence of Reptilia and Mammalia. Therefore, under this scenario, the aspzincin genes probably had useful functions for the common ancestors during this time, but the gene became unnecessary to most life forms and was lost almost entirely from eukaryotes, though preserved in some fungi and the aforementioned vertebrate and sponge species. Because of the taxonomic distribution of aspzincin genes in the sequenced web of life, and the wide phylogenetic distances between these bacterial, fungal, sponge and vertebrate taxa, the hypothesis of three HGT events causing the sparse aspzincin distribution is more parsimonious. Figure 2.15 presents this HGT scenario under the assumption that aspzincins first arose in bacteria, and were horizontally transferred to fungi, sponges and vertebrates; however, as discussed, the origin of the aspzincin gene is open for questioning. Important here, under the vertical inheritance and gene loss hypothesis, Figure 2.15

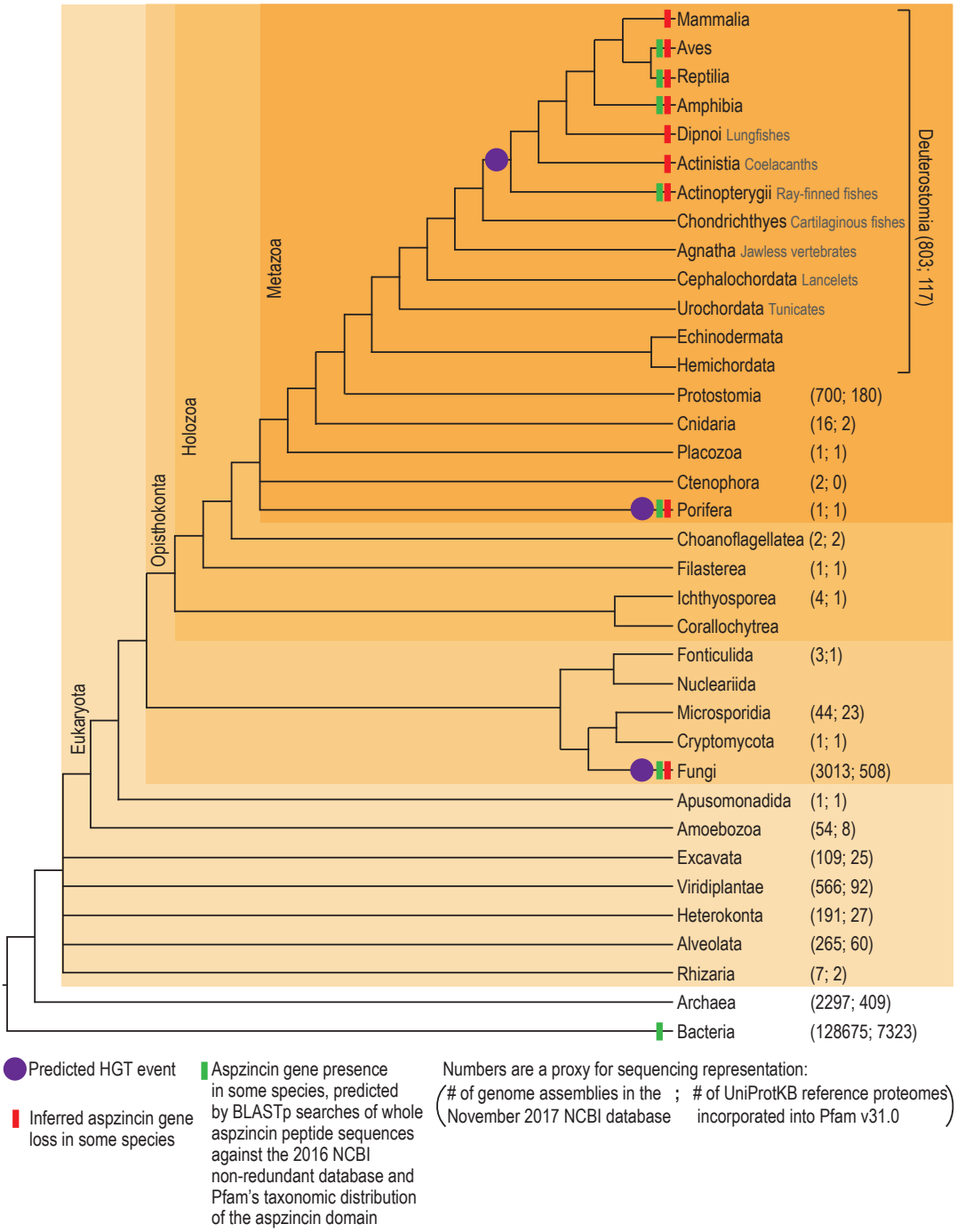


Figure 2.15 The distribution of aspzincin genes and sequenced genomes
Necessary events for the HGT hypothesis are shown, with green markers showing aspzincin genes in seven lineages, red markers showing gene loss in nine lineages and purple circles representing HGT events. The numbers in brackets indicate sequencing representation in the two databases used to find the distribution of the aspzincin gene. With increased sequencing, new aspzincin genes may be discovered in other lineages, which may increase the number of required HGT events, thus making the HGT hypothesis less likely than the alternate hypothesis of vertical inheritance followed by gene loss (if shown on this figure, the alternate hypothesis would require green markers on 20 stems and seven branches, and at least 20 red markers). Note that if the origin of aspzincin genes is not in bacteria as depicted here, three alternative HGT events still explain the contemporary phylogeny and taxonomic distribution of the gene. Groups with uncertain phylogenetic positioning are presented as polytomies. The redacted phylogeny presented is summarised from Baldauf (2003), Niklas and Newman (2013), Torruella et al. (2015), Simion et al. (2017), and Whelan et al. (2017). Not to scale.

would require green markers all through the tree conveying the vertical inheritance of the aspzincin gene from bacteria to vertebrates, as well as red markers in most lineages conveying all the gene loss events. With increased sequencing, the HGT conclusion made here for the aspzincins, and that for other predicted HGTs documented in the literature, may change in favour of the gene loss hypothesis. Regardless of whether they result from HGT or not, the animal aspzincin genes are a fascinating case. Not only have aspzincins expanded in *A. queenslandica*, but also in six of the 16 vertebrate species in which they have been identified. The aspzincin expansions in the vertebrate species are smaller than in *A. queenslandica*; nonetheless, the conservation of gene duplicates signifies that they have an important biological role (Andersson et al. 2015).

The conclusion that the AqAspzs result from a low number transfer events followed by high amounts of duplication is similar to the conclusion made for the AqHGT_PNPs. Some of the identified PNP domains are within *A. queenslandica* genes classified by HGTracker as native, likely contamination and ambiguous. Therefore, phylogenetic analysis of this domain group allowed better testing of classifications made by HGTracker, since the aspzincins were not found in any native genes. In the resulting PNP phylogenetic hypothesis, all the bacterial-like AqPNPs and the bacterial PNPs reliably group together to the exclusion of the native AqPNPs and the animal PNPs, which are predicted in the tree to be more closely related to each other. Therefore, the tree independently supports the HGTracker classifications and the splitting of the AqPNPs into bacterial-like and animal-like. All the bacterial PNPs form one clade. There are four main clades of bacterial-like AqPNPs; two sit within a clade with the bacterial PNPs to the exclusion of the other two bacterial-like AqPNP groups.

Each bacterial-like AqPNP clade of the tree has some members classified by HGTracker as HGT and some as likely contamination. Thus as for the AqAspzs, I infer that all the bacterial-like AqPNPs are in fact incorporated into the sponge genome, and none are contamination. This is rational because if there were truly contaminating bacterial PNPs in the genome, they would have fallen into the bacterial PNP clade. Further, HGTracker does not classify putative contaminants based on positive support; rather, the classification is solely based on the lack of a native gene on the scaffold. As a result, smaller scaffolds, or scaffolds composed entirely of sequence from large transfer events, will be incorrectly predicted as

likely contamination. None of the four main bacterial-like AqPNPs clades and their typology can be explained by some clades being contamination and some being HGTs.

The two groups AqHGT_PNPs that are not as closely related to the bacterial PNPs as the other AqHGT_PNPs have two possible explanations. First, since their best hits clade with the other bacterial PNPs, perhaps their donor bacterium is less related to contemporary bacterial sequences in the NCBI nr database, either because of sampling limitations or a longer extinction time. Alternatively, these more divergent AqHGT_PNPs may be the result of considerable divergence and then duplication. Currently there is no resolution to whether this hypothesised divergence occurred to a transferred PNP, or to a duplicate of an already transferred PNP. Under both scenarios, the lack of donor relatives in the databases or the considerable post transfer and/or duplication divergence, all the AqHGT_PNPs appear to be the result of only two to four HGT events followed by duplication.

Here, I have shown that gene duplication is the mechanistic cause for most of the genes in the two largest domain groups found in the HGT-derived genes of *A. queenslandica*. The duplication of genes, chromosomes and even whole genomes has long been thought a key creator of evolutionary innovation, since it creates new sequence free to evolve without any functional constraint (Haldane 1933; Fisher 1935; Muller 1936; Ohno 1970; Andersson et al. 2015). HGT followed by duplication has already been reported in animals; for instance Crisp et al. (2015) suggest one transfer event followed by duplication in *Homo sapiens* is responsible for each of the following three cases: the three hyaluroninan synthases, the four peptidyl arginine deiminases and nine PRAME family members. Wasps are thought to have thirteen proteins acquired from a transfer from *Wolbachia* and subsequent duplications (Werren et al. 2010). In the pea aphid, five genes are predicted to result from one transfer event followed by duplication (Nikoh et al. 2010). Of 34 HGTs detected in oomycetes, Richards et al. (2011) found that 21 have undergone post-transfer duplication, and that for some genes, there were large numbers of duplicates. Extensive post-transfer duplication resulting in large numbers of HGT-derived genes has also been reported in nematodes (Danchin et al. 2010). Therefore, the extensive post-transfer duplication of HGTs in *A. queenslandica* fits with reports of HGT trajectories in other animals.

The conclusion that many of the apparent AqHGTs are the result of post-transfer duplication, and are thus HGT-derived genes and not HGTs, highlights another source of possible error that contributes to the contention around HGT in animals (Dunning Hotopp 2011; Wijayawardena et al. 2013; Boto 2014). Identifying and providing convincing support for HGT in animals is particularly complicated by the close associations animals have with other life forms (Becq et al. 2010; Dunning Hotopp 2011). Distinguishing between the sequence of an animal and that of their symbionts, environmental contaminants and even experimental contaminants is often difficult but crucial (Dunning Hotopp 2011). Finally, other evolutionary processes that could also explain putative HGTs, such as gene loss and variable rates of evolution, may be too quickly overlooked (Ku and Martin 2016; Salzberg 2017). Reported false positives of HGTs include some controversial and/or unsubstantiated claims, for instance the transfer of glyoxylate cycle genes to non-placental vertebrates (Kondrashov et al. 2006), and the famous photosynthesising sea slug *Elysia chlorotica* (Rumpho et al. 2008; Rumpho et al. 2010; Pierce et al. 2012; Bhattacharya et al. 2013). The reporting of extensive HGT in the tardigrade *Hypsibius dujardini* (Boothby et al. 2015) received fast rebuttal and commentary (Bemm et al. 2016; Richards and Monier 2016; Koutsovoulos et al. 2016). These examples show first, that high quality support for HGT claims is important, such as demonstration of functionality in the recipient organism, perhaps shown by tissue-specific transcription or by correlating a phenotype with the presence of HGT(s) (Dunning Hotopp 2011; Boto 2014). Second, these cases show that careful consideration of outputs from genomic analyses such as those from HGTracker is important. An understanding of the post-transfer trajectories of the AqHGTs has prevented gross overestimation of the extent of HGT in *A. queenslandica.*

2.5.4 Animal genomes are dynamic and flexible: putatively novel genes created from transferred alien and native sequences in A. queenslandica

An evolutionary mechanism that generates novelty is domain fusion, which is thought an important mechanism particularly in the origin and evolution of pathway assembly (Long et al. 2003; Nakamura et al. 2006; Fani et al. 2007; Zhou and Wang 2008; Kaessmann 2010). Domain fusion allows a new physical association of catalysing or regulating domains (Jensen 1987), and reported cases often involve proteins such as enzymes that work in a coordinated manner catalysing sequential steps in a pathway (Yanai et al. 2002; Fani et al. 2007; Nikolaidis et al. 2014). Such fusions are predicted to enable the

evolution of intermediates (Jensen 1987; Yanai et al. 2002). The different biochemical properties and characteristics of different domains result in certain combinations cooperating more successfully and sometimes even have synergistic action with a greater combined affect (Kim et al. 2009; Nikolaidis et al. 2014). Therefore, particular domains are more likely to fuse and cases of the same certain domains fusing together independently in multiple lineages are reported. For instance, horizontally transferred plant expansin domains, involved in cell-wall loosening proteins, have become independently fused to cellulose GH5 domains in at least three new host bacterial species, and independently fused to different types of carbohydrate binding modules in fungal and amoebozoan species (Nikolaidis et al. 2014). In the histidine biosynthesis pathway, it is suggested that several fusions have occurred independently in diverse taxonomic lineages (Fani et al. 2007).

The AqHGTs also show that putatively novel genes may evolve in animals from fusions of HGT-derived and native sequences. While almost all of the here reported cases cannot yet be better explained as either fusions or incorrect gene model predictions, the aspzincin-hemopexin gene model is a well-supported fusion candidate. The Pfam taxonomic distribution of the hemopexin domain suggests it is usually only found in animals (v29.0, accessed April 2016; http://pfam.xfam.org/family/PF00045#tabview=tab7). This distribution increases the intrigue of this novel gene since Pfam and BLASTp results show that aspzincins are typically only found in bacteria and fungi (though see section 2.5.3 for a discussion on the rare cases of animal aspzincins). Hemopexin-like domains are reported in vitronectins, which are cell adhesion factors, and in matrixins, which cut extracellular matrix molecules (Faber et al. 1995; Das et al. 2003). The binding role of the hemopexin domains in these proteins suggests that the aspzincin-hemopexin gene in *A. queenslandica* may bind and cleave a different molecular partner than that of the typical aspzincins.

2.5.5 Animal genomes are dynamic and flexible: domains of great evolutionary distance from animals are expressed in A. queenslandica

Attempts to understand the kinds of genes more likely to be successful HGTs include hypotheses about the functions of transferred genes, their expression levels, their molecular interactive behaviours and their complexity (Rivera et al. 1998; Jain et al. 1999; Gogarten and Townsend 2005; Gophna and Ofran 2011; Park and Zhang 2012). The taxonomic distance between the donor and new host may also

affect the chances of a transferred gene becoming functional. Because of vertical descent, more closely related species are expected to have more similar genomic environments; therefore, a gene transferred to a new but similar host environment may require less adaptation and/or chance for it to become functional, as appropriate regulatory systems may already exist. This logic on taxonomic distance extends to the evolutionary distance between genes. For instance, a bacterial gene that already has a homologue in an animal due to vertical descent may be more likely to become functional if transferred to an animal. Consequently, the host has increased gene dosage or neofunctionalisation flexibility – evolutionary opportunities akin to those that gene duplication offers a genome (Andersson et al. 2015). Accordingly, genes that are more unique to a new host genome are here predicted less likely to become functional because more steps and innovation are necessary. Conversely, many of the reported HGTs in animals involve the transfer of a taxonomically unique phenotype, such as the transferal of a bacterial mannanase gene to a beetle opening a new ecological niche for the animal (Acuña et al. 2012), and the transfer of fungal carotenoid biosynthesis genes to the pea aphid resulting in a new red colour phenotype (Moran and Jarvik 2010). Perhaps though, cases like these were detected because of their large evolutionary impact and do not reflect the overall trend of HGTs in animals. Certainly the AqHGTs suggest that evolutionarily distant genes are less likely to be successful HGTs, since the majority of AqHGTs have putative orthologues in animals already. Only 10% of the HGT-derived genes in *A. queenslandica* are not typically found in animals, though two thirds of these are expressed with a broad range of ontogenetic expression profiles. As already discussed in section 2.5.3, because of the predicted early divergence of the Porifera from the Eumetazoa (Figure 1.1), some of these genes may not be the result of HGT, but of either gene loss or extensive gene divergence along the eumetazoan stem. This alternative hypothesis is most particularly important for the 13 of these 60 genes that are predicted to have been transferred from fungi. These fungal-like putative HGTs are the most likely of all the putative HGTs to result in *A. queenslandica* from vertical inheritance, since fewer gene loss events are required to explain their phylogenetic distribution – one loss in the eumetazoan stem and losses in the sequenced nonmetazoan holozoan species (positioned between fungi and metazoans). Once further verified as HGTs, these genes offer an intriguing entry point of further investigation of HGT in *A. queenslandica* because they may have enabled greater levels of novelty in sponges.

2.6 Conclusion

The similarities in the results of the two independent assessments of HGT in *A. queenslandica* by Conaco et al. (2016) and Fernandez-Valverde et al. (in preparation) offer increased confidence in the occurrence of HGT in *A. queenslandica* and indeed, in animals. But these results also highlight that HGTs have a broad range of sources, trajectories, ages and signatures – consequently some putative HGTs are not as clearly detected and different methods will detect different genes. Therefore, the differences in the two results also support a continued need for cautious and stringent methodologies. I have shown that the AqHGTs identified by HGTracker contain a large variety of predicted domains, that many are differentially expressed through development, and that many of the AqHGTs have unique developmental profiles, though some are co-expressed. In addition, I have found large gene groups and more detailed analyses of the two largest groups show that these groups result from one or few transfer events followed by duplication. This finding highlights the necessity of distinguishing between HGTs and HGT-derivatives to avoid yet another way of misestimating the extent of HGT in animals, in addition to misinterpreting genomic contamination and differential gene losses across species (Ku and Martin 2016; Salzberg 2017). As well as domain conservation, gene fossilisation (shown by poor domain hits) and gene duplication, another post-transfer evolutionary trajectory identified here is domain fusion – involving foreign sequences possibly forming new genes with native or other foreign sequences. Last, while transfers from lesser taxonomic and evolutionary distances are expected more likely to be successful because of greater genomic similarities in donor and host genomes, genes of greater foreignness are more likely to be detected, possibly offer hosts greater opportunity for novelty, and their presence detected here and in other systems (e.g., Acuña et al. 2012; Moran and Jarvik 2010) show the dynamic and opportunistic nature of genomes.

Chapter 3 - The bacterial-like aspzincins in sponges result from ancient horizontal gene transfer

3.1 Abstract

Phylogenetic analyses show that the bacterial-like aspzincin genes discovered in the basal animal *Amphimedon queenslandica* probably result from one or few horizontal gene transfer events followed by extensive duplication of the transferred gene in its new genomic environment. Here, I report aspzincins in representatives of all four of the sponge classes and hence in both of the two major sponge lineages. Therefore, the aspzincin transfer event(s) was probably ancient and deep in the sponge lineage, after sponges diverged from their last common ancestor with the Eumetazoa, but before the contemporary sponge classes emerged. Fifty-four of the total 90 aspzincins in *A. queenslandica* fit into one of four developmental expression profiles, each of which is putatively co-expressed with different suites of *A. queenslandica* genes. Many of the *A. queenslandica* aspzincins have retained the aspzincin-typical signal peptides and key catalytic residues, while others show secretion/extracellular sequence signals yet lack signal peptides, suggestive of either non-classical secretion or misannotation. Based on sequence characteristics and the putatively co-expressed gene groups, I suggest that the aspzincins have maintained proteolytic activity in at least one of their co-opted functions in *A. queenslandica*. The conservation of aspzincins in at least 16 contemporary sponge species through hundreds of millions of years indicates they are functionally important in the biology of the other sponge species too.

3.2 Introduction

The genome of the demosponge *A. queenslandica* contains 90 bacterial-like metalloendopeptidase aspzincin genes that appear to be the result of one or a few horizontal gene transfer (HGT; Conaco et al. 2016) events, followed by extensive duplication (Chapter 2). The presence of similar aspzincins also in another demosponge, *Haliclona amboinensis*, suggests that original transfer event may have occurred in, or prior to, the demosponge last common ancestor (Conaco et al. 2016).

Metallopeptidases are essential metal-ion dependent hydrolytic enzymes that cleave peptide bonds and are ubiquitous in all kingdoms of life (Gomis-Rüth 2003). Currently the MEROPS database has 72 families of metallopeptidases (v11, accessed February 2017; https://www.ebi.ac.uk/merops/; Rawlings et al. 2016). These enzymes play roles in all main physiological processes, including digestion of intake proteins, tissue development and maintenance, specific activating or deactivating cleavage of targets, metabolism, control of signal-transduction pathways, regulation of cell-cell and protein-protein interactions, and so on (Gomis-Rüth 2003; Cerdà-Costa and Gomis-Rüth 2013).

Aspzincin genes are members of the deuterolysin M35 metalloprotease family that belongs to the MEROPS protease MA clan of the metallopeptidase superfamily (Bogdanović et al. 2016; Rawlings et al. 2016). Aspzincins bind an essential catalytic metal ion, usually zinc, in their active site by the common metallopeptidases motif HEXXH, where the two histidines are zinc ligands, and the glutamate acts as a catalytic general base (Fushimi et al. 1999; Schwenteit et al. 2013a). Aspzincins are distinct from other metallopeptidases because their third zinc ligand is an aspartate residue (Fushimi et al. 1999; Schwenteit et al. 2013a). According to Pfam's species distribution in fully sequenced genomes, aspzincin domains are found in bacteria, fungi and two animal species, the sponge *A. queenslandica* and the turtle *Chelonia mydas* (v30.0, accessed February 2017; http://pfam.xfam.org/family/PF14521#tabview=tab7; Finn et al. 2016).

The proteolytic activity of a number of aspzincins has been characterised and, in certain bacterial and fungal pathogenic systems, aspzincin activity is highly toxic to mice and fish (Arnadottir et al. 2009; Yamada et al. 2012). Characterised aspzincins include lysine-specific metalloendopeptidases from the fungi *Grifola frondosa* and *Pleurotus ostreatus* (Nonaka et al. 1998), deuterolysin from fungus *Aspergillus oryzae* (Fushimi et al. 1999), the aspzincin AsaP1 in bacterium *Aeromonas salmonicida* subsp. *achromogenes* (Arnadottir et al. 2009), and penicillolysin from *Penicillium citrinum* (Doi et al. 2004). Some aspzincins do not have proteolytic activity, but their zinc-binding properties facilitate other roles. For instance, aspzincins have an excess metal binding function in the fungus *Botrytis cinerea* (Cherrad et al. 2012), and aspzincins scavenge host zinc for the pathogenic fungus *Candida albicans* (Citiulo et al. 2012).

The large expansion of aspzincins in *A. queenslandica* suggests that they have an important function in this species. To understand the evolutionary history of aspzincins in sponges, here I search for aspzincins in other sponge species and explore the relationship of these newly discovered sponge aspzincins with those of *A. queenslandica,* bacteria, fungi and eumetazoans using phylogenetic analyses. In addition, I further characterise the *A. queenslandica* aspzincins based on their putative co-expression through development with different suites of native genes, on their sequence characteristics, and by exploring the metallopeptidase repertoire of one choanoflagellate, 18 sponges and nine other animals.

3.3 Methods

3.3.1 Data collection

Unless otherwise described, the public sources for all genomic and transcriptomic data used throughout this chapter are detailed in Appendix 3.1. From each transcriptomic dataset, exact transcript duplicates were detected using CD-hit (Li and Godzik 2006; Huang et al. 2010) and removed from further analyses.

3.3.2 Aspzincin distribution in the Porifera

To identify other sponge aspzincins, the available sequence data of twenty-six sponges were searched. Using the hmmsearch program of HMMER v3.0 (default settings including cut-off –E 10.0; Finn et al. 2011) and the Pfam hidden Markov model (HMM; Eddy 1996) for the Aspzincin_M35 domain (accession PF14521; aspzincin hereafter), I interrogated the gene models predicted from the genomes of *Oscarella carmela, Stylissa carteri,* and *Sycon ciliatum,* and the transcripts from the transcriptomes of *Aphrocallistes vastus, Hyalonema populiferum, Rossella fibulata, Sympagella nux, Chondrilla nucula, Crella elegans, Ircinia fasciculata, Petrosia ficiformis, Pseudospongosorites suberitoides, Spongilla lacustris, Kirkpatrickia variolosa, Latrunculia apicalis, Ephydatia muelleri, Tethya wilhelma, Sycon coactum, Corticium candelabrum, Clathria prolifera, Leucosolenia complicata, Xestospongia testudinaria, Cliona varians, H. amboinensis, Haliclona tubifera* and *Niphatidae indet.* All hit alignments were manually checked for length and for conservation of key amino acids, as identified by the HMM. To gain the complete domain compositions and to crosscheck for better-suited domain assignments, all the accepted sequences were submitted to Pfam version 27.0 (cut-off –E 1.0; http://pfam.xfam.org).

3.3.3 Metallopeptidase distribution in animals

To further understand the *A. queenslandica* aspzincins in the broader context of the metallopeptidase repertoire of this species, the metallopeptidase repertoires of *A. queenslandica*, 26 other animals and one choanoflagellate were determined and compared. To identify all metallopeptidase types in whole predicted proteomes, the Ensembl Biomart Tool (http://metazoa.ensembl.org/biomart/martview/f6f40872a4f0933ca309b6a4ee088bc6) was used to filter for both the metallopeptidase GO terms GO0008237 and GO0004222 in the genomes of the following ten animals: *A. queenslandica, Mnemiopsis leidyi, Trichoplax adhaerens, Nematostella vectensis, Lottia gigantean, Capitella teleta, Drosophila melanogaster, D. willistoni, Caenorhabditis elegans,* and *Strongylocentrotus purpuratus.* Then to classify the retrieved proteins into different metallopeptidase families, they were submitted to the Pfam version 27.0 Batch Search and a list of all the metallopeptidase domain types was compiled. The hmmsearch program of HMMER version 3.0 (default settings including cut-off –E 10.0; Finn et al. 2011) was then used to search with the Pfam hidden Markov model (HMM) for each of those metallopeptidase domain types against public datasets not incorporated into the 2014 Ensemble Biomart Tool, in particular against the genomes of *M. brevicollis and O. carmela,* and against the transcriptomes of the following sixteen sponges: *C. nucula, Cr. elegans, I. fasciculata, P. ficiformis, P. suberitoides, S. lacustris, S. coactum, A. vastus, C. candelabrum, H. populiferum, K. variolosa, L. apicalis, R. fibulata, S. nux, T. wilhelma* and *E. muelleri.* All returned sequences with putative hits to any of the HMMs were submitted to the Pfam v27.0 Batch Search tool (cut-off –E 1.0; http://pfam.xfam.org), since the HMMs are all related and thus multiple models hit the same sequences. A generic cut-off point was not used; rather the Pfam v27.0 Batch Search e-value and query coverage results of all 28 species were manually perused for quality control, with any doubt over hit qualities resolved by consideration of the hit alignments and in particular the conservation of amino acids predicted of greater functional significance as determined by the HMM, thereby giving a final metallopeptidase repertoire of each species.

3.3.4 Multiple sequence alignment and phylogenetic analysis

To find similar sequences to the aspzincins of those identified in the sponge species, each sponge aspzincin domain was submitted against the 2017 nonredundant database of the National Centre for Biotechnology Information (NCBI) with the protein Basic Local Alignment Search Tool (BLASTp), using the default settings. The top result was selected for each, but duplicate results were removed.

From Figure 2.10, one eumetazoan sequence was selected from any eumetazoan branch with more than 50% bootstrapping support. Otherwise, all other eumetazoan aspzincins identified in Chapter 2 were included. All retrieved sequences were submitted to the Pfam Batch Search tool to help find the beginning and end of each aspzincin domain.

Using the Geneious Pro 5.1.7 Multiple Alignment using Fast Fourier Transform (MAFFT) 6.814b plug-in (Katoh et al. 2002), I aligned all the sponge, bacterial, fungal and eumetazoan domain sequences. I excluded duplicate sponge sequences, regardless of the differences in the whole proteins since these parts of the proteins were not in the alignment. The multiple alignment was manually trimmed and refined in Geneious Pro 5.1.7 (www.geneious.com; Kearse et al. 2012). ProtTest 2.4 implemented the Bayesian Information Criterion (BIC) to find the appropriate model of evolution for the aspzincin alignment (Darriba et al. 2011).

A Maximum Likelihood (ML) tree was constructed using the PhyML 2.0.12 plug-in in Geneious Pro 5.1.7 (Guindon and Gascuel 2003). Two hundred and fifty bootstrap replicates provided statistical support.

3.3.5 Gene expression and enrichment analyses

All gene expression data presented is from the genome-wide ontogenetic transcript dataset already described in Chapter 2.3.1 (Anavy et al. 2014; Levin et al. 2016; Gene Expression Omnibus accession codes GSE54364 and GPL18214). After averaging the expression data from 82 time points for *A. queenslandica* aspzincin genes (AqAspzs) into 17 developmental stages (as described in Chapter 2.3.1), I graphed the sum of these means for each gene across all stages. For each AqAspzs with a total lower than 50 normalised counts, the average read count of each development stage was perused to distinguish gene models with only low read counts (predicted unexpressed) from those with mostly low read counts, but with higher expression at some stage(s). Thereby, a cut-off point dividing predicted unexpressed and expressed AqAspzs was identified. For the expression of all the other *A. queenslandica* gene models, as in Chapter 2, I used a generic cut-off of at least one developmental stage with at least five normalised counts to distinguish between predicted unexpressed and expressed genes.

To find (1) aspzincins with correlated expression profiles to each other and (2) gene models in the *A. queenslandica* genome that have a correlated gene expression profile with that of any of the aspzincins, I made an expression correlation matrix with significance probabilities of all expressed *A. queenslandica* genes, including the expressed aspzincins, using R (R Core Team 2014). Appendix 3.2 contains the code used for this co-expression analysis, which were already developed by Gaiti et al. (2015). In brief, expression correlations were made by Pearson's correlation and Fisher's exact tests (significance threshold: $p<0.05$ and correlation coefficient >0.95 or <-0.95). Using the Pretty Heatmaps R package in RStudio (Kolde 2012), I generated heat maps for expression correlation groups of aspzincin genes (Euclidean distance row clustering and scaled by row). Using Pfam domain annotations for the whole *A. queenslandica* genome, as determined by BLAST2GO (Gotz et al. 2008; Hatleberg et al. unpublished data), I tested the resulting list of genes for any Pfam domain enrichments. Using the method and script of Chandran et al. (2009; script adjusted by W.L. Hatleberg, as shown in Appendix 3.3), enriched domains in gene groups of interest verses all the *A. queenslandica* first generation gene models were identified, with significance determined by probability values calculated from a hypergeometric distribution. These probability values were corrected for multiple testing with a 10% False Discovery Rate using the Benjamini-Hochberg method. I examined the correlated gene groups for their structural content using the structural information for their contained domains, as annotated by BLAST2GO (Gotz et al. 2008; Hatleberg et al. unpublished data) using the Class Architecture Topology Homologous superfamily database and hierarchal classification system (CATH v4.1; www.cathdb.info; Sillitoe et al. 2015). To detect any HGTs within the correlated gene groups, I extracted the gene classifications and the hypothesised taxonomic source for the genes as determined by HGTracker. SL. Fernandez-Valverde interrogated the first generation Aqu1 gene models against HGTracker (default settings; Fernandez-Valverde et al. in preparation; results were shared via personal communication).

3.3.6 Sequence characteristics analyses

Bacterial and fungal aspzincins have signal peptides targeting them for secretion (Saito et al. 2002; Schwenteit et al. 2013a); therefore, I investigated whether this secretion signature is conserved in the AqAspzs. To detect predicted signal peptides, I submitted the *A. queenslandica* aspzincins to the SignalP 4.1 Server (http://www.cbs.dtu.dk/services/SignalP/; Petersen et al. 2011) and searched against the eukaryotes groups, as well as the Gram-negative and Gram-positive bacterial groups. For prediction of

transmembrane helices, sequences were submitted to the TMHMM Server v2.0 (http://www.cbs.dtu.dk/services/TMHMM-2.0/; Krogh et al. 2001). To detect possible non-classically secreted aspzincins, I submitted sequences without predicted signal peptides and transmembrane helices to SecretomeP 2.0 Server and searched with each of the three available prediction models for Gram-negative bacteria, Gram-positive bacteria and mammalian (http://www.cbs.dtu.dk/services/SecretomeP/; Bendtsen et al. 2004; Bendtsen et al. 2005). To assess if the catalytically important HEXXH+DXXY+NAD motif has been conserved in the aspzincins, the alignments in the HMMER v3.0 hmmsearch results were manually considered.

3.3.7 Hemopexin searches

In Chapter 2, the discovered aspzincin and hemopexin domain combination was hypothesised as unique and the result of post-transfer fusion of HGT-derived and native domains. To further investigate these predictions, the hmmsearch program of HMMER version 3.0 (default settings including cut-off –E 10.0; Finn et al. 2011) was used to search with the Pfam hemopexin HMM (accession PF00045) against the genomes or transcriptomes of all the sponge species already described. Any sequences with HMM hits were submitted to the Pfam v27.0 Batch Search tool for their complete predicted domain architecture (cut-off –E 1.0; http://pfam.xfam.org). Using BLASTp, those sequences containing aspzincin and hemopexin domains were searched against the nonredundant database of NCBI. To see if any bacterial aspzincins also contain hemopexins, I searched for all bacterial hemopexins in the Pfam, NCBI, and SUPERFAMILY (www.supfam.org/SUPERFAMILY/; Gough et al. 2001) databases. To find their predicted domain architectures, these retrieved bacterial hemopexin domain-containing sequences were also submitted to the Pfam v27.0 Batch Search tool. Using the Ensembl Biomart Tool (http://metazoa.ensembl.org/biomart/martview/d83018183678ec07e7a554ae3b89ae60), I filtered for genes with the Pfam hemopexin domain (PF00045) in the genomes of the following 10 animals: *M. leidyi, T. adhaerens, N. vectensis, L. gigantean, C. teleta, D. melanogaster, D. willistoni, C. elegans,* and *S. purpuratus,* and submitted those returned proteins to the Pfam v27.0 Batch Search (cut-off –E 1.0; http://pfam.xfam.org).

3.4 RESULTS

3.4.1 The distribution of aspzincins in poriferans

HMM searching for aspzincins in 26 additional sponge species revealed putative aspzincins in 12 other Demospongiae, one Hexactinellida, one Homoscleromorpha, and one Calcarea (Figure 3.1). Thus at least one species from each of the four poriferan classes contains a predicted aspzincin (Figure 3.1).

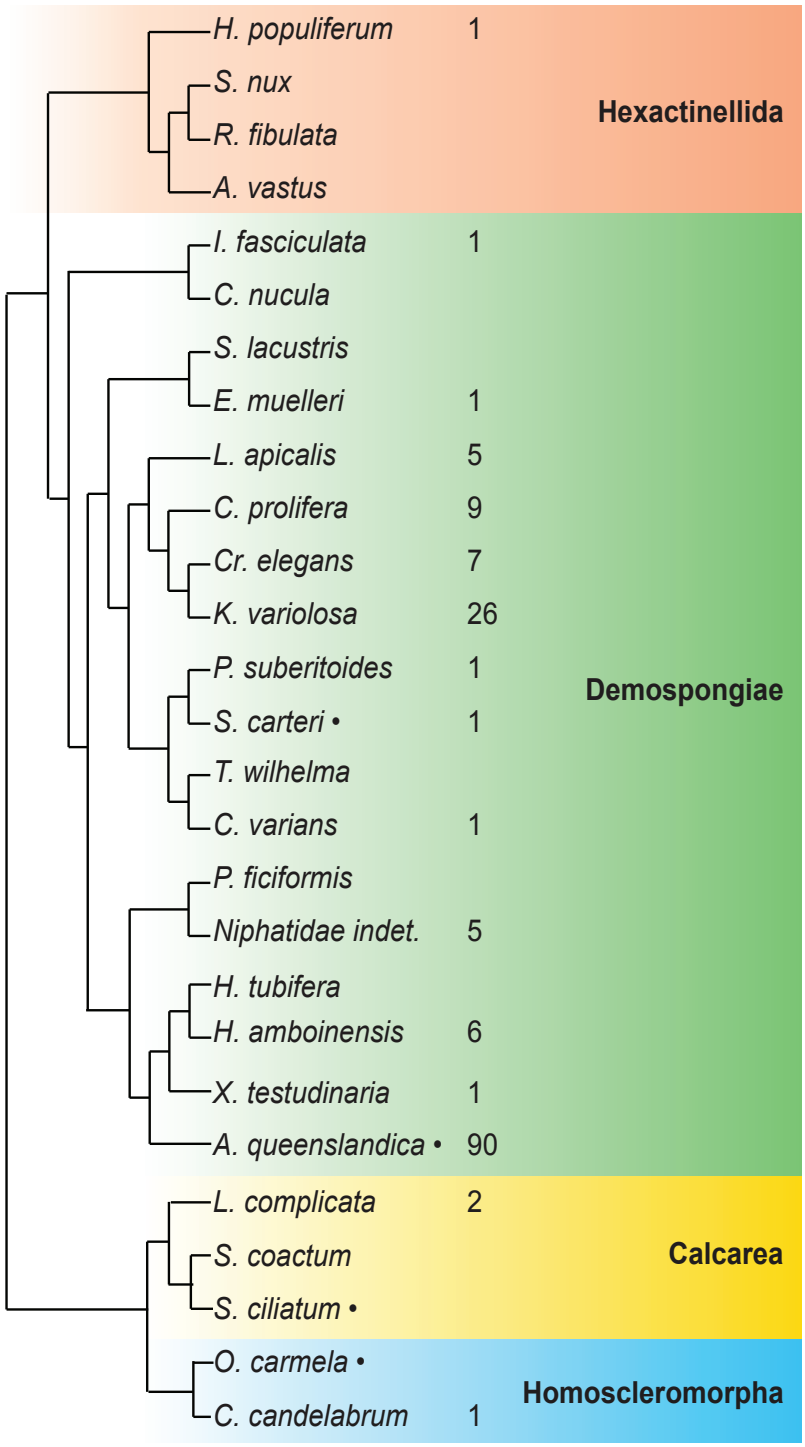


Figure 3.1 Distribution of aspzincins in the sequenced Porifera Counts reflect the number of unique aspzincin domain-containing transcripts as determined by the Pfam aspzincin HMM, or for those demarked by '•', the number of aspzincin domain-containing gene models. Species are arranged by class, with predicted topology adapted from Thacker et al. (2013), Whelan et al. (2015), and Grice et al. (2017).

For most of the species, the datasets searched are transcriptomes, thus absence reflects absence of expression under the conditions tested and not necessarily the absence of an aspzincin. The number of unique aspzincin domain-containing sequences found in each of the 15 other species (all transcripts except for one *S. carteri* gene model) is much less than the 54 transcribed aspzincins of *A. queenslandica* detected in Chapter 2.4.4 and the total 90 gene models detected in *A. queenslandica* (Figure 3.1; mean =4.5, mode and median =1).

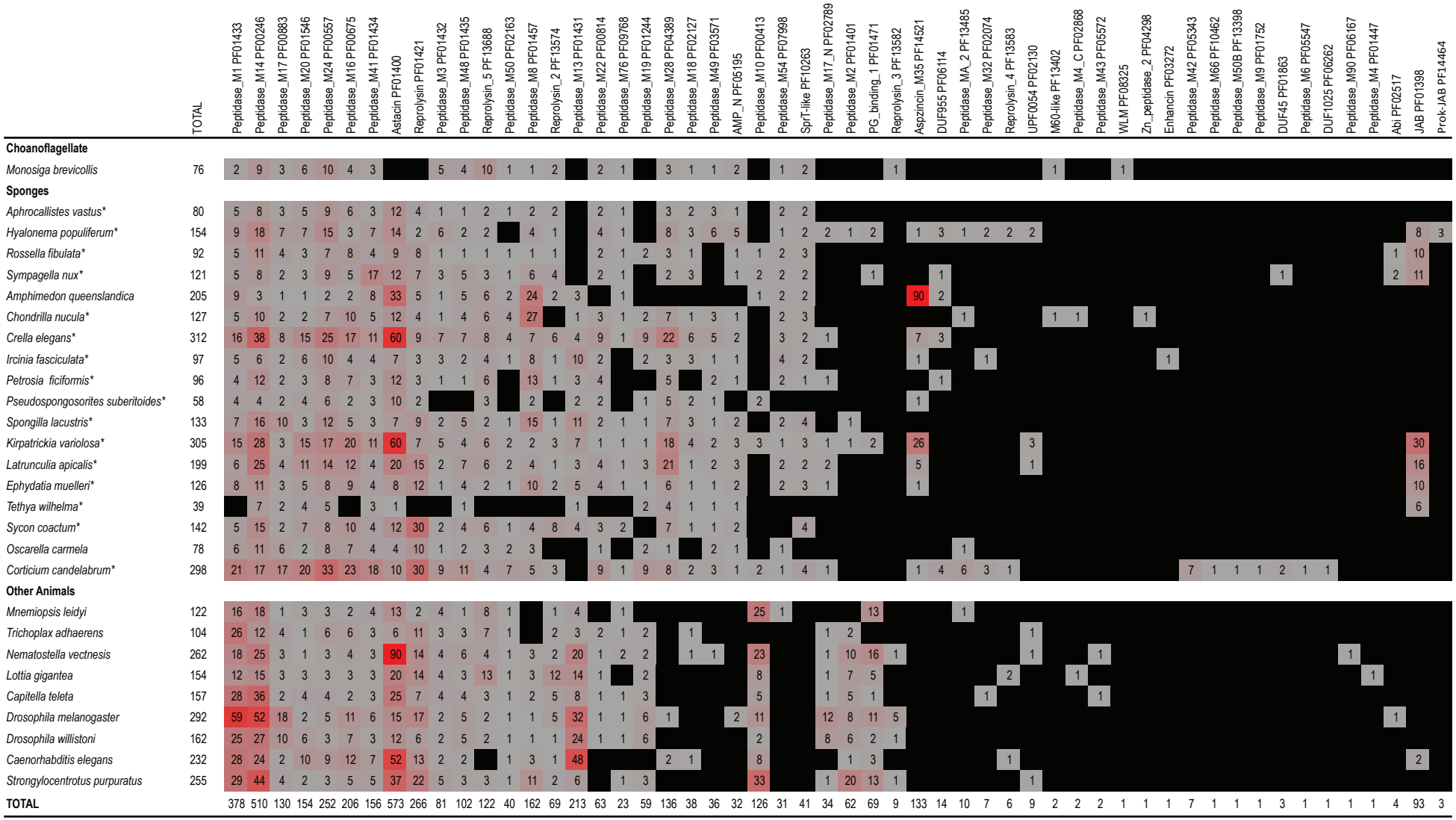
In addition to *A. queenslandica,* the only other genome found to contain an aspzincin is that of the demosponge *S. carteri* – this aspzincin gene (scaffold3889-processed-gene-0.23-mRNA-1) is contiguous to a native gene that contains a typically animal-only Death domain (scaffold3889-snap-gene-0.81-mRNA-1). Therefore, while contamination cannot be dismissed as a possible explanation for the aspzincins found in the transcriptomes of the other species, the *S. carteri* aspzincin appears to be a *bona fide* HGT-derived gene retained in this sponge genome.

3.4.2 The metallopeptidase repertoires of animal representatives

The lack of aspzincins in most animals (Chapter 2), yet their presence and expansion in some sponges, particularly in *A. queenslandica,* raises the hypothesis that aspzincins could be compensating for a deficient metallopeptidase repertoire in some sponges and/or *A. queenslandica.* Therefore, I placed the sponge aspzincins in context by comparing the size and composition of the metallopeptidase repertoires of a choanoflagellate, eighteen sponges, and nine eumetazoans (Table 3.1). Consistent with my previous searches that identified aspzincins in only a small number of eumetazoan species as well as bacteria and fungi, I did not find any aspzincin domains in the predicted proteomes of the choanoflagellate or any of the eumetazoan species.

In the species analysed, the mean number of total metallopeptidases of any kind identified per species is 160; the lowest number (39) is in the sponge *T. wihelma* and the highest (312) is also in a sponge, *Cr. elegans.* Most species have a similar number of metallopeptidase types. Excluding the five sponges *P. suberitoides, T. wihelma, H. populiferum, K. variolosa,* and *C. candelabrum,* all species have between 20 and 29 different types, with an average of 25. The sponge *C. candelabrum* has the third highest total number of metallopeptidases (298) and the highest diversity of metallopeptidase type (38).

Table 3.1 The metallopeptidase repertoires of select representative species



	TOTAL	Peptidase_M1 PF01433	Peptidase_M14 PF00246	Peptidase_M17 PF00883	Peptidase_M20 PF01546	Peptidase_M24 PF00557	Peptidase_M16 PF00675	Peptidase_M41 PF01434	Astacin PF01400	Reprolysin PF01421	Peptidase_M3 PF01432	Peptidase_M48 PF01435	Reprolysin_5 PF13688	Peptidase_M50 PF02163	Peptidase_M8 PF01457	Reprolysin_2 PF13574	Peptidase_M13 PF01431	Peptidase_M22 PF00814	Peptidase_M76 PF09768	Peptidase_M19 PF01244	Peptidase_M28 PF04389	Peptidase_M18 PF02127	Peptidase_M49 PF03571	AMP_N PF05195	Peptidase_M10 PF00413	Peptidase_M54 PF07998	SprT-like PF10263	Peptidase_M17_N PF02789	Peptidase_M2 PF01401	PG_binding_1 PF01471	Reprolysin_3 PF13582	Aspzincin_M35 PF14521	DUF955 PF06114	Peptidase_MA_2 PF13485	Peptidase_M32 PF02074	Reprolysin_4 PF13583	UPF0054 PF02130	M60-like PF13402	Peptidase_M4_C PF02868	Peptidase_M43 PF05572	WLM PF08325	Zn_peptidase_2 PF04298	Enhancin PF03272	Peptidase_M42 PF05343	Peptidase_M66 PF10462	Peptidase_M50B PF13398	Peptidase_M9 PF01752	DUF45 PF01863	Peptidase_M6 PF05547	DUF1025 PF06262	Peptidase_M90 PF06167	Peptidase_M4 PF01447	Abi PF02517	JAB PF01398	Prok-JAB PF14464
Choanoflagellate																																																					
Monosiga brevicollis	76	2	9	3	6	10	4	3			5	4	10	1	1	2		2	1		3	1	1	2		1	2			1							1			1																
Sponges																																																							
*Aphrocallistes vastus**	80	5	8	3	5	9	6	3	12	4	1	1	2	1	2	2		2	1		3	2	3	1		2	2																														
*Hyalonema populiferum**	154	9	18	7	7	15	3	7	14	2	6	2	2		4	1		4	1		8	3	6	5		1	2	2	1	2		1	3	1	2	2	2																8	3
*Rossella fibulata**	92	5	11	4	3	7	8	4	9	8	1	1	1	1	1	1		2	1	2	3	1		1	1	2	3																										1	10	
*Sympagella nux**	121	5	8	2	3	9	5	17	12	7	3	5	3	1	6	4		2	1		2	3		1	2	2	2			1		1														1					2	11	
Amphimedon queenslandica	205	9	3	1	1	2	2	8	33	5	1	5	6	2	24	2	3		1						1	2	2					90	2																			
*Chondrilla nucula**	127	5	10	2	2	7	10	5	12	4	1	4	6	4	27		1	3	1	2	7	1	3	1		2	3							1				1	1		1											
*Crella elegans**	312	16	38	8	15	25	17	11	60	9	7	7	8	4	7	6	4	9	1	9	22	6	5	2		3	2	1				7	3																			
*Ircinia fasciculata**	97	5	6	2	6	10	4	4	7	3	3	2	4	1	8	1	10	2		2	3	3	1	1		4	2					1			1								1									
*Petrosia ficiformis**	96	4	12	2	3	8	7	3	12	3	1	1	6		13	1	3	4			5		2	1		2	1	1					1																			
*Pseudospongosorites suberitoides**	58	4	4	2	4	6	2	3	10	2			3		2		2	2		1	5	2	1		2							1																			
*Spongilla lacustris**	133	7	16	10	3	12	5	3	7	9	2	5	2	1	15	1	11	2	1	1	7	3	1	2		2	4		1																							
*Kirpatrickia variolosa**	305	15	28	3	15	17	20	11	60	7	5	4	6	2	2	3	7	1	1	1	18	4	2	3	3	1	3	1	1	2		26				3															30	
*Latrunculia apicalis**	199	6	25	4	11	14	12	4	20	15	2	7	6	2	4	1	3	4	1	3	21	1	2	3		2	2	2				5				1															16	
*Ephydatia muelleri**	126	8	11	3	5	8	9	4	8	12	1	4	2	1	10	2	5	4	1	1	6	1	1	2		2	3	1				1																			10	
*Tethya wilhelma**	39		7	2	4	5		3	1			1						1			2	4	1	1	1																									6	
*Sycon coactum**	142	5	15	2	7	8	10	4	12	30	2	4	6	1	4	8	4	3	2		7	1	1	2				4																								
Oscarella carmela	78	6	11	6	2	8	7	4	4	10	1	2	3	2	3			1		2	1		2	1		1								1																	
*Corticium candelabrum**	298	21	17	17	20	33	23	18	10	30	9	11	4	7	5	3		9	1	9	8	2	3	1	2	1	4	1				1	4	6	3	1																
Other Animals																																																					
Mnemiopsis leidyi	122	16	18	1	3	3	2	4	13	2	4	1	8	1		1	4		1						25	1			13				1																			
Trichoplax adhaerens	104	26	12	4	1	6	6	3	6	11	3	3	7	1		2	3	2	1	2		1					1	2							1																	
Nematostella vectnesis	262	18	25	3	1	3	4	3	90	14	4	6	4	1	3	2	20	1	2	2		1	1		23			1	10	16	1				1			1									1				
Lottia gigantea	154	12	15	3	3	3	3	3	20	14	4	3	13	1	3	12	14	1		2					8			1	7	5					2		1									1				
Capitella teleta	157	28	36	2	4	4	2	3	25	7	4	4	3	1	2	5	8	1	1	3					5			1	5	1					1			1												
Drosophila melanogaster	292	59	52	18	2	5	11	6	15	17	2	5	2	1	1	5	32	1	1	6	1			2	11			12	8	11	5																	1		
Drosophila willistoni	162	25	27	10	6	3	7	3	12	6	2	5	2	1	1	1	24	1	1	6					2			8	6	2	1																		
Caenorhabditis elegans	232	28	24	2	10	9	12	7	52	13	2	2		1	3	1	48				2	1			8				1	3					1															2	
Strongylocentrotus purpuratus	255	29	44	4	2	3	5	5	37	22	5	3	3	1	11	2	6		1	3					33			1	20	13	1					1														
TOTAL		378	510	130	154	252	206	156	573	266	81	102	122	40	162	69	213	63	23	59	136	38	36	32	126	31	41	34	62	69	9	133	14	10	7	6	9	2	2	2	1	1	1	7	1	1	1	3	1	1	1	1	4	93	3

Domains classified by Pfam hidden Markov models. Numbers reflect the total number of unique gene models containing at least one of the relevant domain, except those demarked with a '*' where the number signifies the number of unique transcripts.

Astacin, M14 and M1 are the most common metallopeptidase domains (total of each from all the species analysed is 573, 510, and 378 respectively). M14 domains are found in all of the analysed species except for *T. wihelma*, M1 domains exist in all of the analysed species, and astacins are found in every animal species considered, but not in the choanoflagellate *M. brevicolis.* Reprolysin is the only other metallopeptidases that is found in all animals analysed, but not in a representative of their closest sister taxa, the choanoflagellate *M. brevicolis.*

There are five metallopeptidase types that are found in all the analysed species, namely the already mentioned M14, as well as M17, M20, M24 and M41 (Table 3.1). Twelve metallopeptidase types are found in only one species (Table 3.1). The WLM (PF08325) domain is the only domain found in choanoflagellate *M. brevicolis* alone and not in any of the animals. Nine of the metallopeptidase types only found in one species are present in one of the sponges, and six of those are in *C. candelabrum.* Two metallopeptidases, DUF955 (accession PF06114) and Aspzincin_M35 (accession PF14521), were found in multiple sponge species but no other species. DUF955 has a Pfam species distribution of bacteria and viruses (v28.0, accessed October 2015; http://pfam.xfam.org/family/PF06114#tabview=tab7); however, I found ten putative domains in six sponges. Strikingly, the 90 aspzincins in *A. queenslandica* are the highest number of one type of metallopeptidase found in a single species, equal with 90 putative astacins in *N. vectnesis.* However, in summary, quartile summaries of both the total number of metallopeptidases and the number of different metallopeptidase types detected in each species analysed suggest that the *A. queenslandica* repertoire is not unusual, with both of these measures for *A. queenslandica* falling between the 25th and 75th percentiles (Appendix 3.4).

3.4.3 Phylogeny of sponge aspzincin domains

I explored the phylogenetic relationships of the aspzincins from 15 other sponge species with the aspzincins of eumetazoans, bacteria, fungi and *A. queenslandica* based on a multiple sequence alignment using a ML analysis. The best model of sequence evolution was predicted to be WAG + G using BIC in ProtTest 2.4 (Darriba et al. 2011). In the resulting phylogenetic hypothesis (Figure 3.2), the only major branch statistically well supported (100%) is the branch separating the vertebrate species from all others, which echoes the result for the vertebrate sequences in Chapter 2. Within the poorly-supported clade containing most of the bacterial and fungal sequences is a well-supported branch (66% bootstrap

Figure 3.2 Phylogeny of the amino acid sequence of aspzincin domains from animals, bacteria and fungi

Phylogeny inferred by ML. Unrooted tree. Topology support was obtained from 250 bootstrap replicates; only support values greater than 50 are shown. Text colour reflects the taxonomy as reported by the NCBI database for the non-sponge domains.

replicates) for the arthropod *H. azteca* aspzincin (XP_018028613) and a bacterial Burkholderiales aspzincin (WP_018058812) (Figure 3.2). Also, there are statistically well-supported branches suggestive of species-specific duplications (e.g., four *Niphatidae indet.* sequences clade together with 98% bootstrap replicates; Figure 3.2). But there is also good support showing that many aspzincins from different demosponge species are more closely related to each other than to aspzincins from the same species (e.g., 95% bootstrap support for a clade containing sequences from demosponges *L. apicalis, K. variolosa, Cr. elegans,* and *C. prolifera* in Figure 3.2).

The phylogenetic analysis clades together the only identified sponge *X. testudinaria* aspzincin and an *A. queenslandica* aspzincin, Aqu1.216638|Aqu2.1.25761_001 (80% bootstrap replicates; Figure 3.2). This is the only well-supported branch predicting an *A. queenslandica* aspzincin to be more closely related to an aspzincin from another species. These two sequences are within a poorly supported clade of three, and the third sequence is the only aspzincin found in the fresh water demosponge *E. muellleri.*

Conclusions from other aspects of the tree are limited since there is poor bootstrapping support (<50%) for the major nodes of interest. Tentatively, some sponge aspzincins may be more similar to bacterial or fungal sequences than they are to other sponge aspzincins (Figure 3.2). Also, excluding *A. queenslandica,* most of the sponge sequences clade together with each other (poorly supported); though some are within the poorly supported *A. queenslandica* clade and possibly reflect orthologues of the originally transferred aspzincin (Figure 3.2). Last, there are six bacterial sequences from six different species that may be more closely related to the sponge sequences (species from the orders: Burkholderiales, accessions WP_020702029, WP_011353905, and WP_041390442; Rhizobiales, accession WP_045023548; Chitinophagales, accession AEW03381; and Oscillatoriales, accession WP_017721273; Figure 3.2).

3.4.4 Many A. queenslandica aspzincins share ontogenetic expression profiles with up to hundreds of other A. queenslandica genes

Sixteen AqAspzs were considered to have no meaningful expression because they have very low levels of expression in all 17 measured developmental stages (Appendix 3.5). The correlation probabilities analysis found 14 aspzincins with unique expression profiles through development that do not

significantly correlate with those of other aspzincins (Figure 3.3). Fifty-four other aspzincins fit into four significantly correlated expression profile groups (p<0.05 and correlation coefficient >0.95 or <-0.95; Figure 3.4). Twelve scaffolds contain more than one aspzincin, with 35 of the 54 aspzincins on a scaffold that has at least one other aspzincin, and often the aspzincins are contiguous (Appendix 2.16). All but two aspzincins (Aqu1.225765|Aqu2.1.38617_001 and Aqu1.225770|Aqu2.1.38622_001) that are on the same scaffold as each other are also within the same correlated expression group.

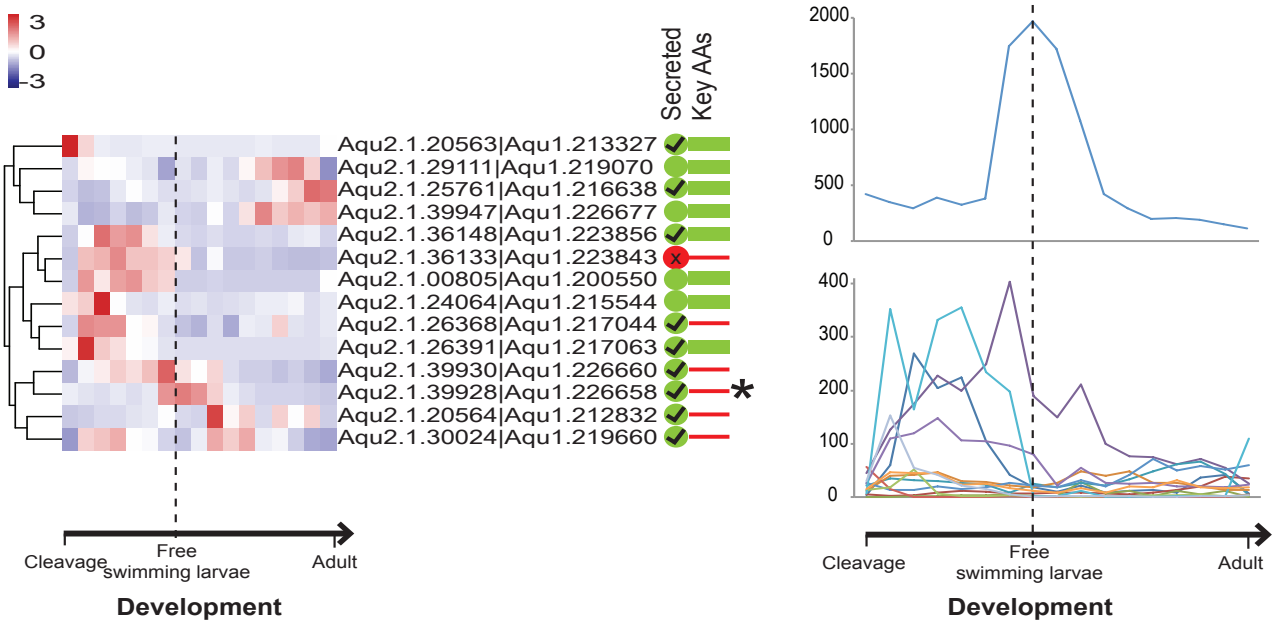


Figure 3.3 AqAspzs with unique expression profiles
Fourteen AqAspzs have expression profiles that do not significantly correlate with that of any other AqAspz. On the left, heat maps scaled by row with genes organised by Euclidean distances represent the expression patterns. The predicted secretion status of each gene is shown (green circle signifies secreted, a tick symbol in the circle shows a predicted SP, red circle signifies no predicted SP or secretion sequence signal). The essential residues in the catalytic motif HEXXH+DXXXY+NAD are conserved (green bar) or not (red line). Graphs on the right show the expression quantities. Note that the gene marked with a '*' is graphed separately (top graph) due to its higher expression levels.

Figure 3.4 AqAspzs with correlated expression profiles

(next page)

The expression profiles of 54 AqAspzs significantly correlate with that of other AqAspzs, with a total of four correlated gene expression patterns. On the left, heat maps scaled by row with genes organised by Euclidean distances represent the expression patterns. The predicted secretion status of each gene is shown (green circle signifies secreted, a tick symbol in the circle shows a predicted SP, blue circle shows predicted transmembrane helices, red circle signifies no predicted SP or secretion sequence signal). The essential residues in the catalytic motif HEXXH+DXXXY+NAD are conserved (green bar) or not (red line). Graphs on the right show the expression quantities. Note that those genes marked with a '*' in the largest group are graphed separately (top graph) due to their higher expression levels.

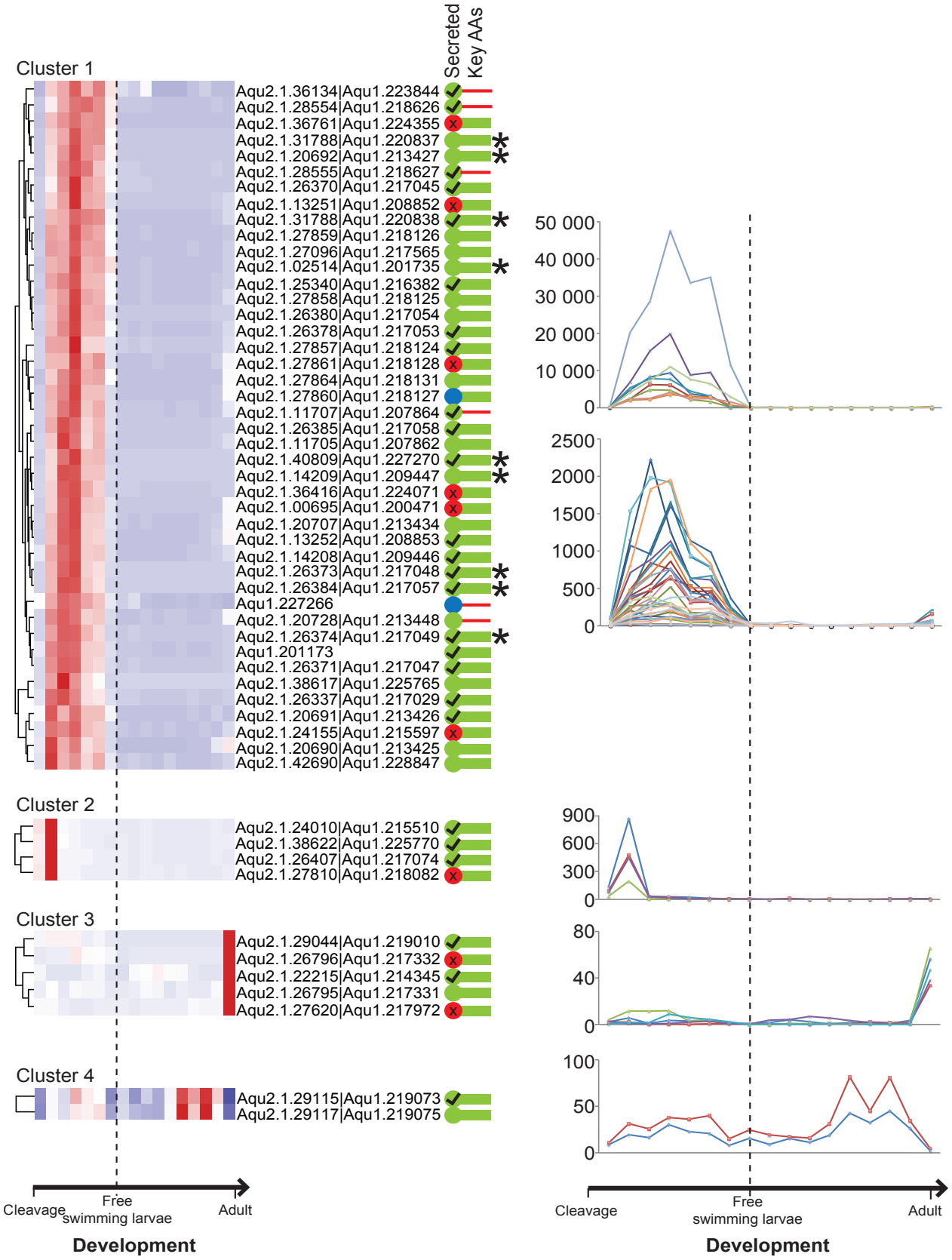


Figure 3.4 AqAspzs with correlated expression profiles
(legend on previous page)

To find hints on the possible function(s) of the *A. queenslandica* aspzincins, I first searched within the whole transcriptome of *A. queenslandica* for non-aspzincin genes that have a significantly correlated expression pattern with that of any of the aspzincins. I found 1512 such genes ($p<0.05$ and correlation coefficient >0.95 or <-0.95; Figure 3.5). The aspzincin profile of low expression until adult stage correlates with the most genes (five aspzincins correlating with 595 other *A. queenslandica* genes). The aspzincin profile of higher expression until the free-swimming larvae stage is slightly less common (54 aspzincins correlating with 410 other genes). The expression profile of Aqu1.213327|Aqu2.1.20563_001, which is not shared by any other aspzincin, is correlated to 278 other genes. Finally, eight aspzincins have a unique expression profile within this *A. queenslandica* dataset, under these clustering parameters (Figure 3.5).

Next, to add meaning to these reported putatively co-expressed genes, and with the aim of possibly finding functional information on the AqApszs, I analysed their domain content (Appendix 3.6) and structural classes (Appendix 3.7). In the CATH database, there are four classes that represent the overall secondary structure of domains: mainly alpha (n_{CATHdb}=48121 domains); mainly beta (n_{CATHdb}=58944 domains); alpha beta (n_{CATHdb}=125772 domains); and few secondary structures (n_{CATHdb}=3021 domains; CATH v4.1; www.cathdb.info). Approximately half of the correlated genes of interest contain domains that have been assigned a CATH class and no major trend emerged, except that the gene groups vary in their proportions of each structural class, and all four structural classes are dominant in at least one gene group (Appendix 3.7).

I also searched for significantly enriched Pfam domains contained by each of the putatively correlated gene groups compared with those of all the first generation gene models (Appendix 3.3; Chandran et al. 2009). These enrichment analyses reveal multiple trends. First and broadly, each correlated gene group contains different gene types than the other groups, as characterised by their domain content (Appendix 3.8). Second, while each group contains mostly different domains, there is enrichment in the correlated genes for structural domains, which are often involved with protein-protein interactions (e.g. leucine rich repeats, ankyrin repeats, Kelch motifs, WD domains, and immunoglobulins; Appendix 3.8). Within this broad structural domain trend, there is diversity amongst the different correlate gene groups.

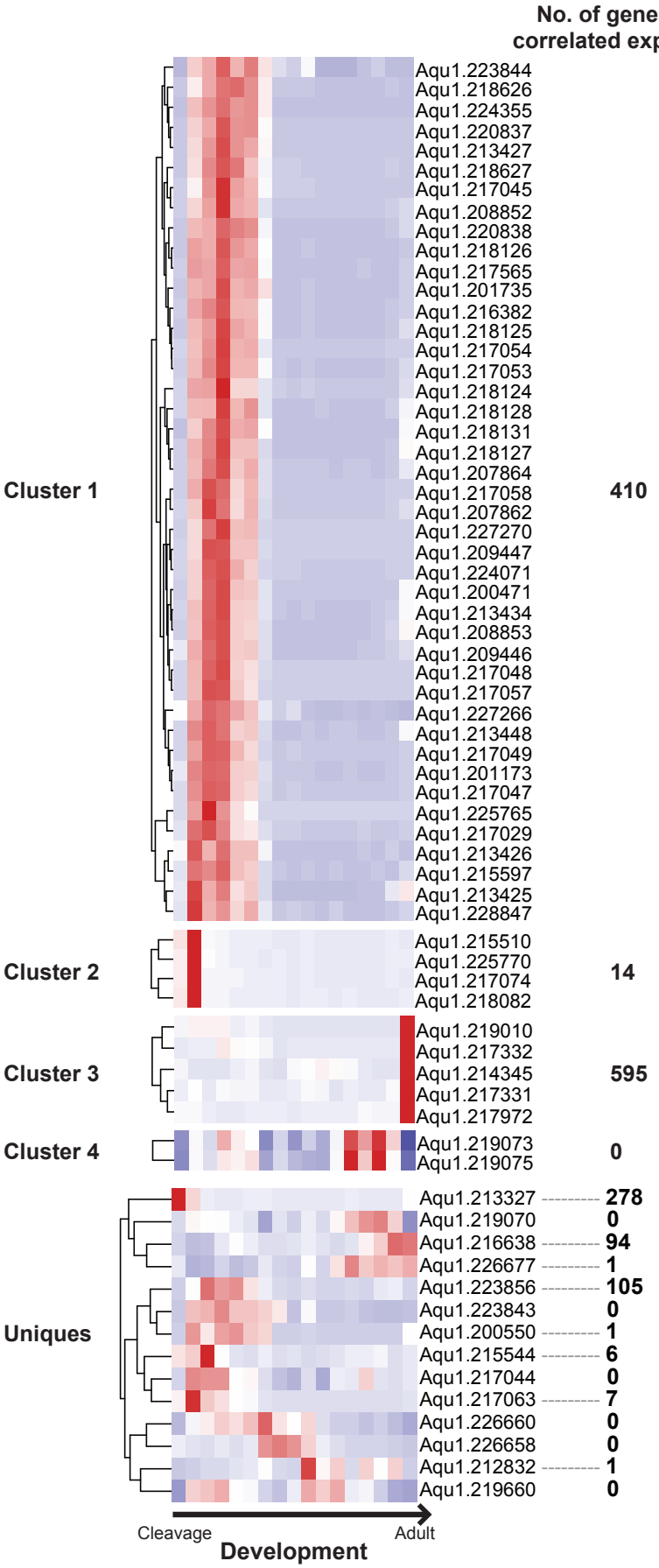


Figure 3.5 Correlation of aspzincin expression profiles with any *A. queenslandica* gene
Heat maps show the expression profiles of expressed AqAspzs through development as shown in Figures 3.3 and 3.4, and are clustered based on significant correlations with each other or demarked as a unique profile. The number of other *A. queenslandica* genes significantly correlated in expression throughout development is displayed on the right.

For instance, four groups of genes are enriched with members of the Beta Propeller clan (CL0186); however, the specific beta propeller domains found within each gene group are unique (Appendix 3.8).

The third trend is the enrichment of tetratricopeptide repeats (TPRs) and leucine rich repeats (LRRs) in the genes that correlate with the most common aspzincin expression profile of higher expression before the free-swimming larval stage (i.e. the profiles of Aqu1.223856| Aqu2.1.36148_001 and Cluster 1 aspzincins; Appendix 3.8). Fourth, there is an enrichment for domains belonging to the P-loop containing nucleoside triphosphate hydrolase superfamily (AAAs) in genes correlating with aspzincins grouped in Clusters 1 and 3 and Aqu1.223856|Aqu2.1.36148_001 (e.g. dynein PF03028, septin PF00735, ATPase PF00004, and AIG1 PF04548; Appendix 3.8). Fifth, ubiquitin domains and Kelch motifs are enriched in the 278 genes correlated with Aqu1.213327|Aqu2.1.20563_001 (Appendix 3.8). Also, there is a marked enrichment of the transmembrane and cell surface domains cadherins, scavenger receptor cysteine-rich (SRCRs), and transcription factor immunoglobulin-like fold (TIGs) in the 94 genes correlated with Aqu1.216638|Aqu2.1.25761_001 (Appendix 3.8). Last, I note another metalloendopeptidase in the enrichment results; the gene groups correlating with aspzincins in Clusters 1 and 3 are enriched with astacins (n= 7 and 6 respectively; Appendix 3.8).

Contained by the 14 genes that correlate with Cluster 2 aspzincins are nine predicted Pfam domains - five of these domains are immunoglobulin domains and two are F-box domains (Appendix 3.6).

Of the 1512 genes correlated with the AqAspzs in ontogenetic expression, two per cent are identified as HGTs (Appendix 3.9), which is the same proportion as the number of HGTs in the total Aqu1 gene models classified by HGTracker (576 of 25303). These 37 putative HGTs detected in the correlated genes contain a range of domains and are mostly bacterial-like (28 are bacterial-like, two are fungal-like and seven are non-animal genes, yet are of unclear origin; Appendix 3.10).

3.4.5 Characteristics of A. queenslandica aspzincins: secretion, transmembrane helices and tight conservation of the catalytic motif

Characterised bacterial and fungal aspzincins have signal peptides and are secreted as a toxin (Saito et al. 2002; Schwenteit et al. 2013a). Forty-three of the 90 *A. queenslandica* aspzincins contain a

predicted signal peptide (SP). There is no apparent relationship of SP presence/absence with expression pattern (Fisher's exact tests, *p*-values>0.4; Figure 3.4). Also, SP presence/absence is not a predictor of expression quantity because there many highly expressed aspzincins with and without a predicted SP (Appendix 3.11), and there is no significant difference between the expression quantity of those with SPs and those without (Student's *t*-test, *p*-value=0.6).

To test for either non-classical secretion or missing SPs resulting from gene prediction errors, I first screened those aspzincins without predicted SPs for predicted transmembrane helices and found two such aspzincins. Then I searched the remaining aspzincins without predicted SPs and transmembrane helices for extracellular properties such as degradation signals, charge, size, and composition (Bendtsen et al. 2004; Bendtsen et al. 2005) and found 24 such aspzincins. Therefore, only 21 of the 90 AqAspzs are not detected as potentially destined for extracellular localisation. Again, a detected signal for secretion does not predict ontogenetic expression profile (Fisher's exact tests, *p*-values>0.3; Figure 3.4). Some AqAspzs that lack detected secretion signals (non-classical or classical) are still expressed and some to relatively high amounts, and vice versa, some AqAspzs with predicted secretion signals are lowly expressed (Appendix 3.11). However, there may be a relationship between secretion signal and expression quantity, since AqAspzs with detected secretion signals have a higher mean expression quantity than those without any secretion signals (Student's *t*-test, *p*-value=0.03, significant when corrected for multiple testing using the Benjamini-Hochberg method with an 11% false discovery rate). No relationship was found between secretion signal presence/absence and the known catalytically-important HEXXH+DXXY+NAD motif; rather all four possible combinations of these two characteristics occur close to proportionally equal levels (Fisher's exact test, *p*-value=0.7; Appendix 3.11). Because no explicit patterns with expression or sequence composition were found, it remains unclear whether the lack of secretion signal in 21 aspzincins is a true signal reflecting functional diversity and/or lack of functional constraint, or gene prediction errors. Certainly though, it appears that the majority of the AqAspzs potentially have an extracellular destination.

Twenty of the 90 AqAspzs lack the catalytically-important HEXXH+DXXY+NAD motif (Appendix 3.11) and there is not a relationship between the lack of this motif with ontogenetic expression profile (Fisher's exact tests, *p*-values>0.3; Figure 3.4). While some AqAspzs without the aforementioned motif

are still expressed and some to relatively high amounts (Appendix 3.11), generally those AqAspzs with the motif are more highly expressed than those without it (Student's *t*-test, *p*-value=0.04, significant when corrected for multiple testing using the Benjamini-Hochberg method with an 11% false discovery rate). Further, a higher proportion of the 16 unexpressed AqAspzs do not have the motif (50% of the unexpressed compared with 16% of the expressed; Fisher's exact test, *p*-value=0.02, significant when corrected for multiple testing using the Benjamini-Hochberg method with an 11% false discovery rate). Therefore, while it appears that some AqAspzs without the motif are expressed and thus may reflect functional divergence, some AqAspzs may not be functional since they are not expressed (albeit in this one transcriptome) and do not seem under functional constraint.

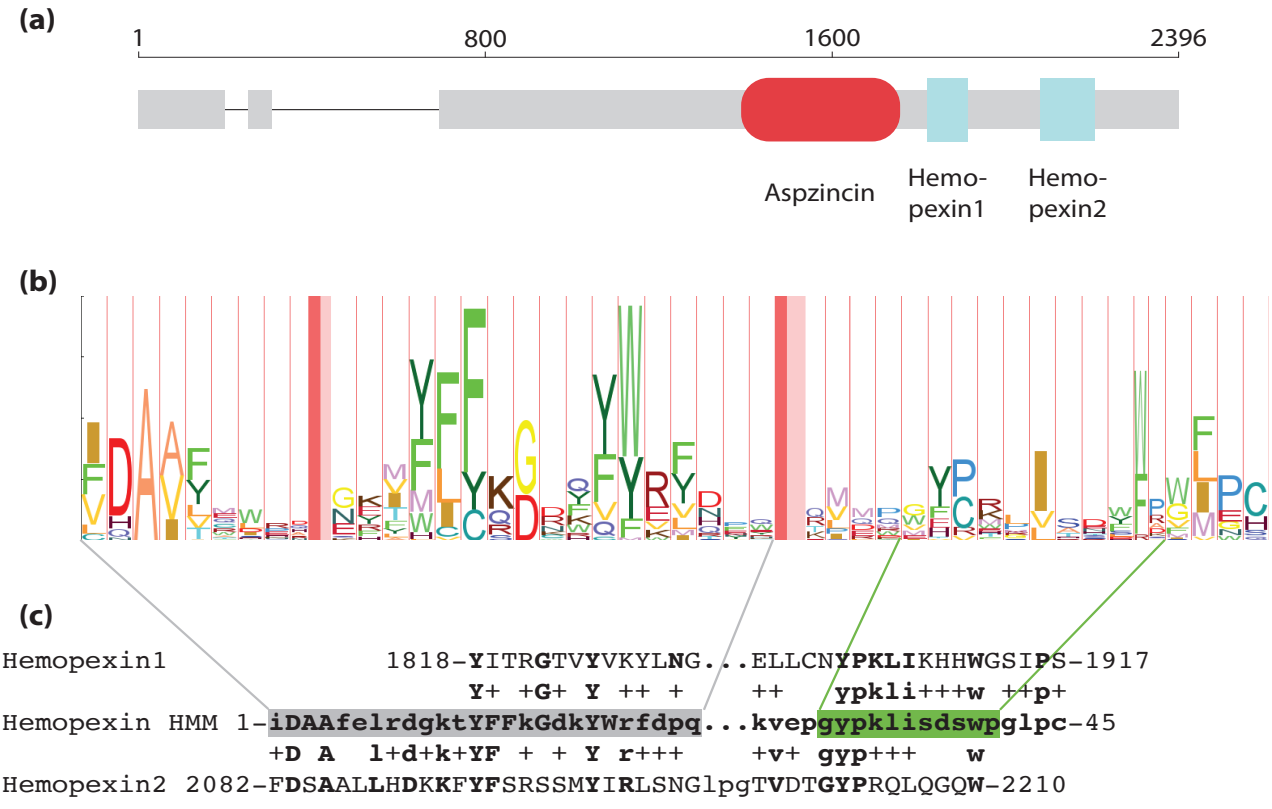
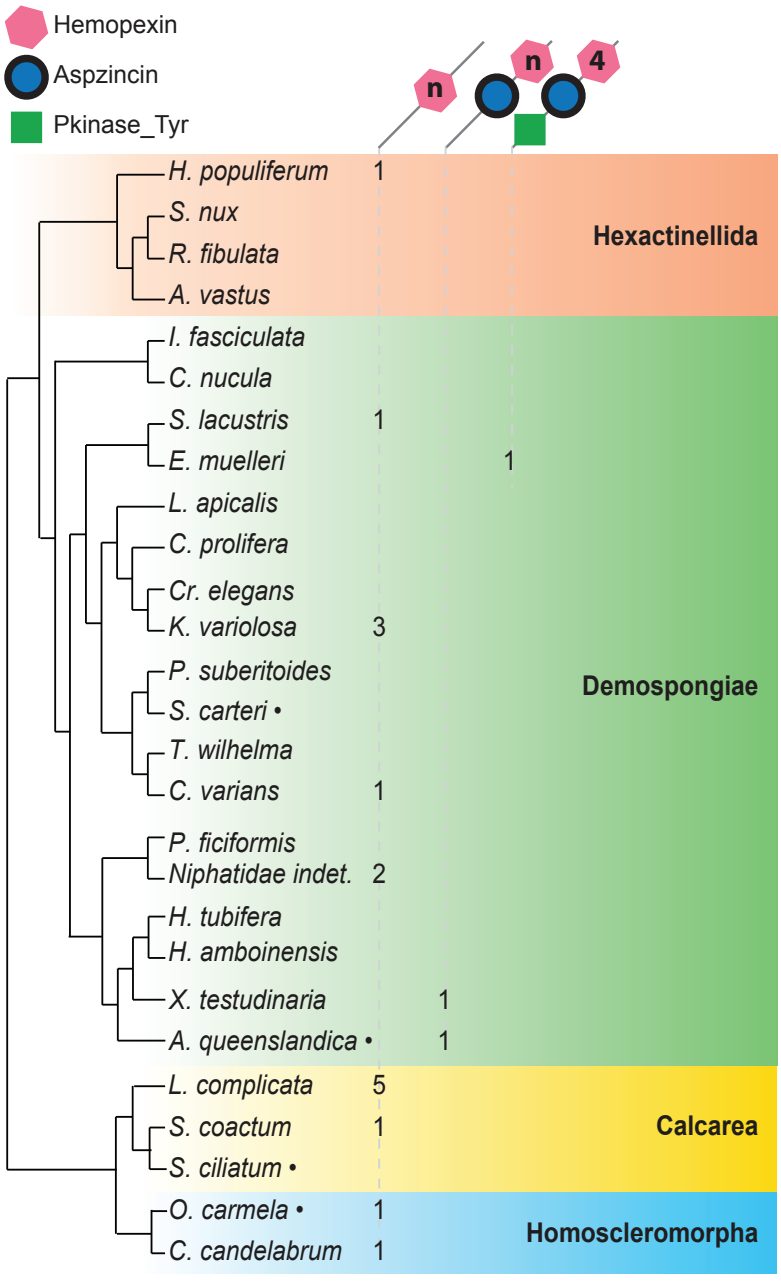


Figure 3.6 The aspzincin-hemopexin domain combination in A. queenslandica
(a) The gene Aqu1.216638/Aqu2.1.25761_001 to scale, exons in grey boxes, introns in grey lines. (b) Pfam's hidden Markov model for the hemopexin domain, letter size reflects the degree of amino acid conservation. (c) The consensus of the hemopexin HMM aligned with the two hemopexins of Aqu1.216638/Aqu2.1.25761_001. '+' in the identity lines reflect synonymous substitutions and capital letters depict highly conserved residues.

3.4.6 The aspzincin and hemopexin domain combination

In Chapter 2 I found that all but one of the AqAspzs only contain a single aspzincin domain; the exception is Aqu1.216638|Aqu2.1.25761_001, which contains two hemopexin domains after the aspzincin domain (Figure 3.6). Most of the hemopexin domain-containing species in the Pfam database are animals, with the exception of five plants, 15 fungi, and 39 bacteria (v31.0, accessed February 2017; http://pfam.xfam.org/family/PF00045#tabview=tab7). Intrigued by a possibly novel gene forming from an HGT-derived domain fusing with host domains, I first checked the integrity of the gene model that predicts these domains together in one gene. Four different gene model prediction methods have all generated the same hypothesis and transcripts cover the complete gene model; therefore, the gene model prediction is well supported.



Next I searched for hemopexins in the 26 other sponges and found 11 other species with hemopexin domain-containing transcripts or gene models (Figure 3.7). All but two of these sponge hemopexin domain-containing sequences contain no other domains,

Figure 3.7 Distribution of the aspzincin-hemopexin domain combination in the sequenced Porifera
Counts reflect the number of unique transcripts containing particular domain architectures, or for those demarked by '•', gene models. Domain architectures were found via submissions to the Pfam database. Species are arranged by class, with predicted topology adapted from Thacker et al. (2013), Whelan et al. (2015), and Grice et al. (2017). Tree and domain architecture cartoons are not to scale.

but the demosponges *X. testudinaria* and *E. muelleri* each encode one hemopexin domain-containing transcript that also contains an aspzincin domain (Figure 3.7). For these two demosponges, the aspzincin-hemopexin transcript is the only aspzincin-containing transcript identified (Figure 3.1).

The aspzincin-hemopexin domain combination could reflect post-transfer fusion within a sponge ancestor, or the originally transferred gene may have also contained hemopexins, which have since been lost in most bacteria. Therefore, using BLASTp, I submitted only the hemopexin-containing sequence of the aspzincin-hemopexin genes to the NCBI nonredundant database and the best hits were bacterial. Next, to see if this domain combination also exists in bacteria, I re-searched the NCBI nonredundant database with the entire genes. Each gene had the same overall result of hits mostly to either the aspzincin or hemopexin domains, but not to both. Only three sequences from three different species cover both the aspzincin and hemopexin domains (Table 3.2). These genes contain an aspzincin and seven or eight hemopexin domains, and are in species from the Actinobacteria genus *Streptomyces* and the Alphaproteobacteria genus *Rhizobium* (Table 3.2). I blasted these genes, but found no other aspzincin-hemopexin genes. In the Pfam, SUPERFAMILY, and NCBI nonredundant databases, I found 102 other hemopexin domain-containing genes in 67 bacterial species, though none also contain an aspzincin domain. However, one of these genes contains a metallopeptidase astacin domain, 18 contain different metal binding hydrolyse domains, and seven contain three types of hydrolyse related domains (Figure 3.8).

Table 3.2 Taxonomic details of the bacterial aspzincin-hemopexin genes

Phylum	Class	Order	Species	Accession	Domain architecture
Proteobacteria	Alphaproteobacteria	Rhizobiales	*Rhizobium sp.* Root1220	WP_056541370	Aspz + 8 hemopexins
Actinobacteria	Actinobacteria	Streptomycetales	*Streptomyces yangpuensis*	WP_052757607	Aspz + 7 hemopexins
Actinobacteria	Actinobacteria	Streptomycetales	*Streptomyces erythrochromogenes*	WP_051892487	Aspz + 7 hemopexins

As of February 2017, publically available databases only contain three bacterial sequences that encode both aspzincin and hemopexin domains. Domain architectures are predicted by Pfam v30.0.

Because of the hemopexin-astacin protein and the possible pattern of hemopexins with metal binding hydrolyse domains, I searched for all hemopexin domain-containing genes in representative animal genomes using the Ensembl BioMart tool. All animals had at least one hemopexin-containing gene

except for *T. adhaerens.* The most common domain architecture of these retrieved animal genes involve another metallopeptidase, the Matrixin M10 domain, followed by variable numbers of hemopexin domains (Figure 3.8).

The aspzincin-hemopexin gene of *A. queenslandica,* Aqu1.216638|Aqu2.1.25761_001, is lowly expressed until late in development, when in juvenile and adult developmental stages the expression peaks (Figure 3.3). The gene has an expression pattern that does not significantly correlate with any other aspzincin, but correlates with 94 other *A. queenslandica* genes. The domain enrichment analysis reveals that these 94 genes are statistically enriched with the transmembrane and cell surface domains cadherins, SRCRs and TIGs (Appendix 3.8). Aqu1.216638|Aqu2.1.25761_001 has a predicted signal peptide and good conservation of the key catalytic residues (HEXXH+DXXY+NAD), though lacks the typically conserved aspartate residue. In the phylogenetic analysis presented in Figure 3.2, Aqu1.216638|Aqu2.1.25761_001 is the only *A. queenslandica* aspzincin predicted with high bootstrapping support to group with an aspzincin of another sponge, as opposed to grouping most closely with other *A. queenslandica* aspzincins. With 80% bootstrapping support, the other sponge aspzincin grouped with Aqu1.216638|Aqu2.1.25761_001 is the only aspzincin of *X. testudinaria,* which also contains hemopexin domains. Further, these two genes are predicted closest to the hemopexin domain-containing aspzincin of *E. muelleri* (albeit with support from only 9% of bootstrap replicates). Aqu1.216638|Aqu2.1.25761_001 is both the only aspzincin and the only HGT on its scaffold, and is not separated by genome assembly gaps from any of its immediate gene neighbours on both sides, which are animal-like/native genes.

Figure 3.8 Distribution of hemopexin domain-containing genes throughout representative genomes

(next page)

Domain architecture is presented by cartoon drawings; a dark box around a domain signifies a metallopeptidase (either aspzincin, matrixin M10 or astacin); “n” reflects that there are different numbers of domain copies. Counts reflect the number of relevant gene models in each species, except for the sponge counts – here numbers reflect the total found in each class and for which transcriptomic, not genomic, data was analysed (see Appendix 3.1). All detected hemopexin domain-containing genes in bacteria were investigated and their domain architectures are shown; however all other taxa shown are representative only and do not show the complete taxonomic distribution and domain architecture diversity of the hemopexin domain, but note that no other aspzincin-hemopexin genes have been identified.

Figure 3.8 Distribution of hemopexin domain-containing genes throughout representative genomes

(legend on previous page)

3.5 Discussion

3.5.1 Aspzincins exist in all sponge classes, indicating an ancient horizontal transfer of functional significance preserved through deep time

Aspzincins are found in bacteria and fungi, where their proteolytic activity has been demonstrated (Nonaka et al. 1998; Fushimi et al. 1999; Doi et al. 2004; Arnadottir et al. 2009), though some have zinc-binding roles (Cherrad et al. 2012; Citiulo et al. 2012). Because of their taxonomic distribution, it was surprising to find bacterial-like aspzincins in 16 diverse sponge species that together are predicted to represent both of the two major sponge lineages (Gazave et al. 2010; Gazave et al. 2012). This distribution of aspzincins in the Porifera and the phylogenetic prediction that they are mostly more similar to each other than to any other aspzincins suggests a shared common origin of the poriferan aspzincins. Therefore, I propose that the sponge aspzincins exist from the horizontal transfer of at least one aspzincin from bacterium to sponge ancestor after the sponges split from all other animals, but before the four sponge classes emerged, followed by vertical inheritance and duplication.

The aspzincins of *A. queenslandica* are well-supported in their HGT-derived status (Chapter 2; Conaco et al. 2016). Therefore, the predicted common origin of the sponge aspzincins supports the HGT-derived status of the aspzincins in the other sponge species too. However, since there is poor bootstrapping support (<50%) for some of the relevant nodes in Figure 3.2, they may not all share a common origin and thus some may be the result of independent HGT events and/or contamination. Important though, there is high statistical support for a clade containing AqAspz Aqu1.216638|Aqu2.1.25761_001, which sits in the genome amongst native genes without any assembly gap separations, with the aspzincin of another sponge *X. testudinaria* (Figure 3.2). This result offers confidence to the proposed shared HGT origin of (most) sponge aspzincins and identifies Aqu1.216638|Aqu2.1.25761_001 as the likely *A. queenslandica* orthologue of the (or one of the) originally transferred aspzincin(s) in the common sponge ancestor.

The previous conclusion that most AqAspzs result from post-transfer duplications (with 59% bootstrapping support) is further supported by the phylogenetic analysis of all the identified sponge aspzincins. The analysis here suggests that many of the aspzincin duplications across all the species are species-specific, since many species have sequences that are most closely related to each other. Sponges

showing species specific duplications include *A. queenslandica, H. amboinensis, Niphatidae indet., K. variolosa,* and *C. prolifera* (Figure 3.2). Along with their transcription, these retained duplicates suggest functionality of the aspzincins in multiple sponge species. While the aspzincin expansions in other sponges do not appear of similar magnitude to that in *A. queenslandica,* this could reflect the difference in data sampling, since I have analysed the genomes of *A. queenslandica, S. carteri, S. ciliatum and O. carmela,* yet only transcriptomic snapshots of the other sponges.

The predicted ancient timing of the aspzincin horizontal transfer event to an early sponge likely explains why analyses to date have not offered taxonomic information on the donor. The placement of six bacterial sequences within the sponge sequences in Figure 3.2 has poor support and further, there is no taxonomic consensus in these sequences, which belong to species from four different orders. The obscurity of the donor is another line of support for the HGT-derived status of the sponge aspzincins and the predicted ancient timing of the original transfer event(s).

The inclusion of the other sponge aspzincins in the phylogenetic prediction did not change the previous general conclusions made for the eumetazoan aspzincins in Chapter 2. Again, the vertebrate sequences are clearly different to all the other aspzincins analysed. The coral and arthropod aspzincins remain predicted most closely related to the bacterial and fungal sequences and are either independent and young HGTs or contaminants. A new outcome here is the reliable grouping of the arthopod *H. azteca* aspzincin with that from a Burkholderiales bacterium; whether this closeness reflects a Burkholderiales origin of an animal HGT or contamination awaits resolution. Certainly, the clarity of this phylogenetic signal and more generally, the placing of the two invertebrates with the bacteria and fungi shows how the alternate hypothesises of young HGTs or contamination would manifest in phylogenetic analyses.

3.5.2 The aspzincins of A. queenslandica probably still have at least one proteolytic function

The aspzincins of *A. queenslandica* are particularly intriguing because their large expansion in this species implies they have an important function. Bacterial and fungal aspzincins with characterised proteolytic activity are synthesised as 37-41 kDa inactive precursors composed of a signal peptide, a propeptide, and the catalytic domain (McAuley et al. 2001; Saito et al. 2002; Doi et al. 2004; Schwenteit et al. 2013a). While the specific function of the aspzincin propeptide is not clear, as proposed for the

propeptides of other bacterial peptidases, it may keep the peptidase inactive inside the cell and thus is an important controller of proteolysis (Häse and Finkelstein 1993). Also, the propeptide may help in the correct folding of the enzyme or be involved in intramolecular autocatalysis (Häse and Finkelstein 1993; Gao et al. 2010). Certainly, the initially synthesised prepropeptide structure seems fundamental to an aspzincin with proteolytic activity, along with tight conservation of a number of key residues in the catalytic domain, specifically the HEXXH+DXXY+NAD residues (Schwenteit et al. 2013a). Therefore, because approximately half of the AqAspzs contain a predicted signal peptide and most of the expressed AqAspzs have conserved those key catalytic residues, I infer that at least one of the original horizontally transferred aspzincins had a prepropeptide structure with the catalytically crucial residues, and thus likely had proteolytic activity. Further, since many of the post-transfer duplicated AqAspzs have retained these features, it seems likely that these are still functioning as peptidases in *A. queenslandica*.

Based on the "use it or lose it" evolutionary principle (Ober 2010; Podlaha and Zhang 2010), I infer that approximately half the AqAspzs have maintained their SPs for peptidase roles. Those without predicted SPs but with secretion sequence signals predicted by SecretomeP (Bendtsen et al. 2004; Bendtsen et al. 2005) may have SPs that are incorrectly omitted from their gene model prediction, since SecretomeP detects sequences with general extracellular protein characteristics and does not distinguish between those proteins with SPs, transmembrane helices or proteins characterised as targets for non-classical secretion (Bendtsen et al. 2004; Bendtsen et al. 2005). Because secreted-like AqAspzs with and without SPs are throughout all of the groups of expression-correlated, uniquely expressed and unexpressed genes, and because I have noted other signs of gene model prediction issues in some AqAspzs (Chapter 2.4.4), it is possible that the majority of the AqAspzs in fact have SPs and thus possible peptidase roles. Alternatively, non-classical secretion pathways for proteins without SPs exist and were first characterised almost 30 years ago (Muesch et al. 1990; Rubartelli et al. 1990; Rubartelli et al. 1992); therefore, the apparent lack of SPs yet extracellular sequence traits could reflect that some AqAspzs are targets of a non-classical secretion pathway, perhaps with a co-opted extracellular role. Co-option of duplicated genes can involve changes in where and when they are expressed (i.e. in their regulation) and/or changes in the encoded protein (True and Carroll 2002). Here, SP-containing and SP-lacking,

but apparently secreted AqAspzs, have shared expression patterns; therefore, their regulation may be conserved, but possibly functional diversification has occurred.

Ten of the 20 AqAspzs that do not have a predicted SP and do not show extracellular sequence traits, that is have no secretion signal, are expressed. Further, all bar one are co-expressed with AqAspzs destined for extracellular roles. This suggests that either (1) again, these proteins have SPs that have been erroneously omitted in the assembly and annotation process or (2) since all of the co-expressed genes with no secretion signal have conserved the key catalytic residues for proteolysis, perhaps they have an intracellular peptidase function, such as protein turnover.

The binding specificity of enzymes can be highly sensitive, with subtle amino acid substitutions impacting on substrate specificity (Schnoes et al. 2009; Gerlt et al. 2011; Amin et al. 2013). Further, some enzymes are catalytically promiscuous, which increases their evolvability since new functions can arise if a secondary adventitious activity becomes useful to the organism (Copley 2003; Mandrich and Manco 2009). Therefore, given the large number of conserved aspzincin duplicates in *A. queenslandica,* they may have evolved specificity for multiple substrates – so called moonlighting functions may use either catalytic or structural components of the original enzyme (Copley 2003). Two fungal species have moonlighting aspzincins that have other predicted functions; their zinc binding properties enable their roles in excess metal binding in *B. cinerea* (Cherrad et al. 2012) and host zinc scavenging by the pathogen *C. albicans* (Citiulo et al. 2012). However, co-option of the zinc binding properties is just one possible evolutionary trajectory of the AqAspzs, since moonlighters can also have functions quite distant and unconnected to the normal function or structure of the original peptide (Copley 2003).

Speculations on the putative functions of the AqAspzs include the possibility that the zinc-binding properties of aspzincins may have been harnessed for a zinc nutritional immunity system. Metal homeostasis is a crucial and fundamental biological requirement since many metals are both essential to organisms, but also toxic at certain levels (Honsa et al. 2013; Hao et al. 2015; Palmer and Skaar 2016). Bacteria and eukaryotes have developed different mechanisms of maintaining such metal homeostasis (Hao et al. 2015), and as such, those processes are often challenged in host-pathogen interactions (Hood and Skaar 2012; Hao et al. 2015). For example, as defense from bacterial infection, hosts can

take advantage of the crucial nature of zinc to cellular functioning and can restrict zinc availability to bacteria (Honsa et al. 2013; Palmer and Skaar 2016). Alternatively, hosts can switch their innate immune response and can overload zinc availability to bacteria, thereby using the toxicity of high levels of zinc (Honsa et al. 2013; Palmer and Skaar 2016). In general, bacteria have systems for managing environmental instabilities in zinc, including zinc importers and exporters (Roosa et al. 2014; Hao et al. 2015), which have sometimes further evolved to cope with host-induced lethal zinc fluctuations (Guilhen et al. 2013; Honsa et al. 2013; Braymer and Giedroc 2014; Djoko et al. 2015; Palmer and Skaar 2016). Increased capacities for both resistance to or scavenging of host zinc have been linked to virulence in bacteria (Shafeeq et al. 2013; Hao et al. 2015) and often the systems involved exist on plasmids, pathogenicity islands and/or near transposable elements, thereby enabling their efficient transferal among bacteria (Roosa et al. 2014). Similar to how manganese and zinc chelation by host calprotectin inhibits growth of the invading *Staphylococcus aureus* in mouse abscessed tissue protein (Corbin et al. 2008), *A. queenslandica* may use the AqAspzs to restrict zinc availability to microbes as a form of nutritional immunity. The density of symbionts is an important and fine balance that can also be regulated through host regulation of nutrient availability (Wilkinson et al. 2007). Therefore, any possible nutritional regulation of bacteria by the AqAspzs could be targeting *A. queenslandica* symbionts, or perhaps the symbionts have evolved increased resistance to the system, thereby allowing *A. queenslandica* to target pathogenic bacteria.

Generally, aspzincins are extracellular peptidases and in certain systems, they are well-characterised bacterial virulence factors (Arnadottir et al. 2009; Yamada et al. 2012). For instance, the fish pathogenic bacterium *A. salmoncida* subsp. *achromogenes* secretes the aspzincin AsaP1, which is highly toxic to the fish host and causes an immune response and disease (Arnadottir et al. 2009; Schwenteit et al. 2011; Schwenteit et al. 2013a; Schwenteit et al. 2013b). The effects of AsaP1 on the host fish are eliminated with the removal of the aspzincin via antibodies or gene knockout (Schwenteit et al. 2013a; Schwenteit et al. 2013b). Further, aspzincins isolated from fungi are lethally toxic to mice (Yamada et al. 2012). Therefore, Conaco et al. (2016) speculate that the *A. queenslandica* aspzincins may be a defence mechanism against bacteria. Some sponges may have harnessed and co-opted these putative toxins as a direct defence mechanism. Schwenteit et al. (2013a) show that certain AsaP1 mutants created by site directed mutagenesis were inactive and non-toxic to Arctic charr host *Salvelinus alpinus* L.;

further, the mutants stimulated a specific antibody response against AsaP1 and are thus toxoids. Here, I speculate that amino acid substitutions may have rendered some AqAspzs as similarly inactive and as toxoids – no longer toxic, but antigenic and thus acting as a vaccine – possibly as an immunity defence against toxic aspzincins produced by pathogenic bacteria and/or from the other active AqAspzs.

The aspzincins belong in the metallopeptidase superfamily, which also contains the large matrix metalloproteinase (MMP) family that remodels and degrades the extracellular matrix (ECM) in animals (Ra and Parks 2007; Mittal et al. 2016). This metallopeptidase activity of MMPs is vital for proper ECM remodelling, which in turn is imperative for correct ECM functioning, including the roles of the ECM in signalling pathways governing the growth and differentiation of cells (Lukashev and Werb 1998; Ra and Parks 2007; Mittal et al. 2016). It has therefore been speculated that perhaps some of the aspzincins contribute to ECM remodelling in sponges (Conaco et al. 2016), and this is explored further below.

3.5.3 A putative role of some sponge aspzincins in spiculogenesis

At least two *A. queenslandica* aspzincins appear to play a role in spiculogenesis, the process by which sponge silica skeleton forms (Wang et al. 2012a). K. Roper and S. Degnan (personal communication) used *in situ* hybridisation to analyse the spatial expression of seven AqAspzs of varied ontogenetic expression patterns in *A. queenslandica* embryos. The analysis revealed two AqAspzs that show co-localisation with silicatein, a protein involved in siliceous spicule formation (Figure 3.9; K. Roper, personal communication; Shimizu et al. 1998; Cha et al.1999; Cha et al. 2000; Krasko et al. 2000). These two AqAspzs (AqAspzScs) share an ontogenetic expression profile that is not shared by any other *A. queenslandica* gene in the analysed dataset (Cluster 4 in Figures 3.4 and 3.5). The pattern of AqAspzsSc-Silicatein co-localisation varies: some cells have stronger AqAspzSc signal than that of silicatein or vice versa, and though silicatein is expressed in some cells that have no AqAspzSc expression, AqAspzSc expression appears to be always accompanied with silicatein expression (K. Roper, personal communication).

Silicatein putatively evolved from cathepsin L, a type of cysteine protease enzyme, and the original cathepsin proteolytic activity has changed to one that polycondensates enzymatically dissolved silicon

into amorphous biogenic silica, which creates the skeleton of some demosponges and hexactinellids (Shimizu et al. 1998; Cha et al. 1999; Gröger et al. 2008; Wang et al. 2012a; Riesgo et al. 2015). Thus far, silicatein is exclusively found in sponges, though not all sponges have silicatein – even some siliceous sponges lack silicatein and thus an independent pathway for silica production may exist in some sponges (Maldonado and Riesgo 2007). *A. queenslandica* encodes six silicatein genes (AqSilicateins)

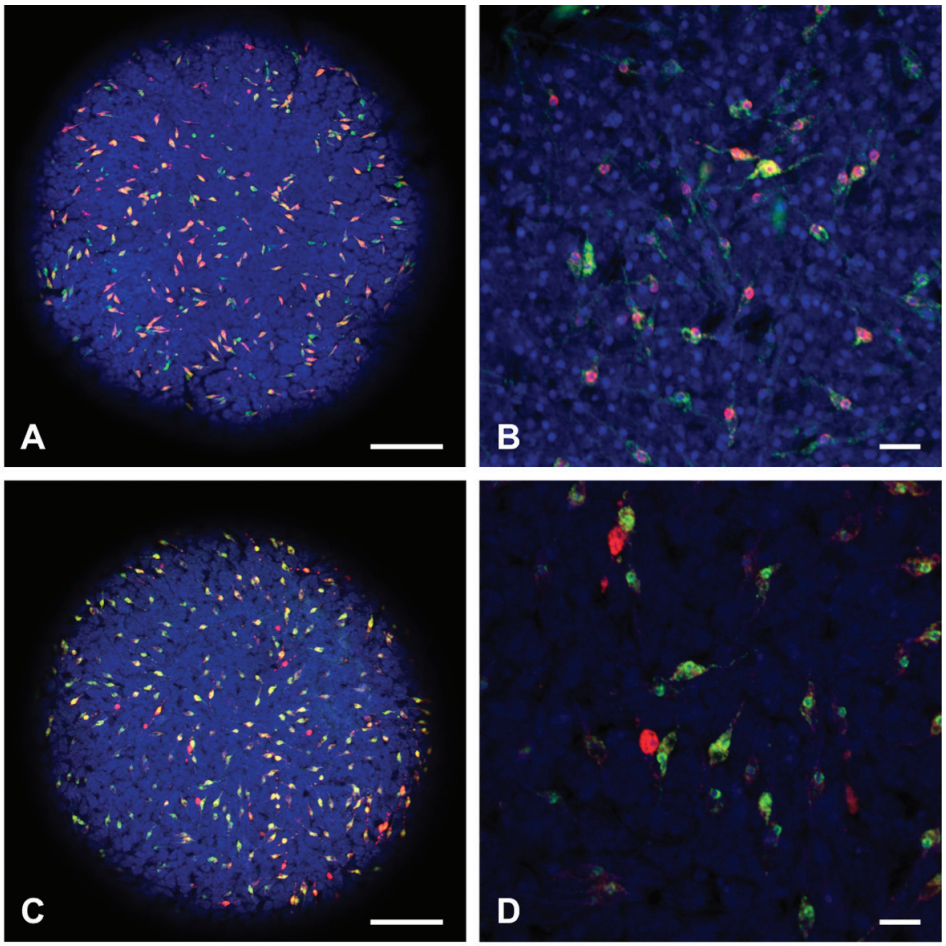


Figure 3.9 Co-localisation of two candidate aspzincins with silicatein in *A. queenslandica* embryos as exhibited by double fluorescent *in situ* hybridisation
(A+B) Expression of aspzincin gene Aqu1.219073|Aqu2.129115_001 (red), silicatein gene Aqu1.228663|Aqu2.1.42494_001 (green), and co-localised expression (yellow). (C+D) Expression of aspzincin gene Aqu1.219075|Aqu2.1.29117_001 (red), silicatein gene Aqu1.228663|Aqu2.1.42494_001 (green), and co-localised expression (yellow). Data and images from K. Roper and S. Degnan (unpublished data). Scale bars: (A+C) 100 µm; (B+D) 20 µm.

and through development the expression of the three most highly expressed AqSilicateins does not relate to the number of spicules (Gauthier 2014). In fact, these genes are highly expressed before spicule formation begins, so the apparent lack of any relationship between silicatein expression and spicule

number may reflect a lag time due to the number of stages of spiculogenesis, as well as an incomplete understanding of the process (Gauthier 2014).

While the pathways governing spicule formation, growth and morphology are not entirely understood and differ among sponge species (Riesgo et al. 2015), in addition to the central role of silicatein in spiculogenesis, collagen is also important for spicule growth by promoting proper arrangement of the immature lamellar structures that form mature spicules (Müller et al. 2005; Schröder et al. 2006; Kozhemyako et al. 2010; Wang et al. 2010; Wiens et al. 2011). Cathepsin L, the evolutionary origin of silicatein, functions in various aspects of collagen cleavage in animals (Kakegawa et al. 1993; Felbor et al. 2000; Dalton et al. 2003; Lustigman et al. 2004) and as such, is speculated to be involved in the reconstruction of the collagen templates that are crucial for spicule formation (Kozhemyako et al. 2010). Spicule formation begins in the specialised cells named sclerocytes (Wilkinson and Garrone 1980; Simpson 1984); these cells (1) absorb and accumulate silica; (2) produce silicatein; and (3) secrete both silicatein and immature spicules (Schröder et al. 2004; Schröder et al. 2007; Wang et al. 2012b). The immature spicules are filaments composed of silicatein on which the first layer of silica has been deposited (Müller et al. 1999; Müller et al. 2005; Wang and Müller 2011). Once secreted, the extracellular growth, and possibly morphology, of the immature spicule is promoted by the surrounding collagen arrangement (Müller et al. 2005; Schröder et al. 2006; Kozhemyako et al. 2010; Wang et al. 2010; Wiens et al. 2011).

Aspzincins belong in the large metallopeptidase superfamily, which also contains the matrix metalloproteinase family that remodels and degrades the ECM, including collagen, in animals (Ra and Parks 2007; Mittal et al. 2016). Therefore, the expression of the AqAspzScs in sclerocytes along with silicatein suggests not just the co-option of some AqAspzs for a role in spiculogenesis, but prompts further speculation that this role could be involved in the reconstruction of the collagen that promotes correct spicule development. Alternatively, similar to how the proteolytic activity of cathepsin L changed to an anabolic activity in silicatein (Shimizu et al. 1998; Cha et al. 1999), perhaps the AqAspzScs have become co-opted to polycondensate silicon for spiculogenesis. The evolution of silicatein from cathepsin L involved a mutation replacing the catalytic cysteine to a serine in a duplicated cathepsin L gene (Cha et al. 1999). Further, the addition of long clusters of serine residues near the catalytic

histidine is hypothesised as necessary for effective spicule formation (Müller et al. 2007), and possibly other changes too, including those that enable an association with galectin (Fairhead et al. 2008). Since subtle residue changes can markedly alter enzyme functionality (Schnoes et al. 2009; Gerlt et al. 2011; Amin et al. 2013), the putative role(s) of AqAspzScs in spiculogenesis could be varied and future work clarifying the role(s) could help develop further understanding of spiculogenesis and of the impact of HGT on animal evolution.

3.5.4 Rare aspzincin-hemopexin genes discovered in just three bacteria and three sponges

Genes with the aspzincin and hemopexin domain combination exist in at least three demosponge and three bacterial species. Hemopexin domains have the capacity to bind a variety of cells and molecules because they have a disk shaped structure with a large adhesive recognition surface area and they can bind cations and anions (Faber et al. 1995). As such, the hemopexin domain functions in substrate recognition in a variety of cell adhesion factors and MMPs (Faber et al. 1995; Das et al. 2003; Ra and Parks 2007), and target molecules including haem, hyaluronan, triple-helical collagen, and tissue inhibitor of metallopeptidases (TIMPs; Zhu et al. 1994; Gomis-Rüth et al. 1996; Overall et al. 1999; Faber et al. 1995; Tolosano et al. 2010). Because of the flexible binding properties and substrate recognition function of the hemopexin domain, the hemopexin domain might enable the aspzincin-hemopexin proteins to bind and cleave a different substrate than other aspzincins. The versatile binding nature of the hemopexin domain makes speculations on specific binding targets difficult, yet supports the notion that aspzincin-hemopexin proteins have a different role to that of other aspzincins.

The seemingly rare and patchy taxonomic distribution of aspzincin-hemopexin genes in bacteria could result from: (1) biased sampling and database limitations; (2) domain, gene and/or species extinctions; (3) the horizontal transfer of a sponge innovation to bacteria; or (4) an independent fusion event of these domains in bacteria. A novel gene evolving from a fusion event creating a unique domain combination is a well-documented evolutionary process (Nakamura et al. 2006). Further, certain domains have a biochemical propensity for each other (Fani et al. 2007; Kim et al. 2009; Nikolaidis et al. 2014). Hemopexins are often found in MMPs (Das et al. 2003), and I found in the NCBI nonredundant database one bacterial gene that contains a hemopexin and an astacin metallopeptidase domain. Also, more generally, I found eighteen bacterial genes containing both at least one hemopexin and a metal

binding hydrolase domain. Therefore, these metallopeptidase and hemopexin domains may have a biological affinity for each other and this combination could have independently arisen multiple times in diverse taxa through domain fusion events.

The phylogenetic hypothesis presented in Figure 3.2 reliably supports that the aspzincin domains from the aspzincin-hemopexin genes of *A. queenslandica* and *X. testudinaria* are most closely related to each other and, with poor support, also to the aspzincin domain of the only other sponge aspzincin-hemopexin gene, that of *E. muelleri.* It seems likely therefore, that these are orthologues, but it remains unclear if they are orthologues of an original transferred aspzincin-hemopexin HGT or orthologues of an original transferred aspzincin gene that post-transfer fused with hemopexin domains in the sponge ancestor.

In *A. queenslandica* the aspzincin-hemopexin gene is assembled amongst clearly native genes with no assembly gap separation; therefore, the gene is well supported as truly incorporated into the sponge genome. The gene is not correlated in expression profile with other aspzincins, but is significantly correlated in ontogenetic expression with 94 other *A. queenslandica* genes that are more highly expressed in juvenile and adult developmental stages. These genes are enriched with transmembrane and cell surface domains, namely cadherins (function in cell-adhesion; Takeichi 1988), SRCRs (function in nonself-recognition, immunity, host-microbiome interactions; Blumbach et al. 1998; Steindler et al. 2007; Hentshel et al. 2012; Buckley and Rast 2015; Fieth et al. 2016) and TIGs (function in ECM dissociation and movement; Collesi et al. 1996). Along with the typical MMP ECM remodelling role of hemopexin domains (Ra and Parks 2007; Mittal et al. 2016), the described putative co-expression of the aspzincin-hemopexin gene with these extracellular genes supports speculation that the sponge aspzincin-hemopexin genes have a role in ECM remodelling.

3.5.5 Each aspzincin ontogenetic expression profile correlates with different suites of genes

The large expansion of aspzincins within *A. queenslandica* is intriguing and suggests functionality of those genes. Indeed, 68 of the 90 AqAspzs are differentially expressed through the developmental life cycle and many have high levels of expression in relation to many other genes in this transcriptome-wide dataset. In addition, the ontogenetic expression profiles of many AqAspzs are not random, but appear tightly controlled, since most of them fit into one of four co-regulated groups of similar expression

patterns through developmental time. Overall, their dominant role(s) may be in the embryo and larvae life stages, before individuals leave the maternal environment as free-swimming larvae, and large amounts of gene product may be required, at least in the individuals of this transcription experiment.

Each of the gene groups putatively co-expressed with the AqAspzs are different in terms of their domain content, suggestive that the aspzincin gene family in *A. queenslandica* has been co-opted and diversified into groups that might play roles in different developmental processes. Broadly, many of the enriched domains of the putatively co-expressed genes are domains involved in protein-protein interactions. Often these domains are present in multiples and form solenoid structures such as beta-propeller domains (e.g. leucine rich repeats, ankyrin repeats, Kelch motifs, WD domains, and so on). Integration of HGTs (or their derivatives) into host protein-protein interactions and cellular networks is not well understood, even in bacteria (Davids and Zhang 2008; Ragan and Beiko 2009; Gophna and Ofan 2011). However, the putative co-expression of AqAspzs with genes containing domains that are implicated in protein-protein interactions could indicate the means by which the HGT-derived aspzincins interact with native sponge genes. Proteins with more interaction sites and thus increased capacity for protein-protein interactions have been suggested as more likely to become successfully functioning HGTs because they have greater chances of integrating into host protein networks (Gophna and Ofran 2011).

The majority of the domains detected in those 14 genes correlating with Cluster 2 are immunoglobulins and F-box domains, which are found in a wide range of proteins, including extracellular and cell surface proteins with communication, adhesion and degradation roles in processes such as immunity, cell development, proliferation and death (Kipreos and Pagano 2000; Vogel and Chothia 2003; Tskhovrebova and Trinick 2004; Huang et al. 2009). Again, this putative co-expression with protein-protein interacting genes may reflect the mechanism through which aspzincins interact with native proteins and influence their behaviour. Further, while the specific domains are different, there are broad similarities in the domains found in eumetazoan aspzincins (i.e., the vertebrate apolipoprotein-aspzincin combination presented in Chapter 2 and the aspzincin-hemopexin combination discussed here) and in the genes putatively associated with the aspzincins as predicted by putative co-expression, such as

these immunoglobulin and F-box domain-containing genes. The similarities include an extracellular or cell-surface localisation, ECM remodelling roles, and immunity roles including apoptosis.

The vastly different expression patterns of Clusters 1 (mostly embryonic) and 3 (mostly adult) are reflected by enrichments of entirely different domains, with one notable exception. The exception is that both groups of correlated genes are enriched with astacins. This is particularly interesting since like aspzincins, astacins are metalloendopeptidases. According to the MEROPS database, *A. queenslandica* has 44 astacins (v11, accessed February 2017; https://www.ebi.ac.uk/merops/cgi-bin/genome_dist?family=M12). Astacins are found in proteins with a variety of proteolytic functions including degradation of proteins, activation of growth factors, and processing of extracellular proteins (Bong and Beynon 1995). While highly speculative, the co-expression of some aspzincins with native astacins may reflect the co-option or adaptation of the gene regulatory network that controls astacins as a pathway for the functionalisation and regulation of some AqAspzs, since co-expression can reflect gene network co-option (Monteiro 2012)

3.5.6 The metallopeptidase repertoire of A. queenslandica is not deficient

As a frame of reference for the metallopeptidase repertoire of *A. queenslandica*, I considered the metallopeptidase repertoires of a choanoflagellate, seventeen additional sponges, and nine eumetazoans. Based on size and diversity of metallopeptidase types, the metallopeptidase repertoire of *A. queenslandica* is not unusual, in comparison to the other animals analysed. Further, this analysis did not reveal a metallopeptidase deficiency that may have given clues on the sponge aspzincin function(s). Strikingly though, the 90 aspzincins in *A. queenslandica* are the highest number of one type of metallopeptidase found in a single species, equal with 90 putative astacins in *N. vectensis.*

The sponge *C. candelabrum* has the third highest total number of metallopeptidases and the highest diversity of metallopeptidase type. Further, nine of the metallopeptidase types only found in one species are present in one of the sponges, and six of those are in *C. candelabrum.* It is possible that contamination or alternative splicing may explain these figures, since most of the sponge sequences analysed are transcriptomic, and not genomic as for all the eumetazoan species, *M. brevicollis, O. carmela* and *A. queenslandica.* Last, this metallopeptidase analysis has revealed another class of metallopeptidases

that may also result from an ancient bacterium to sponge-ancestor HGT: domain DUF955 (accession PF06114) was found in multiple sponge species, but none of the other species considered. DUF955 is documented in bacteria and viruses (Pfam database v28.0, accessed October 2015; http://pfam.xfam.org/family/PF06114#tabview=tab7); however, I found ten putative domains in six sponges.

3.6 Conclusion

Here, I have identified HGT-derived aspzincins in at least sixteen sponges species that together represent all four of the sponge classes. I conclude that these aspzincins result from an ancient transfer deep in the sponge lineage before the four sponge classes diverged, which pre-dates the demosponge lineage timing of the transfer previously suggested by Conaco et al. (2016). Further, I demonstrate that the aspzincin-hemopexin domain combination is rare in contemporary sequence databases and could represent an ancient bacterial protein that is almost extinct, or a sponge innovation back transferred to bacteria, or convergent evolution. In *A. queenslandica*, the aspzincins have various different ontogenetic expression profiles, each of which correlates with different suites of genes. Further, these aspzincins are most likely secreted for extracellular functioning and are likely employed for at least one proteolytic role, though some may remain intracellular. The sponge lineage diverged into the four contemporary sponge classes between 450 and 700 million years ago (Erwin et al. 2011). The conservation of aspzincins derived from an ancient HGT event from bacterium to sponge ancestor hundreds of millions of years ago suggests they are functionally important in the biology of these sponges.

Chapter 4 - The associations of mobile elements with horizontal gene transfers in the sponge *Amphimedon queenslandica*

4.1 Abstract

Mobile genetic elements (MEs) are speculated as possible mechanistic players in horizontal gene transfer (HGT) in animals because of their ability for genomic excision and integration. While some reports of animal HGTs are accompanied by a description of a nearby ME, few studies have systematically tested for an association of HGTs with MEs in animal genomes, and those that have did not account for the confounding factor of unequal rates of gene duplication in the gene groups compared. *Amphimedon queenslandica* is a sea sponge with 576 previously described putative HGT-derived genes of mostly bacterial, plant or fungal origin. These genes are predicted to result from ancient transfer events followed by duplications of some genes. Here, I find similar densities of repeats surrounding unduplicated HGTs and native genes; however, the repeats content surrounding unduplicated HGTs has slightly increased proportions of the *helitron* DNA transposon and simple repeats. These sequence elements may be associated with the genomic integration of HGTs; further, the regulatory elements often harboured by *helitrons* may increase the chances of nearby HGTs becoming transcribed, functional, and fixed in populations. Using Repeat Masker, RepeatModeler and Pfam domain content results, 168 HGT-derived genes are classified as transposable elements. Strikingly, half of these are unknown in class, a quarter are *copia* long terminal repeat retrotransposons and the other quarter are *helitrons*. Phylogenetic analyses suggest that some HGT-derived *A. queenslandica helitrons* are most similar to each other, probably because of post-transfer replication. None of those analysed HGT-derived *helitrons* are expressed in an ontogenetic transcription dataset, indicating that they may not have been domesticated and conserved from functional constraint, but result from recent transpositional activity. Finally, 41 of the HGT-derived gene models have high sequence similarity to proteins present on bacterial plasmids, suggestive that plasmids were a transferring vector for some of the HGTs. Overall, these results support the mechanistic involvement of MEs in HGT to animals, though alternative scenarios of duplications and transposon-enabled functionality of HGTs can lead to similar genomic signatures.

4.2 INTRODUCTION

Genomes throughout all domains of life contain mobile elements (MEs) that are dynamic and highly diverse in terms of copy numbers and genomic locations, even among individuals of the same species (McClintock 1950; Lederberg 1952; Zinder and Lederberg 1952; Lerat 2010; Keane et al. 2013; Fiston-Lavier et al. 2015). In prokaryotes, such MEs include plasmids, bacteriophage, genomic islands (GIs), integrons, group I and II introns, and transposable elements; the latter of which include insertion sequences, conjugative transposons, and mobilisable transposons (Table 4.1; Lederberg 1952; Zinder and Lederberg 1952; Brügger et al. 2002; Osborn and Böltner 2002; Toussaint and Merlin 2002; Makarova et al. 2014). In eukaryotes, MEs are classified into two distinct classes of transposable elements (TEs) based on their transposition mechanism: class I retroelements replicate via an RNA intermediate with a copy-and-paste type method, while class II DNA transposons do not function through an RNA intermediate, but have either a copy-and-paste method via DNA or have a cut-and-paste method (Table 4.1; Ivancevic et al. 2013; Keane et al. 2013; Fiston-Lavier et al. 2015). In prokaryotes and eukaryotes, MEs move both intra- and inter- cellularly, including nonsexual movement between species (Lederberg 1952; Zinder and Lederberg 1952; Allen et al. 2009; Gilbert et al. 2013; Walsh et al. 2013; Makarova et al. 2014). Such mobility is achieved by a diversity of MEs through a range of mechanisms; however, different combinations of the same functional building blocks make up much of the diversity (Osborn and Böltner 2002; Toussaint and Merlin 2002). For instance, a range of MEs have one of three types of recombinase coupled with different mobilising transfer components that originate from plasmids (Osborn and Böltner 2002). These independent and exchangeable functional modules create a continuum collage of MEs and a seemingly flexible system (Osborn and Böltner 2002; Toussaint and Merlin 2002; Zaneveld et al. 2008; Wozniak and Waldor 2010).

Distinct yet inherently related to MEs is horizontal gene transfer (HGT), the nonsexual interspecies transfer of genetic material (Kidwell 1993). Both HGT and MEs were discovered in the mid 20th century, with the discovery of HGT in bacteria (Griffith 1928; Avery et al. 1944; Tatum and Lederberg 1947) leading to the discovery of bacteriophage and plasmid MEs because of their mechanistic role in HGT (Lederberg 1952; Zinder and Lederberg 1952). In eukaryotes, McClintock (1950) demonstrated the activity of TEs in maize and, 25 years later, the first well-supported case of eukaryotic horizontal

transfer was documented; this was the transfer of a retroviral TE from primates to certain feline species (Benveniste and Todaro 1974).

Further cases of the horizontal transfer of TEs (HTTs) from eukaryote to eukaryote have been well characterised and often the enabling interspecies transfer vectors have been identified. A classical example is the extensive interspecies horizontal movement of the *P* element among *Drosophila,* enabled by the feeding behaviour of a semi-parasitic mite vector, *Proctolaelaps regalis* (Daniels et al. 1990; Houck et al. 1991; Clark et al. 1995; Clark and Kidwell 1997; Loreto et al. 2001). The ticks *Bothriocroton hydrosauri* and *Amblyomma limbatum* are vectors of the non-long terminal repeats (LTR) retrotransposon *BovB* among reptiles and into marsupials and ruminants (Walsh et al. 2013). Other arthropods are identified HTT vectors too (de Almeida and Carareto 2005; Gilbert et al. 2010), and even a freshwater snail *Lymnaea stagnalis* is the vector of the DNA transposon *SPIN* (Gilbert et al. 2010). Identified viral vectors include a poxvirus, baculoviruses and a double-stranded DNA virus; these viruses have enabled the interspecies transfer in animals of TEs such as *Short Interspersed Nuclear Element (SINE*), *Transposon Ellen Dempsey (TED*), *TCp3.2*, and *mariner* (Friesen and Nissen 1990; Jehle et al. 1998; Yoshiyama et al. 2001; Turnbull and Webb 2002; Piskurek and Okada 2007). Routh et al. (2012) found that the virus-like particles that package the single-stranded RNA flock house virus (FHV) also contain large amounts of host sequence, thus FHV virions could facilitate HGT between eukaryotes that are infected by the same viral pathogen. In butterflies and moths, the DNA transposon *mariner* may act as a vector for the non-LTR retrotransposon *CR1*, a possibility inferred by the finding of several *CR1*/*mariner* fusion sequences found in the genomes of species from *Maculinea* and *Bombyx* (Novikova et al. 2007; Sormacheva et al. 2012). Finally, some LTR retrotransposons such as the *gypsy* element in *Drosophila* species are independently infectious and do not need a vector (Song et al. 1994). These TEs have an additional open reading frame in the same position as the functional envelop-like gene *env* of retroviruses – *env* facilitates the interspecies mobility of retroviruses by recognising host surface receptors, thus allowing penetration of the membrane and infection of new cells (Malik et al. 2000; Vicient et al. 2001). Despite the numerous reports of HTTs, which involve a broad range of TEs, enabling vectors and animals, the precise mechanisms involved remain largely unknown (Schaack et al. 2010; Ivancevic et al. 2013).

Table 4.1 Summary of mobile elements from the three domains of life

(Part 1 of 2)

Class/type	Order	Superfamily	Structural components	Autonomous?	Taxonomic distribution	Comments	References
Class I retroelements	*LTR*	*Copia*	LTR, *gag, pol* (AP, INT, RT, RH)	Y	P, M, F, O		Wicker et al. 2007
		Gypsy	LTR, *gag, pol* (AP, RT, RH, INT)	Y	P, M, F, O		Wicker et al. 2007
		Bel-Pao	LTR, *gag, pol* (AP, RT, RH, INT)	Y	M		Wicker et al. 2007
		Retrovirus	LTR, *gag, pol* (AP, RT, RH, INT), *env*	Y	M		Wicker et al. 2007
		ENV	LTR, *gag, pol* (AP, RT, RH, INT), *env*	Y	M		Wicker et al. 2007
	YR	*DIRS*	IRT, *gag, pol* (RT, RNase H, DNA N-6-adenine-methyltransferase), tyrosine recombinase, internal complementary region	Y	P, M, F, O		Muszewska et al. 2013
		Ngaro	Terminal repeats, *gag, pol* (RT, RH), tyrosine recombinase	Y	M, F		Muszewska et al. 2013
		VIPER	*gag*, tyrosine recombinase, RT, RH	Y	O		Lorenzi et al. 2006
	PLE	*Penelope*	ITR, RT, endonuclease, LTR	Y	P, M, F, O		Wicker et al. 2007
	LINE	*R2*	RT, endonuclease	Y	M		Wicker et al. 2007
		RTE	Apurinic endonuclease, RT	Y	M		Wicker et al. 2007
		Jockey	ORF1, apurinic endonuclease, RT	Y	M		Wicker et al. 2007
		L1	ORF1, apurinic endonuclease, RT	Y	P, M, F, O		Wicker et al. 2007
		I	ORF1, apurinic endonuclease, RT, RH	Y	P, M, F		Wicker et al. 2007
	SINE	*tRNA*	(Noncoding)	N	P, M, F		Wicker et al. 2007
		7SL	(Noncoding)	N	P, M, F		Wicker et al. 2007
		5S	(Noncoding)	N	M, O		Wicker et al. 2007
Class II DNA transposons	*TIR*	*Mariner*	IRT, DDE transposase	Y	P, M, F, O		Wicker et al. 2007
		P	ITR, Transposase	Y	P, M		Wicker et al. 2007
		hAT	IRT, DDE transposase	Y	P, M, F, O		Wicker et al. 2007
		Mutator	IRT, DDE transposase	Y	P, M, F, O		Wicker et al. 2007
		Merlin	IRT, DDE transposase	Y	M, O		Wicker et al. 2007
		Transib	IRT, DDE transposase	Y	M, F		Wicker et al. 2007
		PiggyBac	IRT, Transposase	Y	M, O		Wicker et al. 2007
		PIF	IRT, DDE transposase, ORF2	Y	P, M, F, O		Wicker et al. 2007
		CACTA	IRT, Transposoase, ORF2	Y	P, M, F		Wicker et al. 2007
		IS	IRT, ORF1, ORF2, Transposase (DDE transposase, tyrosine recombinase or serine recombinase)	Y	M, B, A	Active transposition unit of larger composite Tns. Copy-and-paste, cut-and-paste, co-integrate & rolling-circle modes of intracellular movement.	Kröger and Hobom 1982; Brügger et al. 2002; Zhang and Saier 2009; Siguier et al. 2014
	Crypton	*Crypton*	Tyrosine recombinase	Y	F, M, O		Kojima and Jurka 2011
	Transposons Tns		IRT, transposases, passenger genes, resolvases	Y	B	Originally distinguished from ISs because they carry passenger genes. Most eukaryotic DNA transposons have relatives to bacterial ISs.	Hickman et al. 2010
	Helitron	*Helitron*	Replicon protein, ORFs, YY tyrosine recombinase	Y	P, M, F		Wicker et al. 2007
	Maverick	*Maverick*	C-integrase, ATPase, cysteine protease, DNA polymerase B	Y	M, F, O		Wicker et al. 2007

Table 4.1 Summary of mobile elements from the three domains of life

(Part 2 of 2)

Class/type	Order	Superfamily	Structural components	Autonomous?	Taxonomic distribution	Comments	References
Phages			DDE recombinases, resolvases, short region of sequence identity between element and target site, components for lysis, head, tail, cell-cell contact	Y	V, B	Intercellular movement.	Hickman et al. 2010; Toussaint and Merlin 2002
Genomic islands			Direct repeats, integrases, transposases, IS, genes of putative selective advantage (fitness, resistance, metabolic, symbiosis) or virulence	N	B	Sometimes encompass conjugative transposons, phage & plasmids.	Juhas et al. 2009
Plasmids			Transposase (DDE transposase, tyrosine recombinase or serine recombinase), passenger genes; mating pair formation components (*mpf* or type IV secretion system)	Y	B	Intercellular movement.	Toussaint and Merlin 2002
Integrons			Tyrosine recombinase, promoter *Pc*	Y	B		Toussaint and Merlin 2002; Mazel 2006
Group I introns			Endonuclease	Y	V, B, O, P		Lambowitz and Belfort 1993
Group II introns			Reverse transcriptase	Y	A, B, O, F, P		Lambowitz and Belfort 1993
Integrative conjugative elements (ICEs)			Tyrosine recombinase, passenger genes, mating pair formation components (*mpf* or type IV secretion system)	Y	B		Toussaint and Merlin 2002; Christie and Vogel 2000; Wozniak and Waldor 2010)
Mobilisaable transposons			Transposase, resolvases, integrase, *oriT* site	Y	B	Intercellular movement facilitated by co-resident ICEs.	Adams et al. 2002; Osborn and Böltner 2002

Elements within classes I and II are presented based on Wicker's classification (Wicker et al. 2007). Autonomy reflects whether an element encodes proteins necessary for transposition and thus can move by itself ("Y") or if it lacks those proteins, but has the cis sequences needed for transposition, along with the required presence of another element ("N"). Taxonomic distribution: "P" (plants), "M" (metazoan), "F" (fungal), "O" (other eukaryotes such as protists), "B" (bacteria), and "A" (Archaea). Structural abbreviations: LTR long terminal repeats, AP aspartic proteinase, INT integrase, RT reverse transcriptase, RH RNase H, ORF open reading frame, IRT inverted terminal repeats.

The occurrence of HGT in prokaryotes and of HTTs among eukaryotes is widely accepted because the mechanisms for ME excision, intra- and inter-cellular movement, transportation of immobile genes, and genomic integration have been characterised to some extent. However, the notion of bacterium to eukaryote transfer remains controversial because of the sequestered germline of many animals and the apical meristem in plants, and because the currently understood mechanisms of HGT/HTT do not explain how genetic material is transferred across wide taxonomic distances and beyond shared vector boundaries (Glansdorff et al. 2009; Wijayawardena et al. 2013; Matveeva and Lutova 2014; Jensen et al. 2016; Koutsovoulos et al. 2016; Ku and Martin 2016). Cases of bacterium to eukaryote transfers first appeared in the literature in the 1980s, with the discovery that a stable plasmid segment from *Agrobacterium* integrates into some plants (White et al. 1982; White et al. 1983; Furner et al. 1986; Zambryski et al. 1989), and the detection of bacterial aldose type II genes in yeast (Schwelberger et al. 1989; Smith et al. 1992). However, reports of interdomain HGT only became common in the 21st century with the advent of whole genome sequencing (reviewed in Chapter 1). While some cases are well-supported, their transfer mechanisms remain mysterious; however, MEs are speculated to facilitate such HGT to animals (Jiang et al. 2004; Friesen et al. 2006; Piskurek and Okada 2007; Keeling and Palmer 2008; Gladyshev et al. 2008; Klasson et al. 2009; Schaack et al. 2010; Acuña et al. 2012; Paganini et al. 2012; Flot et al. 2013; Pauchet et al. 2013; Walsh et al. 2013; Gilbert and Cordaux 2013; Boto 2014; Schönknecht et al. 2014).

Currently, there is correlative support for the mechanistic involvement of MEs in transdomain HGT to eukaryotes. Dunning Hotopp et al. (2007) found horizontally-transferred *Wolbachia* genes contiguous to host retrotransposons in *Drosophila ananassae.* In the rotifer *Adineta vaga*, Gladyshev et al. (2008) discovered HGTs in contigs with the retrovirus-like *env*-containing retrotransposon *Juno* and, since then, Flot et al. (2013) have further characterised this TE-HGT co-occurrence, finding significantly greater densities of TEs around HGTs than around native genes. Similarly, a significantly greater density of TEs around HGTs than around native genes was detected in the genome of the nematode *Meloidogyne incognita* (Paganini et al. 2012). Further, 24% of the identified HGTs of nematodes *M. incognita* and *M. hapla* have close sequence similarity to genes found on bacterial MEs (Paganini et al. 2012). A horizontally transferred mannanase gene in the coffee berry borer beetle sits between two eukaryotic TEs (Acuña et al. 2012). Despite the predicted old age of two contiguous and transcribed

bacterial-like HGTs identified in mosquito *Aedes aegypti,* a neighbouring prophage is still detectable (Klasson et al. 2009). These reported associations of HGTs and MEs, along with the mobile nature of MEs and their flexible mobility mechanisms, indicate that MEs may mediate the incorporation of transferring sequence into a genome (Kidwell 1993; Syvanen and Kado 2002).

Amphimedon queenslandica is a sea sponge that has symbiotic relationships with bacteria and relies on bacteria as a food source (Fieth et al. 2016; Gauthier et al. 2016). Hundreds of predicted HGTs have been detected in the genome of *A. queenslandica,* and are distinct from foreign contaminating sequence erroneously included into the genome assembly (Conaco et al. 2016; Fernandez-Valverde et al. in preparation). The chances of HGTs becoming heritable are possibly greater in sponges because they segregate germ cells from adult stem cells recurrently, while most other animals typically sequester germ cells only early in embryogenesis (Juliano and Wessel 2010). Therefore, sponge germ cells can descend from an adult cell line that has been less protected from foreign DNA than the germ cells of many animals; that higher exposure of adult cells may result in a HGT that in most animals would not become heritable since their germ cells are already sequestered and protected (Tanaka-Ichiara and Watanabe 1990; Tsurumi and Reiswig 1997; Ereskovsky 2010; Funayama 2010; Juliano and Wessel 2010; Nakanishi et al. 2014; Sogabe et al. 2016). The HGTs already identified by Conaco et al. (2016) and Fernandez-Valverde et al. (in preparation) in *A. queenslandica* offer an opportunity to test the hypothesis of a ME mediated mechanism of nonanimal to animal HGT.

Precise detection and analysis of TEs is challenging. First, their highly repetitive nature both in sequence and copy number causes false positives in genome assemblies because of mapping artefacts (Lerat 2010; Ewing 2015; Mir et al. 2015; Fiston-Lavier et al. 2015; Hoen et al. 2015). Second, TEs are highly diverse in their replication rates and their differential activity within individuals creates within-species variation in their copy numbers and genomic locations. As a result, fully-sequenced genomes only contain a reference set of TEs – a TE snapshot for that species (Lerat 2010; Ewing 2015; Mir et al. 2015). For instance, typically two human individuals will vary at 285 genomic locations in regards to the presence or absence of the TE *Long Interspersed Nuclear Element-1* (*LINE-1*; Ewing and Kazazian 2010). While the mobilome of *A. queenslandica* remains largely unexplored, recently insights have been attained on the TE content of the genome. Fifty per cent of *A. queenslandica* protein-coding genes

have exonic sequences at least partially (≥10 bp) originating from TEs and the total TE coverage of long-noncoding RNAs is less than that of protein-coding exons in *A. queenslandica* (22% and 31% respectively; Gaiti et al. 2015).

Most of the HGT-derived genes of *A. queenslandica* have host-like sequence characteristics (Conaco et al. 2016; Fernandez-Valverde et al. in preparation), and many have a similar sequence in at least one other sponge species (Conaco et al. 2016; Chapter 3). These features imply that the HGT-derived genes of *A. queenslandica* result from ancient transfer events and have thus spent a long time in the sponge lineage (Conaco et al. 2016; Chapter 3). Therefore, the detectability of any associated TEs may be compromised since TEs have highly variable but often short lifespans (Feschotte 2008). Nonetheless, traces of TE relics may still exist, particularly since some TEs are maintained through deep time because of their fast replication times (Feschotte 2008; Gilbert et al. 2013). The conserved noncoding element family *LF-SINE* has approximately 245 copies in the human genome and has diverged from an ancient TE that was until recently still active in the coelacanth fish – these data suggest that the TE family has stayed active for more than 400 my (Bejerano et al. 2006). Further, the so-called molecular domestication of TEs results in their sequence being preserved through deep time because their acquired function is under selective pressures (Miller et al. 1999; Xie et al. 2006; Lowe et al. 2007; Feschotte 2008; Chalopin et al. 2015). In Chapter 2, I found ME-related domains in some of the HGT-derived genes of *A. queenslandica* (AqHGT-derivatives), domains such as the reverse transcriptase RVT_2, the integrase core domain rve, and the rolling circle Helitron_like_N. The detectability of these offers promise for finding other MEs putatively associated with the AqHGT-derivatives. In addition, Conaco et al. (2016) assessed HGTs for traits associated with time spent in *A. queenslandica* and the putatively younger HGTs may have stronger TE signals than HGTs that have resided in the sponge lineage for longer.

While there are many speculations about TEs mediating HGT in animals, only two studies have systematically tested for a genomic co-occurrence of HGTs and TEs in animals, as summarised above (Paganini et al. 2012; Flot et al. 2013). However, neither of these studies considered the possibly confounding effect of gene duplication on their analysis, despite reporting extensive duplications of HGTs (Paganini et al. 2012; Flot et al. 2013). Here, I investigate and compare the densities and content

of bacterial and eukaryotic TEs around putatively unduplicated HGTs and native genes in the genome of *A. queenslandica.* Further, I assess and characterise the ME content of the AqHGT-derivatives and I search the AqHGT-derivatives against plasmid-borne genes. Together, these analyses offer support for the involvement of MEs in HGT to animals.

4.3 METHODS

4.3.1 Data

Two versions of gene model predictions for *A. queenslandica* were used; the Aqu1s from the original genome assembly (described in Chapter 1.4; Srivastava et al. 2010; available from Ensembl Genomes http://metazoa.ensembl.org/Amphimedon_queenslandica/Info/Index and the JGI Genome Portal http://genome.jgi.doe.gov/AmpqueaRenierasp/AmpqueaRenierasp.download.html) and the more recent genome and transcriptome based Aqu2.1s (Fernandez-Valverde et al. 2015; available from Ensembl Genomes http://metazoa.ensembl.org/Amphimedon_queenslandica/Info/Index and from QCloud http://amphimedon.qcloud.qcif.edu.au/downloads.html). Whole scaffold and CDS sequences from the *A. queenslandica* genome project were used and are publically available (Srivastava et al. 2010; available from the JGI Genome Portal http://metazoa.ensembl.org/Amphimedon_queenslandica/Info/Index). S.L. Fernandez-Valverde interrogated the first generation Aqu1 gene models against HGTracker (default settings) and the results were shared via personal communication (Fernandez-Valverde et al. in preparation). From the public data of Conaco et al. (2016), a list of putatively younger HGTs was extracted, as outlined in Chapter 2.3.1. All gene expression data mentioned is from the genome-wide ontogenetic transcript dataset already described in Chapter 2.3.1 (Anavy et al. 2014; Levin et al. 2016; Gene Expression Omnibus accession codes GSE54364 and GPL18214). The *A. queenslandica* reference TE catalogue was generated by S.L. Fernandez-Valverde using both a sequence similarity classification approach with Repeat Masker (Smit et al. 1996-2010) and a *de novo* approach for uncharacteristed TEs using Repeat Modeler (Smit and Hubley 2008-2015; available from QCloud http://amphimedon.qcloud.qcif.edu.au/downloads.html).

4.3.2 Searching for bacterial insertion sequences (ISs)

To test if any of the AqHGT-derivatives originated from bacterial insertion sequences (ISs), or were transferred by an IS, 48 HGT-containing whole *A. queenslandica* scaffolds were submitted to the

bacterial IS reference database ISFinder (www-is.biotoul.fr; Siguier et al. 2006). Because this analysis consists of animal scaffolds interrogating a bacterial sequence database with the aim of detecting possibly decaying and old ISs, a pilot test search was used to find an appropriate cut-off point for hit e-values. The default settings for four e-value cut-offs (10, 1, 0.1 and 0.01) were used for the BLASTn searches of all 48 AqHGT-derivative containing scaffolds. Based on the results, an e-value cut-off point of 0.01 was used for further *A. queenslandica* sequence submissions to ISFinder (Siguier et al. 2006): these were (1) all 275 scaffolds classified by HGTracker as likely contamination; (2) the remaining 272 HGT-containing scaffolds not already searched; and (3) 51 sample scaffolds classified by HGTracker as only containing native genes (out of the 11790 total native scaffolds). To avoid scaffold size bias in the native scaffold sampling, the scaffolds were ordered by size and every 230th scaffold was selected, including both the smallest and largest scaffolds.

4.3.3 Investigating the A. queenslandica putative type IV secretion protein Rhs

The detected *A. queenslandica* gene model with high sequence similarity to type IV secretion protein Rhs was further investigated to explore if it is connected to the type IV secretion system (T4SS) that is involved in some bacterial HGT. The gene model, Aqu1.224342|Aqu2.1.36742_001, was submitted against the Pfam v31.0 database (cut-off –E 1.0; http://pfam.xfam.org; Finn et al. 2016) and against the National Centre for Biotechnology Information (NCBI) Conserved Domain Database v3.16 (CDD; default settings with cut-off –E 1.0; www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi; Marchler-Bauer et al. 2017). The gene model was also submitted to the 2018 NCBI nonredundant (nr) database using the protein Basic Local Alignment Search Tool (BLASTp; default settings with organism exclude: *Amphimedon queenslandica* (taxid:400682)). The domain architecture of the five best BLASTp hits was considered using the Pfam and CDD databases (as described already in this section for Aqu1.224342|Aqu2.1.36742_001). The CDS of Aqu1.224342|Aqu2.1.36742_001 and of the five best BLASTp hits were each searched against the predicted T4SS database 2.0 using the BLAST For T4SS Online Program (e-value cut-off 0.1; www.secretion.org/navigateBlast.action; Han et al. 2016). The HGTracker classifications of the three closest genes on each side of Aqu1.224342|Aqu2.1.36742_001 were retrieved and these genes were submitted to the 2018 NCBI nr database using BLASTp (default settings with organism exclude: *Amphimedon queenslandica* (taxid:400682)).

4.3.4 Selection of gene subsets for comparative analyses

To test for a possible association of HGT with eukaryotic TEs in *A. queenslandica*, I compared the TE content that surrounds gene models categorised into five groups based on their HGT status (native or HGT), their duplication status (strictly unduplicated or likely unduplicated) and for the HGTs, their relative age (likely younger or older). Figure 4.1 summarises the group-defining criteria through which gene models were filtered. More specific details are as follows. For the native gene groups, only native genes on scaffolds containing just native genes were used, to remove the possible effect from HGTs existing in the surrounding sequence of the native genes. To avoid the confounding factor of gene duplication and pseudoreplication, only gene models less likely to be gene duplicates were considered. To find "strictly unduplicated" (SUD) genes, the cDNA of Aqu1 HGTs and native genes were separately blasted against the cDNA of all Aqu1 gene models using Geneious Pro 5.1.7 (www.geneious.com; Kearse et al. 2012, e-value cut-off 1e-20). Those models that received no hits were further filtered based on their domain content. Genes containing either no predicted Pfam domains or Pfam domains contained only by one HGT (for the HGTs), or by one native (for natives) were classified as either SUD HGTs or SUD natives. Pfam domain architectures were predicted using the Pfam v28.0 Batch Search tool (cut-off –E 1.0; http://pfam.xfam.org; Finn et al. 2016). Because these strict criteria may create a bias in the type of genes selected thus resulting in unrepresentative gene groups, a second less stringent category was included. HGTs and native genes that received no hits in the cDNA blasting were classified as "likely unduplicated" (LUD) HGTs or LUD natives. Each gene in the resulting five groups were further filtered based on TE content, to avoid bias in the results of the downstream TE analyses. Genes were excluded if they contain a domain related to MEs, as determined by Pfam. Further, since the total TE coverage of *A. queenslandica* protein coding genes is 31% (Gaiti et al. 2015), the *A. queenslandica* reference TE catalogue (described in section 4.3.1) was intersected with the positions of each of the candidate gene models and genes with a combined TE, low complexity or simple and unknown repeat content greater than 31% across the exons and introns were excluded.

4.3.5 Analyses of the TE content surrounding subsets of A. queenslandica genes

Reports of specific TEs near HGTs in animals range in distance from approximately 10 to 20 kbp (Dunning Hotopp et al. 2007; Gladyshev et al. 2008; Acuña et al. 2012). Flot et al. (2013) found an increased TE content in 5000 bp windows surrounding HGTs in comparison to that surrounding native

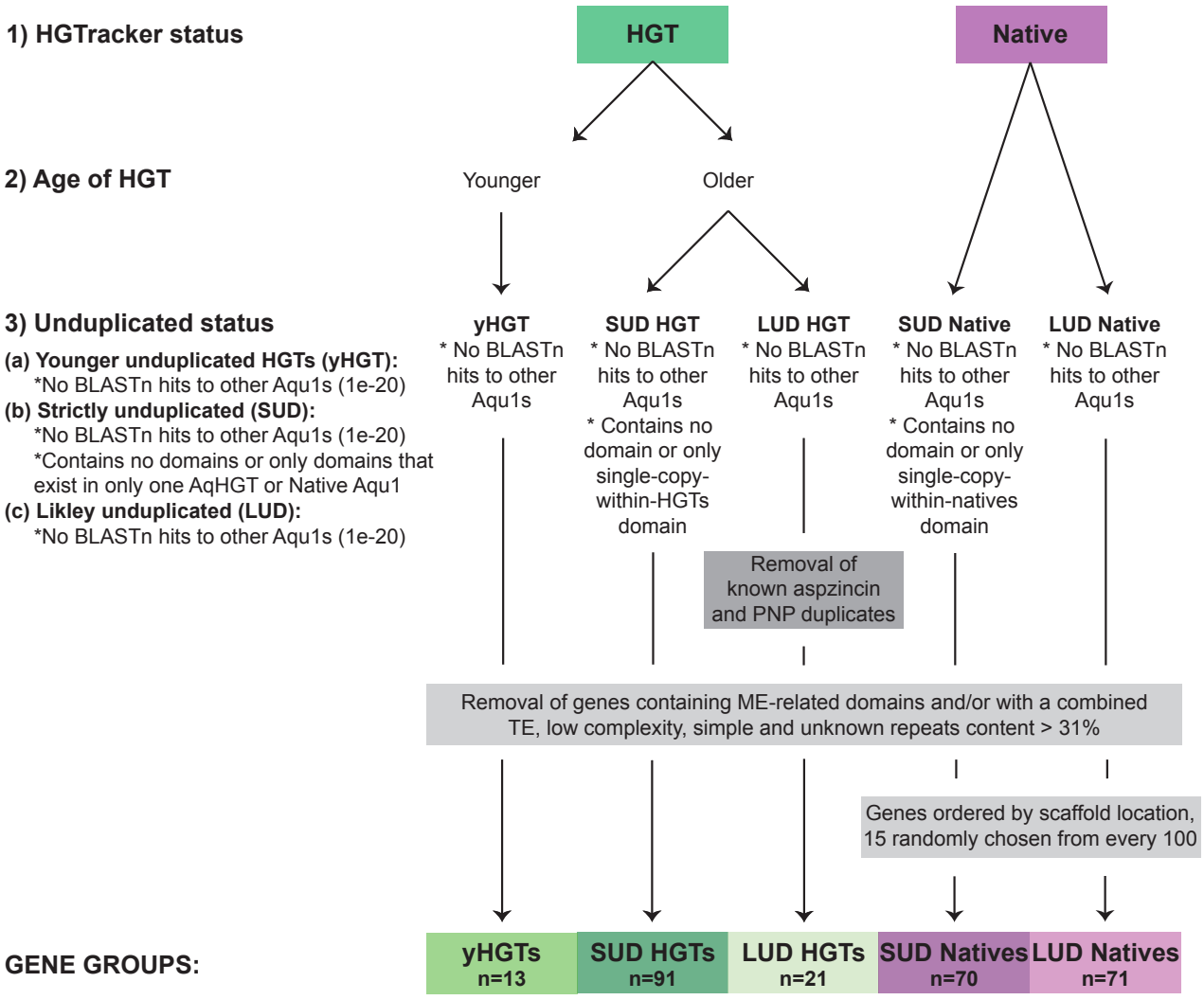


Figure 4.1 Methodology for the selection of gene groups for comparative analyses of the TE content surrounding genes
Flowchart depicts the group-defining criteria through which gene models were filtered. Only genes that are classified by HGTracker (Fernandez-Valverde et al. in preparation) as either native genes or HGTs were considered; further, only native genes on scaffolds containing just natives genes were selected. HGT gene candidates were split by age, as determined by Conaco et al. (2016). Genes that filtered into the SUD and LUD native groups were separately ordered based on scaffold location and for every 100 genes, 15 genes of nonoverlapping search spaces (see section 4.3.4) were randomly selected using a random number generator.

genes in *A. vaga*, and Paganini et al. (2012) found a similar results in all total search spaces of 500, 1000 and 2000 bp surrounding HGTs in *M. incognita.* Therefore, I analysed two search spaces for each selected gene, the gene +/- 2500 bp and the gene +/- 20 kbp, with each search window hereafter referred to as ~5 kbp and ~40 kbp respectively.

The TE and repeats content of each genomic window was extracted from the *A. queenslandica* reference TE catalogue (described in section 4.3.1). In the reference TE catalogue, six repeat classes were characterised by Repeat Masker (Smit et al. 1996-2010) and Repeat Modeler (Smit and Hubley 2008-2015), these are DNA transposons, retroelements, unknown TEs, low complexity sequence, simple repeats, and satellite repeats. For each repeat class of each genomic window, the lengths of each hit were summed and analysed as a proportion of that search window. Using R (R Core Team 2014), between gene group differences in proportions were tested for significance with analyses of deviance with quasibinomial errors, an empirical scale parameter used to account for overdispersed binomial errors arising from much larger residual scaled deviances than residual degrees of freedom (Crawley 2005). The raw probability values are reported for all the statistical tests that analysed the content of these genomic search spaces; however, they were all also corrected for multiple testing using the Benjamini-Hochberg method with a 20% false discovery rate.

4.3.6 TE content of A. queenslandica HGT-derivatives

To compare the TE content of *A. queenslandica* HGT-derivatives, protein-coding genes and long noncoding RNAs, the approach of Gaiti et al. (2015) was used to find the TE content of the AqHGT-derivatives, as follows. The TE and repeats content of each HGT-derivative was extracted from the reference TE catalogue. The extracted annotations were parsed to remove simple and satellite repeats, low complexity sequence and TE annotations under 10 bp in length, both for consistency with the approach of Gaiti et al. (2015) and to reduce possible over-counting of single TEs that can be falsely fragmented in the RepeatMasker analysis (Kapusta et al. 2013). Therefore, these percentage based measures of the TE content of HGT-derivatives incorporate only TE annotations at least 10bps in length from known and unknown classes. The percentage of HGT-derivatives with at least one exon overlapping a TE by at least 10 bp was found using BEDTools v2.26.0 (Quinlan and Hall 2010) to intersect the genomic coordinates of the HGT-derived exons with the genomic coordinates of the known and unknown TEs. The total TE coverage of HGT-derived exons was also determined from the intersection results as the proportion of exonic nucleotides that overlap with a known or unknown class of TE of at least 10 bp out of the total number of exonic nucleotides.

4.3.7 Searching A. queenslandica HGT-derivatives for genes from bacterial MEs

Often bacterial sequence data is generated from whole genomic DNA extracted from isolates and therefore contains both chromosomal and plasmid DNA as well as phage sequences (Edwards and Holt 2013). However, these different DNA sources are often not distinguished from each other in downstream analyses (Edwards and Holt 2013). Consequently, the HGTracker sequence similarity classification step of searching against the NCBI nr database identified bacterial-like HGTs, but more specific information on the likely chromosome or plasmid source is lacking. Therefore, to test if any of the AqHGT-derivatives may have been transferred to the sponge by a plasmid or phage, all Aqu2.1 HGT gene models were searched against a set of proteins present on bacterial MEs. The sequences from all known proteins from bacterial plasmids, prophages and phages were downloaded from the ACLAME database (version 4.0; http://aclame.ulb.ac.be/perl/Aclame/Tools/exporter.cgi; Leplae et al. 2010) and formatted as a custom database in Geneious Pro 5.1.7 (www.geneious.com; Kearse et al. 2012). The Aqu2.1 gene models of the 576 AqHGT-derivatives were searched against this custom database using BLASTp in Geneious Pro 5.1.7 (www.geneious.com; Kearse et al. 2012), with the same cut-off thresholds used by Paganini et al. (2012) in their similar search (e-value <0.001, with at least 30% identity on at least 50% of both the query and subject proteins).

4.3.8 Multiple sequence alignment and phylogenetic analysis

The analyses of this chapter identify first, a possible association of AqHGTs with surrounding host *helitron* DNA transposons, and second, that *helitrons* make up a large proportion of the AqHGT-derived TEs. To explore whether *helitrons* may be more prone to HGT than other genes, a phylogenetic analysis tested whether the HGT-derived *helitrons* result from post-transfer TE activity or from multiple transfer events. To find *helitron* sequences for alignment, the Pfam hidden Markov model (HMM) for the conserved helitron domain "Helitron helicase-like domain at N terminus" (accession PF14214, helitron domain hereafter) was used to interrogate the Aqu1 gene models using the hmmsearch program of HMMER v3.0 (default settings including cut-off –E 10.0; Finn et al. 2011). All hit alignments were manually checked for length and for conservation of key amino acids, as identified by the HMMs. The detected domains were submitted against the 2013 NCBI nr database with BLASTp (default settings with organism exclude: *Amphimedon queenslandica* (taxid:400682)). The top result for each was selected, though duplicates were removed. The helitron domains from these selected sequences were

aligned with those of *A. queenslandica* using the Geneious Pro 5.1.7 Multiple Alignment using Fast Fourier Transform (MAFFT) plug-in (Katoh et al. 2002). The multiple alignment was manually trimmed and refined in Geneious Pro 5.1.7 (www.geneious.com; Kearse et al. 2012). ProtTest 2.4 implemented the Bayesian Information Criterion to find the appropriate model of evolution (Darriba et al. 2011). A Maximum Likelihood (ML) tree was constructed using the PhyML 2.0.12 plug-in in Geneious Pro 5.1.7 (Guindon and Gascuel 2003). Five hundred bootstrap replicates provided statistical support.

4.4 RESULTS

4.4.1 Only short hits to bacterial ISs

The initial e-value cut-off of 0.01 for BLASTn searches of the initial test 48 HGT-containing scaffolds against the IS reference database ISFinder returned nine hits to six different loci (Appendix 4.1). These hits are between 23 and 33 nucleotides in length, which are short relative to the lengths of the ISs to which they hit (mean length of the five involved ISs is 3137 nucleotides). Therefore, to place these results in context, I repeated the searches with more relaxed e-value cut-off values. The number of hits and their mean e-values increased with larger e-value cut-offs; however, the hit lengths did not markedly increase (Appendix 4.2); therefore, a cut-off of 0.01 was used for subsequent searches. Comparisons of ISFinder results for the HGT, native and putative contamination scaffolds are unremarkable, with all hits below 40 nucleotides in length. All hits from all scaffolds have a similar mean e-value and hit length (Table 4.2). While manually investigating these hits, I noticed that one result was a hit within the bacterial-like HGT Aqu1.224342|Aqu2.1.36742_001. When submitted to the 2018 NCBI nr database using BLASTp (default search settings), this gene receives hits to type IV secretion protein Rhs in different bacterial species (e.g., Pseudomonas sp. B28; 90% query coverage, 29% identity,

Table 4.2 Comparison of the ISFinder hits to native, likely contamination and HGT-derivative classified scaffolds of *A. queenslandica*

Scaffold status	Search space (no. of scaffolds; total number of nucleotides)	No. of IS hit containing scaffolds (total nucleotides in hits)	IS hit proportion (nucleotides in hits/nucleotides searched)	Mean hit e-value	Mean hit length
Likely contamination	273; 1700075	17 (416)	0.0245%	0.0037	24
HGT-derivative	320; 64336461	21 (548)	0.0009%	0.0036	26
Native	51; 1377659	2 (47)	0.0034%	0.003	24

Scaffold status was determined by HGTracker (Fernandez-Valverde et al. in preparation); e-value cut-off =0.01; the likely contamination scaffolds exclude the two that contain at least one aspzincin, which have been shown as a *bono fide* part of the genome of *A. queenslandica* (Chapter 2).

e-value=4e-140). Because a type IV secretion system (T4SS) mediates the transport of transferring bacterial DNA into plant cells (Alvarez-Martinez and Christie 2009; Bhatty et al. 2013), this gene was explored in more detail.

4.4.2 Exploration of the A. queenslandica putative type IV secretion protein Rhs

Submission of the *A. queenslandica* putative type IV secretion protein Rhs (Aqu1.224342|Aqu2.1.36742_001) to the Pfam database v31.0 (cut-off –E 1.0; http://pfam.xfam.org; Finn et al. 2016) returned no predicted protein domains. Searching the gene model against the CDD v3.16 (cut-off –E 1.0; www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi; Marchler-Bauer et al. 2017) revealed two predicted RhsA domains (accession COG3209; e-values=3.79e-10 and 0.11), with a Rhs repeat region in between (accession PF05593; e-value=0.05). When submitted to the 2018 NCBI nr database using BLASTp (default search settings), Aqu1.224342|Aqu2.1.36742_001 received 148 hits to 78 bacterial species (141 of the hits are to 72 species belonging to the Gammaproteobacteria class) and five hits to four fungal species. All the hits are to proteins labelled as hypothetical, type IV secretion protein Rhs, Rhs repeat-associated core domain-containing protein, or Rhs family protein. According to the CDD, these proteins contain a variety of Rhs related domains, some also have a secretion system effector C (SseC; Pfam accession PF04888) domain and at least one in the best five hits also contains a toxin domain (Tox-HDC; Pfam accession PF15656). Appendix 4.3 contains the BLASTp details and domain architecture of the five best hits of Aqu1.224342|Aqu2.1.36742_001, each of which is similar to Aqu1.224342|Aqu2.1.36742_001 in length (1613 aa) and has high query coverage (82-90%). Because many of the BLASTp hits are to type IV secretion protein Rhs, the CDS of Aqu1.224342|Aqu2.1.36742_001 and of the best five BLASTp hits (detailed in Appendix 4.3) were submitted against the predicted T4SS database 2.0 (cut-off –E 0.1; www.secretion.org/navigateBlast.action; Han et al. 2016), but none received hits from predicted T4SS genes. The three genes on either side of Aqu1.224342|Aqu2.1.36742_001 on Contig13474 are all native genes and do not offer further clues on the Rhs-like AqHGT-derivative.

4.4.3 Similar repeats densities surrounding HGTs and native genes in A. queenslandica

Comparisons of the densities of the repeats in genomic windows around five groups of *A. queenslandica* genes revealed mostly no significant differences or patterns between the groups (Figures 4.2 and 4.3). The seven repeat classes identified are DNA transposons, retroelements, low complexity, simple repeats,

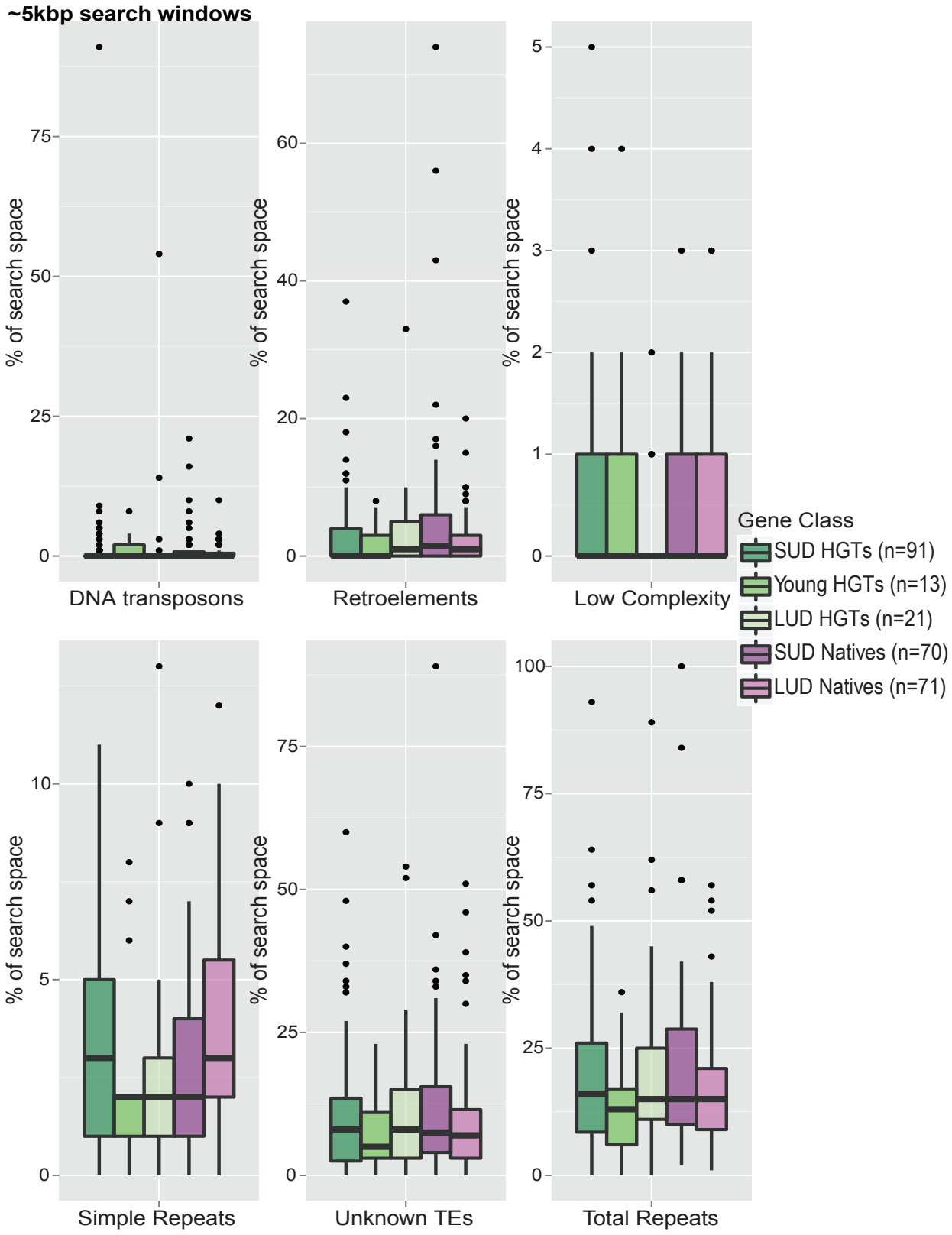


Figure 4.2 The classes of repetitive sequences in 5 kbp windows surrounding unduplicated HGT and native genes

Repeats retrieved from the TE reference library of *A. queenslandica* curated using RepeatMasker and RepeatModeler (Smit et al. 1996-2010; Smit and Hubley 2008-2015). Each data point represents the repeats in one gene +/- 2500bp either side. SUD refers to strict criteria for a gene's putative unduplicated status, LUD reflects less strict criteria, and the young HGTs are 13 genes identified by Conaco et al. (2016) as resulting from more recent transfer events.

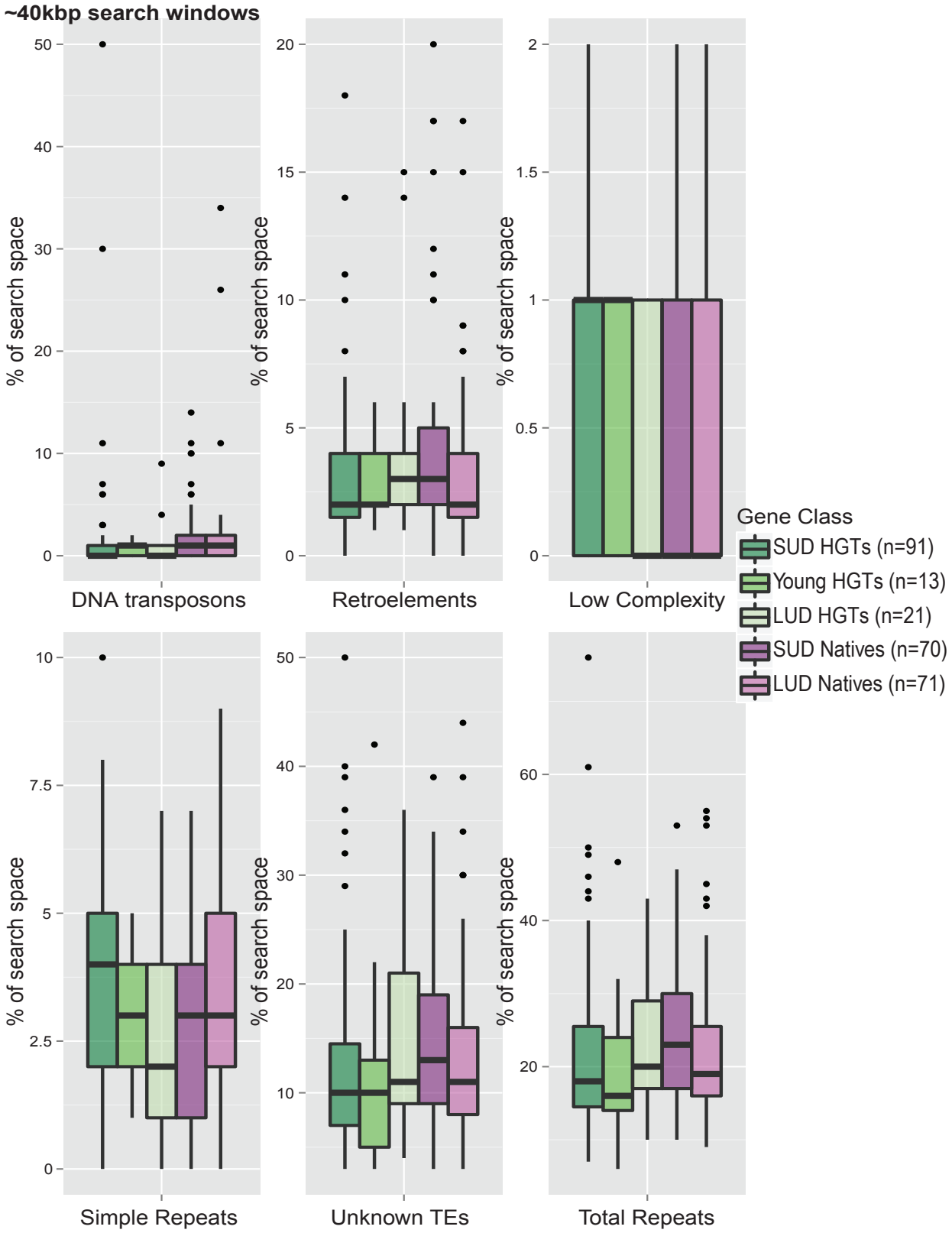


Figure 4.3 The classes of repetitive sequences in 40 kbp windows surrounding unduplicated HGT and native genes

Repeats retrieved from the TE reference library of *A. queenslandica* curated using RepeatMasker and RepeatModeler (Smit et al. 1996-2010; Smit and Hubley 2008-2015). Each data point represents the repeats in one gene +/- 20 kbp either side. SUD refers to strict criteria for a gene's putative unduplicated status, LUD reflects less strict criteria, and the young HGTs are 13 genes identified by Conaco et al. (2016) as resulting from more recent transfer events.

satellites, and unknowns. Because of differences in gene sizes and thus search spaces, the proportions of the search space composed of each of the repeat classes were compared across the gene groups and were mostly not significantly different (raw *p*-values>0.05; Appendices 4.4 and 4.5). All analysis of deviance tests had overdispersed errors with the residual scaled deviance much greater than the residual degrees of freedom, which was accounted for using an empirical scale parameter implemented by quasibinomial errors (Crawley 2005). Finally, all the test results in this section were corrected for multiple testing using the Benjamini-Hochberg method with a 20% false discovery rate. While the raw probability values are reported throughout this section, none are significant after multiple testing corrections and thus the few associations cautiously reported below may be false positives. Nonetheless, the associations are still suggested because both multiple testing and correcting for multiple testing can increase the risk of false negatives, as well as false positives (Vadillo et al. 2016). Since this is early exploration, the disadvantage of missing potential relationships is greater than the disadvantage of reporting possible false positives.

Of the seven exceptions where differences were found, only two suggest that MEs and/or repetitive elements may be associated with HGTs. First, in the ~5 kbp search windows, the LUD HGTs have higher proportions of DNA transposons than the LUD native genes (*p*-value=0.006). On closer inspection, two of the 21 LUD HGTs are contiguous to a DNA transposon (both TEs have top BLASTp hits to other animal sequences, and one belongs to the *CMS-EnSpm* family while the other to the *sola* family). The only other DNA transposon within the search space of any LUD gene is a short hit to a DNA transposon of the *hAT-Ac* family near one of the 72 LUD native genes, hence a difference in proportions was detected. Second, in the ~40 kbp search windows, the SUD HGTs have greater proportions of simple repeats than the SUD native genes (Figure 4.3; *p*-value=0.041).

The other results with possible differences do not support the ME-HGT association hypothesis. In both ~5 kbp and ~40 kbp search windows, the SUD HGTs have lower proportions of retroelements than the SUD native genes (*p*-values=0.045 and 0.032 respectively). The remaining three different results involve comparisons between the SUD and LUD native genes. In the ~5 kbp search windows, the SUD native genes have greater proportions of retroelements, but lower proportions of simple repeats, than the LUD native genes (*p*-values=0.017 and 0.041 respectively). Similarly, in the ~40 bp search

windows, the SUD native genes have lower proportions of simple repeats than the LUD native genes (*p*-value=0.045).

Approximately half of the TE content of each subset of genes comprises of unknown TEs, the next most common TEs are retroelements, followed by simple repeats, then DNA transposons, with much smaller proportions of low complexity sequences and satellite repeats (Figure 4.4). While the TE content of most of the analysed genomic windows does not differ in amount or broad types of repetitive elements, there are a few noteworthy differences in their content of more specific TE classes. The proportion of the TE content composed of *helitrons* for each ~5 kbp search window of the SUD HGTs is higher that that of the SUD natives (testing for between group differences with analysis of deviance with quasibinomial errors, *p*-value=0.0125). The ~5 kbp search windows of the 91 SUD HGTs contain five rolling circle *helitron* TEs (6% of the total TE nucleotide content; hit lengths are 2714, 2274, 371, 260, and 198 nucleotides), while none of the other gene groups have *helitrons* in their 5 kbp search windows, except for three short hits in the 70 SUD natives (0.4% of the total TE nucleotide content; mean hit length=117 nucleotides; Figure 4.5). Meanwhile, the ~5 kbp search windows of the 70 SUD natives contain 15 LTR retrotransposons (13% of the total TE nucleotide content; mean hit length=770 nucleotides), but the other groups have either none or much lower proportions of this TE family (Figure 4.5). Specifically, the 91 SUD HGTs have five short LTR retrotransposon hits and the 72 LUD natives have four short hits (2% and 0.7% of the total TE nucleotide content of each group respectively; mean hit lengths=345 and 132 nucleotides respectively; Figure 4.5). In the ~40 kbp search windows, the younger HGTs are the only group without LTR retrotransposons (Figure 4.5). More precisely, in the ~40 kbp search windows, the 91 SUD HGTs have 48 LTR retrotransposon hits (3% of the total TE nucleotide content), the 21 LUD HGTs have five (4%), the 70 SUD natives have 82 (7%), and the 72 LUD natives have 42 LTR retrotransposon hits (4%; Figure 4.5).

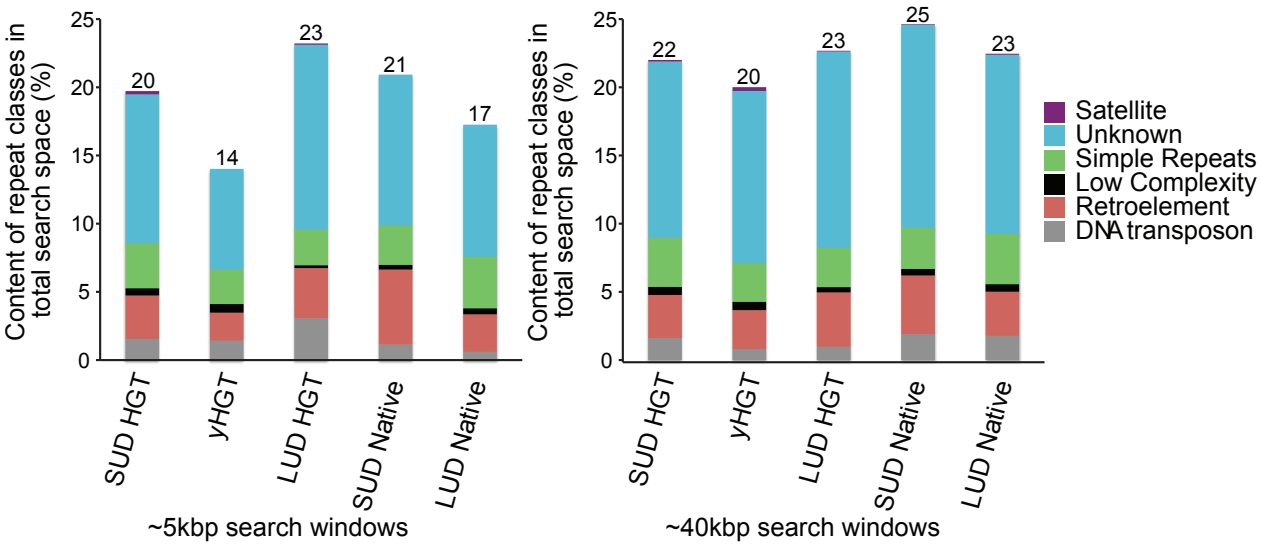


Figure 4.4 The total amount of each repeat class as a percentage of the total amount of sequence searched
The total hits and total search space of each gene are pooled within gene groups. Repeats retrieved from the TE reference library of *A. queenslandica*. SUD refers to strict criteria for a gene's putative unduplicated status, LUD reflects less strict criteria, and the younger HGT gene group consists of 13 genes identified by Conaco et al. (2016) as resulting from more recent transfer events.

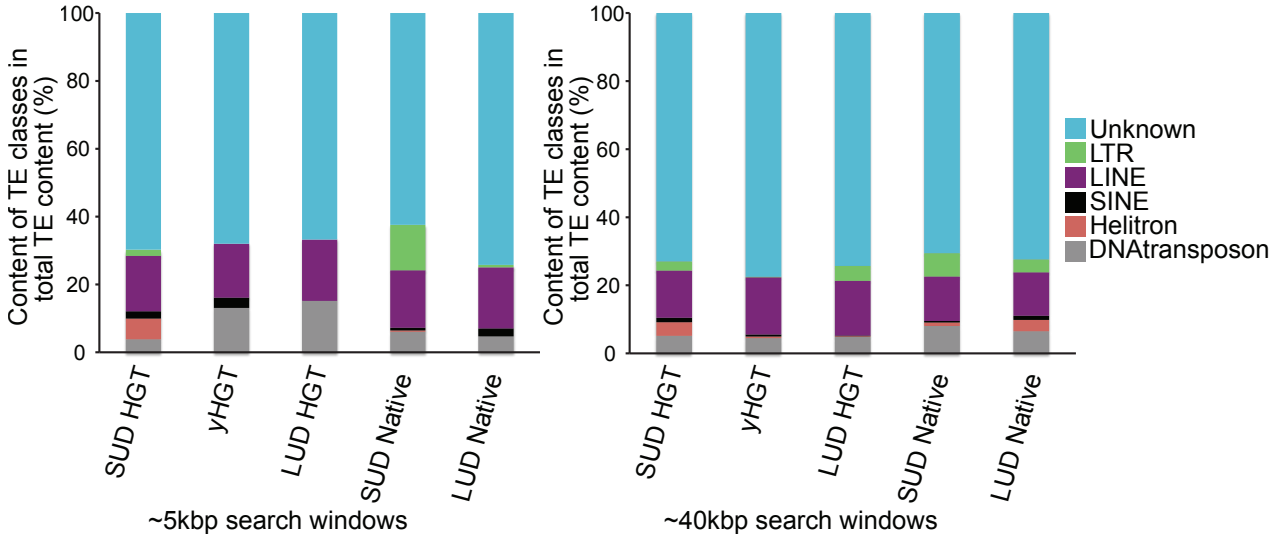


Figure 4.5 The proportion of each TE order within the total TE content of each gene group
These data exclude low complexity, simple and satellite repeats. Repeats retrieved from the TE reference library of *A. queenslandica.* SUD refers to strict criteria for a gene's putative unduplicated status, LUD reflects less strict criteria, and the younger HGT gene group consists of 13 genes identified by Conaco et al. (2016) as resulting from more recent transfer events.

4.4.4 TE content of A. queenslandica AqHGT-derivatives

In Chapter 2, TE-related domains were detected in some of the AqHGT-derivatives. Therefore, here I considered the ME content of all AqHGT-derived gene models. Using the reference TE catalogue

generated for *A. queenslandica*, 43% (245 of 576) of AqHGT-derivatives have at least one exon overlapped by a TE by at least 10 bp. Total AqHGT-derivative exonic coverage by TE-derived sequences is 26% (188290 bp of 729426 bp).

The majority of AqHGT-derivatives have less than 6% exonic coverage by TE sequences (61%, n=350), a further 19% have between 6 and 94% TE exonic coverage (n=108), and finally, 20% have at least 95% TE exonic coverage (n=113). Figure 4.6 shows a consistent spread of genes with an exonic TE content between 1 and 99%. Therefore, because of no clear cut-off point in these data, a conservative and arbitrary cut-off was chosen based on (1) the criterion that the majority of each putative TE gene model comprises of TE-derived sequences and (2) the number of putative TE gene models excluded/included depending on the cut-off point (Appendix 4.6). Those genes with at least 75% of their exons predicted as TE-derived sequences and/or that contain a ME-related domain, determined by Pfam annotations, are here considered putative MEs. Using these criteria, 29% of the AqHGT-derivatives (168 of 576) are categorised as MEs, either because they have at least 75% exonic TE coverage (n=103), or they contain a ME-related domain (n=24), or they meet both these criteria (n=42). HGTracker has classified 68 of these as bacterial-like (62 identified by exonic TE% and four by Pfam domains), 46 as plant-like (38 identified by exonic TE% and eight by Pfam domains), 38 as fungal-like (34 identified by exonic TE% and four by Pfam domains), and 16 as definitely not animal-like but otherwise ambiguous (eight identified by exonic TE% and eight by Pfam domains).

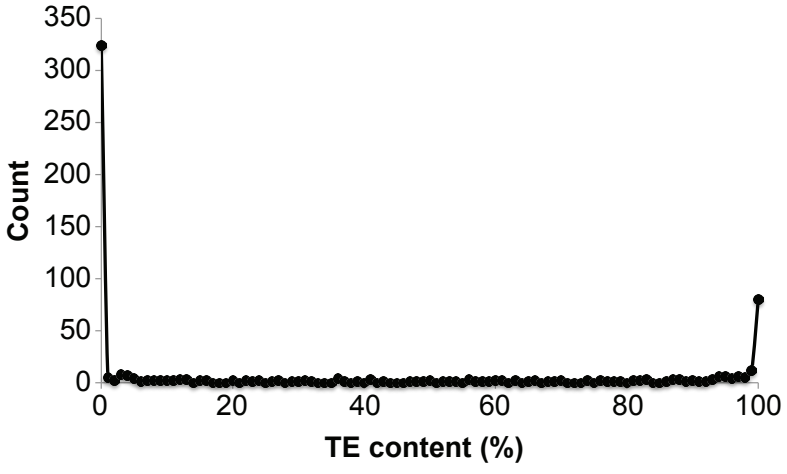


Figure 4.6 TE content of the 576 AqHGT-derived genes
TE content retrieved from the TE reference library of *A. queenslandica* and excludes low complexity sequences, and simple and satellite repeats.

Based on both the Repeat Masker hit classifications and the Pfam domain content, approximately half of the AqHGT-derived MEs are unknown in TE class. They are identifiable based on TE components that are not diagnostic, such as reverse transcriptases, integrases, and transposase endonucleases (e.g., the DDE_3, RVT_2, RVT_1, and rve domains), which are all found in a wide range of TEs across different TE classes. Of the known 73 AqHGT-derived MEs, 42 are Class I retroelements, 29 are Class II DNA transposons and two are baculoviral proteins. Most of the retroelements belong to the LTR superfamily *copia* (n=39), while two are *LINEs* and one is a *SINE*. All but one of the 29 DNA transposons are *helitrons*, with the exception belonging to the *hAT* superfamily. In sum, most of the AqHGT-derived MEs are either unclassified (57%), *copia* LTR retroelements (23%) or *helitron* DNA transposons (17%).

Phylogenetic predictions on the relationships between helitron domains from *A. queenslandica,* other animals, and fungi suggest that at least some of the large proportion of *helitrons* in the AqHGT-derived MEs result from post-transfer replications (Figure 4.7). Searching the Aqu1 predicted proteome with Pfam's HMM for the helitron domain identified 120 *A. queenslandica* helitron domain-containing proteins. HGTracker cannot classify 73 of these genes due to ambiguous BLASTp results, but classifies 20 as native genes and 12 as HGTs. Excluding 27 of the helitron domain-containing genes that had very low query coverage, the sequences were aligned with the corresponding domain region of animal and fungal *helitrons*, which were identified via BLASTp searching of the *A. queenslandica* helitron domains against the NCBI nr database. In the ML tree constructed from this multiple alignment, all six of the included *A. queenslandica* HGT helitron domains form a clade with four other *A. queenslandica* sequences that are unclassified by HGTracker (94% bootstrap support for clade; Figure 4.7). With the exception of one animal sequence from *Nematostella vectensis*, these *A. queenslandica* sequences and all but one of the fungal sequences form a robustly supported clade, separate from all the remaining *A. queenslandica* (none identified as HGTs) and other animal sequences (87% bootstrap support; Figure 4.8). Therefore, the ten analysed fungal-like *A. queenslandica helitrons* that are most closely related to each other are inferred the product of post-transfer replications. Since the fungal-like *N. vectensis* sequence has poor bootstrapping support for its predicted phylogenetic position and is at the end of a long branch, it is unclear whether this animal sequence is a HGT, a contaminant, or the result of vertical evolution and rate variation.

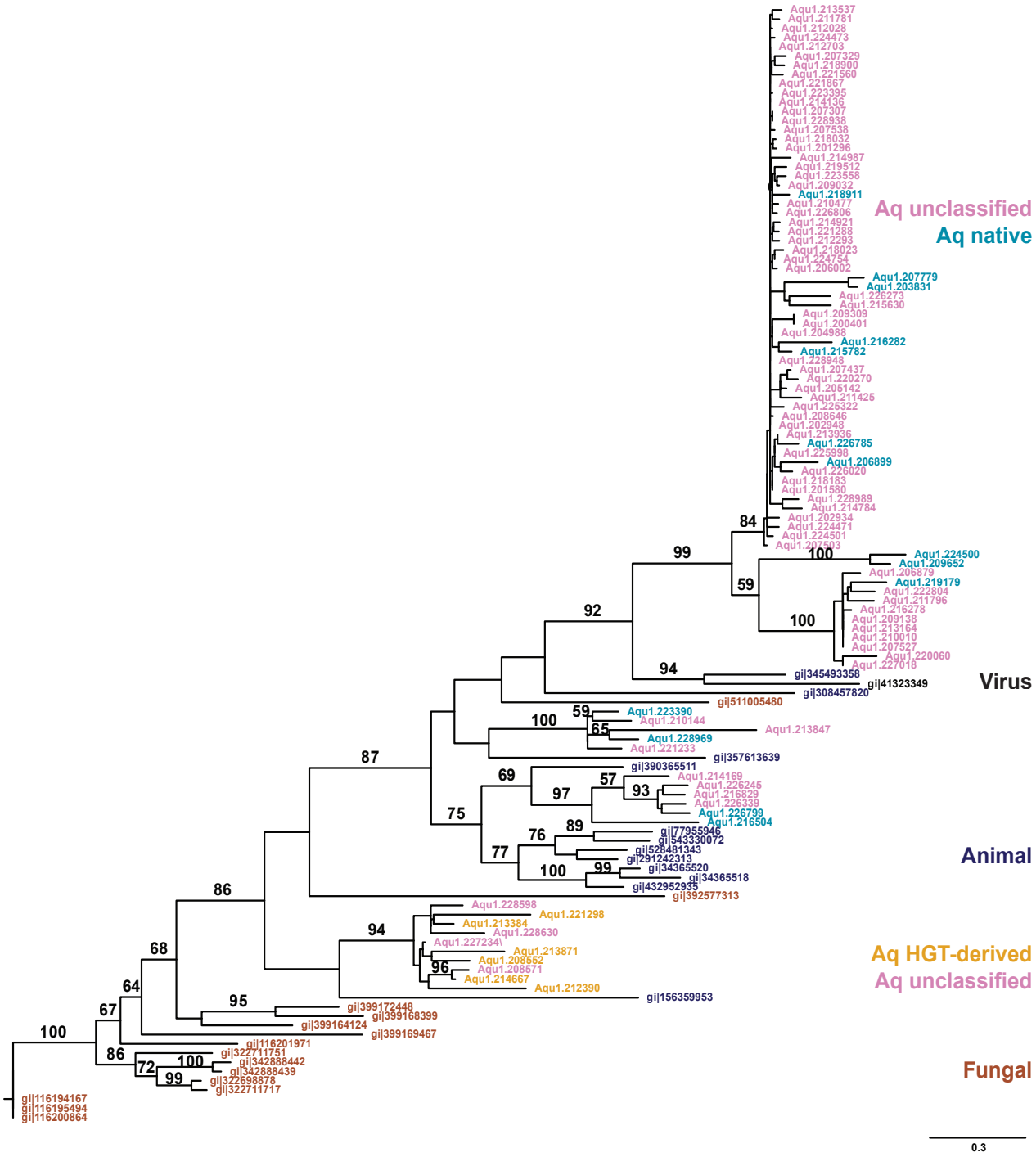


Figure 4.7 Phylogeny of the amino acid sequence for the helitron domain from animals, fungi, and *A. queenslandica*

Phylogeny inferred from a multiple alignment of 145 amino acids by ML. Topology support was obtained from 500 bootstrap replicates; only support values greater than 50 and on major branches are shown. Unrooted tree. "Aq" refers to *A. queenslandica*. The best model of sequence evolution was predicted to be WAG+G+F using Bayesian Information Criterion in ProtTest 2.4 (Darriba et al. 2011). Text colour reflects the taxonomy as reported by the NCBI database for the non-sponge domains. For the sponge domains, the text colour reflects the taxonomic classification assigned by HGTracker based on sequence similarities from BLASTp searches (Fernandez-Valverde et al. in preparation).

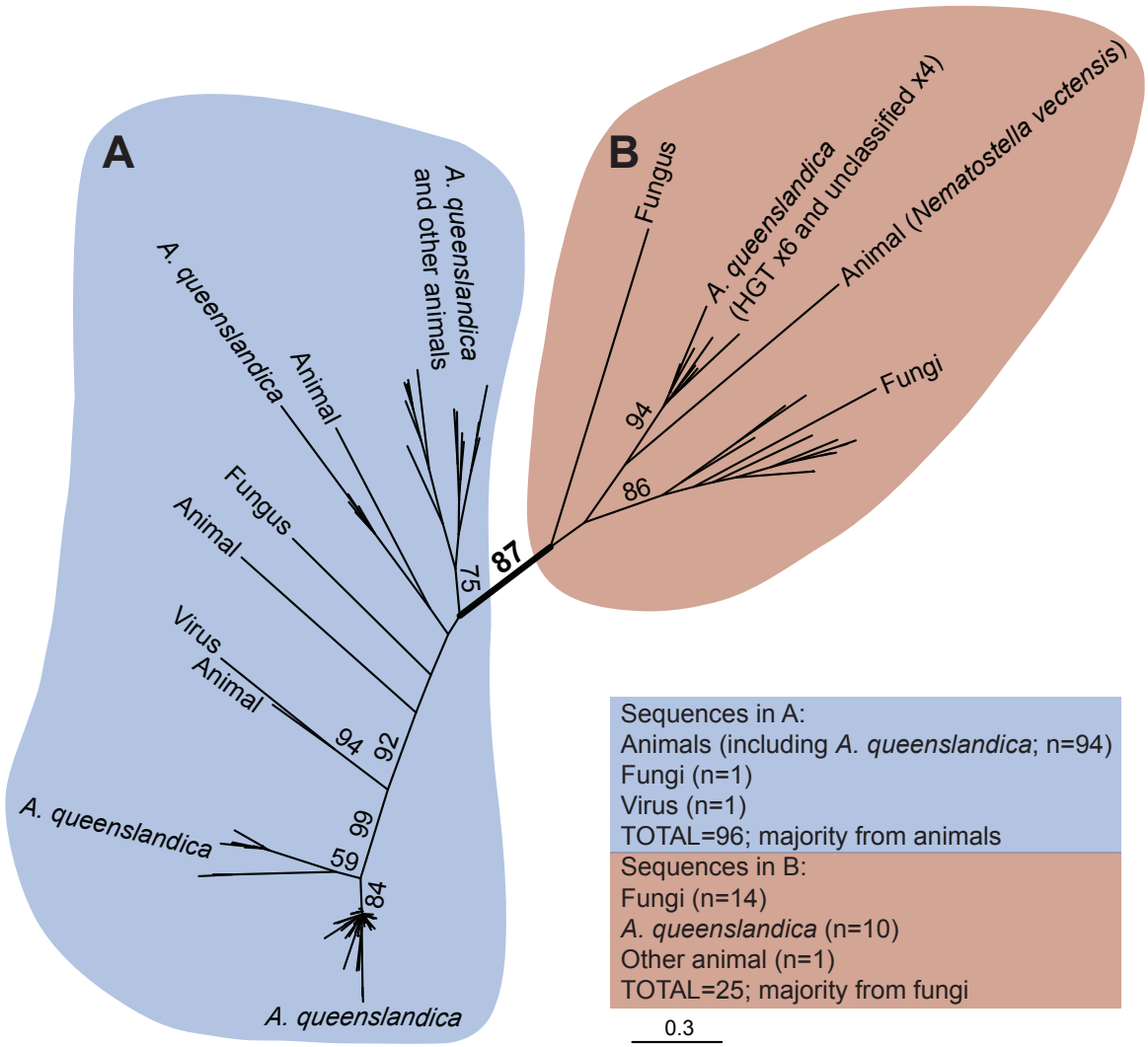


Figure 4.8 Summary of the helitron domain phylogenetic analysis
A radial phylogram illustrating the same result presented in Figure 4.7, though here minimal sequence labels are shown. A branch supported by 87% of bootstrap replicates separates mostly animal sequences (A) from mostly fungal (or fungal-like) sequences (B). In (A) are helitron domain sequences from *A. queenslandica* gene models that are predicted by HGTracker to have greatest sequence similarity to either animal genes or an ambiguous mix of taxa. In (B) are six helitron domain sequences from *A. queenslandica* gene models that are predicted by HGTracker to be likely HGTs and a further four are from gene models that are unclassified by HGTracker because of ambiguous sequence similarity results. The anomalies of the analysis are labelled and include a virus and a fungal sequence in (A) and an animal sequence in (B).

4.4.5 49 AqHGTs have high sequence similarity to proteins from bacterial MEs

Plasmids and bacteriophages are commonly involved in bacterial HGT, and many of the AqHGT-derivatives are from a bacterial source. Therefore, the AqHGT-derivatives were interrogated against a custom database of genes from plasmids and bacteriophages. Forty-nine of the 576 AqHGT-derivatives retrieved BLAST hits with e-values less than 0.001 and with at least 30% identity to at least half of both the *A. queenslandica* query and the subject plasmid or bacteriophage protein (Appendix 4.7). Forty-one

of these AqHGT-derivatives hit to proteins from bacterial plasmids, six to proteins from prophages and two to proteins from bacterial viruses. The best hits for 35 of these 49 AqHGTs are all unique; therefore, the majority of these genes are not likely the result of post-transfer duplication. The other 14 genes could result from seven originally transferred genes being duplicated, since they can be split into seven pairs that best match seven bacterial genes. The 49 AqHGT-derivatives were also searched against the NCBI nr database using the quick BLASTp algorithm and the best hit for each was compared to the best hit from the custom bacterial MEs database. The NCBI hits have higher identity and coverage values for the majority of AqHGT-derivatives (n=30), though the corresponding identity and query coverage results from both databases are within 10% of each other for some (n=14), and for five, the bacterial MEs database hit is better. Based on the information from both searches as well as the Pfam domain content, 27 of the AqHGT-derivatives are putative enzymes, including five biosynthesis enzymes and five dehydrogenases. Five of the proteins are putative transposases. Finally, these genes also include a putative beta-lactamase enzyme involved in resistance to beta-lactam antibiotics, and two putative transmembrane protein channel proteins. Of the 49 AqHGT-derivatives, 42 are not separated from a native gene by a gap in the *A. queenslandica* genome assembly and the other seven are assembled on native gene-containing scaffolds. The mean GC content (42.3) and mean intron number (3) of these 49 AqHGT-derivatives is comparable to that of both the other AqHGT-derivatives and the native *A. queenslandica* genes (Appendix 4.8). One of these AqHGT-derivatives is a putatively younger HGT. In an ontogenetic transcriptomic dataset, 38 of the 49 AqHGT-derivatives are expressed, that is, they have at least one developmental stage with at least five counts. Together, these data suggest the AqHGT-derivatives that possibly originate from bacterial MEs are well supported in their HGT-derived status and many are likely to be functioning.

4.5 DISCUSSION

Mobile elements (MEs) are diverse types of DNA able to move within and among genomes and are ubiquitous throughout all forms of life (Keane et al. 2013; Fiston-Lavier et al. 2015). Because of their ability for genomic movement, their characterised role in bacterial HGT, and because they have been reported near some HGTs in animals, MEs are hypothesised to play a role in HGT to animals from nonanimal sources (Kidwell 1993; Syvanen and Kado 2002; Gladyshev et al. 2008; Acuña et al. 2012; Paganini et al. 2012; Flot et al. 2013). Previous studies have tested this hypothesis by analysing the TE

content surrounding native genes and HGTs (Paganini et al. 2012; Flot et al. 2013). The present study extends these analyses by taking into account gene duplication as a potentially confounding factor. While this analysis did not reveal an increased density of TEs around the HGTs of *A. queenslandica* (AqHGTs) than around native genes, possible differences in the TE classes of the TE sequences were detected. Further, the large number of AqHGT-derived MEs and the high similarity of some AqHGT-derivatives to proteins from bacterial MEs support theories that MEs are mechanistically involved in some cases of interdomain HGT.

4.5.1 Similar TE densities around HGTs and native genes

Overall, there is a similar TE density surrounding the sampled HGT and native genes of *A. queenslandica.* This contrasts to increased densities around HGTs in the nematode *M. incognita* (Paganini et al. 2012) and in the rotifer *A. vaga* (Flot et al. 2013). The lack of difference found in *A. queenslandica* could reflect the predicted ancient timing of the transfer events (Conaco et al. 2016; Chapter 2) – HGT-implicated TEs may have decayed during the long time since their integration into the ancient sponge genome. The sponge lineage diverged from other animals at least 700 mya, and the contemporary demosponges *A. queenslandica* and *Petrosia ficiformis* diverged at least 450 mya (Erwin et al. 2011). If indeed many of the AqHGT-derivatives have resided in a sponge host for such a lengthy time of 450 to 700 million years, any associated TEs simply may no longer be identifiable. In a survey for bacterial ISs in 430 eukaryote genomes, Gilbert and Cordaux (2013) found only 80 such sequences within just 14 of the study species. Many of those detected ISs are under purifying selection, indicating their detectability results from either their recent arrival or molecular domestication (Gilbert and Cordaux 2013). Similarly, while eukaryotic TEs can remain active for hundreds of millions of years (Bejerano et al. 2006; Feschotte 2008; Gilbert et al. 2013) or can be preserved through domestication (Miller et al. 1999; Xie et al. 2006; Lowe et al. 2007; Feschotte 2008; Chalopin et al. 2015), it is possible that if TEs were involved in the transfers of AqHGTs, they may since have decayed.

Alternatively, the similar TE densities found around both the HGTs and natives of *A. queenslandica* could reflect that TEs were not a major mechanism underlying HGT into the sponge. The diversity of HGT cases in a broad range of animals of different body plans, lifestyles and environments (reviewed in Chapter 1) suggest that many mechanisms may be at play in HGT processes in animals. For example,

transfer events may arise from stress, phagocytosis and/or symbiosis (Westcott et al. 1976; Houck et al. 1991; Doolittle 1998; Kondo et al. 2002; Dunning Hotopp et al. 2007; Nikoh et al. 2008; Werren et al. 2010; Danchin et al. 2010; McNulty et al. 2010; Husnik et al. 2013; Sloan et al. 2014; Schönknecht et al. 2014; Jensen et al. 2016). Therefore, TEs may have played a role in HGT in *M. incognita* and *A. vaga* (Paganini et al. 2012; Flot et al. 2013), but not in *A. queenslandica*. Alternatively, the different findings in these studies could result from the confounding factor of gene duplications in the analyses of *M. incognita* and *A. vaga*. Genes near a TE are more likely to be duplicated, from the proliferation of the TE also duplicating hitchhiking genes (Morgante et al. 2005; Xiao et al. 2008; Jiang et al. 2009) or from the repetitive TE sequences conferring to nearby genes an increased likelihood of duplication by nonallelic/ectopic homologous recombination (Kaessmann 2010). Therefore, if the HGT group has higher rates of duplication than that of the native gene group, including duplicates in the analysis skews results towards increased TE proportions in the HGT group. In all three animals from each study, extensive post-transfer duplications are reported (Chapter 2; Paganini et al. 2012; Flot et al. 2013) – if the rates of duplication are higher in the HGTs than in the native genes compared, then the inclusion of duplicates in the analysis is pseudoreplication. Consequently, conclusions on HGT-TE associations cannot be made confidently without the removal of the confounding duplication factor.

4.5.2 Helitrons and simple repeats may increase the chances of genomic integration and post-transfer assimilation for HGTs

Although the overall TE density surrounding HGTs and native genes is similar, there are differences in the constituents of the TE content compared that may be signatures of the HGT process in sponges. First, two of the 21 likely unduplicated HGTs are each contiguous to a DNA transposon, one belonging to the *CMS-EnSpm* family and the other to the *sola* family. In the literature, the only other specifically reported two DNA transposons close to an animal HGT belong to the *hAT* and *Tc1* mariner families, which flank a HGT in the beetle *H. hampei* (Acuña et al. 2012). If TEs are mechanistically involved in some transfers to animals, this variety of possibly implicated DNA transposons suggests that common features of a broad range of TEs enable HGT.

Second, in regards to the TE classes surrounding the analysed *A. queenslandica* genes, in the ~5 kbp search windows, only the strictly unduplicated HGTs are associated with the DNA transposon *helitron*,

disregarding three very short hits near the strictly unduplicated natives. Note this result does not arise from the *helitron* AqHGT-derivatives, since all identified TEs were not included as sample genes in this analysis. *Helitrons* vary hugely in size because they frequently capture gene fragments (Kapitonov and Jurka 2001; Han et al. 2013). This acquisition of gene fragments, coupled with their mobility, make *helitrons* not only prime candidates for horizontal transfer, which has been found to occur repeatedly among some animals (Thomas et al. 2010), but also for acting as vectors for other genes (Thomas and Pritham 2015). The importance of the possible association of *helitrons* with AqHGTs is unclear, since it results from five *helitrons* within the search space of 91 AqHGTs and is not statistically significant after corrections for multiple testing using the Benjamini-Hochberg method with a 20% false discovery rate (testing for differences between the SUD HGT and native groups with analysis of deviance with quasibinomial errors, raw *p*-value=0.0125). However, it is tempting to speculate that the possible association could reflect a vector role of *helitrons* in a mechanism of HGT in animals. Furthermore, because they accumulate gene fragments, *helitrons* carry a range of regulatory elements and in plants and animals these have enabled transcription initiation of nearby genes by contributing regulatory elements such as the CArG motif, polyadenylation sites, promoters and TATA box sites (Miller et al. 1995; Miller et al. 1997; Miller et al. 2000; Berger et al. 2011; Thomas et al. 2014; Thomas and Pritham 2015). HGTs close to a *helitron* may thus be more likely to be conserved by selection due to functionality enabled to the HGT by *helitron* regulatory elements, thereby leading to the conservation of an association between HGTs and *helitrons*. In contrast, HGTs distant from *helitrons* may have less chances of becoming transcribed and functional, and are thus are more likely to become fossilised. In sum, *helitrons* may have been a transferring vector for some HGTs in the sponge and/or may have enabled some HGTs to become functional post-transfer.

Finally, in the ~40 kbp search windows, there are higher proportions of simple repeats (SRs) around the strictly unduplicated HGTs than around the corresponding native genes, though the significance of the difference in unclear. Greater amounts of simple repeats around the AqHGTs may signify a role of repeats in the integration of HGTs into new host genomes through recombination. Increased proportions of SRs are correlated with increased local rates of recombination in a wide spectrum of organisms, possibly because SRs may have a mechanistic and/or regulatory role in recombination initiation (Murphy and Stringer 1986; Treco and Arnheim 1986; Schultes and Szostak 1991; Gendrel

et al. 2000; Beye et al. 2006; Niehuis et al. 2010; George et al. 2015) or because non-homologous recombination may increase the abundance of SRs (Beye et al. 2006). Repetitive sequences can cause hairpin, G-quartet and triplex structures that produce a pause in the polymerase replication complex, thereby generating single-strand gaps with ends that may provoke illegitimate recombination (George et al. 2015). In the transfer of DNA from *A. tumefaciens* to plants, the bacterial single-stranded DNA is integrated into the plant genome by illegitimate recombination (Ghai and Das 1989; Gelvin 2000; Tzfira et al. 2000; Lacroix and Citovsky 2013). If recombination is one process by which foreign sequences integrate into new genomes, then HGTs are expected to accumulate in genomic areas of higher recombination rates. This theory is supported here, since I have found a possible association of HGTs and SRs in *A. queenslandica*, and SRs correlate with localised higher recombination rates in numerous animals (Jensen-Seaman et al. 2004; Beye et al. 2006; Niehuis et al. 2010). It is also further supported by HGTs in the rotifer *A. vaga,* as they are more frequently located in telomeric areas, which are also well-documented areas of higher recombination rates (Hey and Kliman 2002; Jensen-Seaman et al. 2004; Prachumwat et al. 2004; Beye et al. 2006; Gladyshev et al. 2008).

The association of SRs with the AqHGTs could alternatively be a red flag for assembly issues in these genomic regions, since DNA amplification and assembly technologies struggle with repetitive regions (Baker 2012). Nevertheless, the *A. queenslandica* genome is fully sequenced and annotated (see Chapter 1.4 for quality measures) and all the AqHGT-derivatives are assembled into scaffolds also containing native genes and in many cases, in contigs with native genes; thus the AqHGT-derivatives are well-supported as *bono fide* HGTs (or derivatives of). Furthermore, they have acquired host-like sequence characteristics including GC content and increased predicted introns (Appendix 4.8; Fernandez-Valverde et al. in preparation), many have polyadenylated transcripts, which is more common for eukaryotic mRNA than for bacterial mRNA (Régnier and Marujo 2013), and the spatial expression of some have been demonstrated through *in situ* hybridisation (K. Roper and S. Degnan, unpublished data). Separate to genome assembly aspects, it is also acknowledged that some genomic regions may be more prone to accumulating *helitrons*, SRs and HGTs (Niehuis et al. 2010; Wijayawardena et al. 2013).

4.5.3 One third of the HGT-derived genes are unknown, copia or helitron TEs

The percentage of AqHGT-derivatives with exonic sequences containing at least 10 bp of TE (43%) is less than that of both the *A. queenslandica* protein-coding genes (50%) and long noncoding RNAs (46%; Gaiti et al. 2015). Similarly, the total TE content of the AqHGT-derivatives together as a group (26%) is less than that of *A. queenslandica* protein-coding genes (31%), but greater than that of *A. queenslandica* long noncoding RNAs (22%; Gaiti et al. 2015). Although most AqHGTs are generally characterised by a low TE content, almost one third of them are enriched in TE-derived sequences and are here classified as predicted TEs (AqHGT-derived TEs).

Just over half the AqHGT-derived TEs are unclassified, approximately one quarter are *copia* LTR retrotransposons, and the remainder consists almost entirely of the DNA transposon *helitrons*. Hence, while there is a degree of diversity, the large proportion of AqHGT-derivatives that are predicted TEs could result from the very fast replicative nature of TEs creating many copies after the original transfer event(s). Indeed, phylogenetic analyses of ten HGT-derived helitron domains of *A. queenslandica* suggest they arise from one transfer event followed by subsequent duplications. None of these HGT-derived *helitrons* are expressed in the ontogenetic transcription dataset analysed in Chapter 2 (Anavy et al. 2014; Levin et al. 2016). While this lack of expression is only under one set of conditions tested in one dataset, it could imply that the duplicated *helitrons* result from continued transpositional activity, and not from positive selection following domestication. These results show that while TEs may be more prone to HGT than nonmobile sequences, and thus the same type of TE may be independently transferred multiple times, their propensity for replication may lead to increased levels of post-transfer duplications, regardless of functionality or lack thereof. The replication and genomic movement of HGT-derived TEs is another way that HGTs can have evolutionary impact on their hosts, beyond offering functional potential (Schaack et al. 2010; Wheeler 2013). TE activity is well recognised as affecting genome size and structure, causing major chromosomal rearrangements and creating or modifying gene regulatory networks (Gray 2000, Eichler and Sankoff 2003; Coghlan et al. 2005, Feschotte 2008; Herpin et al. 2010; Schaack et al. 2010; Wheeler 2013). In sum, TEs are a large proportion of the AqHGT-derivatives and have influenced the genome of *A. queenslandica* through their increased propensity for either or both HGT and post-transfer replication.

Horizontal movement of TEs between animal hosts followed by transpositional activity is well documented and demonstrates nonspecific host-TE compatibility for TE transcription and/or transposase targeting (Schaack et al. 2010; Ivancevic et al. 2013; Walsh et al. 2013). Remaining in one lineage risks host silencing and eventual TE extinction, thus successful host-transfer and subsequent transposition by TEs may increase their evolutionary longevity, resulting in selection for host jumping (Kaplan et al. 1985; Marshall 2008). Almost 60% of the AqHGT-derived TEs are predicted by HGTracker (Fernandez-Valverde et al. in preparation) to come from fungal or plant donors. Therefore, by extension, perhaps the compatibilities between varied hosts and TEs that enables the extensive host-transfers described for animals also exist to some extent at the eukaryotic level. Further, a large proportion of the transferred TEs in *A. queenslandica* are *helitrons*, which unlike other DNA transposons, do not use a transposase to replicate (Thomas and Pritham 2015). Rather, *helitrons* transpose through a rolling circle mechanism of copy-and-paste via single-stranded DNA (Kapitonov and Jurka 2001; Han et al. 2013) and thus may be more flexible in successful host transfers.

The predicted bacterial-origin of 40% of the AqHGT-derived TEs is surprisingly high since only a low number of bacterial ISs were found in a large survey of 430 eukaryotes (Gilbert and Cordaux 2013). Further, to date the only bacterial IS detected in an animal genome, that of *A. vaga,* appears not able to replicate and is inferred to be a recent arrival, "caught in the act" and not yet deteriorated (Gladyshev and Arkhipova 2009). These findings, and the lack of ISs in *A. queenslandica,* suggest that bacterial ISs are incompatible with eukaryotic host cells, possibly from transcription issues or from the compartments of the eukaryotic host cell causing issues with transposase targeting (Gilbert and Cordaux 2013). Therefore, with transposition inhibited, any transfers deteriorate unless they are first domesticated for a new function (Gilbert and Cordaux 2013). Exceptions exist, for instance Gilbert and Cordaux (2013) found a total of 80 ISs within the 430 eukaryotes and all belong to the same *IS607* family, which appears more flexible in transposition requirements. Most of the AqHGT-derivatives are ancient, therefore it seems that the persisting AqHGT-derived bacterial TEs either have a similar level of flexibility in host requirements for their transposition, and/or they have been co-opted for other roles. Gilbert and Cordaux (2013) were searching specifically for ISs, which are DNA segments that encode the enzymes necessary for their transposition, usually with one or two open reading frames and between 0.7 and 2.5 kb in length (Berg and Howe 1989; Craig et al. 2002). The bacterial-like AqHGT-derived

TEs reported here are usually shorter and may consist of only one component of a TE; therefore, they possibly result from the co-option of smaller components from within the original larger units.

4.5.4 Bacterial MEs are possible vectors for some A. queenslandica HGTs

To date, there is one well-characterised mechanism of natural interdomain DNA transfer. Bacterium *A. tumefaciens* uses the bacterial type IV secretion system (T4SS) to transfer a specific segment from its plasmid, which contains most of the genes necessary for the transfer, to the nucleus of plant cells where the sequence is integrated into the host genome by illegitimate recombination (White et al. 1982; Ghai and Das 1989; Gelvin 2000; Tzfira et al. 2000; Gelvin 2003; Lacroix and Citovsky 2013; Lacroix and Citovsky 2016). While not known to occur naturally, with physical cell contact, plasmids from bacterium *Escherichia coli* can be transferred to yeast (Heinemann and Sprague 1989), cultured human cells (Waters 2001), and diatoms (Karas et al. 2015) through a conjugation-like pathway.

Based on the high sequence similarity of 42 AqHGT-derivatives to plasmid-borne genes plus the known role of plasmids in bacterium to plant gene transfer, it appears that a similar process could facilitate gene transfer from bacteria to sponges. Sequence similarities to plasmid-borne genes have been reported for 32 HGTs in the nematodes *M. incognita* and *M. hapla* (Paganini et al. 2012); therefore, this gene transfer process appears not restricted between some bacteria and plants, but possibly occurs in sponges and other animals too. In both the natural bacterium to plant gene transfer and the experimental systems of bacterial plasmid transfer to yeast, diatoms and human cells, the T4SS is necessary for creating the molecular structure through which DNA is transported between the donor and recipient cells (Alvarez-Martinez and Christie 2009; Bhatty et al. 2013; Lacroix and Citovsky 2016). While the transport of molecules though the T4SS pilus across bacterial membranes and cells walls is characterised (Alvarez-Martinez and Christie 2009; Christie et al. 2014), it is not understood how the transported DNA would cross the recipient eukaryotic cell wall and membrane. Based on the type III secretion system, the pilus structure might penetrate through the eukaryotic cell and deliver the DNA to the cytoplasm (Galán et al. 2014; Lacroix and Citovsky 2016). Alternatively, host receptors or endocytosis might facilitate the internalisation of DNA molecules deposited on the host cell surface, as has been shown in yeast (Kawai et al. 2004; Lacroix and Citovsky 2016).

Because the bacterial T4SS is the only currently verified mechanism of DNA transfer from bacteria to eukaryotes and also because of the possible plasmid origin of 42 AqHGT-derivatives, it is compelling that *A. queenslandica* encodes a bacterial-derived gene apparently connected to the T4SS. This gene model, Aqu1.224342|Aqu2.1.36742_001, has high sequence similarity to proteins named type IV secretion protein Rhs, Rhs repeat-associated core domain-containing protein, or Rhs family protein. However, to my knowledge there is no reference of type IV secretion protein Rhs in the literature, nor are Rhs proteins described in any T4SS processes (e.g., not in Guglielmini et al. 2013; Christie et al. 2014; Darbari and Waksman 2015; Gillespie et al. 2015), bar the following two irrelevant exceptions. First, the rearrangement hotspot (Rhs) proteins (Hill et al. 1994) relevant here are different to the short motif named RGD helper sequence (RHS) that is found within the T4SS protein CagL (Conradi et al. 2012). Second, Rhs repeats are found in a variety of bacterial toxins and all but type V secretion systems transport at least one type of these toxins (Zhang et al. 2012). More relevantly, T4SS secretes the Rhs repeat-containing 16S rRNA endonuclease CdiA (Zhang et al. 2012); however, this is a different protein to the Rhs proteins that are highly similar to the *A. queenslandica* gene model in question. Rhs proteins are found in a diverse range of eukaryotes and bacteria, with functions that are not fully understood (Koskiniemi et al. 2013). However, many bacterial Rhs proteins are known toxins that are secreted by type VI secretion systems (T6SSs) and play a role in intercellular competition (Koskiniemi et al. 2013; Steele et al. 2017). Some of the most similar proteins in NCBI's nr database to Aqu1.224342|Aqu2.1.36742_001 contain SseC or Tox-HDC domains, along with Rhs domains. SseC proteins are secreted by the type III secretion system and are vital for the pathogenicity and survival of some intracellular bacteria (Bhowmick et al. 2011; Cooper et al. 2013). Tox-HDC domain-containing proteins are toxins found in pathogens (Zhang et al. 2012). These proteins are secreted via the type II secretion system and some also contain Rhs repeats (Zhang et al. 2012). In summary, it is likely that the *A. queenslandica* gene derives from a possibly toxic protein that is secreted via one of the effector translocator secretion systems, which delivers proteins (Christie et al. 2014), and does not derive from an apparatus component of a conjugation secretion system that transports DNA. Indeed, additional work is required to clarify the actual function of this intriguing gene and to identify which secretion system was relevant to its original function. However, important here, since the *A. queenslandica* gene is probably not derived from part of the T4SS apparatus, the gene offers no support for the hypothesis that a T4SS was part of the transferal mechanism of some AqHGTs.

Certainly, the transferal processes at play remain obscure and the speculation that plasmids may be mechanistically involved in animal HGTs requires further support. Nevertheless, my findings indicate that *A. queenslandica* has multiple HGTs of putative plasmid origin. Coupled with the putative plasmid-borne HGTs of *M. incognita* and *M. hapla* (Paganini et al. 2012) and the characterised plasmid system for bacterium to eukaryote gene transfers, together these results suggest that plasmids are likely involved in some animal HGTs. Indeed, the molecular mechanisms of characterised MEs for genomic movement are a mosaic continuum comprised of flexible and exchangeable modules that are probably further adaptable beyond our current understanding (Osborn and Böltner 2002; Toussaint and Merlin 2002; Zaneveld et al. 2008; Wozniak and Waldor 2010).

4.6 Conclusion

Taken together, these results offer support for the possible involvement of MEs and simple repeats in the HGT mechanisms in animals and show a high proportion of either MEs or genes from MEs within the sponge HGT-derivatives. However, these results also caution about interpretations around TE associations with HGTs in animals; while they may indicate TE-mediated transfer mechanisms, they may be signatures of alternative scenarios such as post-transfer duplication. Here, I find that such associations could also arise from post-transfer mechanisms enabling HGTs to become functional, as may be the case with *helitrons* increasing the chances of nearby HGTs becoming transcribed, functionally conserved and fixed in their new host.

Chapter 5 - General discussion

5.1 Overview

Horizontal gene transfer (HGT), the nonsexual genomic gain of exogenous genetic material (Kidwell 1993; Andersson 2005), is now not only commonly accepted as important in bacterial evolution, but as a factor of animal evolution too. However, the nature, extent and mechanisms of HGT in animals are not well understood, and these research areas continue to receive attention and debate. In this thesis I have expanded research on HGT in animals by characterising 576 predicted horizontally transferred genes (HGTs) that have previously been computationally identified in the animal *Amphimedon queenslandica*. Specifically, this work seeks to improve our understanding of three subjects: (1) the extent and nature of HGT in *A. queenslandica*; (2) the possible involvement of mobile elements (MEs) in mechanisms for HGT in *A. queenslandica*; and (3) the post-transfer evolutionary trajectory of the HGT-derived aspzincin genes in animals.

A large range of gene types has been transferred into the *A. queenslandica* genome from a broad range of taxonomic sources (Chapter 2). Many of these genes are differentially expressed throughout development and are inferred to be functional (Chapters 2 and 3). Some HGT-derivatives are enriched and I show that the two largest groups, the aspzincins and the PNP_UDP_1 phosphorylases, each result from a small number of transfer events followed by duplication (Chapter 2). The HGTs of *A. queenslandica* have some associations with MEs despite similar repeat densities surrounding predicted unduplicated HGTs and native genes (Chapter 4). Many of the HGT-derivatives are MEs themselves and regions surrounding unduplicated HGTs have slightly increased proportions of both simple repeats and the *helitron* transposable element (Chapter 4). Further, 41 of the HGT-derived genes of *A. queenslandica* are highly similar to bacterial plasmid borne genes. Currently the most understood mechanism of natural interdomain HGT is the *Agrobacterium tumefaciens* plasmid-facilitated transfer system into plants (Lacroix and Citovsky 2016). The putatively plasmid borne HGTs of *A. queenslandica*, and also those in the nematodes *Meloidogyne incognita* and *M. hapla* (Paganini et al. 2012), suggest that a plasmid-facilitated transfer system may be a mechanism of HGT to animals also.

The HGT-derived aspzincin genes here documented in *A. queenslandica,* 15 other sponge species, and 16 vertebrate species highlight the evolutionary impact that can result from a small number of HGT events (Chapters 2 and 3). Thus far, the 90-member aspzincin gene family of *A. queenslandica* represents the largest reported animal aspzincin expansion. The *A. queenslandica* aspzincins have multiple different ontogenetic expression profiles, though 54 fall into one of four developmental expression profiles, each of which is putatively co-expressed with different suites of *A. queenslandica* genes (Chapter 3). The co-expressed gene groups and sequence characteristics of the *A. queenslandica* aspzincins indicate that some have maintained the proteolytic activity that is typical of bacterial and fungal aspzincins (Chapter 3; Nonaka et al. 1998; Fushimi et al. 1999; Doi et al. 2004; Arnadottir et al. 2009). However, there is a degree of variation between the sequence characteristics and expression profiles of different *A. queenslandica* aspzincins and thus they may have been co-opted into a number of different roles.

In this final chapter, I will discuss three outstanding issues: (1) is HGT from nonanimal sources significant to animal evolution? (2) how does HGT occur in animals? and (3) what happens after HGT has occurred in animals?

5.2 IS HGT FROM NONANIMAL SOURCES SIGNIFICANT TO ANIMAL EVOLUTION?

The nature of HGT in *A. queenslandica* is varied, with a large range of gene types predicted to have been transferred from bacteria, fungi, and plants (Chapter 2). These HGTs have various post-transfer trajectories, including evolved functionality inferred from domain conservation, differential ontogenetic expression and co-expression with native *A. queenslandica* genes, as well as gene fossilisation, and gene chimeras/fusions (Chapters 2 and 3). Further, many of the apparent HGTs of *A. queenslandica* are in fact HGT-derivatives resulting from post-transfer duplications (Chapters 2 and 4). Therefore, since there are 350 different domain types predicted in the 576 originally predicted HGTs, the extent of HGT in *A. queenslandica* is extensive, but it is not as significant as originally hypothesised by HGTracker (Fernandez-Valverde et al. in preparation). Coupled with other evolutionary processes, most particularly gene duplication, HGT has expanded this animal genome by contributing large amounts of sequence including inferred functional genes.

The aspzincin case study of this thesis documents aspzincin domain-containing genes in 16 vertebrate species and not in any other eumetazoan species that have data in publically available sequence databases (Chapter 2.4.4). The vertebrate aspzincins have two different predicted domain architectures; these genes either contain a sole aspzincin domain or they contain an aspzincin domain and at least one apolipoprotein (ApoL) domain. No pattern was identified between the two domain architectures and the phylogenetic signal of the aspzincin domain since phylogenetic analyses of the vertebrate aspzincin domains reliably group together domains from genes of both domain architectures (Figure 2.10). Further, there is no pattern in the key catalytic residues in the aspzincins belonging to both domain architecture gene groups (e.g., the motifs HEASH and HEVSH are each found in aspzincins of genes of both domain architectures). Therefore, I speculate that these vertebrate aspzincins result from an old transfer along the vertebrate stem and that a unique domain combination was created post-transfer, but before the last common ancestor (LCA) of these species diverged. This predicted timing of the domain fusion event is further supported by the phylogenetic distributions of both the aspzincin domain (typically bacteria and fungi only; Chapter 2) and the ApoL domain (animal only; Vanhollebeke and Pays 2006; Limou et al. 2015), which suggest that the ApoL domain was unlikely to have arrived with the originally transferred gene. Thus vertebrate species inherited by descent both aspzincin and aspzincin-ApoL genes, both of which were differentially lost in some species, but which still persist in these few contemporary vertebrate species.

The vertebrate aspzincins are an intriguing case that provides multiple lines of support for the HGT-derived status of these genes and for further solidifying the notion of HGT in animals. Ku and Martin (2016) undertook a large-scale analysis of prokaryotic HGTs in eukaryote genomes and argue that prokaryotic sequences in animal genomes, excluding mitochondrial and chloroplast DNA, are usually the result either of genome assembly or annotation errors, or of differential gene loss. Here, multiple details presented for the aspzincin proteins found in 16 vertebrate species suggest they are not the result of contamination, assembly or annotation errors, or of differential gene loss. First, these genes have high conservation of typical aspzincin amino acid residues and are clearly aspzincin genes, not a related animal peptidase incorrectly identified. Second, they are unlikely to be sequence contamination because their phylogenetic signal is different to all other aspzincins analysed in Figure 2.10; this probably reflects vertebrate-specific divergence after an old transfer event from an ancient donor, before

these vertebrate species diverged from their LCA. Third, they are not likely to be contamination from a different, rare and/or unsampled donor, since approximately half of the vertebrate aspzincin genes from ten of the species also contain an ApoL domain that is not found outside of the animal kingdom. Fourth, these aspzincin-ApoL genes are not likely the result of genome assembly or annotation problems since they have been independently predicted in ten different species. Finally, the great phylogenetic distance between bacteria, fungi, sponges and these vertebrates suggests that these animal aspzincins are not the result of differential gene loss, but of independent horizontal transfer events in the sponge and vertebrate lineages. Under an alternative hypothesis of gene loss, the aspzincins would have been inherited through vertical descent by both the metazoan and vertebrate LCAs and only since then, would have undergone extensive gene loss, not just in the majority of animals, but in most other eukaryotes too. In time, as more genomes are sequenced, the known phylogenetic distribution of aspzincins may change and thus the support for these hypotheses may require reevaluation; however, currently the gene loss hypothesis appears to be unlikely. These five factors make the vertebrate aspzincins a strong case for HGT-derived genes in animals.

The conservation of aspzincins through millions of years in 16 vertebrate species and in at least 16 contemporary sponge species from each of the four sponge classes strongly suggests that the aspzincins are functionally important in the biology of these animals. Regardless of the continuing debate on the frequency of HGT in animals in time (e.g., Crisp et al. 2015; Ku and Martin 2016; Salzberg 2017), the animal aspzincins demonstrate that HGT can and does impact animal evolution over long periods of time.

5.3 HOW DOES HGT OCCUR IN ANIMALS?

My conclusion that HGT has significantly affected the genome of *A. queenslandica* does not imply that HGTs are continually fixed in animal genomes, since most of the bacterial HGT-derived genes of this animal likely result from ancient transfer events (Conaco et al. 2016). While HGTs may slowly accumulate in animal genomes over time, it seems more likely that occasional bursts of HGT in times of environmental stress cause the majority of animal HGTs, since some mechanisms of HGT to animals may be more likely under cellular stress (Gladyshev et al. 2008; Rybarczyk-Mydlowska et al. 2012; Flot et al. 2013; Wolf and Koonin 2013). For instance, MEs possibly play mechanistic roles in animal

HGT (Chapter 4) and since MEs are more active in times of host stress (Oliver and Greene 2009; Casacuberta and González 2013; Belyayev 2014), HGT may be more common then too. Further, if individuals carrying a HGT survive a population bottleneck, population genetics predicts that the HGT has increased chances of reaching fixation in the smaller population through inbreeding and subsequent genetic homogeneity (Lynch et al. 1995; Lovatt and Hoelzel 2013).

Partially fueling some of the contention surrounding HGT in animals are the unknown mechanisms by which foreign DNA enter cells destined for reproduction, enter the nucleus, integrate into a chromosome and become functional. Outcomes of this thesis add to already-documented biological circumstances that collectively show each of those necessary steps for HGT to animals as possible and provide clues on the putative mechanisms involved in animal HGT.

To date, there are two characterised mechanisms of natural interdomain DNA transfer systems, the most understood of which is the DNA transfer from *A. tumefaciens* to plants. *A. tumefaciens* uses the bacterial type IV secretion system (T4SS) to transfer to plant cells a specific 10-30 kbp DNA segment (T-DNA) that is flanked by short direct repeats and contained in a large tumor-inducing (Ti) plasmid (White et al. 1982; Gelvin 2003; Lacroix and Citovsky 2016). The Ti plasmid contains most of the bacterial genes required for the transfer: plant signals activate expression of the bacterial virulence genes *(vir)* and the Vir proteins facilitate the transfer of the single-stranded T-DNA molecule to the plant cell nucleus (Ghai and Das 1989; Gelvin 2000; Lacroix and Citovsky 2016). The transfer is integrated into the host genome by illegitimate recombination via host double-strand break DNA repair pathways and subsequently can be transcribed due to eukaryotic promoters in the transferred segment (Ghai and Das 1989; Gelvin 2000; Tzfira et al. 2000; Tzfira 2004; Lacroix and Citovsky 2013).

The other described mechanism of interdomain transfers to eukaryotes arises from intimate endosymbiotic relationships where transfer barriers such cell membranes and sequestered germlines are eliminated. Plastid or mitochondrion gene transfers to the nucleus are prevalent in plants and animals (Meyer-Gauen et al. 1994; Mourier et al. 2001; Hazkani-Covo et al. 2003; Richly and Leister 2004; Timmis et al. 2004; Reyes-Prieto and Bhattacharya 2007; Deschamps and Moreira 2009; Qiu et al. 2013). Such endosymbiotic gene tranfers (EGTs; transfers to the nucleus specifically from the chloroplast

or mitochondrion; Katz 2015; Ku et al. 2015; Lacroix and Citovsky 2016) are widespread - of the analysed 85 fully sequenced eukaryotic genomes (47 animals, 20 fungi, 11 protists and seven plants), 77 contain EGTs (Hazkani-Covo et al. 2010). Organelle DNA might enter the nucleus in a number of ways; to date, the most supported hypothesis involves the targeting and degradation of irregular mitochondria through mitophagy, which increases organelle DNA escape to the nucleus (Campbell and Thorsness 1998). Alternatively, direct contact between the nuclear and organelle membranes followed by membrane fusions or organelle encapsulation by the nucleus might also occur (Mota 1963; Jensen et al. 1976). Once inside the nucleus, the DNA inserts into double-stranded breaks of open chromatin regions by the nonhomologous end joining machinery (Blanchard and Schmidt 1996; Ricchetti et al. 2004; Lloyd and Timmis 2011; Wang and Timmis 2013). The inserts can be incorporated in chromosome ends with 1-7 bp of microhomology, termed microhomology-mediated repair, or with no homology through blunt-end repair (Hazkani-Covo and Covo 2008; Wang and Timmis 2013).

Related to EGTs, transfers from contemporary bacterial endosymbionts are also reported in animals, most particularly in insects where endosymbionts can reside within germ cells or even in stem germ cells (Fast et al. 2001; Kondo et al. 2002; Dunning Hotopp et al. 2007; Nikoh et al. 2008; Werren et al. 2010; Husnik et al. 2013; McFall-Ngai et al. 2013; Robinson et al. 2013; Boto 2014; Sloan et al. 2014). While the mechanisms involved in these transfers are not known, it is plausible that they may be similar to those described for EGT. HGTs involving symbiotic sequences may be more common not only because their intimate proximity to the heritable host nuclear genome means they may be more likely, but also because these transfers might be highly selected for since they could further facilitate the symbiont-host relationship (Degnan 2014; Fieth et al. 2016). Further, transfers to the nuclear genome may reduce the burden of Muller's ratchet, the asexual problem of deleterious mutations accumulating (Muller 1964) in endosymbiotic bacteria and mitochondrial genomes (Lynch 1996; Moran 1996). Thus, fixation of genes in the nuclear genome could be favoured because of greater recombination rates and lower genetic load (Martin et al. 1998).

5.3.1 Step by step mechanisms for HGT from nonanimal sources to animals

a. Accessing heritable animal cells

MEs not only have the machinery for genomic excision, integration, and movement between different cells and sometimes even hosts, but they also have evolutionary pressure for changing hosts (host-transfers). Host-transfers by transposable elements (TEs) may increase the evolutionary longevity of TEs since remaining in one lineage risks host silencing and eventual TE extinction (Kaplan et al. 1985; Marshall 2008). Thus host-transfers may be under selection (Kaplan et al. 1985; Marshall 2008). TEs can be the HGTs themselves and they can also transfer non-self-mobile hitchhiking genes (Keeling and Palmer 2008; Gilbert and Cordaux 2013; Wijayawardena et al. 2013; Thomas and Pritham 2015). Within animals, TEs can reach new hosts via an animal or viral vector (reviewed in Chapter 4.2). To date, transdomain vectors are not reported in the literature; however, some TEs are independently infectious and do not need a vector, for instance the *gypsy* element in *Drosophila* species (Song et al. 1994). These TEs have an additional open reading frame in the same position as the functional envelop-like gene *env* of retroviruses - *env* facilitates the interspecies mobility of retroviruses by recognising host surface receptors, thus allowing penetration of the membrane and infection of new cells (Malik et al. 2000; Vicient et al. 2001). Further, MEs including plasmids and bacteriophages have been identified as possible vectors for the delivery of foreign DNA to recipient animal cells, thereby facilitating interdomain HGT to animals (Chapter 4; Klasson et al. 2009; Paganini et al. 2012).

Other scenarios that can lead to the close contact of nonmetazoan DNA with metazoan cells include phagocytosis and feeding behaviours (Westcott et al. 1976; Houck et al. 1991; Doolittle 1998; Sloan et al. 2014; Jensen et al. 2016), stress (Gladyshev et al. 2008; Rybarczyk-Mydlowska et al. 2012; Flot et al. 2013), symbiosis (Kondo et al. 2002; Dunning Hotopp et al. 2007; Nikoh et al. 2008; Danchin et al. 2010; McNulty et al. 2010; Werren et al. 2010; Husnik et al. 2013; Sloan et al. 2014) and parasitism (Mower et al. 2004; Danchin et al. 2010; Schaack et al. 2010; Paganini et al. 2012; Gilbert et al. 2010). Therefore, there is a broad range of circumstances under which donor cell and animal host cell contact occurs.

To have evolutionary impact, HGTs must be heritable and thus HGT events must occur in reproductive cells (Keeling 2009; Danchin et al. 2010; Pauchet and Heckel 2013; Boto 2014; Crisp et al. 2015).

Therefore, chances of heritable HGT are expected more likely under certain conditions, such as in animals that have endosymbionts inhabiting their germ cells (Dunning Hotopp 2011), or in animals that do not have a sequestered germline and are more fluid in their germline determination (Zhaxybayeva and Doolittle 2011; Jensen et al. 2016). The latter scenario includes animals that continually recruit germ cells from pluripotent adult stem cells, such as sponges (Ereskovsky 2010; Juliano and Wessel 2010), and animals, such as cnidarians and sponges, that have a lifecycle with a vegetative stage mediated by budding or excision of pluripotent somatic cells (Boto 2014). Therefore, the germline is a hurdle of varying heights to HGT in animals, that sometimes is not even present and regardless, can be overcome.

b. Entering heritable animal cells

Once a foreign piece of DNA has made contact with the cell surface of a heritable animal cell, it must enter the cell and avoid degradation by the host. Excluding transfers from sources already inside the animal cell, such as endosymbionts and intracellular parasites, the transferring DNA must pass through the cell membrane to enter the recipient cell. The only demonstrated mechanism of bacterial sequences entering eukaryotic cells for HGT in nature and laboratory experiments is the T4SS of *A. tumefaciens* (Lacroix and Citovsky 2016). While the T4SS is well-characterised for its export of macromolecules through bacterial cells and membranes (Alvarez-Martinez and Christie 2009; Christie et al. 2014), in this transdomain transfer case, it is not understood how the transferring complex continues through the eukaryotic cell wall and membrane (Lacroix and Citovsky 2016). Because yeast can internalise foreign DNA molecules deposited on the yeast cell surface (Kawai et al. 2004), the T4SS may deposit the transferring complex on the recipient cell surface where a distinct mechanism possibly involving endocytosis and/or host receptors delivers the complex into the cytoplasm (Lacroix and Citovsky 2016). Alternatively, the T4SS pilus may continue through the recipient cell wall and membrane (Lacroix and Citovsky 2016), akin to the injection mechanism of type III secretion systems (Galán et al. 2014). While these details remain obscure, the described system demonstrates plasmids as at least partial facilitators of bacterial genetic material entering cells of plants (White et al. 1982), fungi (Heinemann and Sprague 1989), diatoms (Karas et al. 2015), and cultured human cells (Waters 2001). Further, this plasmid-facilitated mechanism is possibly relevant to the mechanisms responsible for some animal HGTs, since seven per cent of the HGT-derivatives in *A. queenslandica* have high sequence similarity

to plasmid-borne genes (Chapter 4). The same sequence similarities to plasmid-born genes have been reported for five per cent of the HGTs in the nematodes *M. incognita* and *M. hapla* (Paganini et al. 2012); therefore, this gene transfer mechanism is likely involved in HGT to animals.

Although viruses have specific and restricted host ranges that to date are not known to span wide taxonomic distances (Glansdorff et al. 2009; Bilewitch and Degnan 2011; Nilsson 2014; Mihara et al. 2016), the same virus need not infect both donor and recipient to be a HGT vector since the process could involve a ME as an intermediary vector. Through environmental contact within the extracellular space of an animal, a mobilised donor ME could move or copy genetic material to a viral genome that then infects an animal cell via membrane fusion, penetration or endocytosis (Schönknecht et al. 2014). Viral genomes are known to (1) accumulate host cellular genes (Moreira 2000; Bratke and McLysaght 2008; Moreira and Brochier-Armanet 2008; Moreira and Lopez-Garcia 2009), (2) exchange genes with other viruses (Hendrix et al. 1999; Awadalla 2003; Koonin and Dolja 2006; Moreira and Lopez-Garcia 2009), and (3) insert sequence into host genomes (Bejarano et al. 1996; Filée et al. 2003; Linial et al. 2005; Katzourakis et al. 2007; Bertsch et al. 2009; Monier et al. 2009; Geuking et al. 2009; Liu et al. 2010). Therefore, the possible three-way movement of genes between MEs, viruses and animals could be convoluted and could result in extensive gene mixing across wide taxonomic distances. Eukaryotic HGTs, including animal HGTs, are sometimes MEs, virus-like, or are found close to TEs in the host genome (Chapter 4; Jiang et al. 2004; Piskurek and Okada 2007; Gladyshev et al. 2008; Liu et al. 2010; Schaack et al. 2010; Acuña et al. 2012; Paganini et al. 2012; Flot et al. 2013; Gilbert and Cordaux 2013; Walsh et al. 2013; Pauchet Heckel 2013; Schönknecht et al. 2014). This signature could reflect a ME-virus mediated transfer mechanism.

Host membrane disruption from stress may allow foreign genetic material to enter animal cells (Schönknecht et al. 2014) and by extension, possibly germ cells too. Stress from desiccation is suspected to facilitate HGT in asexual belloid rotifers (Gladyshev et al. 2008; Flot et al. 2013), and in nematode species that can withstand desiccation by developing into Dauer larva (Gladyshev et al. 2008; Rybarczyk-Mydlowska et al. 2012).

Finally, bacteria themselves can enter animal cells in a variety of ways. Through pathogenic pathways, bacteria can alter professional phagocyte activity after being phagocytised, thereby allowing intracellular survival, or they can become internalised by and survive in non-professional phagocytes by so-called zipper or trigger mechanisms (Ribet and Cossart 2015). In addition, many animals have intracellular bacterial symbionts (Dunning Hotopp et al. 2007; Wernegreen 2012; McFall-Ngai et al. 2013). Furthermore, many bacterial symbionts are vertically transmitted between generations through reproductive cells (Krueger et al. 1996; Moran and Baumann 2000; Webster et al. 2010; Fan et al. 2012). For instance, *A. queenslandica* embryos are provisioned with nurse cells containing bacteria to ensure the inheritance of specific symbionts (Fieth et al. 2016). Additionally, animal cells can store bacteria as food, for instance the nurse cells in *A. queenslandica* also provide a bacterial food source (Fieth et al. 2016). Phagocytosis for digestion (and/or defense) creates intracellular foreign and free DNA sequences in animal cells; for example, the primary food source of *A. queenslandica* is bacteria and consequently, there is naked bacterial DNA within sponge cells (Srivastava et al. 2010; Degnan et al. 2015; Fieth et al. 2016). Furthermore, in some sponges the feeding choanocyte cells that catch and ingest bacteria can transdifferentiate into spermatocytes (Tanaka-Ichiara and Watanabe 1990; Tsurumi and Reiswig 1997). Related, *A. queenslandica* feeding choanocytes can dedifferentiate into archeocytes (Nakanishi et al. 2014; Sogabe et al. 2016), which are pluripotent cells that move throughout the body, are continuously exposed to bacteria, and give rise to oocyte germ cells (Funayama 2010). Therefore, not only are there biological circumstances in which some animals have ample whole bacteria or free and foreign DNA within cells, but because of cell fate specification, an abundant supply of foreign sequences can exist within evolutionarily relevant cells.

c. Invading the nucleus

While the outer and inner membranes of the nuclear envelop are a major barrier for nucleus invasion, such invasions occur by bacteria, bacterial molecules and organelle DNA (Campbell and Thorsness 1998; Bierne and Cossart 2012; Schulz and Horn 2015). The known mechanisms by which these entities enter the nucleus are introduced below, though they largely remain elusive (Hazkani-Covo et al. 2010; Bierne and Cossart 2012; Schulz and Horn 2015). Important here though, these cases show that nucleus invasion is possible and that it occurs via various processes. Further, these processes can

lead to HGT opportunity by organelle DNA or bacteria having direct access to host chromosomes, and they may be further adaptable for the facilitation of HGT beyond our current knowledge.

In the process of *Agrobacterium* species transforming plants, interactions of bacterial Vir proteins, which are encoded by the Ti plasmid, with host factors allow the T-DNA to use several host pathways for nuclear import (Lacroix and Citovsky 2016). Bacterial VirD2 is covalently bound to the 5' end of the T-DNA and binds to importin alpha, thereby localising the T-DNA to a nuclear pore (Howard et al. 1992; Bierne and Cossart 2012; Lacroix and Citovsky 2016). Bacterial VirE2 interactions with host VIP1 and a bacterial mimic of VIP1, VirE3, enable the T-DNA to pass through the pore using the importin-α-dependent nuclear import pathway (Citovsky et al. 2007; Magori and Citovsky 2011; Bierne and Cossart 2012). By contrast, the processes by which EGTs and transfers from contemporary bacterial endosymbionts enter the host nucleus are less characterised. Organelle DNA might enter the nucleus in a number of ways, including through mitophagy, membrane fusions, or organelle encapsulation by the nucleus (Mota 1963; Jensen et al. 1976; Campbell and Thorsness 1998; Hazkani-Covo et al. 2010).

Intranuclear bacteria can be stable over long periods of time in protozoa, dinoflagellates, free-living amoebae, and arthropods (Grandi et al. 1997; Görtz 2001; Alverca et al. 2002; Fujishima and Kodama 2012; Schulz et al. 2014), but they can also be detrimental to host cells in protozoa, euglenoids, free-living amoebae, and marine invertebrates by ultimately causing lysis (Roth 1957; Leedale 1969; Elston 1986; Azevedo 1989; Görtz 2001; Zielinski et al. 2009; Jensen et al. 2010). Intranuclear bacteria have also been observed in the marine sponge *Aplysina cavernicola,* though have not been further characterised (Friedrich et al. 1999), and while thought rare, nuclei have been invaded by bacteria in mammalian cells *in vitro* (Burgdorfer et al. 1968; Urakami et al. 1982; Pongponratn et al. 1998; Ogata et al. 2006). Currently, four mechanisms by which bacteria enter host nuclei are described in the literature. First, '*Candidatus* Nucleicultrix amoebiphila' enters the nucleus when the nuclear envelop of the amoeba host is disintegrated during open mitosis (Schulz et al. 2014). Second, *Holospora* species have a macromolecular structure that facilitates binding to and penetration of the nuclear envelope, whence the bacterium enters the nucleus leaving behind the structure (Fujishima and Kodama 2012). Third, phagocytised bacteria in *Euglena* have been shown to survive phagosome-lysosome fusion and when the membrane of the phagolysosome fuses with the outer nuclear membrane, the bacteria

are released into the perinuclear space (Shin et al. 2003). The bacteria are then inferred to enter the nucleus by invagination of the inner nuclear membrane (Shin et al. 2003). Finally, in the leafhopper *Nephotettix cincticeps,* intranuclear bacteria are inherited paternally from the nuclei of sperm cells (Watanabe et al. 2014); and while the pathway is unknown, they can also infect nuclei of silkworm and mosquito cells *in vitro* (Watanabe et al. 2014).

Such sophisticated strategies enabling bacteria to enter nuclei may be a secondary adaptation of intracellular bacteria that allows them to modify host expression and/or exploit a new niche that provides nutrients and an escape from host defense mechanism (Huang and Brumell 2014; Schulz and Horn 2015). Not all intranuclear bacteria are pathogenic; for instance, dinoflagellate *Gyrodinium instriatum* has no observable harm from intranuclear bacteria and in fact, single bacteria are released from the nucleus into the cytoplasm, perhaps as food for starving host cells (Alverca et al. 2002). Further, in protozoa, intranuclear bacteria can enhance host survival under environmental stresses including changes in temperatures and salinity, by modifying host expression of heat shock proteins (Hori and Fujishima 2003; Fujishima et al. 2005; Hori and Fujishima 2008). Because stress has been implicated in possible HGT mechanisms in some animals (Gladyshev et al. 2008; Rybarczyk-Mydlowska et al. 2012; Flot et al. 2013), this proposed selective advantage of intranuclear bacteria in times of stress may also facilitate increased HGT opportunity. Increased HGT may be often detrimental, but akin to increased TE activity in times of stress (Oliver and Greene 2012; Casacuberta and González 2013), it may sometimes offer beneficial variation. In sum, while many details remain unknown, living bacteria exist within eukaryotic nuclei and offer potential for HGT in both directions.

Some animal and plant bacterial pathogens produce molecules named nucleomodulins that enter host nuclei and directly affect host RNA splicing, transcription, DNA replication and repair, and chromatin-remodeling (Lebreton et al. 2011; Rolando et al. 2013). The entry of nucleomodulins into eukaryotic nuclei is not fully understood. Some nucleomodulins possess classical nuclear localisation signals (NLSs) that mediate their entry to the nucleus (Boch et al. 2009; Weinthal et al. 2011), though many lack a classical NLS and must have either non-classical NLSs or might be transferred by host nuclear proteins (Bierne and Cossart 2012). While nucleomodulins are a class of effectors and are not naked DNA and possible HGTs, nucleomodulins demonstrate that the frontiers of knowledge on nuclear

invasion are likely to shift and that our current knowledge, or lack thereof, is not a valid argument against the feasibility of HGT to animals.

d. Genomic integration

To be passed on during reproduction, the foreign DNA must integrate into the host genome. In both of the only two mechanistically understood circumstances of natural interdomain HGT, that of *A. tumefaciens* and of EGTs, the transferring sequence is integrated into the host genome by recombination (Ghai and Das 1989; Blanchard and Schmidt 1996; Gelvin 2000; Tzfira et al. 2000; Hazkani-Covo and Covo 2008; Lacroix and Citovsky 2013). Important for the heritability of animal HGTs, in eukaryotes, homologous and illegitimate recombination occurs for DNA repair in any cell, including germ cells (Helle 2012), and homologous recombination occurs during meiosis, the process creating haploid reproductive gametes (Gerton and Hawley 2005; Lenormand et al. 2016). While not known to occur naturally, plasmids can be transferred from bacteria to various yeast species and depending if the plasmid has sequence homology (0.8-1.4 kb region of homology) with the yeast genome, or not, the plasmid recombines through illegitimate or homologous recombination (Bundock et al. 1999). Therefore, the reported possible association of TEs and animal HGTs (Gladyshev et al. 2008; Klasson et al. 2009; Acuña et al. 2012; Paganini et al. 2012; Flot et al. 2013) could result from the numerous duplicate TE copies providing increased areas of homology for integrating foreign sequences that have the same TE modules, thereby increasing the chances of homologous recombination (Zaneveld et al. 2008; Shapiro et al. 2016). The proposed TE-HGT association may also reflect TE transposition as a mechanism of genomic integration for HGTs, where TEs could be the HGT themselves, as shown in Chapter 4.4.4, and/or could also integrate hitchhiking HGTs into the genome, as discussed in Chapter 4.2.

The illegitimate recombination that occurs in transfers from *A. tumefaciens* to plants and in EGTs requires a host double-stranded break (DSB) before nonhomologous end joining pathways can repair the break, and in the process, incorporate the HGT into the host genome (Tzfira 2004). Because simple repeats are correlated with increased local rates of recombination (Beye et al. 2006), the possible association of HGTs and SRs in *A. queenslandica* (Chapter 4.4.3) supports a role for illegitimate recombination via DSB repair pathways in the genomic integration of HGTs in this animal. Repetitive sequences can cause hairpin, G-quartet and triplex structures that produce a pause in the polymerase replication

complex, thereby generating recombination breakpoints with ends that may provoke illegitimate recombination (George et al. 2015).

e. Becoming transcribed, regulated, and functional

After genomic integration, successful HGTs must not be lost from genome rearrangements occurring in cell divisions and must become fixed in populations. Further to the roles of time and chance, both the gene type and its expression, or lack thereof, impacts HGT fixation by affecting the strength of positive or negative selection (reviewed in Chapter 2). Interdomain HGTs may be expressed and regulated if regulatory elements are also transferred, as in the *A. tumefaciens* to plants transfer system (Ghai and Das 1989; Gelvin 2000; Tzfira et al. 2000; Tzfira 2004; Lacroix and Citovsky 2013). Alternatively, host regulatory elements may facilitate the evolution of HGT expression and functionality, and TEs may increase the chances of this facilitation (Chapter 4). In Chapter 4, a possible relationship was detected between the rolling-circle *helitron* TE and the HGT-derivatives of *A. queenslandica*, many of which are transcribed (Conaco et al. 2016; Chapter 2). *Helitrons* are a reservoir of regulatory elements, which have been shown responsible for altering the transcription of nearby genes in plants and animals, including enabling transcription initiation (Miller et al. 1995; Miller et al. 1997; Miller et al. 2000; Berger et al. 2011; Thomas et al. 2014; Thomas and Pritham 2015). Therefore, bacterial genes may become functional in animals post-transfer due to host regulatory elements, of which active TEs such as *helitrons* may offer an increased supply through their genomic movement and insertions (Chapter 4).

5.4 WHAT HAPPENS AFTER HGT HAS OCCURRED IN ANIMALS?

The aspzincin catalytic domain is predicted to have been transferred independently into both the sponge and vertebrate lineages and subsequently has been conserved in some species, which indicates functional constraint. Further, putatively novel domain architectures have evolved from the aspzincin domain separately combining with the hemopexin and ApoL domains, possibly enabling novel roles for these proteins (Chapters 2 and 3). The evolution of functional diversity in *A. queenslandica* aspzincins is also inferred from their separation into groups that have different ontogenetic expression profiles, with each aspzincin group correlating in expression with different suites of other *A. queenslandica* genes (Chapter 3). Also, they have variation in sequence characteristics like secretion signals and in the conservation of key catalytic residues (Chapter 3). While aspzincin gene loss is evident in both

sponges and vertebrates (Chapters 2 and 3), the aspzincin catalytic domain appears to be functionally significant to some animals. Speculations on putative aspzincin functions in *A. queenslandica* include roles in nutritional immunity, in direct defense as toxins, in antigenic immune defense as toxoids against other toxic aspzincins, and in ECM remodeling (Chapter 3; Conaco et al. 2016). However, the actual roles of these HGT-derived genes remain to be properly explored. K. Roper and S. Degnan (personal communication) show with *in situ* hybridisation that at least two aspzincin genes are likely to be involved in spiculogenesis (Chapter 3). The large expansion of the aspzincin genes in *A. queenslandica* indicates that they play important functions in this animal and demonstrates the significance that HGT can have on animal evolution.

5.4.1 A putative role of vertebrate aspzincin-ApoL proteins in apoptosis and/or immunity

The putative role(s) of the vertebrate aspzincin-ApoL proteins are unknown, though based on the properties of both domains, they could have an apoptosis and/or immunity role. The ApoL domain is a member of the lipoprotein family and while the various functions of ApoL domain-containing proteins are not fully clear, some are involved in innate immunity (Limou et al. 2015), in apoptosis (Smith and Malik 2009; Uzureau et al. 2016), and in autophagy (Zhaorigetu et al. 2008; Hu et al. 2012). Human ApoL-I binds to parasitic trypanosomes and generates ionic pores, which cause osmotic swelling that kills the parasite (Vanhamme et al. 2003; Vanhollebeke et al. 2006; Namangala 2011; Kuriakose et al. 2016). ApoL-I also inhibits bacterial and yeast growth (Pérez-Morga et al. 2005). ApoL domain-containing proteins are implicated in cancer (Chidiac et al. 2016), many diseases (Genovese et al. 2010; Parsa et al. 2013; Freedman et al. 2015; Hu and Ray 2016), and HIV suppression (Taylor et al. 2014). There are 11 domain architectures reported in the Pfam database involving the ApoL domain, with by far the most common being just one ApoL domain (v30.0, accessed January 2017; http://pfam.xfam.org/family/PF05461#tabview=tab1). In accordance with ApoL role(s) in apoptosis, three ApoL domain-containing sequences within the Pfam database also contain a Peptidase C14 domain (accession PF00656), which belongs to the caspase family of cysteine peptidases (Nicholson 1999; Salvesen and Abrams 2004). These peptidases have crucial roles in programmed cell death, including apoptosis, where the activated caspases cleave targets and also activate other proteases (Nicholson 1999; Vanhollebeke and Pays 2006). Caspases may act independently or together with members of the Bcl-2 proteins, which are another family essential for programmed cell death (Nicholson 1999;

Salvesen and Abrams 2004; Vanhollebeke and Pays 2006; Limou et al. 2015). ApoL proteins are likened to Bcl-2 proteins because of structural and functional similarities (Vanhollebeke and Pays 2006; Limou et al. 2015). Like caspases, aspzincin are peptidases and are lethal toxins in some systems (Fushimi et al. 1999; Arnadottir et al. 2009; Schwenteit et al. 2013a). The structural and functional similarities between aspzincins and caspases, and between ApoLs and Bcl-2s, coupled with the links between them (Figure 5.1), suggest that the cleaving properties of the aspzincin domains might be harnessed by the ApoL proteins for an apoptosis and/or immunity role. Whether the vertebrate aspzincins without an ApoL domain have a related role remains another interesting question.

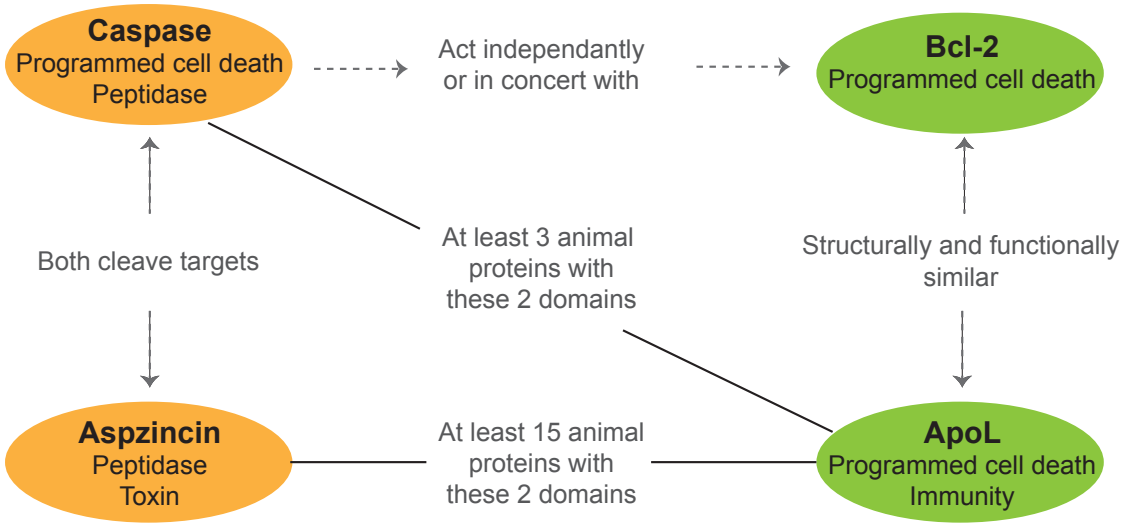


Figure 5.1 Schematic of functional domain components involved with the apolipoprotein L domain
ApoL domain-containing proteins usually contain one ApoL domain only; some of these proteins have programmed cell death and immunity roles, and are similar in structure and function to Bcl-2 proteins, which also play a crucial role in programmed cell death. Bcl-2 proteins sometimes work in concert with caspases, which are peptidases that are also involved in programmed cell death. A few ApoL domain-containing proteins also contain a caspase domain, and since both these components (caspases and ApoLs) are independently involved in programmed cell death, the unique vertebrate aspzincin-ApoL proteins also may be involved in programmed cell death and possibly immunity. The peptidase aspzincin domain may play a convergent function in the aspzincin-ApoL proteins to the peptidase caspase domain in the caspase-ApoL proteins.

5.5 CONCLUSIONS AND LOOKING FORWARD

This study characterises HGT-derived genes and their ME genomic signatures in the animal *A. queenslandica*. Also presented here is the first reported analysis of animal aspzincin genes, which further reveals some intricacies and impact of HGT in animals. This work contributes to the fast-growing but still contentious research field of animal HGT by demonstrating that the extent of HGT

in animals can be significant and the nature varied, since the genome of *A. queenslandica* has been expanded in size and sequence diversity by HGTs arising from bacteria, fungi and plants and of diverse putative functions. This suggestion that HGT is not an insignificant rarity of animal evolution is further supported by my detection of independent horizontal aspzincin transfers into the stems of the sponge and vertebrate lineages, with both cases well-supported as the result of ancient HGT and not other scenarios such as differential gene loss, contamination, or assembly issues. I speculate that bacterial plasmids, TEs, simple repeats and *helitrons* were possibly involved in some of the processes necessary for HGT in *A. queenslandica* through roles as transfer vectors, in genomic integration and in the post-transfer evolution of HGT functionality. Finally, I show that the post-transfer evolutionary trajectories of HGTs are varied; the aspzincin case study alone shows HGT followed by differential gene loss, duplication, conservation, domain fusions, and inferred functionality.

In agreement with the growing awareness of the limitations of large-scale automated analyses (Bemm et al. 2016; Richards and Monier 2016; Salzberg 2017; Sieber et al. 2017), this work has highlighted the necessity for distinguishing between HGTs and HGT-derived genes to avoid overestimation of HGT in animals, a research topic that continues to be debated. Overall, I conclude that the verification and expansion of several findings and inferences presented in this thesis will provide further valuable insights. Specifically, I contend that the three most revealing future lines of inquiry arising from this study are as follows. First, further characterisation of the reported plasmid, simple repeats and *helitron* findings is needed to more precisely pinpoint their roles and significance in the different transfer processes. Second, a wealth of information and other fascinating cases remain hidden within a large group of *A. queenslandica* genes currently labeled as "unclassified" by HGTracker (Fernandez-Valverde *et al.* in preparation) – these await and require manual inspection. Last and most intriguingly, I propose that the *A. queenslandica* aspzincins have varied functions, but much more experimental work is required to elucidate those roles. Characterising the spatial expression and the proteolytic activity of the sponge aspzincins are exciting future research avenues. Here, the genome of *A. queenslandica* is shown to be versatile in the acquisition and use of genetic variation arising from horizontal gene transfer.

Reference List

Acuña, R., Padilla, B.E., Florez-Ramos, C.P., Rubio, J.D., Herrera, J.C., Benavides, P., Lee, S.J., Yeats, T.H., Egan, A.N., Doyle, J.J. and Rose, J.K.C., 2012. Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee. *Proceedings of the National Academy of Sciences*, 109(11), pp.4197-4202.

Adams, V., Lyras, D., Farrow, K.A. and Rood, J.I., 2002. The clostridial mobilisable transposons. *Cellular and Molecular Life Sciences*, 59(12), pp.2033–2043.

Aikawa, T., Anbutsu, H., Nikoh, N., Kikuchi, T., Shibata, F. and Fukatsu, T., 2009. Longicorn beetle that vectors pinewood nematode carries many *Wolbachia* genes on an autosome. *Proceedings of the Royal Society B: Biological Sciences*, 276(1674), pp.3791–3798.

Akiba, T., Koyama, K., Ishiki, Y., Kimura, S. and Fukushima, T., 1960. On the mechanism of the development of multiple-drug-resistant clones of *Shigella*. *Japanese Journal of Microbiology*, 4(2), pp.219–227.

Alegado, R.A. and King, N., 2014. Bacterial influences on animal origins. *Cold Spring Harbor Perspectives in Biology*, 6(11), a016162.

Allen, M.A., Lauro, F.M., Williams, T.J., Burg, D., Siddiqui, K.S., De Francisci, D., Chong, K.W.Y., Pilak, O., Chew, H.H., De Maere, M.Z., Ting, L., Katrib, M., Ng, C., Sowers, K.R., Galperin, M.Y., Anderson, I.J., Ivanova, N., Dalin, E., Martinez, M., Lapidus, A., Hauser, L., Land, M., Thomas, T. and Cavicchioli, R., 2009. The genome sequence of the psychrophilic archaeon, *Methanococcoides burtonii*: the role of genome evolution in cold adaptation. *The ISME Journal*, 3(9), pp.1012–1035.

Alvarez-Martinez, C.E. and Christie, P.J., 2009. Biological diversity of prokaryotic type IV secretion systems. *Microbiology and Molecular Biology Reviews*, 73(4), pp.775–808.

Alverca, E., Biegala, I., Kennaway, G., Lewis, J. and Franca, S., 2002. *In situ* identification and localization of bacteria associated with *Gyrodinium instriatum* (Gymnodiniales, Dinophyceae) by electron and confocal microscopy. *European Journal of Phycology*, 37(4), pp.523–530.

Amin, S.R., Erdin, S., Ward, R.M., Lua, R.C. and Lichtarge, O., 2013. Prediction and experimental validation of enzyme substrate specificity in protein structures. *Proceedings of the National Academy of Sciences*, 110(45), pp.E4195–4202.

Anavy, L., Levin, M., Khair, S., Nakanishi, N., Fernandez-Valverde, S.L., Degnan, B.M. and Yanai, I., 2014. BLIND ordering of large-scale transcriptomic developmental timecourses. *Development*, 141(5), pp.1161–1166.

Anderson, M.T. and Seifert, H.S., 2011. Opportunity and means: horizontal gene transfer from the human host to a bacterial pathogen. *mBio*, 2(1), e5–11.

Andersson, J.O., 2005. Lateral gene transfer in eukaryotes. *Cellular and Molecular Life Sciences*, 62(11), pp.1182–1197.

Andersson, D.I., Jerlström-Hultqvist, J. and Näsvall, J., 2015. Evolution of new functions *de novo* and from preexisting genes. *Cold Spring Harbor Perspectives in Biology*, 7(6), a017996.

Aravind, L., Iyer, L.M., Leipe, D.D. and Koonin, E.V., 2004. A novel family of P-loop NTPases with an unusual phyletic distribution and transmembrane segments inserted within the NTPase domain. *Genome Biology*, 5(5), R30.

Arnadottir, H., Hvanndal, I., Andresdottir, V., Burr, S.E., Frey, J. and Gudmundsdottir, B.K., 2009. The AsaP1 peptidase of *Aeromonas salmonicida* subsp. *achromogenes* is a highly conserved deuterolysin metalloprotease (family M35) and a major virulence factor. *Journal of Bacteriology*, 191(1), pp.403–410.

Avery, O.T., Macleod, C.M. and McCarty, M., 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus Type III. *The Journal of Experimental Medicine*, 79(2), pp.137–158.

Awadalla, P., 2003. The evolutionary genomics of pathogen recombination. *Nature Reviews Genetics*, 4(1), pp.50–60.

Azad, R.K. and Lawrence, J.G., 2011. Towards more robust methods of alien gene detection. *Nucleic Acids Research*, 39(9), pp.e56–e56.

Azevedo, C., 1989. Fine structure of endonucleobiotic bacteria in the gill epithelium of *Ruditapes decussatus*. *Marine Biology*, 100, pp.339–341.

Baker, M., 2012. De novo genome assembly: what every biologist should know. *Nature Methods*, 9(4), pp.333–337.

Baldauf, S.L., 2003. The deep roots of eukaryotes. *Science*, 300(5626), pp.1703–1706.

Baltrus, D.A., 2013. Exploring the costs of horizontal gene transfer. *Trends in Ecology & Evolution*, 28(8), pp.489–495.

Baños, R.C., Vivero, A., Aznar, S., Garcia, J., Pons, M., Madrid, C. and Juárez, A., 2009. Differential regulation of horizontally acquired and core genome genes by the bacterial modulator H-NS. *PLoS Genetics*, 5(6), e1000513.

Barrett, R. and Schluter, D., 2008. Adaptation from standing genetic variation. *Trends in Ecology & Evolution*, 23(1), pp.38–44.

Becq, J., Churlaud, C. and Deschavanne, P., 2010. A benchmark of parametric methods for horizontal transfers detection. *PLoS ONE*, 5(4), e9989.

Beiko, R.G., Harlow, T.J. and Ragan, M.A., 2005. Highways of gene sharing in prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, 102(40), pp.14332–14337.

Bejarano, E.R., Khashoggi, A., Witty, M. and Lichtenstein, C., 1996. Integration of multiple repeats of geminiviral DNA into the nuclear genome of tobacco during evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 93(2), pp.759–764.

Bejerano, G., Lowe, C.B., Ahituv, N., King, B., Siepel, A., Salama, S.R., Rubin, E.M., James Kent, W. and Haussler, D., 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature*, 441(7089), pp.87–90.

Belyayev, A., 2014. Bursts of transposable elements as an evolutionary driving force. *Journal of Evolutionary Biology*, 27(12), pp.2573–2584.

Bemm, F., Weiß, C.L., Schultz, J. and Förster, F., 2016. Genome of a tardigrade: horizontal gene transfer or bacterial contamination? *Proceedings of the National Academy of Sciences*, 113(22), pp.E3054–E3056.

Bendtsen, J.D., Jensen, L.J., Blom, N., Heijne, von, G. and Brunak, S., 2004. Feature-based prediction of non-classical and leaderless protein secretion. *Protein Engineering Design and Selection*, 17(4), pp.349–356.

Bendtsen, J.D., Kiemer, L., Fausbøll, A. and Brunak, S., 2005. Non-classical protein secretion in bacteria. *BMC Microbiology*, 5(58).

Benveniste, RE, Todaro, G.J., 1974. Evolution of C-Type viral genes - inheritance of exogenously acquired genes. *Nature*, 252(5483), pp.456-459.

Berg D.E. and Howe M.M., 1989. Mobile DNA. *American Society for Microbiology*, Washington D.C.

Berger, N., Dubreucq, B., Roudier, F., Dubos, C. and Lepiniec, L., 2011. Transcriptional regulation of *Arabidopsis* LEAFY COTYLEDON2 involves RLE, a cis-element that regulates trimethylation of histone H3 at lysine-27. *The Plant Cell*, 23(11), pp.4065–4078.

Bertsch, C., Beuve, M., Dolja, V.V., Wirth, M., Pelsy, F., Herrbach, E. and Lemaire, O., 2009. Retention of the virus-derived sequences in the nuclear genome of grapevine as a potential pathway to virus resistance. *Biology Direct*, 4(1).

Beye, M., Gattermeier, I., Hasselmann, M., Gempe, T., Schioett, M., Baines, J.F., Schlipalius, D., Mougel, F., Emore, C., Rueppell, O., Sirvio, A., Guzman-Novoa, E., Hunt, G., Solignac, M. and Page, R.E., 2006. Exceptionally high levels of recombination across the honey bee genome. *Genome Research*, 16(11), pp.1339–1344.

Bhattacharya, D., Pelletreau, K.N., Price, D.C., Sarver, K.E. and Rumpho, M.E., 2013. Genome analysis of *Elysia chlorotica* egg DNA provides no evidence for horizontal gene transfer into the germ line of this kleptoplastic mollusc. *Molecular Biology and Evolution*, 30(8), pp.1843–1852.

Bhattacharya, D., Agrawal, S., Aranda, M., Baumgarten, S., Belcaid, M., Drake, J.L., Erwin, D., Foret, S., Gates, R.D., Gruber, D.F., Kamel, B., Lesser, M.P., Levy, O., Liew, Y.J., MacManes, M., Mass, T., Medina, M., Mehr, S., Meyer, E., Price, D.C., Putnam, H.M., Qiu, H., Shinzato, C., Shoguchi, E., Stokes, A.J., Tambutté, S., Tchernov, D., Voolstra, C.R., Wagner, N., Walker, C.W., Weber, A.P., Weis, V., Zelzion, E., Zoccola, D. and Falkowski, P.G., 2016. Comparative genomics explains the evolutionary success of reef-forming corals. *eLife*, 5(5741).

Bhatty, M., Gomez, J.A.L. and Christie, P.J., 2013. The expanding bacterial type IV secretion lexicon. *Research in Microbiology*, 164(6), pp.620–639.

Bhowmick, P.P., Devegowda, D., Ruwandeepika, H.A.D., Karunasagar, I. and Karunasagar, I., 2011. Presence of *Salmonella* pathogenicity island 2 genes in seafood-associated *Salmonella serovars* and the role of the sseC gene in survival of *Salmonella enterica* serovar Weltevreden in epithelial cells. *Microbiology*, 157(1), pp.160–168.

Bierne, H., Hamon, M. and Cossart, P., 2012. Epigenetics and bacterial infections. *Cold Spring Harbor Perspectives in Medicine*, 2(a010272).

Bilewitch, J.P. and Degnan, S.M., 2011. A unique horizontal gene transfer event has provided the octocoral mitochondrial genome with an active mismatch repair gene that has potential for an unusual self-contained function. *BMC Evolutionary Biology*, 11(228).

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Stamatoyannopoulos, J.A., Thurman, R.E., Kuehn, M.S., Taylor, C.M., Neph, S., Koch, C.M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J.A., Andrews, R.M., Flicek, P., Boyle, P.J., Cao, H., Carter, N.P., Clelland, G.K., Davis, S., Day, N., Dhami, P., Dillon, S.C., Dorschner, M.O., Fiegler, H., Giresi, P.G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K.D., Johnson, B.E., Johnson, E.M., Frum, T.T., Rosenzweig, E.R., Karnani, N., Lee, K., Lefebvre, G.C., Navas, P.A., Neri, F., Parker, S.C.J., Sabo, P.J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu1, M., Collins, F.S., Dekker, J., Lieb, J.D., Tullius, T.D., Crawford, G.E., Sunyaev, S., Noble, W.S., Dunham, I., Dutta, A., Guigó, R., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I.L., Baertsch, R., Keefe, D., Flicek, P., Dike, S., Cheng, J., Hirsch, H.A., Sekinger, E.A., Lagarde, J., Abril, J.F., Shahab, A., Flamm, C., Fried, C., Hackermüller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J.S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M.C., Thomas, D.J., Weirauch, M.T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K.G., Sung, W.-K., Ooi, H.S., Chiu, K.P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M.L., Valencia, A., Choo, S.W., Choo, C.Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T.G., Brown, J.B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C.N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X., Bickel, P., Mattick, J.S., Carninci, P., Hayashizaki, Y., Weissman, S., Dermitzakis, E.T., Margulies, E.H., Hubbard, T.,

Myers, R.M., Rogers, J., Stadler, P.F., Lowe, T.M., Wei, C.-L., Ruan, Y., Snyder, M., Birney, E., Struhl, K., Gerstein, M., Antonarakis, S.E., Gingeras, T.R., Brown, J.B., Flicek, P., Fu, Y., Keefe, D., Birney, E., Denoeud, F., Gerstein, M., Green, E.D., Kapranov, P., Karaöz, U., Myers, R.M., Noble, W.S., Reymond, A., Rozowsky, J., Struhl, K., Siepel, A., Stamatoyannopoulos, J.A., Taylor, C.M., Taylor, J., Thurman, R.E., Tullius, T.D., Washietl, S., Zheng, D., Liefer, L.A., Wetterstrand, K.A., Good, P.J., Feingold, E.A., Guyer, M.S., Collins, F.S., Margulies, E.H., Cooper, G.M., Asimenos, G., Thomas, D.J., Dewey, C.N., Siepel, A., Birney, E., Keefe, D., Hou, M., Taylor, J., Nikolaev, S., Montoya-Burgos, J.I., Löytynoja, A., Whelan, S., Pardi, F., Massingham, T., Brown, J.B., Huang, H., Zhang, N.R., Bickel, P., Holmes, I., Mullikin, J.C., Ureta-Vidal, A., Paten, B., Seringhaus, M., Church, D., Rosenbloom, K., Kent, W.J., Stone, E.A., Gerstein, M., Antonarakis, S.E., Batzoglou, S., Goldman, N., Hardison, R.C., Haussler, D., Miller, W., Pachter, L., Green, E.D., Sidow, A., Weng, Z., Trinklein, N.D., Fu, Y., Zhang, Z.D., Karaöz, U., Barrera, L., Stuart, R., Zheng, D., Ghosh, S., Flicek, P., King, D.C., Taylor, J., Ameur, A., Enroth, S., Bieda, M.C., Koch, C.M., Hirsch, H.A., Wei, C.-L., Cheng, J., Kim, J., Bhinge, A.A., Giresi, P.G., Jiang, N., Liu, J., Yao, F., Sung, W.-K., Chiu, K.P., Vega, V.B., Lee, C.W.H., Ng, P., Shahab, A., Sekinger, E.A., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M.J., Inman, D., Singer, M.A., Richmond, T.A., Munn, K.J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Clelland, G.K., Wilcox, S., Dillon, S.C., Andrews, R.M., Fowler, J.C., Couttet, P., James, K.D., Lefebvre, G.C., Bruce, A.W., Dovey, O.M., Ellis, P.D., Dhami, P., Langford, C.F., Carter, N.P., Vetrie, D., Kapranov, P., Nix, D.A., Bell, I., Patel, S., Rozowsky, J., Euskirchen, G., Hartman, S., Lian, J., Wu, J., Urban, A.E., Kraus, P., Van Calcar, S., Heintzman, N., Hoon Kim, T., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C.K., Rosenfeld, M.G., Aldred, S.F., Cooper, S.J., Halees, A., Lin, J.M., Shulha, H.P., Zhang, X., Xu, M., Haidar, J.N.S., Yu, Y., Birney, E., Weissman, S., Ruan, Y., Lieb, J.D., Iyer, V.R., Green, R.D., Gingeras, T.R., Wadelius, C., Dunham, I., Struhl, K., Hardison, R.C., Gerstein, M., Farnham, P.J., Myers, R.M., Ren, B., Snyder, M., Thomas, D.J., Rosenbloom, K., Harte, R.A., Hinrichs, A.S., Trumbower, H., Clawson, H., Hillman-Jackson, J., Zweig, A.S., Smith, K., Thakkapallayil, A., Barber, G., Kuhn, R.M., Karolchik, D., Haussler, D., Kent, W.J., Dermitzakis, E.T., Armengol, L., Bird, C.P., Clark, T.G., Cooper, G.M., de Bakker, P.I.W., Kern, A.D., Lopez-Bigas, N., Martin, J.D., Stranger, B.E., Thomas, D.J., Woodroffe, A., Batzoglou, S., Davydov, E., Dimas, A., Eyras, E., Hallgrímsdóttir, I.B., Hardison, R.C., Huppert, J., Sidow, A., Taylor, J., Trumbower, H., Zody, M.C., Guigó, R., Mullikin, J.C., Abecasis, G.R., Estivill, X., Birney, E., Bouffard, G.G., Guan, X., Hansen, N.F., Idol, J.R., Maduro, V.V.B., Maskeri, B., McDowell, J.C., Park, M., Thomas, P.J., Young, A.C., Blakesley, R.W., Muzny, D.M., Sodergren, E., Wheeler, D.A., Worley, K.C., Jiang, H., Weinstock, G.M., Gibbs, R.A., Graves, T., Fulton, R., Mardis, E.R., Wilson, R.K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D.B., Chang, J.L., Lindblad-Toh, K., Lander, E.S., Koriabine, M., Nefedov, M., Osoegawa, K., Yoshinaga, Y., Zhu, B. and de Jong, P.J., 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146), pp.799–816.

Bishop, J.D.D. and Pemberton, A.J., 2006. The third way: spermcast mating in sessile marine invertebrates. *Integrative and Comparative Biology*, 46(4), pp.398–406.

Blanchard, J.L. and Schmidt, G.W., 1996. Mitochondrial DNA migration events in yeast and humans: integration by a common end-joining mechanism and alternative perspectives on nucleotide substitution patterns. *Molecular Biology and Evolution*, 13(6), pp.893–548.

Blencowe, B.J., 2006. Alternative splicing: new insights from global analyses. *Cell*, 126(1), pp.37–47.

Blumbach, B., Pancer, Z., Diehl-Seifert, B., Steffen, R., Münkner, J., Müller, I. and Müller, W.E., 1998. The putative sponge aggregation receptor. Isolation and characterization of a molecule composed of scavenger receptor cysteine-rich domains and short consensus repeats. *Journal of Cell Science*, 111, pp.2635–2644.

Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., Nickstadt, A. and Bonas, U., 2009. Breaking the code of DNA binding specificity of TAL-type III effectors. *Science*, 326(5959), pp.1509–1512.

Bogdanović, X., Palm, G.J., Schwenteit, J., Singh, R.K., Gudmundsdottir, B.K. and Hinrichs, W., 2016. Structural evidence of intramolecular propeptide inhibition of the aspzincin metalloendopeptidase AsaP1. *FEBS letters*, 590(18), pp.3280–3294.

Bong, J.S., and Beynon, R.J., 1995. The astacin family of metalloendopeptidases. *Protein Science*, 4(7), pp.1247-1261.

Boothby, T.C., Tenlen, J.R., Smith, F.W., Wang, J.R., Patanella, K.A., Osborne Nishimura, E., Tintori, S.C., Li, Q., Jones, C.D., Yandell, M., Messina, D.N., Glasscock, J. and Goldstein, B., 2015. Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proceedings of the National Academy of Sciences*, 112(52), pp.15976–15981.

Bornberg-Bauer, E., Huylmans, A.K. and Sikosek, T., 2010. How do new proteins arise? *Current Opinion in Structural Biology*, 20(3), pp.390–396.

Boschetti, C., Carr, A., Crisp, A., Eyres, I., Wang-Koh, Y., Lubzens, E., Barraclough, T.G., Micklem, G. and Tunnacliffe, A., 2012. Biochemical diversification through foreign gene expression in bdelloid rotifers. *PLoS Genetics*, 8(11), e1003035.

Boto, L., 2010. Horizontal gene transfer in evolution: facts and challenges. *Proceedings of the Royal Society B: Biological Sciences*, 277(1683), pp.819–827.

Boto, L., 2014. Horizontal gene transfer in the acquisition of novel traits by metazoans. *Proceedings of the Royal Society B: Biological Sciences*, 281(20132450).

Boto, L., 2015. Evolutionary change and phylogenetic relationships in light of horizontal gene transfer. *Journal of Biosciences*, 40(2), pp.465–472.

Böttger, A., Doxey, A.C., Hess, M.W., Pfaller, K., Salvenmoser, W., Deutzmann, R., Geissner, A., Pauly, B., Altstätter, J., Münder, S., Heim, A., Gabius, H.-J., McConkey, B.J. and David, C.N., 2012. Horizontal gene transfer contributed to the evolution of extracellular surface structures: the freshwater polyp hydra is covered by a complex fibrous cuticle containing glycosaminoglycans and proteins of the PPOD and SWT (sweet tooth) families. *PLoS ONE*, 7(12), e52278.

Boucher, Y., Douady, C.J., Papke, R.T., Walsh, D.A., Boudreau, M.E.R., Nesbø, C.L., Case, R.J. and Doolittle, W.F., 2003. Lateral gene transfer and the origins of prokaryotic groups. *Annual Review of*

Genetics, 37, pp.283–328.

Brakefield, P.M., 2011. Evo-devo and accounting for Darwin's endless forms. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1574), pp.2069–2075.

Bratke, K.A. and McLysaght, A., 2008. Identification of multiple independent horizontal gene transfers into poxviruses using a comparative genomics approach. *BMC Evolutionary Biology*, 8(67).

Braymer, J.J. and Giedroc, D.P., 2014. Recent developments in copper and zinc homeostasis in bacterial pathogens. *Current Opinion in Chemical Biology*, 19, pp.59–66.

Brügger, K., Redder, P., She, Q., Confalonieri, F., Zivanovic, Y. and Garrett, R.A., 2002. Mobile elements in archaeal genomes. *FEMS Microbiology Letters*, 206(2), pp.131–141.

Bruins, M.R., Kapil, S. and Oehme, F.W., 2000. Microbial resistance to metals in the environment. *Ecotoxicology and Environmental Safety*, 45(3), pp.198–207.

Buchanan-Wollaston, V., Passiatore, J.E. and Cannon, F., 1987. The *mob* and *oriT* mobilization functions of a bacterial plasmid promote its transfer to plants. *Nature,* 328, pp.172-175.

Buckley, K.M., and Rast, J.P., 2015. Developmental and comparative iummunology. *Developmental & Comparative Immunology*, 49(1), pp.179–189.

Bundock, P., Mróczek, K., Winkler, A.A., Steensma, H.Y. and Hooykaas, P.J., 1999. T-DNA from *Agrobacterium tumefaciens* as an efficient tool for gene targeting in *Kluyveromyces lactis*. *Molecular & General Genetics*, 261(1), pp.115–121.

Burgdorfer, W., Anacker, R.L., Bird, R.G. and Bertram, D.S., 1968. Intranuclear growth of *Rickettsia rickettsii*. *Journal of Bacteriology*, 96(4), pp.1415–1418.

Campbell, C.L. and Thorsness, P.E., 1998. Escape of mitochondrial DNA to the nucleus in yme1 yeast is mediated by vacuolar-dependent turnover of abnormal mitochondrial compartments. *Journal of Cell Science*, 111, pp.2455–2464.

Carroll, S.B., 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell*, 134(1), pp.25–36.

Casacuberta, E. and González, J., 2013. The impact of transposable elements in environmental adaptation. *Molecular Ecology*, 22(6), pp.1503–1517.

Cerdà-Costa, N. and Xavier Gomis-Rüth, F., 2013. Architecture and function of metallopeptidase catalytic domains. *Protein Science*, 23(2), pp.123–144.

Cha, J.N., Shimizu, K., Zhou, Y., Christiansen, S.C., Chmelka, B.F., Stucky, G.D. and Morse, D.E., 1999. Silicatein filaments and subunits from a marine sponge direct the polymerization of silica and

silicones in vitro. *Proceedings of the National Academy of Sciences of the United States of America*, 96(2), pp.361–365.

Cha, J.N., Stucky, G.D., Morse, D.E. and Deming, T.J., 2000. Biomimetic synthesis of ordered silica structures mediated by block copolypeptides. *Nature*, 40(20).

Chalopin, D., Naville, M., Plard, F., Galiana, D. and Volff, J.N., 2015. Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome Biology and Evolution*, 7(2), pp.567–580.

Chandran, D., Tai, Y.C., Hather, G., Dewdney, J., Denoux, C., Burgess, D.G., Ausubel, F.M., Speed, T.P. and Wildermuth, M.C., 2009. Temporal global expression data reveal known and novel salicylate-impacted processes and regulators mediating powdery mildew growth and reproduction on *Arabidopsis*. *Plant Physiology*, 149(3), pp.1435–1451.

Chandran Darbari, V. and Waksman, G., 2015. Structural biology of bacterial type IV secretion systems. *Annual Review of Biochemistry*, 84(1), pp.603–629.

Chapman, J.A., Kirkness, E.F., Simakov, O., Hampson, S.E., Mitros, T., Weinmaier, T., Rattei, T., Balasubramanian, P.G., Borman, J., Busam, D., Disbennett, K., Pfannkoch, C., Sumin, N., Sutton, G.G., Viswanathan, L.D., Walenz, B., Goodstein, D.M., Hellsten, U., Kawashima, T., Prochnik, S.E., Putnam, N.H., Shu, S., Blumberg, B., Dana, C.E., Gee, L., Kibler, D.F., Law, L., Lindgens, D., Martinez, D.E., Peng, J., Wigge, P.A., Bertulat, B., Guder, C., Nakamura, Y., Ozbek, S., Watanabe, H., Khalturin, K., Hemmrich, G., Franke, A., Augustin, R., Fraune, S., Hayakawa, E., Hayakawa, S., Hirose, M., Hwang, J.S., Ikeo, K., Nishimiya-Fujisawa, C., Ogura, A., Takahashi, T., Steinmetz, P.R.H., Zhang, X., Aufschnaiter, R., Eder, M.-K., Gorny, A.-K., Salvenmoser, W., Heimberg, A.M., Wheeler, B.M., Peterson, K.J., Böttger, A., Tischler, P., Wolf, A., Gojobori, T., Remington, K.A., Strausberg, R.L., Venter, J.C., Technau, U., Hobmayer, B., Bosch, T.C.G., Holstein, T.W., Fujisawa, T., Bode, H.R., David, C.N., Rokhsar, D.S. and Steele, R.E., 2010. The dynamic genome of Hydra. *Nature*, 464(7288), pp.592–596.

Chen, F., Mackey, A.J., Stoeckert, C.J. and Roos, D.S., 2006. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Research*, 34(90001), pp.D363–D368.

Cherrad, S., Girard, V., Dieryckx, C., Gonçalves, I.R., Dupuy, J.-W., Bonneu, M., Rascle, C., Job, C., Job, D., Vacher, S. and Poussereau, N., 2012. Proteomic analysis of proteins secreted by *Botrytis cinerea* in response to heavy metal toxicity. *Metallomics*, 4(8), pp.835-846.

Chidiac, M., Fayyad-Kazan, M., Daher, J., Poelvoorde, P., Bar, I., Maenhaut, C., Delrée, P., Badran, B. and Vanhamme, L., 2016. ApolipoproteinL1 is expressed in papillary thyroid carcinomas. *Pathology - Research and Practice*, 212(7), pp.631–635.

Chor, B., Horn, D., Goldman, N., Levy, Y. and Massingham, T., 2009. Genomic DNA k-mer spectra: models and modalities. *Genome Biology*, 10(10).

Christie, P.J. and Vogel, J.P., 2000. Bacterial type IV secretion: conjugation systems adapted to deliver effector molecules to host cells. *Trends in Microbiology*, 8(8), pp.354–360.

Christie, P.J., Whitaker, N. and González-Rivera, C., 2014. Mechanism and structure of the bacterial type IV secretion systems. *Biochimica et Biophysica Acta*, 1843(8), pp.1578–1591.

Citiulo, F., Jacobsen, I.D., Miramón, P., Schild, L., Brunke, S., Zipfel, P., Brock, M., Hube, B. and Wilson, D., 2012. *Candida albicans* scavenges host zinc via Pra1 during endothelial invasion. *PLoS Pathogens*, 8(6), e1002777.

Citovsky, V., Kozlovsky, S.V., Lacroix, B., Zaltsman, A., Dafny-Yelin, M., Vyas, S., Tovkach, A. and Tzfira, T., 2007. Biological systems of the host cell involved in *Agrobacterium* infection. *Cellular Microbiology*, 9(1), pp.9–20.

Clark, J.B. and Kidwell, M.G., 1997. A phylogenetic perspective on *P* transposable element evolution in *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America,* 94(21), pp.11428–11433.

Clark, J.B., Altheide, T.K., Schlosser, M.J. and Kidwell, M.G., 1995. Molecular evolution of *P* transposable elements in the genus *Drosophila I.* the *Saltans* and *Willistoni* species groups. *Molecular Biology and Evolution*, 12(5), pp.902–913.

Cleland, C.E., 2013. Pluralism or unity in biology: could microbes hold the secret to life? *Biology & Philosophy*, 28(2), pp.189–204.

Cobbs, C., Heath, J., Stireman, J.O., III and Abbot, P., 2013. Carotenoids in unexpected places: gall midges, lateral gene transfer, and carotenoid biosynthesis in animals. *Molecular Phylogenetics and Evolution*, 68(2), pp.221–228.

Coghlan, A., Eichler, E.E., Oliver, S.G., Paterson, A.H. and Stein, L., 2005. Chromosome evolution in eukaryotes: a multi-kingdom perspective. *Trends in Genetics*, 21(12), pp.673–682.

Collesi, C., Santoro, M.M., Gaudino, G. and Comoglio, P.M., 1996. A splicing variant of the RON transcript induces constitutive tyrosine kinase activity and an invasive phenotype. *Molecular and Cellular Biology*, 16(10), pp.5518–5526.

Conaco, C., Neveu, P., Zhou, H., Arcila, M.L., Degnan, S.M., Degnan, B.M. and Kosik, K.S., 2012. Transcriptome profiling of the demosponge *Amphimedon queenslandica* reveals genome-wide events that accompany major life cycle transitions. *BMC Genomics*, 13(1).

Conaco, C., Tsoulfas, P., Sakarya, O., Dolan, A., Werren, J. and Kosik, K.S., 2016. Detection of prokaryotic genes in the *Amphimedon queenslandica* genome. *PLoS ONE*, 11(3), e0151092.

Conant, G.C. and Wolfe, K.H., 2008. Turning a hobby into a job: how duplicated genes find new functions. *Nature Reviews Genetics*, 9(12), pp.938–950.

Confalonieri, F. and Duguet, M., 1995. A 200-amino acid ATPase module in search of a basic function. *BioEssays*, 17(7), pp.639-650.

Conradi, J., Tegtmeyer, N., Woźna, M., Wissbrock, M., Michalek, C., Gagell, C., Cover, T.L., Frank, R., Sewald, N. and Backert, S., 2012. An RGD helper sequence in CagL of *Helicobacter pylori* assists in interactions with integrins and injection of CagA. *Front Cellular and Infection Microbiology*, 2(70).

Cooper, C.A., Mulder, D.T., Allison, S.E., Pilar, A. and Coombes, B.K., 2013. The SseC translocon component in *Salmonella enterica* serovar Typhimurium is chaperonedby SscA. *BMC Microbiology*, 13(221), pp.1–8.

Cooper, E.D., 2014. Horizontal gene transfer: accidental inheritance drives adaptation. *Current Biology*, 24(12), pp.R562–R564.

Copley, S., 2003. Enzymes with extra talents: moonlighting functions and catalytic promiscuity. *Current Opinion in Chemical Biology*, 7(2), pp.265–272.

Corbin, B.D., Seeley, E.H., Raab, A., Feldmann, J., Miller, M.R., Torres, V.J., Anderson, K.L., Dattilo, B.M., Dunman, P.M., Gerads, R., Caprioli, R.M., Nacken, W., Chazin, W.J. and Skaar, E.P., 2008. Metal chelation and inhibition of bacterial growth in tissue abscesses. *Science*, 319(5865), pp.962–965.

Craig, N.L., Craigie, R., Gellert, M. and Lambowitz, A., 2002. Mobile DNA II. *American Society of Microbiology*, Washington.

Crawley, M.J., 2005. Statistics: An Introduction Using R. *Wiley*, West Sussex.

Crisp, A., Boschetti, C., Perry, M., Tunnacliffe, A. and Micklem, G., 2015. Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. *Genome Biology*, 16(50).

Dalton, J.P., Neill, S.O., Stack, C., Collins, P., Walshe, A., Sekiya, M., Doyle, S., Mulcahy, G., Hoyle, D., Khaznadji, E., Moiré, N., Brennan, G., Mousley, A., Kreshchenko, N., Maule, A.G. and Donnelly, S.M., 2003. *Fasciola hepatica* cathepsin L-like proteases: biology, function, and potential in the development of first generation liver fluke vaccines. *International Journal for Parasitology*, 33(11), pp.1173–1181.

Danchin, E.G.J., Rosso, M.-N., Vieira, P., de Almeida-Engler, J., Coutinho, P.M., Henrissat, B. and Abad, P., 2010. Multiple lateral gene transfers and duplications have promoted plant parasitism ability in nematodes. *Proceedings of the National Academy of Sciences*, 107(41), pp.17651–17656.

Danchin, E.G.J., 2016. Lateral gene transfer in eukaryotes: tip of the iceberg or of the ice cube? *BMC Biology*, 14(101).

Daniels, S.B., Peterson, K.R., Strausbaugh, L.D., Kidwell, M.G. and Chovnick, A., 1990. Evidence for horizontal transmission of the *P* transposable element between *Drosophila* species. *Genetics*, 124(2),

pp.339–355.

Das, S., Mandal, M., Chakraborti, T., Mandal, A. and Chakraborti, S., 2003. Structure and evolutionary aspects of matrix metalloproteinases: a brief overview. *Molecular and Cellular Biochemistry*, 253, pp.31–40.

Darriba, D., Taboada, G.L., Doallo, R. and Posada, D., 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, 27(8), pp.1164–1165.

Davids, W. and Zhang, Z., 2008. The impact of horizontal gene transfer in shaping operons and protein interaction networks – direct evidence of preferential attachment. *BMC Evolutionary Biology*, 8(23).

Davies, J. and Davies, D., 2010. Origins and evolution of antibiotic resistance. *Microbiology and Molecular Biology Reviews*, 74(3), pp.417–433.

de Almeida, L.M. and Carareto, C.M.A., 2005. Multiple events of horizontal transfer of the *Minos* transposable element between *Drosophila* species. *Molecular Phylogenetics and Evolution*, 35(3), pp.583–594.

Degnan, B.M., Adamska, M., Craigie, A., Degnan, S.M., Fahey, B., Gauthier, M., Hooper, J.N.A., Larroux, C., Leys, S.P., Lovas, E. and Richards, G.S., 2008. The demosponge *Amphimedon queenslandica*: reconstructing the ancestral metazoan genome and deciphering the origin of animal multicellularity. *Cold Spring Harbor Protocols*, 2008(12).

Degnan, B.M., Adamska, M., Richards, G.S., Larroux, C., Leininger, S., Bergum, B., Calcino, A., Taylor, K., Nakanishi, N. and Degnan, S.M., 2015. Porifera. In Wanninger A. (ed.), *Evolutionary Developmental Biology of Invertebrates 1: Introduction, Non-bilateria, Acoelomorpha, Xenoturbellida, Chaetognatha.* Vienna, Austria: Springer Verlag, pp.65-106.

Degnan, S.M., 2014. Think laterally: horizontal gene transfer from symbiotic microbes may extend the phenotype of marine sessile hosts. *Frontiers in Microbiology*, 5(638).

Deschamps, P. and Moreira, D., 2009. Signal conflicts in the phylogeny of the primary photosynthetic eukaryotes. *Molecular Biology and Evolution*, 26(12), pp.2745–2753.

Ding, Y., Zhou, Q. and Wang, W., 2012. Origins of new genes and evolution of their novel functions. *Annual Review of Ecology, Evolution, and Systematics*, 43(1), pp.345–363.

Djoko, K.Y., Ong, C.L.Y., Walker, M.J. and McEwan, A.G., 2015. The role of copper and zinc toxicity in innate immune defense against bacterial pathogens. *Journal of Biological Chemistry*, 290(31), pp.18954–18961.

Doi, Y., Akiyama, H., Yamada, Y., Ee, C.E., Lee, B.R., Ikeguchi, M. and Ichishima, E., 2004. Thermal stabilization of penicillolysin, a thermolabile 19 kDa Zn2+-protease, obtained by site-directed mutagenesis. *Protein Engineering Design and Selection*, 17(3), pp.261–266.

Domínguez, A. and Albornoz, J., 1996. Rates of movement of transposable elements in *Drosophila melanogaster*. *Molecular & General Genetics,* 251(2), pp.130–138.

Doolittle, W.F., 1998. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends in Genetics*, 14(8), pp.307–311.

Driscoll, T., Gillespie, J.J., Nordberg, E.K., Azad, A.F. and Sobral, B.W., 2013. Bacterial DNA sifted from the *Trichoplax adhaerens* (Animalia:Placozoa) genome project reveals a putative rickettsial endosymbiont. *Genome Biology and Evolution*, 5(4), pp.621–645.

Dunn, C.W. and Ryan, J.F., 2015. The evolution of animal genomes. *Current Opinion in Genetics & Development*, 35, pp.25–32.

Dunning Hotopp, J.C., Clark, M.E., Oliveira, D.C.S.G., Foster, J.M., Fischer, P., Torres, M.C.M., Giebel, J.D., Kumar, N., Ishmael, N., Wang, S., Ingram, J., Nene, R.V., Shepard, J., Tomkins, J., Richards, S., Spiro, D.J., Ghedin, E., Slatko, B.E., Tettelin, H. and Werren, J.H., 2007. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science*, 317, pp.1753–1756.

Dunning Hotopp, J.C., 2011. Horizontal gene transfer between bacteria and animals. *Trends in Genetics*, 27(4), pp.157–163.

Eddy, S.R., 1996. Hidden markov models. *Current Opinion in Structural Biology*, 6(3), pp.361–365.

Edwards, D.J. and Holt, K.E., 2013. Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. *Microbial Informatics and Experimentation*, 3(2).

Eichler, E.E. and Sankoff, D., 2003. Structural dynamics of eukaryotic chromosome evolution. *Science*, 301(5634), pp.793–797.

Eldredge, N. and Gould, S.J., 1972. Punctuated equilibria: an alternative to phyletic gradualism. *Models in Paleobiology*, pp.82–115.

Elston, R.A., 1986. An intranuclear pathogen [nuclear inclusion X (NIX)] associated with massive mortalities of the Pacific razor clam, *Siliqua patula*. *Journal of Invertebrate Pathology*, 47, pp.93–104.

Ereskovsky A.V., 2010. The Comparative Embryology of Sponges. New York: Springer.

Erwin, D.H., Laflamme, M., Tweedt, S.M., Sperling, E.A., Pisani, D. and Peterson, K.J, 2011. The Cambrian conundrum: early divergence and later ecological success in the early history of animals. *Science*, 334, pp.1091–1097.

Ewing, A.D., 2015. Transposable element detection from whole genome sequence data. *Mobile DNA*, 6(24).

Ewing, A.D. and Kazazian, H.H., 2010. High-throughput sequencing reveals extensive variation in

human-specific L1 content in individual human genomes. *Genome Research*, 20(9), pp.1262–1270.

Faber, H.R., Groom, C.R., Baker, H.M., Morgan, W.T., Smith, A. and Baker, E.N., 1995. 1.8 A crystal structure of the C-terminal domain of rabbit serum haemopexin. *Structure*, 3(6), pp.551–559.

Fairhead, M., Johnson, K.A., Kowatz, T., McMahon, S.A., Carter, L.G., Oke, M., Liu, H., Naismith, J.H. and van der Walle, C.F., 2008. Crystal structure and silica condensing activities of silicatein α–cathepsin L chimeras. *Chemical Communications*, 2(15), pp.1765-1767.

Fan, L., Reynolds, D., Liu, M., Stark, M., Kjelleberg, S., Webster, N.S. and Thomas, T., 2012. Functional equivalence and evolutionary convergence in complex communities of microbial sponge symbionts. *Proceedings of the National Academy of Sciences*, 109(27), pp.E1878–1887.

Fani, R., Brilli, M., Fondi, M. and Lió, P., 2007. The role of gene fusions in the evolution of metabolic pathways: the histidine biosynthesis case. *BMC Evolutionary Biology*, 7(S4).

Fast, N.M., Kissinger, J.C., Roos, D.S. and Keeling, P.J., 2001. Nuclear-encoded, plastid-targeted genes suggest a single common origin for Apicomplexan and Dinoflagellate plastids. *Molecular Biology and Evolution*, 18(3), pp.418–426.

Felbor, U., Dreier, L., Bryant, R.A., Ploegh, H.L., Olsen, B.R. and Mothes, W., 2000. Secreted cathepsin L generates endostatin from collagen XVIII. *The EMBO Journal*, 19(6), pp.1187–1194.

Fernandez-Valverde, S.L., Higgie, S.S. and Degnan, S.M., in preparation. A pipeline for the detection of interkingdom gene transfer events reveals horizontal gene transfer contributes to lineage-specific genome composition in basal marine animals.

Fernandez-Valverde, S.L., Calcino, A.D. and Degnan, B.M., 2015. Deep developmental transcriptome sequencing uncovers numerous new genes and enhances gene annotation in the sponge *Amphimedon queenslandica*. *BMC Genomics*, 16(720).

Feschotte, C., 2008. Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics*, 9(5), pp.397–405.

Fieth, R.A., Gauthier, M.E.A., Bayes, J., Green, K.M. and Degnan, S.M., 2016. Ontogenetic changes in the bacterial symbiont community of the tropical demosponge *Amphimedon queenslandica*: metamorphosis is a new beginning. *Frontiers in Marine Science*, 3(228).

Filée, J., Forterre, P. and Laurent, J., 2003. The role played by viruses in the evolution of their hosts: a view based on informational protein phylogenies. *Research in Microbiology*, 154(4), pp.237–243.

Finn, R.D., Clements, J. and Eddy, S.R., 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, 39, pp.W29–W37.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M.,

Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J. and Bateman A., 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research,* 44, pp.D279-285.

Finn, R.D., Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S.R., Sonnhammer, E.L.L. and Bateman, A., 2006. Pfam: clans, web tools and services. *Nucleic Acids Research*, 34, pp.D247–51.

Fisher, R., 1935. The sheltering of lethals. *American Naturalist.* 69, pp.446–455.

Fiston-Lavier, A.S., Barron, M.G., Petrov, D.A. and Gonzalez, J., 2015. T-lex2: genotyping, frequency estimation and re-annotation of transposable elements using single or pooled next-generation sequencing data. *Nucleic Acids Research*, 43(4), e22.

Flot, J.F., Hespeels, B., Li, X., Noel, B., Arkhipova, I., Danchin, E.G.J., Hejnol, A., Henrissat, B., Koszul, R., Aury, J.-M., Barbe, V., Barthélémy, R.-M., Bast, J., Bazykin, G.A., Chabrol, O., Couloux, A., Da Rocha, M., Da Silva, C., Gladyshev, E., Gouret, P., Hallatschek, O., Hecox-Lea, B., Labadie, K., Lejeune, B., Piskurek, O., Poulain, J., Rodriguez, F., Ryan, J.F., Vakhrusheva, O.A., Wajnberg, E., Wirth, B., Yushenova, I., Kellis, M., Kondrashov, A.S., Welch, D.B.M., Pontarotti, P., Weissenbach, J., Wincker, P., Jaillon, O. and Van Doninck, K., 2013. Genomic evidence for ameiotic evolution in the bdelloid rotifer *Adineta vaga*. *Nature*, 500(7463), pp.453–457.

Fortunato, S.A.V., Adamski, M., Ramos, O.M., Leininger, S., Liu, J., Ferrier, D.E.K. and Adamska, M., 2014. Calcisponges have a ParaHox gene and dynamic expression of dispersed NK homeobox genes. *Nature*, 514(7524), pp.620–623.

Fraser, J.S., Yu, Z., Maxwell, K.L. and Davidson, A.R., 2006. Ig-like domains on bacteriophages: A tale of promiscuity and deceit. *Journal of Molecular Biology*, 359(2), pp.496–507.

Freedman, B.I., Langefeld, C.D., Lu, L., Palmer, N.D., Carrie Smith, S., Bagwell, B.M., Hicks, P.J., Xu, J., Wagenknecht, L.E., Raffield, L.M., Register, T.C., Jeffrey Carr, J., Bowden, D.W. and Divers, J., 2015. APOL1 associations with nephropathy, atherosclerosis, and all-cause mortality in African Americans with type 2 diabetes. *Kidney International*, 87(1), pp.176–181.

Friedrich, A.B., Merkert, H., Fendert, T., Hacker, J., Proksch, P. and Hentschel, U., 1999. Microbial diversity in the marine sponge *Aplysina cavernicola* (formerly *Verongia cavernicola*) analyzed by fluorescence in situ hybridization (FISH). *Marine Biology*, 134, pp.461–470.

Friesen, P.D. and Nissen, M.S., 1990. Gene organization and transcription of *TED,* a lepidopteran retrotransposon integrated within the baculovirus genome. *Molecular and Cellular Biology*, 10(6), pp.3067–3077.

Friesen, T.L., Stukenbrock, E.H., Liu, Z., Meinhardt, S., Ling, H., Faris, J.D., Rasmussen, J.B., Solomon, P.S., McDonald, B.A. and Oliver, R.P., 2006. Emergence of a new disease as a result of interspecific virulence gene transfer. *Nature Genetics*, 38(8), pp.953–956.

Fujishima, M. and Kodama, Y., 2012. Endosymbionts in *Paramecium*. *European Journal of Protistology*, 48, pp.124–137.

Fujishima, M., Kawai, M. and Yamamoto, R., 2005. *Paramecium caudatum* acquires heat-shock resistance in ciliary movement by infection with the endonuclear symbiotic bacterium *Holospora obtusa*. *FEMS Microbiology Letters*, 243(1), pp.101–105.

Funayama, N., 2010. The stem cell system in demosponges: insights into the origin of somatic stem cells. *Development, Growth & Differentiation*, 52(1), pp.1–14.

Furner, I.J., Huffman, G.A., Amasino, R.M., Garfinkel, D.J., Gordon, M.P. and Nester, E.W., 1986. An *Agrobacterium* transformation in the evolution of the genus *Nicotiana*. *Nature*, 319, pp.422-427.

Fushimi, N., Ee, C.E., Nakajima, T. and Ichishima, E., 1999. Aspzincin, a family of metalloendopeptidases with a new zinc-binding motif: identification of new zinc-binding sites (His128, His132, and Asp164) and three catalytically crucial residues (Glu129, Asp143, and Tyr106) of deuterolysin from *Aspergillus oryzae* by site-directed mutagenesis. *Journal of Biological Chemistry*, 274(34), pp.24195–24201.

Gaiti, F., Fernandez-Valverde, S.L., Nakanishi, N., Calcino, A.D., Yanai, I., Tanurdzic, M. and Degnan, B.M., 2015. Dynamic and widespread lncRNA expression in a sponge and the origin of animal complexity', *Molecular Biology and Evolution*, 32(9), pp.2367–2382.

Galán, J.E., Lara-Tejero, M., Marlovits, T.C. and Wagner, S., 2014. Bacterial type III secretion systems: specialized nanomachines for protein delivery into target cells. *Annual Review of Microbiology*, 68(1), pp.415–438.

Gao, X., Wang, J., Yu, D.Q., Bian, F., Xie, B.B., Chen, X.L., Zhou, B.C., Lai, L.H., Wang, Z.X., Wu, J.W. and Zhang, Y.Z., 2010. Structural basis for the autoprocessing of zinc metalloproteases in the thermolysin family. *Proceedings of the National Academy of Sciences*, 107(41), pp.17569–17574.

Gauthier, A., 2014. Analysis of silicatein gene expression and spicule formation in the demosponge *Amphimedon queenslandica.* University of Queensland, Brisbane.

Gauthier, M.E.A, Watson, J.R. and Degnan, S.M., 2016. Draft genomes shed light on the dual bacterial symbiosis that dominates the microbiome of the coral reef sponge *Amphimedon queenslandica*. *Frontiers in Marine Science*, 3(196).

Gazave, E., Lapébie, P., Ereskovsky, A.V., Vacelet, J., Renard, E., Cárdenas, P. and Borchiellini, C., 2012. No longer Demospongiae: Homoscleromorpha formal nomination as a fourth class of Porifera. *Hydrobiologia*, 687, pp.3–10.

Gazave, E., Lapébie, P., Renard, E., Vacelet, J., Rocher, C., Ereskovsky, A.V., Lavrov, D.V. and Borchiellini, C., 2010. Molecular Phylogeny Restores the Supra-Generic Subdivision of Homoscleromorph Sponges (Porifera, Homoscleromorpha). *PLoS ONE*, 5(12), e14290.

Gelvin, S.B., 2000. *Agrobacterium* and plant genes involved in T-DNA transfer and integration. *Annual Review of Plant Physiology and Plant Molecular Biology*, 51, pp.223–256.

Gelvin, S.B., 2003. *Agrobacterium*-mediated plant transformation: the biology behind the 'Gene-Jockeying' tool. *Microbiology and Molecular Biology Reviews*, 67(1), pp.16–37.

Gendrel, C.G., Boulet, A. and Dutreix, M., 2000. (CA/GT)(n) microsatellites affect homologous recombination during yeast meiosis. *Genes & Development*, 14(10), pp.1261–1268.

Genovese, G., Friedman, D.J., Ross, M.D., Lecordier, L., Uzureau, P., Freedman, B.I., Bowden, D.W., Langefeld, C.D., Oleksyk, T.K., Uscinski Knob, A.L., Bernhardy, A.J., Hicks, P.J., Nelson, G.W., Vanhollebeke, B., Winkler, C.A., Kopp, J.B., Pays, E. and Pollak, M.R., 2010. Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science*, 329(5993), pp.841–845.

George, B., Alam, C.M., Kumar, R.V., Gnanasekaran, P. and Chakraborty, S., 2015. Potential linkage between compound microsatellites and recombination in geminiviruses: evidence from comparative analysis. *Virology*, 482, pp.41–50.

Gerlt, J.A., Allen, K.N., Almo, S.C., Armstrong, R.N., Babbitt, P.C., Cronan, J.E., Dunaway-Mariano, D., Imker, H.J., Jacobson, M.P., Minor, W., Poulter, C.D., Raushel, F.M., Sali, A., Shoichet, B.K. and Sweedler, J.V., 2011. The enzyme function initiative. *Biochemistry*, 50(46), pp.9950–9962.

Gerton, J.L. and Hawley, R.S., 2005. Homologous chromosome interactions in meiosis: diversity amidst conservation. *Nature Reviews Genetics*, 6(6), pp.477–487.

Geuking, M.B., Weber, J., Dewannieux, M., Gorelik, E., Heidmann, T., Hengartner, H., Zinkernagel, R.M. and Hangartner, L., 2009. Recombination of retrotransposon and exogenous RNA virus results in nonretroviral cDNA integration. *Science*, 323(5912), pp.393–396.

Ghai, J. and Das, A., 1989. The virD operon of *Agrobacterium tumefaciens* Ti plasmid encodes a DNA-relaxing enzyme. *Proceedings of the National Academy of Sciences of the United States of America*, 86(9), pp.3109–3113.

Gilbert, C., Hernandez, S.S., Flores-Benabib, J., Smith, E.N. and Feschotte, C., 2012. Rampant horizontal transfer of SPIN transposons in squamate reptiles. *Molecular Biology and Evolution*, 29(2), pp.503–515.

Gilbert, C. and Cordaux, R., 2013. Horizontal transfer and evolution of prokaryote transposable elements in eukaryotes. *Genome Biology and Evolution*, 5(5), pp.822–832.

Gilbert, C., Schaack, S., Pace, J.K., II, Brindley, P.J. and Feschotte, C., 2010. A role for host-parasite interactions in the horizontal transfer of transposons across phyla. *Nature*, 464(29), pp.1347–1350.

Gilbert, C., Waters, P., Feschotte, C. and Schaack, S., 2013. Horizontal transfer of OC1 transposons in the Tasmanian devil. *BMC Genomics*, 14(134).

Gillespie, J.J., Phan, I.Q.H., Scheib, H., Subramanian, S., Edwards, T.E., Lehman, S.S., Piitulainen, H., Sayeedur Rahman, M., Rennoll-Bankert, K.E., Staker, B.L., Taira, S., Stacy, R., Myler, P.J., Azad, A.F. and Pulliainen, A.T., 2015. Structural insight into how bacteria prevent interference between multiple divergent type IV secretion systems. *mBio*, 6(6), pp.e01867-15.

Gladyshev, E.A. and Arkhipova, I.R., 2009. A single-copy *IS5*-like transposon in the genome of a bdelloid rotifer. *Molecular Biology and Evolution*, 26(8), pp.1921–1929.

Gladyshev, E.A., Meselson, M. and Arkhipova, I.R., 2008. Massive horizontal gene transfer in bdelloid rotifers. *Science*, 320(5880), pp.1210–1213.

Glansdorff, N., Xu, Y. and Labedan, B., 2009. The conflict between horizontal gene transfer and the safeguard of identity: origin of meiotic sexuality. *Journal of Molecular Evolution*, 69(5), pp.470–480.

Goebl, M. and Yanagida, M., 1991. The TPR snap helix: a novel protein repeat motif from mitosis to transcription. *Trends in Biochemical Sciences*, 16(5), pp.173–177.

Gogarten, J.P., Doolittle, W.F. and Lawrence, J.G., 2002. Prokaryotic evolution in light of gene transfer. *Molecular Biology and Evolution*, 19(12), pp.2226–2238.

Gogarten, J.P. and Townsend, J.P., 2005. Horizontal gene transfer, genome innovation and evolution. *Nature Reviews Microbiology*, 3(9), pp.679–687.

Gomis-Rüth, F.X., Gohlke, U., Betz, M., Knäuper, V., Murphy, G., López-Otín, C. and Bode, W., 1996. The helping hand of collagenase-3 (MMP-13): 2.7 Å crystal structure of its C-terminal haemopexin-like domain. *Journal of Molecular Biology*, 264(3), pp.556–566.

Gomis-Rüth, F.X., 2003. Structural aspects of the metzincin clan of metalloendopeptidases. *Molecular Biotechnology*, 24(2), pp.157–202.

Gophna, U. and Ofran, Y., 2011. Lateral acquisition of genes is affected by the friendliness of their products. *Proceedings of the National Academy of Sciences*, 108(1), pp.343–348.

Gotz, S., García-Gómez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J., Robles, M., Talón, M., Dopazo, J. and Conesa1, A., 2008. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research*, 36(10), pp.3420–3435.

Gough, J., Karplus, K., Hughey, R. and Chothia, C., 2001. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *Journal of Molecular Biology*, 313(4), pp.903–919.

Gould, S.J. and Eldredge, N., 1977. Punctuated equilibria: the tempo and mode of evolution reconsidered. *Paleobiology*, 3, pp.115–151.

Görtz, H.D., 2001. Intracellular bacteria in ciliates. *International Microbiology*, 4(3), pp.143–150.

Grbić, M., Van Leeuwen, T., Clark, R.M., Rombauts, S., Rouzé, P., Grbić, V., Osborne, E.J., Dermauw, W., Ngoc, P.C.T., Ortego, F., Hernández-Crespo, P., Diaz, I., Martinez, M., Navajas, M., Sucena, É., Magalhães, S., Nagy, L., Pace, R.M., Djuranović, S., Smagghe, G., Iga, M., Christiaens, O., Veenstra, J.A., Ewer, J., Villalobos, R.M., Hutter, J.L., Hudson, S.D., Velez, M., Yi, S.V., Zeng, J., Pires-daSilva, A., Roch, F., Cazaux, M., Navarro, M., Zhurov, V., Acevedo, G., Bjelica, A., Fawcett, J.A., Bonnet, E., Martens, C., Baele, G., Wissler, L., Sanchez-Rodriguez, A., Tirry, L., Blais, C., Demeestere, K., Henz, S.R., Gregory, T.R., Mathieu, J., Lou Verdon, Farinelli, L., Schmutz, J., Lindquist, E., Feyereisen, R. and Van de Peer, Y., 2011. The genome of *Tetranychus urticae* reveals herbivorous pest adaptations. *Nature*, 479(7374), pp.487–492.

Grandi, G., Guidi, L. and Chicca, M., 1997. Endonuclear bacterial symbionts in two termite species: an ultrastructural study. *Journal of Submicroscopic Cytology and Pathology*, 29, pp.281-292.

Gray, Y.H., 2000. It takes two transposons to tango: transposable-element-mediated chromosomal rearrangements. *Trends in Genetics*, 16(10), pp.461–468.

Grice, L.F., Gauthier, M.E.A., Roper, K.E., Fernàndez-Busquets, X., Degnan, S.M. and Degnan, B.M., 2017. Origin and evolution of the sponge Aggregation Factor gene family. *Molecular Biology and Evolution*, 34(5), pp.1083–1099.

Griffith, F., 1928. The significance of pneumococcal types. *The Journal of Hygiene*, 27(2), pp.113–159.

Gröger, C., Sumper, M. and Brunner, E., 2008. Silicon uptake and metabolism of the marine diatom *Thalassiosira pseudonana*: Solid-state 29Si NMR and fluorescence microscopic studies. *Journal of Structural Biology*, 161(1), pp.55–63.

Guglielmini, J., la Cruz, de, F. and Rocha, E.P.C., 2013. Evolution of conjugation and type IV secretion systems. *Molecular Biology and Evolution*, 30(2), pp.315–331.

Guilhen, C., Taha, M.-K. and Veyrier, F.J., 2013. Role of transition metal exporters in virulence: the example of *Neisseria meningitidis*. *Frontiers in Cellular and Infection Microbiology*, 3(102).

Guindon, S. and Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5), pp.696–704.

Guzman, C. and Conaco, C., 2016. Comparative transcriptome analysis reveals insights into the streamlined genomes of haplosclerid demosponges. *Scientific Reports*, 6(18774).

Gyles, C. and Boerlin, P., 2014. Horizontally transferred genetic elements and their role in pathogenesis of bacterial disease. *Veterinary Pathology*, 51(2), pp.328–340.

Haldane, J., 1933. The part played by recurrent mutation by evolution. *America Naturalist.* 67, pp.5–9.

Hall, C. and Dietrich, F.S., 2007. The reacquisition of biotin prototrophy in *Saccharomyces cerevisiae* involved horizontal gene transfer, gene duplication and gene clustering. *Genetics*, 177(4), pp.2293–2307.

Hall, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*, 41, pp.95–98.

Han, M.J., Shen, Y.H., Xu, M.S., Liang, H.Y., Zhang, H.H. and Zhang, Z., 2013. Identification and evolution of the silkworm helitrons and their contribution to transcripts. *DNA Research*, 20(5), pp.471–484.

Han, N., Yu, W., Qiang, Y. and Zhang, W., 2016. T4SP Database 2.0: an improved database for type IV secretion systems in bacterial genomes with new online analysis tools. *Computational and Mathematical Methods in Medicine*, 2016(9415459).

Hao, X., Lüthje, F.L., Qin, Y., McDevitt, S.F., Lutay, N., Hobman, J.L., Asiani, K., Soncini, F.C., German, N., Zhang, S., Zhu, Y.-G. and Rensing, C., 2015. Survival in amoeba—a major selection pressure on the presence of bacterial copper and zinc resistance determinants? Identification of a 'copper pathogenicity island'. *Applied Microbiology and Biotechnology*, 99(14), pp.5817–5824.

Hashimshony, T., Wagner, F., Sher, N. and Yanai, I., 2012. CEL-Seq: Single-cell RNA-seq by multiplexed linear amplification. *Cell Reports*, 2(3), pp.666–673.

Honsa, E.S., Johnson, M.D.L. and Rosch, J.W., 2013. The roles of transition metals in the physiology and pathogenesis of *Streptococcus pneumoniae*. *Frontiers in Cellular and Infection Microbiology*, 3(92).

Hood, M.I. and Skaar, E.P., 2012. Nutritional immunity: transition metals at the pathogen–host interface. *Nature Reviews Microbiology*, 10(8), pp.525–537.

Häse, C.C. and Finkelstein, R.A., 1993. Bacterial extracellular zinc-containing metalloproteases. *Microbiological Reviews*, 57(4), pp.823–837.

Hazkani-Covo, E. and Covo, S., 2008. Numt-mediated double-strand break repair mitigates deletions during primate genome evolution. *PLoS Genetics*, 4(10), e1000237.

Hazkani-Covo, E., Sorek, R. and Graur, D., 2003. Evolutionary dynamics of large numts in the human genome: rarity of independent insertions and abundance of post-insertion duplications. *Journal of Molecular Evolution*, 56(2), pp.169–174.

Hazkani-Covo, E., Zeller, R.M. and Martin, W., 2010. Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genetics*, 6(2), e1000834.

Heinemann, J.A. and Sprague, G.F., 1989. Bacterial conjugative plasmids mobilize DNA transfer between bacteria and yeast. *Nature*, 340(6230), pp.205–209.

Helle, F., 2012. Germ cell DNA-repair systems—possible tools in cancer research? *Cancer Gene Therapy*, 19(4), pp.299–302.

Hendrix, R.W., Smith, M.C.M., Burns, R.N., Ford, M.E. and Hatfull, G.F., 1999. Evolutionary

relationships among diverse bacteriophages and prophages: all the world's a phage. Proceedings of the National Academy of Sciences of the United States of America. 96, pp.2192–2197.

Hentschel, U., Piel, J., Degnan, S.M. and Taylor, M.W., 2012. Genomic insights into the marine sponge microbiome. *Nature Reviews Microbiology*, 10, pp.641-654.

Herpin, A., Braasch, I., Kraeussling, M., Schmidt, C., Thoma, E.C., Nakamura, S., Tanaka, M. and Schartl, M., 2010. Transcriptional rewiring of the sex determining dmrt1 gene duplicate by transposable elements. *PLoS Genetics*, 6(2), e1000844.

Hey, J. and Kliman, R.M., 2002. Interactions between natural selection, recombination and gene density in the genes of *Drosophila*. *Genetics*, 160(2), pp.595–608.

Hickman, A.B., Chandler, M. and Dyda, F., 2010. Integrating prokaryotes and eukaryotes: DNA transposases in light of structure. *Critical Reviews in Biochemistry and Molecular Biology*, 45(1), pp.50–69.

Hill, C.W., Sandt, C.H. and Vlazny, D.A., 1994. Rhs elements of *Escherichia coli*: a family of genetic composites each encoding a large mosaic protein. *Molecular Microbiology*, 12(6), pp.865–871.

Hoen, D.R., Hickey, G., Bourque, G., Casacuberta, J., Cordaux, R., Feschotte, C., Fiston-Lavier, A.-S., Hua-Van, A., Hubley, R., Kapusta, A., Lerat, E., Maumus, F., Pollock, D.D., Quesneville, H., Smit, A., Wheeler, T.J., Bureau, T.E. and Blanchette, M., 2015. A call for benchmarking transposable element annotation methods. *Mobile DNA*, 6(13).

Hoffmann, A.A. and Sgrò, C.M., 2011. Climate change and evolutionary adaptation. *Nature*, 470(7335), pp.479–485.

Hooper, J. and Van Soest, R., 2006. A new species of *Amphimedon* (Porifera, Demospongiae, Haplosclerida, Niphatidae) from the Capricorn-Bunker Group of Islands, Great Barrier Reef, Australia: target species for the 'sponge genome project'. *Zootaxa*, 1314, pp.31–39.

Hori, M. and Fujishima, M., 2003. The endosymbiotic bacterium *Holospora obtusa* enhances heat-shock gene expression of the host *Paramecium caudatum*. *The Journal of Eukaryotic Microbiology*, 50(4), pp.293–298.

Hori, M., F, K. and Fujishima, M., 2008. Micronucleus-specific bacterium *Holospora elegans* irreversibly enhances stress gene expression of the host *Paramecium caudatum*. *Journal of Eukaryotic Microbiology*, 55(6), pp.515–521.

Houck, M.A., Clark, J.B., Peterson, K.R. and Kidwell, M.G., 1991. Possible horizontal transfer of *Drosophila* genes by the mite *Proctolaelaps regalis*. *Science*, 253(5024), pp.1125–1128.

Howard, E.A., Zupan, J.R., Citovsky, V. and Zambryski, P.C., 1992. The VirD2 protein of *A. tumefaciens* contains a C-terminal bipartite nuclear localization signal: implications for nuclear uptake of DNA in

plant cells. *Cell*, 68(1), pp.109–118.

Hu, C.-A.A. and Ray, P.E., 2016. How complicated can it be? The link between APOL1 risk variants and lipoprotein heterogeneity in kidney and cardiovascular diseases. *Nephrology Dialysis Transplantation*, 31(4), pp.509–511.

Hu, C.-A.A., Klopfer, E.I. and Ray, P.E., 2012. Human Apolipoprotein L1 (ApoL1) in cancer and chronic kidney disease. *FEBS letters*, 586(7), pp.947–955.

Huang, J. and Brumell, J.H., 2014. Bacteria–autophagy interplay: a battle for survival. *Nature Reviews Microbiology*, 12(2), pp.101–114.

Huang, C.R.L., Burns, K.H. and Boeke, J.D., 2012. Active transposition in genomes. *Annual Review of Genetics*, 46(1), pp.651–675.

Huang, L., Cheng, T., Xu, P., Duan, J., Fang, T., and Xia, Q., 2009. Immunoglobulin superfamily is conserved but evolved rapidly and is active in the silkworm, *Bombyx mori*. *Insect Molecular Biology,* 18(4), pp.517-530.

Huang, Y., Niu, B., Gao, Y., Fu, L. and Li, W., 2010. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, 26(5), pp.680–682.

Husnik, F., Nikoh, N., Koga, R., Ross, L., Duncan, R.P., Fujie, M., Tanaka, M., Satoh, N., Bachtrog, D., Wilson, A.C.C., Dohlen, von, C.D., Fukatsu, T. and McCutcheon, J.P., 2013. Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. *Cell*, 153(7), pp.1567–1578.

Ito, K., Bick, A.G., Flannick, J., Friedman, D.J., Genovese, G., Parfenov, M.G., Depalma, S.R., Gupta, N., Gabriel, S.B., Taylor, H.A., Fox, E.R., Newton-Cheh, C., Kathiresan, S., Hirschhorn, J.N., Altshuler, D.M., Pollak, M.R., Wilson, J.G., Seidman, J.G. and Seidman, C., 2014. Increased burden of cardiovascular disease in carriers of APOL1 genetic variants. *Circulation Research*, 114(5), pp.845–850.

Ivancevic, A.M., Walsh, A.M., Kortschak, R.D. and Adelson, D.L., 2013. Jumping the fine LINE between species: Horizontal transfer of transposable elements in animals catalyses genome evolution. *BioEssays*, 35(12), pp.1071–1082.

Iyer, L.M., Aravind, L., Coon, S.L., Klein, D.C. and Koonin, E.V., 2004. Evolution of cell–cell signaling in animals: did late horizontal gene transfer from bacteria have a role? *Trends in Genetics*, 20(7), pp.292–299.

Jackson, D.J., Macis, L., Reitner, J. and Wörheide, G., 2011. A horizontal gene transfer supported the evolution of an early metazoan biomineralization strategy. *BMC Evolutionary Biology*, 11(238).

Jain, R., Rivera, M.C. and Lake, J.A., 1999. Horizontal gene transfer among genomes: the complexity hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, 96(7),

pp.3801–3806.

Jehle, J.A., Nickel, A., Vlak, J.M. and Backhaus, H., 1998. Horizontal escape of the novel Tc1-like Lepidopteran transposon TCp3.2 into *Cydia pomonella* granulovirus. *Journal of Molecular Evolution*, 46(2), pp.215–224.

Jensen, H., Engedal, H. and Saetersdal, T.S., 1976. Ultrastructure of mitochondria-containing nuclei in human myocardial cells. *European Journal of Pathology,* 21, pp.1–12.

Jensen, L., Grant, J.R., Laughinghouse, H.D., IV and Katz, L.A., 2016. Assessing the effects of a sequestered germline on interdomain lateral gene transfer in Metazoa. *Evolution*, 70(6), pp.1322–1333.

Jensen, R.A., 1987. Evolution of metabolic pathways in enteric bacteria. In: Neidhardt, F.C., Ingraham, J.L., Low, K.B., Magasanik, B., Schaechter, M., Humbarger, H.D., (ed.), *Escherichia coli and Salmonella typhimurium: cellular and molecular biology*. ASM Press, Washington DC, pp.2649–2662.

Jensen, S., Duperron, S., Birkeland, N.-K. and Hovland, M., 2010. Intracellular *Oceanospirillales* bacteria inhabit gills of *Acesta bivalves*. *FEMS Microbiology Ecology*, 74(3), pp.523–533.

Jensen-Seaman, M.I., Furey, T.S., Payseur, B.A., Lu, Y., Roskin, K.M., Chen, C.-F., Thomas, M.A., Haussler, D. and Jacob, H.J., 2004. Comparative recombination rates in the rat, mouse, and human genomes. *Genome Research*, 14(4), pp.528–538.

Jiang, N., Bao, Z., Zhang, X., Eddy, S.R. and Wessler, S.R., 2004. *Pack-MULE* transposable elements mediate gene evolution in plants. *Nature*, 431(7008), pp.569–573.

Jiang, N., Gao, D., Xiao, H. and van der Knaap, E., 2009. Genome organization of the tomato sunlocus and characterization of the unusual retrotransposon Rider. *The Plant Journal*, 60(1), pp.181–193.

Juhas, M., van der Meer, J.R., Gaillard, M., Harding, R.M., Hood, D.W. and Crook, D.W., 2009. Genomic islands: tools of bacterial horizontal gene transfer and evolution. *FEMS Microbiology Reviews*, 33(2), pp.376–393.

Juliano, C. and Wessel, G., 2010. Versatile Germline Genes. *Science*, 329(5992), pp.640–641.

Kaessmann, H., 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Research*, 20(10), pp.1313–1326.

Kakegawa, H., Nikawa, T., Tagami, K., Kamioka, H., Sumitani, K., Kawata, T., Drobnic-Kosorok, M., Lenarcic, B., Turk, V. and Katunuma, N., 1993. Participation of cathepsin L on bone resorption. *FEBS letters*, 321(2-3), pp.247–250.

Kapitonov, V.V. and Jurka, J., 2001. Rolling-circle transposons in eukaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, 98(15), pp.8714–8719.

Kaplan, N., Darden, T. and Langley, C.H., 1985. Evolution and extinction of transposable elements in Mendelian populations. *Genetics*, 109(2), pp.459–480.

Kapusta, A., Kronenberg, Z., Lynch, V.J., Zhuo, X., Ramsay, L., Bourque, G., Yandell, M. and Feschotte, C., 2013. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genetics*, 9(4), e1003470.

Karas, B.J., Diner, R.E., Lefebvre, S.C., McQuaid, J., Phillips, A.P.R., Noddings, C.M., Brunson, J.K., Valas, R.E., Deerinck, T.J., Jablanovic, J., Gillard, J.T.F., Beeri, K., Ellisman, M.H., Glass, J.I., Hutchison, C.A., III, Smith, H.O., Venter, J.C., Allen, A.E., Dupont, C.L. and Weyman, P.D., 2015. Designer diatom episomes delivered by bacterial conjugation. *Nature Communications*, 6, pp.1–10.

Katoh, K., Misawa, K., Kuma, K.-I. and Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14), pp.3059–3066.

Katz, L.A., 2015. Recent events dominate interdomain lateral gene transfers between prokaryotes and eukaryotes and, with the exception of endosymbiotic gene transfers, few ancient transfer events persist. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1678), 20140324.

Katzourakis, A., Tristem, M., Pybus, O.G. and Gifford, R.J., 2007. Discovery and analysis of the first endogenous lentivirus. *Proceedings of the National Academy of Sciences of the United States of America*, 104(15), pp.6261–6265.

Kawai, S., Pham, T.A., Nguyen, H.T., Nankai, H., Utsumi, T., Fukuda, Y. and Murata, K., 2004. Molecular insights on DNA delivery into *Saccharomyces cerevisiae*. *Biochemical and Biophysical Research Communications*, 317(1), pp.100–107.

Keane, T.M., Wong, K. and Adams, D.J., 2013. RetroSeq: transposable element discovery from next-generation sequencing data. *Bioinformatics*, 29, pp.389–390.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P. and Drummond, A., 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data', *Bioinformatics*, 28(12). pp.1647–1649.

Keeling, P.J. and Palmer, J.D., 2008. Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, 9(8), pp.605–618.

Keeling, P.J., 2009. Functional and ecological impacts of horizontal gene transfer in eukaryotes. *Current Opinion in Genetics & Development*, 19(6), pp.613–619.

Khersonsky, O., Roodveldt, C. and Tawfik, D., 2006. Enzyme promiscuity: evolutionary and mechanistic aspects. *Current Opinion in Chemical Biology*, 10(5), pp.498–508.

Kidwell, M.G., 1993. Lateral transfer in natural populations of eukaryotes. *Annual Review of Genetics,*

27, pp.235-256.

Kim, E.S., Lee, H.J., Bang, W.-G., Choi, I.-G. and Kim, K.H., 2009. Functional characterization of a bacterial expansin from *Bacillus subtilis* for enhanced enzymatic hydrolysis of cellulose. *Biotechnology and Bioengineering*, 102(5), pp.1342–1353.

Kim, S., Jeon, T.-J., Oberai, A., Yang, D., Schmidt, J.J. and Bowie, J.U., 2005. Transmembrane glycine zippers: physiological and pathological roles in membrane proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 102(40), pp.14278–14283.

King, M.C. and Wilson, A.C., 1975. Evolution at two levels in humans and chimpanzees. *Science*, 188(4184), pp.107–116.

King, N., Westbrook, M.J., Young, S.L., Kuo, A., Abedin, M., Chapman, J., Fairclough, S., Hellsten, U., Isogai, Y., Letunic, I., Marr, M., Pincus, D., Putnam, N., Rokas, A., Wright, K.J., Zuzow, R., Dirks, W., Good, M., Goodstein, D., Lemons, D., Li, W., Lyons, J.B., Morris, A., Nichols, S., Richter, D.J., Salamov, A., Sequencing, J., Bork, P., Lim, W.A., Manning, G., Miller, W.T., McGinnis, W., Shapiro, H., Tjian, R., Grigoriev, I.V. and Rokhsar, D., 2008. The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature*, 451(7180), pp.783–788.

Kipreos, E.T. and Pagano, M., 2000. The F-box protein family. *Genome Biology*, 1(5).

Klasson, L., Kambris, Z., Cook, P.E., Walker, T. and Sinkins, S.P., 2009. Horizontal gene transfer between *Wolbachia* and the mosquito *Aedes aegypti*. *BMC Genomics*, 10(33).

Kojima, K.K. and Jurka, J., 2011. Crypton transposons: identification of new diverse families and ancient domestication events. *Mobile DNA*, 2(12).

Kolde, R., 2012. Pheatmap: pretty heatmaps. *R package version* 61.

Kondo, N., Nikoh, N., Ijichi, N., Shimada, M. and Fukatsu, T., 2002. Genome fragment of *Wolbachia* endosymbiont transferred to X chromosome of host insect. *Proceedings of the National Academy of Sciences*, 99(22), pp.14280–14285.

Kondrashov, F.A., Koonin, E.V., Morgunov, I.G., Finogenova, T.V. and Kondrashova, M.N., 2006. Evolution of glyoxylate cycle enzymes in Metazoa: evidence of multiple horizontal transfer events and pseudogene formation. *Biology Direct*, 1(31).

Koonin, E.V., 2009a. Darwinian evolution in the light of genomics. *Nucleic Acids Research*, 37(4), pp.1011–1034.

Koonin, E.V., 2009b. The Origin at 150: is a new evolutionary synthesis in sight? *Trends in Genetics*, 25(11), pp.473–475.

Koonin, E.V. and Dolja, V.V., 2006. Evolution of complexity in the viral world: the dawn of a new

vision. *Virus Research*, 117(1), pp.1–4.

Koskiniemi, S., Lamoureux, J.G., Nikolakakis, K.C., t'Kint de Roodenbeke, C., Kaplan, M.D., Low, D.A. and Hayes, C.S., 2013. Rhs proteins from diverse bacteria mediate intercellular competition. *Proceedings of the National Academy of Sciences*, 110(17), pp.7032–7037.

Koutsovoulos, G., Kumar, S., Laetsch, D.R., Stevens, L., Daub, J., Conlon, C., Maroon, H., Thomas, F., Aboobaker, A.A. and Blaxter, M., 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), pp.5053–5058.

Kozhemyako, V.B., Veremeichik, G.N., Shkryl, Y.N., Kovalchuk, S.N., Krasokhin, V.B., Rasskazov, V.A., Zhuravlev, Y.N., Bulgakov, V.P. and Kulchin, Y.N., 2010. Silicatein genes in spicule-forming and nonspicule-forming Pacific demosponges. *Marine Biotechnology*, 12(4), pp.403–409.

Krasko, A., Lorenz, B., Batel, R., Schröder, H.C., Müller, I.M. and Müller, W.E., 2000. Expression of silicatein and collagen genes in the marine sponge *Suberites domuncula* is controlled by silicate and myotrophin. *European Journal of Biochemistry*, 267(15), pp.4878–4887.

Kröger, M. and Hobom, G., 1982. Structural analysis of insertion sequence IS5. *Nature*, 297(5862), pp.159–162.

Krogh, A., Larsson, B., Heijne, von, G. and Sonnhammer, E.L.L., 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology*, 305(3), pp.567–580.

Krueger, D.M., Gustafson, R.G. and Cavanaugh, C.M., 1996. Vertical transmission of chemoautotrophic symbionts in the bivalve *Solemya velum* (Bivalvia: Protobranchia). *The Biological Bulletin*, 190(2), pp.195–202.

Ku, C. and Martin, W.F., 2016. A natural barrier to lateral gene transfer from prokaryotes to eukaryotes revealed from genomes: the 70 % rule. *BMC Biology*, 14(89).

Ku, C., Nelson-Sathi, S., Roettger, M., Garg, S., Hazkani-Covo, E. and Martin, W.F., 2015. Endosymbiotic gene transfer from prokaryotic pangenomes: Inherited chimerism in eukaryotes. *Proceedings of the National Academy of Sciences*, 112(33), pp.10139–10146.

Kuriakose, S.M., Singh, R. and Uzonna, J.E., 2016. Host intracellular signaling events and pro-inflammatory Cytokine production in African Trypanosomiasis. *Frontiers in Immunology*, 7(181).

Lacroix, B. and Citovsky, V., 2013. The roles of bacterial and host plant factors in *Agrobacterium*-mediated genetic transformation. *The International Journal of Developmental Biology*, 57(6-7-8), pp.467–481.

Lacroix, B. and Citovsky, V., 2016. Transfer of DNA from Bacteria to Eukaryotes. *mBio*, 7(4).

Lambowitz, A.M. and Belfort, M., 1993. Introns as mobile genetic elements. *Annual Review of Biochemistry*, 62, pp.587–622.

Lebreton, A., Lakisic, G., Job, V., Fritsch, L., Tham, T.N., Camejo, A., Matteï, P.-J., Regnault, B., Nahori, M.-A., Cabanes, D., Gautreau, A., Ait-Si-Ali, S., Dessen, A., Cossart, P. and Bierne, H., 2011. A bacterial protein targets the BAHD1 chromatin complex to stimulate type III interferon response. *Science*, 331(6022), pp.1319–1321.

Lederberg, J., 1952. Cell genetics and hereditary symbiosis. *Physiological Reviews*, 32(4), pp.403–430.

Leedale, G.F., 1969. Observations on endonuclear bacteria in Euglenoid flagellates. *Plant Systematics and Evolution*, 116, pp.279–294.

Leipe, D.D., Koonin, E.V. and Aravind, L., 2003. Evolution and classification of P-loop kinases and related proteins. *Journal of Molecular Biology*, 333(4), pp.781-815.

Lenormand, T., Engelstädter, J., Johnston, S.E., Wijnker, E. and Haag, C.R., 2016. Evolutionary mysteries in meiosis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(20160001).

Leplae, R., Lima-Mendez, G. and Toussaint, A., 2010. ACLAME: A CLAssification of Mobile genetic Elements, update 2010. *Nucleic Acids Research*, 38, pp.D57–D61.

Lerat, E., 2010. Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Heredity*, 104(6), pp.520–533.

Levin, M., Anavy, L., Cole, A.G., Winter, E., Mostov, N., Khair, S., Senderovich, N., Kovalev, E., Silver, D.H., Feder, M., Fernandez-Valverde, S.L., Nakanishi, N., Simmons, D., Simakov, O., Larsson, T., Liu, S.-Y., Jerafi-Vider, A., Yaniv, K., Ryan, J.F., Martindale, M.Q., Rink, J.C., Arendt, D., Degnan, S.M., Degnan, B.M., Hashimshony, T. and Yanai, I., 2016. The mid-developmental transition and the evolution of animal body plans. *Nature*, 531(7596), pp.637–641.

Levin, T.C., Greaney, A.J., Wetzel, L. and King, N., 2014. The rosetteless gene controls development in the choanoflagellate *S. rosetta*. *eLife*, 3(946).

Leys, S.P. and Degnan, B.M., 2002. Embryogenesis and metamorphosis in a haplosclerid demosponge: gastrulation and transdifferentiation of larval ciliated cells to choanocytes. *Invertebrate Biology*, 121(3), pp.171–189.

Li, W. and Godzik, A., 2006. CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13), pp.1658–1659.

Limou, S., Dummer, P.D., Nelson, G.W., Kopp, J.B. and Winkler, C.A., 2015. APOL1 toxin, innate immunity, and kidney injury. *Kidney International*, 88(1), pp.28–34.

Linial, M.L., Fan, B., Hahn, R., Lwer, J., Neil, S., Quackenbush, A., Rethwilm, P., Sonigo, J., Stoye

and Tristem, M., 2005. Retroviridae. In Fauquet, C.M, Mayo, M.A., Maniloff, Desselberger, U. and Ball, L.A., (ed.), *Eighth report of the International Committee on Taxonomy of Viruses. Elsevier Academic Press*, London, United Kingdom, pp.421-440.

Liu, H., Fu, Y., Jiang, D., Li, G., Xie, J., Cheng, J., Peng, Y., Ghabrial, S.A. and Yi, X., 2010. Widespread horizontal gene transfer from double-stranded RNA viruses to eukaryotic nuclear genomes. *Journal of Virology*, 84(22), pp.11876–11887.

Lloyd, A.H. and Timmis, J.N., 2011. The origin and characterization of new nuclear genes originating from a cytoplasmic organellar genome. *Molecular Biology and Evolution*, 28(7), pp.2019–2028.

Long, M., Betrán, E., Thornton, K. and Wang, W., 2003. The origin of new genes: glimpses from the young and old. *Nature Reviews Genetics*, 4(11), pp.865–875.

Lorenzi, H.A., Robledo, G. and Levin, M.J., 2006. The VIPER elements of trypanosomes constitute a novel group of tyrosine recombinase-enconding retrotransposons. *Molecular and Biochemical Parasitology*, 145(2), pp.184–194.

Loreto, E.L., Valente, V.L., Zaha, A., Silva, J.C. and Kidwell, M.G., 2001. Drosophila mediopunctata P elements: a new example of horizontal transfer. *The Journal of Heredity*, 92(5), pp.375–381.

Lovatt, F.M. and Hoelzel, A.R., 2013. Impact on reindeer (*Rangifer tarandus*) genetic diversity from two parallel population bottlenecks founded from a common source. *Evolutionary Biology*, 41(2), pp.240–250.

Lowe, C.B., Bejerano, G. and Haussler, D., 2007. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proceedings of the National Academy of Sciences*, 104(19), pp.8005–8010.

Lowe, C.B. and Haussler, D., 2012. 29 mammalian genomes reveal novel exaptations of mobile elements for likely regulatory functions in the human genome. *PLoS ONE*, 7(8), e43128.

Lukashev, M.E. and Werb, Z., 1998. ECM signalling: orchestrating cell behaviour and misbehaviour. *Trends in Cell Biology*, 8(11), pp.437–441.

Lustigman, S., Zhang, J., Liu, J., Oksov, Y. and Hashmi, S., 2004. RNA interference targeting cathepsin L and Z-like cysteine proteases of *Onchocerca volvulus* confirmed their essential function during L3 molting. *Molecular and Biochemical Parasitology*, 138(2), pp.165–170.

Lynch, M., Conery, J. and Burger, R., 1995. Mutation accumulation and the extinction of small populations. *The American Naturalist*, 146, pp.489–518.

Lynch, M., 1996. Mutation accumulation in transfer RNAs: molecular evidence for Muller's ratchet in mitochondrial genomes. *Molecular Biology Evol*ution, 13(1), pp.209-220.

Magori, S. and Citovsky, V., 2011. Epigenetic control of *Agrobacterium* T-DNA integration. *Biochimica et Biophysica Acta*, 1809(8), pp.388–394.

Makarova, K.S., Wolf, Y.I., Forterre, P., Prangishvili, D., Krupovic, M. and Koonin, E.V., 2014. Dark matter in archaeal genomes: a rich source of novel mobile elements, defense systems and secretory complexes. *Extremophiles*, 18(5), pp.877–893.

Maldonado, M. and Riesgo, A., 2007. Intra-epithelial spicules in a homosclerophorid sponge. *Cell and Tissue Research*, 328(3), pp.639–650.

Malik, H.S., Henikoff, S. and Eickbush, T.H., 2000. Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Research*, 10(9), pp.1307–1318.

Mallet, L.V., Becq, J. and Deschavanne, P., 2010. Whole genome evaluation of horizontal transfers in the pathogenic fungus *Aspergillus fumigatus*. *BMC Genomics*, 11(171).

Mandrich, L. and Manco, G., 2009. Evolution in the Amidohydrolase Superfamily: substrate-assisted gain of function in the E183K mutant of a phosphotriesterase-like metal-carboxylesterase. *Biochemical and Biophysical Research Communications*, 48(24), pp.5602–5612.

Marcet-Houben, M. and Gabaldón, T., 2010. Acquisition of prokaryotic genes by fungal genomes. *Trends in Genetics*, 26(1), pp.5–8.

Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C.J., Lu, S., Chitsaz, F., Derbyshire, M.K., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Lu, F., Marchler, G.H., Song, J.S., Thanki, N., Wang, Z., Yamashita, R.A., Zhang, D., Zheng, C., Geer, L.Y. and Bryant, S.H., 2017. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Research*, 45, pp.D200–D203.

Marshall, J.M., 2008. A branching process for the early spread of a transposable element in a diploid population. *Journal of Mathematical Biology*, 57(6), pp.811–840.

Martin, W., Stoebe, B., Goremykin, V., Hapsmann, S., Hasegawa, M. and Kowallik, K.V., 1998. Gene transfer to the nucleus and the evolution of chloroplasts. *Nature*, 393(6681), pp.162–165.

Martinson, E.O., Martinson, V.G., Edwards, R., Mrinalini and Werren, J.H., 2016. Laterally transferred gene recruited as a venom in parasitoid wasps. *Molecular Biology and Evolution*, 33(4), pp.1042–1052.

Maritz, K., Calcino, A., Fahey, B., Degnan, B. and Degnan, S.M., 2010. Remarkable consistency of larval supply in the spermcast-mating demosponge *Amphimedon queenslandica* (Hooper and van Soest). *Open Marine Biology*, 4, pp.57-64.

Masel, J. and Trotter, M.V., 2010. Robustness and evolvability. *Trends in Genetics*, 26(9), pp.406–414.

Matlin, A.J., Clark, F. and Smith, C.W.J., 2005. Understanding alternative splicing: towards a cellular

code. *Nature Reviews Molecular Cell Biology*, 6(5), pp.386–398.

Matveeva, T.V., Bogomaz, D.I., Pavlova, O.A., Nester, E.W. and Lutova, L.A., 2012. Horizontal gene transfer from genus *Agrobacteriumto* the plant *Linaria* in nature. *Molecular Plant-Microbe Interactions*, 25(12), pp.1542–1551.

Matveeva, T.V. and Lutova, L.L,, 2014. Horizontal gene transfer from *Agrobacterium* to plants. *Frontiers in Plant Science,* 5(326).

Mazel, D., 2006. Integrons: agents of bacterial evolution. *Nature Reviews Microbiology*, 4(8), pp.608–620.

McAuley, K.E., Jia-Xing, Y., Dodson, E.J., Lehmbeck, J., Østergaard, P.R. and Wilson, K.S., 2001. A quick solution: ab initio structure determination of a 19 kDa metalloproteinase using ACORN. *Biological crystallography*, 57, pp.1571–1578.

McClintock, B., 1950. The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences of the United States of America*, 36(6), pp.344–355.

McClintock, B., 1984. The significance of responses of the genome to challenge. *Science*, 226(4676), pp.792–801.

McCue, A.D., Nuthikattu, S., Reeder, S.H. and Slotkin, R.K., 2012. Gene expression and stress response mediated by the epigenetic regulation of a transposable element small RNA. *PLoS Genetics*, 8(2), e1002474.

McDonald, J.F., 1995. Transposable elements: possible catalysts of organismic evolution. *Trends in Ecology & Evolution*, 10(3), pp.123–126.

McFall-Ngai, M., Hadfield, M.G., Bosch, T.C.G., Carey, H.V., Domazet-Loso, T., Douglas, A.E., Dubilier, N., Eberl, G., Fukami, T., Gilbert, S.F., Hentschel, U., King, N., Kjelleberg, S., Knoll, A.H., Kremer, N., Mazmanian, S.K., Metcalf, J.L., Nealson, K., Pierce, N.E., Rawls, J.F., Reid, A., Ruby, E.G., Rumpho, M., Sanders, J.G., Tautz, D. and Wernegreen, J.J., 2013. Animals in a bacterial world, a new imperative for the life sciences. *Proceedings of the National Academy of Sciences*, 110(9), pp.3229–3236.

McFall-Ngai, M.J., 2015. Giving microbes their due - animal life in a microbially dominant world. *Journal of Experimental Biology*, 218(12), pp.1968–1973.

McNulty, S.N., Foster, J.M., Mitreva, M., Dunning Hotopp, J.C., Martin, J., Fischer, K., Wu, B., Davis, P.J., Kumar, S., Brattig, N.W., Slatko, B.E., Weil, G.J. and Fischer, P.U., 2010. Endosymbiont DNA in endobacteria-free filarial nematodes indicates ancient horizontal genetic transfer. *PLoS ONE*, 5(6), e11029.

Meyer-Gauen, G., Schnarrenberger, C., Cerff, R. and Martin, W., 1994. Molecular characterization of

a novel, nuclear-encoded, NAD(+)-dependent glyceraldehyde-3-phosphate dehydrogenase in plastids of the gymnosperm Pinus sylvestris L. *Plant Molecular Biology*, 26(4), pp.1155–1166.

Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., Hingamp, P., Goto, S. and Ogata, H., 2016. Linking virus genomes with host taxonomy. *Viruses*, 8(66).

Miller, W.J., McDonald, J.F. and Pinsker, W., 1997. Molecular domestication of mobile elements. *Genetica*, 100, pp.261–270.

Miller, W.J., McDonald, J.F., Nouaud, D. and Anxolabéhère, D., 1999. Molecular domestication--more than a sporadic episode in evolution. *Genetica*, 107, pp.197–207.

Miller, W.J., Nagel, A., Bachmann, J. and Bachmann, L., 2000. Evolutionary dynamics of the SGM transposon family in the *Drosophila obscura* species group. *Molecular Biology and Evolution*, 17(11), pp.1597–1609.

Miller, W.J., Paricio, N., Hagemann, S., Martínez-Sebastián, M.J., Pinsker, W. and de Frutos, R., 1995. Structure and expression of clustered *P* element homologues in *Drosophila subobscura* and *Drosophila guanche*. *Gene*, 156(2), pp.167–174.

Mir, A.A., Philippe, C. and Cristofari, G., 2015. euL1db: the European database of L1HS retrotransposon insertions in humans. *Nucleic Acids Research*, 43, pp.D43–D47.

Mittal, R., Patel, A.P., Debs, L.H., Nguyen, D., Patel, K., Grati, M., Mittal, J., Yan, D., Chapagain, P. and Liu, X.Z., 2016. Intricate functions of matrix metalloproteinases in physiological and pathological conditions. *Journal of Cellular Physiology*, 231(12), pp.2599–2621.

Miura, A., Yonebayashi, S., Watanabe, K., Toyama, T., Shimada, H. and Kakutani, T., 2001. Mobilization of transposons by a mutation abolishing full DNA methylation in *Arabidopsis*. *Nature*, 411(6834), pp.212–214.

Moczek, A.P., Sears, K.E., Stollewerk, A., Wittkopp, P.J., Diggle, P., Dworkin, I., Ledon-Rettig, C., Matus, D.Q., Roth, S., Abouheif, E., Brown, F.D., Chiu, C.-H., Cohen, C.S., Tomaso, A.W.D., Gilbert, S.F., Hall, B., Love, A.C., Lyons, D.C., Sanger, T.J., Smith, J., Specht, C., Vallejo-Marin, M. and Extavour, C.G., 2015. The significance and scope of evolutionary developmental biology: a vision for the 21st century. *Evolution & Development*, 17(3), pp.198–219.

Monier, A., Pagarete, A., de Vargas, C., Allen, M.J., Read, B., Claverie, J.M. and Ogata, H., 2009. Horizontal gene transfer of an entire metabolic pathway between a eukaryotic alga and its DNA virus. *Genome Research*, 19(8), pp.1441–1449.

Monteiro, A., 2012. Gene regulatory networks reused to build novel traits. *BioEssays*. 34(3), pp.181–186.

Moore, A.D., Held, A., Terrapon, N., Weiner, J. and Bornberg-Bauer, E., 2014. DoMosaics: software for domain arrangement visualization and domain-centric analysis of proteins. *Bioinformatics*, 30(2),

pp.282–283.

Moran, N.A., 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 93(7), pp.2873–2878.

Moran, N.A. and Baumann, P., 2000. Bacterial endosymbionts in animals. *Current Opinion in Microbiology*, 3(3), pp.270–275.

Moran, Y., Fredman, D., Szczesny, P., Grynberg, M. and Technau, U., 2012. Recurrent horizontal transfer of bacterial toxin genes to eukaryotes. *Molecular Biology and Evolution*, 29(9), pp.2223–2230.

Moran, N.A. and Jarvik, T., 2010. Lateral transfer of genes from fungi underlies carotenoid production in aphids. *Science*, 328(5978), pp.624–627.

Moreira, D. and Brochier-Armanet, C., 2008. Giant viruses, giant chimeras: the multiple evolutionary histories of Mimivirus genes. *BMC Evolutionary Biology*, 8(12).

Moreira, D. and López-García, P., 2009. Ten reasons to exclude viruses from the tree of life. *Nature Reviews Microbiology*, 7(4), pp.306–311.

Moreira, D., 2000. Multiple independent horizontal transfers of informational genes from bacteria to plasmids and phages: implications for the origin of bacterial replication machinery. *Molecular Microbiology*, 35(1), pp.1–5.

Moret, B.M., Roshan U. and Warnow, T., 2002. Sequence-length requirements for phylogenetic methods. In: Guigó, R. and Gusfield, D., (eds), *WABI 2002: Algorithms in Bioinformatics*, Springer, Berlin, Heidelberg. 2452, pp343-356.

Morgante, M., Brunner, S., Pea, G., Fengler, K., Zuccolo, A. and Rafalski, A., 2005. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nature Genetics*, 37(9), pp.997–1002.

Mota, M., 1963. Electron microscope study of the relationship between the nucleus and mitochondria in *Chlorophytum capense* (L.) Kuntze. *Cytologia*, 28, pp.409–416.

Mourier, T., Hansen, A.J., Willerslev, E. and Arctander, P., 2001. The Human Genome Project reveals a continuous transfer of large mitochondrial fragments to the nucleus. *Molecular Biology and Evolution*, 18(9), pp.1833–1837.

Mower, J., Stefanovic, S., Young, G.J. and Palmer, J.D., 2004. Gene transfer from parasitic to host plants. *Nature*, 432, pp.165-166.

Muesch, A., Hartmann, E., Rohde, K., Rubartelli, A., Sitia, R. and Rapoport, T.A., 1990. A novel pathway for secretory proteins? *Trends in Biochemical Sciences*, 15(3), pp.86–88.

Muller, H.J., 1936. Bar Duplication. *Science*, 83(2161), pp.528–530.

Muller, H.J., 1964. The relation of recombination to mutational advance. *Mutation Research*, 106, pp.2–9.

Müller, W., Wiens, M., Batel, R., Steffen, R., Schröder, H.C., Borojevic, R. and Custodio, M.R., 1999. Establishment of a primary cell culture from a sponge: primmorphs from *Suberites domuncula*. *Marine Ecology Progress Series*, 178, pp.205–219.

Müller, W.E.G., Boreiko, A., Wang, X., Belikov, S.I., Wiens, M., Grebenjuk, V.A., Schloβmacher, U. and Schröder, H.C., 2007. Silicateins, the major biosilica forming enzymes present in demosponges: protein analysis and phylogenetic relationship. *Gene*, 395, pp.62–71.

Müller, W.E.G., Rothenberger, M., Boreiko, A., Tremel, W., Reiber, A. and Schröder, H.C., 2005. Formation of siliceous spicules in the marine demosponge *Suberites domuncula*. *Cell and Tissue Research*, 321(2), pp.285–297.

Murphy, K.E. and Stringer, J.R., 1986. RecA independent recombination of poly[d (GT)-d (CA)] in pBR322. *Nucleic Acids Research*, 14(8), 7325-7340.

Muszewska, A., Steczkiewicz, K. and Ginalski, K., 2013. *DIRS* and *Ngaro* retrotransposons in fungi. *PLoS ONE*, 8(9), e76319.

Nakamura, Y., Itoh, T. and Martin, W., 2006. Rate and polarity of gene fusion and fission in *Oryza sativa* and *Arabidopsis thaliana*. *Molecular Biology and Evolution*, 24(1), pp.110–121.

Nakanishi, N., Sogabe, S. and Degnan, B.M., 2014. Evolutionary origin of gastrulation: insights from sponge development. *BMC Biology*, 12(26).

Nakashima, K., Yamada, L., Satou, Y., Azuma, J.-I. and Satoh, N., 2004. The evolutionary origin of animal cellulose synthase. *Development Genes and Evolution*, 214(2), pp.81–88.

Namangala, B., 2011. Contribution of innate immune responses towards resistance to African trypanosome infections. *Scandinavian Journal of Immunology*, 75(1), pp.5–15.

Neuwald, A.F., Aravind, L., Spouge, J.L. and Koonin, E.V., 1999. AAA+: A class of chaperone-like ATPases associated with the assembly, operation, and disassembly of protein complexes. *Genome Research*, 9(1), pp.27–43.

Nichols, S.A., Roberts, B.W., Richter, D.J., Fairclough, S.R., King, N., 2012. Origin of metazoan cadherin diversity and the antiquity of the classical cadherin/β-catenin complex. *Proceedings of the National Academy of Sciences*, 109(32), pp.13046–13051.

Nicholson, D.W., 1999. Caspase structure, proteolytic substrates, and function during apoptotic cell death. *Cell Death and Differentiation*, 6(11), pp.1028–1042.

Niehuis, O., Gibson, J.D., Rosenberg, M.S., Pannebakker, B.A., Koevoets, T., Judson, A.K., Desjardins, C.A., Kennedy, K., Duggan, D., Beukeboom, L.W., van de Zande, L., Shuker, D.M., Werren, J.H. and Gadau, J., 2010. Recombination and its impact on the genome of the haplodiploid parasitoid wasp Nasonia. *PLoS ONE*, 5(1), e8597.

Nies, D.H., 2003. Efflux-mediated heavy metal resistance in prokaryotes. *FEMS Microbiology Reviews*, 27(2-3), pp.313–339.

Niklas, K.J. and Newman, S.A., 2013. The origins of multicellular organisms. *Evolution & Development*, 15(1), pp.41–52.

Nikoh, N., McCutcheon, J.P., Kudo, T., Miyagishima, S.-Y., Moran, N.A. and Nakabachi, A., 2010. Bacterial genes in the aphid genome: absence of functional gene transfer from *Buchnera* to its host. *PLoS Genetics*, 6(2), e1000827.

Nikoh, N., Tanaka, K., Shibata, F., Kondo, N., Hizume, M., Shimada, M. and Fukatsu, T., 2008. *Wolbachia* genome integrated in an insect chromosome: evolution and fate of laterally transferred endosymbiont genes. *Genome Research*, 18(2), pp.272–280.

Nikolaidis, N., Doran, N. and Cosgrove, D.J., 2014. Plant expansins in bacteria and fungi: evolution by horizontal gene transfer and independent domain fusion. *Molecular Biology and Evolution*, 31(2), pp.376–386.

Nilsson, A.S., 2014. Phage therapy - constraints and possibilities. *Upsala Journal of Medical Sciences*, 119(2), pp.192–198.

Nonaka, T., Hashimoto, Y. and Takiot, K., 1998. Kinetic characterization of lysine-specific metalloendopeptidases from *Grifola frondosa* and *Pleurotus ostreatus* fruiting bodies. *Journal of Biochemistry*, 124(1), pp.157–162.

Novikova, O., Śliwińska, E., Fet, V., Settele, J., Blinov, A. and Woyciechowski, M., 2007. *CR1* clade of non-LTR retrotransposons from *Maculinea* butterflies (Lepidoptera: Lycaenidae): evidence for recent horizontal transmission. *BMC Evolutionary Biology*, 7(93).

Nuzhdin, S.V. and Mackay, T.F., 1995. The genomic rate of transposable element movement in *Drosophila melanogaster*. *Molecular Biology and Evolution*, 12(1), pp.180–181.

Ober, D., 2010. Gene duplications and the time thereafter - examples from plant secondary metabolism. *Plant Biology*, 12, pp.570-577.

Ochman, H., Lawrence, J.G. and Groisman, E.A., 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784), pp.299–304.

Ogata, H., La Scola, B., Audic, S., Renesto, P., Blanc, G., Robert, C., Fournier, P.-E., Claverie, J.-M. and Raoult, D., 2006. Genome sequence of *Rickettsia bellii* illuminates the role of amoebae in gene

exchanges between intracellular pathogens. *PLoS Genetics*, 2(5), e76.

Ohno, S., 1970. Evolution by gene duplication. Springer-Verlag, Berlin.

Oliver, K.R. and Greene, W.K., 2009. Transposable elements: powerful facilitators of evolution. *BioEssays*, 31(7), pp.703–714.

Oliver, K.R. and Greene, W.K., 2012. Transposable elements and viruses as factors in adaptation and evolution: an expansion and strengthening of the TE-Thrust hypothesis. *Ecology and Evolution*, 2(11), pp.2912–2933.

Osborn, A.M. and Böltner, D., 2002. When phage, plasmids, and transposons collide: genomic islands, and conjugative- and mobilizable-transposons as a mosaic continuum. *Plasmid*, 48(3), pp.202–212.

Overall, C.M., King, A.E., Sam, D.K., Ong, A.D., Lau, T.T., Wallon, U.M., DeClerck, Y.A. and Atherstone, J., 1999. Identification of the tissue inhibitor of metalloproteinases-2 (TIMP-2) binding site on the hemopexin carboxyl domain of human gelatinase A by site-directed mutagenesis. The hierarchical role in binding TIMP-2 of the unique cationic clusters of hemopexin modules III and IV. *The Journal of Biological Chemistry*, 274(7), pp.4421–4429.

Pace, J.K., Gilbert, C., Clark, M.S. and Feschotte, C., 2008. Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods. *Proceedings of the National Academy of Sciences*, 105(44), pp.17023–17028.

Paganini, J., Campan-Fournier, A., Da Rocha, M., Gouret, P., Pontarotti, P., Wajnberg, E., Abad, P. and Danchin, E.G.J., 2012. Contribution of lateral gene transfers to the genome composition and parasitic ability of root-knot nematodes. *PLoS ONE*, 7(11), e50875.

Palmer, L.D. and Skaar, E.P., 2016. Transition metals and virulence in bacteria. *Annual Review of Genetics*, 50(1), pp.67–91.

Paquin, C.E. and Williamson, V.M., 1984. Temperature effects on the rate of Ty transposition. *Science*, 226(4670), pp.53–55.

Park, C. and Zhang, J., 2012. High expression hampers horizontal gene transfer. *Genome Biology and Evolution*, 4(4), pp.523–532.

Parsa, A., Kao, W.H.L., Xie, D., Astor, B.C., Li, M., Hsu, C.-Y., Feldman, H.I., Parekh, R.S., Kusek, J.W., Greene, T.H., Fink, J.C., Anderson, A.H., Choi, M.J., Wright, J.T., Jr., Lash, J.P., Freedman, B.I., Ojo, A., Winkler, C.A., Raj, D.S., Kopp, J.B., He, J., Jensvold, N.G., Tao, K., Lipkowitz, M.S. and Appel, L.J., 2013. APOL1 risk variants, race, and progression of chronic kidney disease. *The New England Journal of Medicine*, 369(23), pp.2183–2196.

Pauchet, Y. and Heckel, D.G., 2013. The genome of the mustard leaf beetle encodes two active xylanases originally acquired from bacteria through horizontal gene transfer. *Proceedings of the Royal Society*

B: Biological Sciences, 280(20131021).

Pérez-Morga, D., Vanhollebeke, B., Paturiaux-Hanocq, F., Nolan, D.P., Lins, L., Homblé, F., Vanhamme, L., Tebabi, P., Pays, A., Poelvoorde, P., Jacquet, A., Brasseur, R. and Pays, E., 2005. Apolipoprotein L-I promotes trypanosome lysis by forming pores in lysosomal membranes. *Science*, 309(5733), pp.469–472.

Pérez-Porro, A.R., Navarro-Gómez, D., Uriz, M.J. and Giribet, G., 2013. A NGS approach to the encrusting Mediterranean sponge *Crella elegans* (Porifera, Demospongiae, Poecilosclerida): transcriptome sequencing, characterization and overview of the gene expression along three life cycle stages. *Molecular Ecology Resources*, 13(3), pp.494–509.

Petersen, T.N., Brunak, S., Heijne, von, G. and Nielsen, H., 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Publishing Group*, 8(10), pp.785–786.

Pierce, S.K., Fang, X., Schwartz, J.A., Jiang, X., Zhao, W., Curtis, N.E., Kocot, K.M., Yang, B. and Wang, J., 2012. Transcriptomic evidence for the expression of horizontally transferred algal nuclear genes in the photosynthetic sea slug, *Elysia chlorotica*. *Molecular Biology and Evolution*, 29(6), pp.1545–1556.

Piskurek, O. and Jackson, D.J., 2012. Transposable Elements: From DNA Parasites to Architects of Metazoan Evolution. *Genes*, 3(4), pp.409–422.

Piskurek, O. and Okada, N., 2007. Poxviruses as possible vectors for horizontal transfer of retroposons from reptiles to mammals. *Proceedings of the National Academy of Sciences of the United States of America*, 104(29), pp.12046–12051.

Podlaha, O. and Zhang, J., 2010. Pseudogenes and their evolution. *Encyclopaedia of Life Sciences (eLS)*, John Wiley & Sons, Ltd: Chichester.

Polz, M.F., Alm, E.J. and Hanage, W.P., 2013. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends in Genetics*, 29(3), pp.170–175.

Pongponratn, E., Maneerat, Y., Chaisri, U., Wilairatana, P., Punpoowong, B., Viriyavejakul, P. and Riganti, M., 1998. Electron-microscopic examination of *Rickettsia tsutsugamushi*-infected human liver. *Tropical Medicine & International Health*, 3(3), pp.242–248.

Porter, T.M. and Golding, G.B., 2011. Are similarity- or phylogeny-based methods more appropriate for classifying internal transcribed spacer (ITS) metagenomic amplicons? *New Phytologist*, 192(3), pp.775–782.

Prachumwat, A., DeVincentis, L. and Palopoli, M.F., 2004. Intron size correlates positively with recombination rate in *Caenorhabditis elegans. Genetics*, 166(3), pp.1585–1590.

Qiu, H., Yoon, H.S. and Bhattacharya, D., 2013. Algal endosymbionts as vectors of horizontal gene

transfer in photosynthetic eukaryotes. *Frontiers in Plant Science*, 4(366).

Quinlan, A.R. and Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6), pp.841–842.

R Core Team, 2014. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria.

Ra, H.-J. and Parks, W.C., 2007. Control of matrix metalloproteinase catalytic activity. *Matrix Biology*, 26(8), pp.587–596.

Ragan, M.A. and Beiko, R.G., 2009. Lateral genetic transfer: open issues. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1527), pp.2241–2251.

Raoult, D., 2010. The post-Darwinist rhizome of life. *The Lancet*, 375(9709), pp.104–105.

Rawlings, N.D., Barrett, A.J. and Finn, R.D., 2016. Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Research*, 44, pp.D343–D350.

Redrejo-Rodríguez, M., Muñoz-Espín, D., Holguera, I., Mencía, M. and Salas, M., 2012. Functional eukaryotic nuclear localization signals are widespread in terminal proteins of bacteriophages. *Proceedings of the National Academy of Sciences*, 109(45), pp.18482–18487.

Régnier, P. and Marujo, P.E., 2013. Polyadenylation and degradation of RNA in prokaryotes. *Madame Curie Bioscience Database*. Austin (TX), Landes Bioscience.

Reyes-Prieto, A. and Bhattacharya, D., 2007. Phylogeny of nuclear-encoded plastid-targeted proteins supports an early divergence of plaucophytes within Plantae. *Molecular Biology and Evolution*, 24(11), pp.2358–2361.

Ribet, D. and Cossart, P., 2015. How bacterial pathogens colonize their hosts and invade deeper tissues. *Microbes and Infection*, 17(3), pp.173–183.

Ricchetti, M., Tekaia, F. and Dujon, B., 2004. Continued colonization of the human genome by mitochondrial DNA. *PLoS Biology*, 2(9), e273.

Rice, P., Longden, I. and Bleasby, A., 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics*, 16(6), pp.276–277.

Richards GS., 2010. The origins of cell communication in the animal kingdom: Notch signalling during embryogenesis and metamorphosis of the demosponge *Amphimedon queenslandica*. University of Queensland.

Richards, T.A. and Monier, A., 2016. A tale of two tardigrades. *Proceedings of the National Academy of Sciences*, 113(18), pp.4892–4894.

Richards, T.A., Soanes, D.M., Jones, M.D.M., Vasieva, O., Leonard, G., Paszkiewicz, K., Foster, P.G., Hall, N. and Talbot, N.J., 2011. Horizontal gene transfer facilitated the evolution of plant parasitic mechanisms in the oomycetes. *Proceedings of the National Academy of Sciences*, 108(37), pp.15258–15263.

Richly, E. and Leister, D., 2004. NUMTs in sequenced eukaryotic genomes. *Molecular Biology and Evolution*, 21(6), pp.1081–1084.

Riesgo, A., Farrar, N., Windsor, P.J., Giribet, G., et al., 2014a. The analysis of eight transcriptomes from all poriferan classes reveals surprising genetic complexity in sponges. *Molecular Biology and Evolution*, 31(5), pp.1102–1120.

Riesgo, A., Peterson, K., Richardson, C., Heist, T., Strehlow, B., McCauley, M., Cotman, C., Hill, M. and Hill, A., 2014b. Transcriptomic analysis of differential host gene expression upon uptake of symbionts: a case study with Symbiodinium and the major bioeroding sponge *Cliona varians*. *BMC Genomics*, 15(376).

Riesgo, A., Maldonado, M., López-Legentil, S. and Giribet, G., 2015. A proposal for the evolution of cathepsin and silicatein in sponges. *Journal of Molecular Evolution*, 80(5), pp.278–291.

Rivera, M.C., Jain, R., Moore, J.E. and Lake, J.A., 1998. Genomic evidence for two functionally distinct gene classes. *Proceedings of the National Academy of Sciences of the United States of America*, 95(11), pp.6239–6244.

Robinson, K.M., Sieber, K.B. and Dunning Hotopp, J.C., 2013. A review of bacteria-animal lateral gene transfer may inform our understanding of diseases like cancer. *PLoS Genetics*, 9(10), e1003877.

Rogozin, I.B., Carmel, L., Csuros, M. and Koonin, E.V., 2012. Origin and evolution of spliceosomal introns. *Biology Direct*, 7, pp.1–28.

Rolando, M., Sanulli, S., Rusniok, C., Gomez-Valero, L., Bertholet, C., Sahr, T., Margueron, R. and Buchrieser, C., 2013. *Legionella pneumophila* effector RomA uniquely modifies host chromatin to repress gene expression and promote intracellular bacterial replication. *Cell Host and Microbe*, 13(4), pp.395–405.

Roosa, S., Wattiez, R., Prygiel, E., Lesven, L., Billon, G. and Gillan, D.C., 2014. Bacterial metal resistance genes and metal bioavailability in contaminated sediments. *Environmental Pollution*, 189, pp.143–151.

Ros, V.I. and Hurst, G.D., 2009. Lateral gene transfer between prokaryotes and multicellular eukaryotes: ongoing and significant? *BMC Biology*, 7(20).

Roth, L.E., 1957. An electron microscope study of the cytology of the protozoan *Euplotes patella*. *The Journal of Biophysical and Biochemical Cytology*, 3(6), pp.985–1000.

Routh, A., Domitrovic, T. and Johnson, J.E., 2012. Host RNAs, including transposons, are encapsidated by a eukaryotic single-stranded RNA virus. *Proceedings of the National Academy of Sciences*, 109(6), pp.1907–1912.

Rubartelli, A., Bajetto, A., Allavena, G., Wollman, E. anmd Sitia, R., 1992. Secretion of thioredoxin by normal and neoplastic cells through a leaderless secretory pathway. *The Journal of Biological Chemistry*, 267(34), pp.24161–24164.

Rubartelli, A., Cozzolino, F., Talio, M. and Sitia, R., 1990. A novel secretory pathway for interleukin-1 beta, a protein lacking a signal sequence. *The EMBO Journal*, 9(5), pp.1503–1510.

Rumpho, M.E., Pelletreau, K.N., Moustafa, A. and Bhattacharya, D., 2010. The making of a photosynthetic animal. *Journal of Experimental Biology*, 214(2), pp.303–311.

Rumpho, M.E., Worful, J.M., Lee, J., Kannan, K., Tyler, M.S., Bhattacharya, D., Moustafa, A. and Manhart, J.R., 2008. Horizontal gene transfer of the algal nuclear gene *psbO* to the photosynthetic sea slug *Elysia chlorotica*. *Proceedings of the National Academy of Sciences*, 105(46), pp.17867–17871.

Ryan, J.F., Pang, K., Schnitzler, C.E., Nguyen, A.D., Moreland, R.T., Simmons, D.K., Koch, B.J., Francis, W.R., Havlak, P., NISC Comparative Sequencing Program, Smith, S.A., Putnam, N.H., Haddock, S.H.D., Dunn, C.W., Wolfsberg, T.G., Mullikin, J.C., Martindale, M.Q. and Baxevanis, A.D., 2013. The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. *Science*, 342(6164), pp.1242592–1242592.

Rybarczyk-Mydłowska, K., Maboreke, H.R., Megen, H., Elsen, S., Mooyman, P., Smant, G., Bakker, J. and Helder, J., 2012. Rather than by direct acquisition via lateral gene transfer, GHF5 cellulases were passed on from early Pratylenchidae to root-knot and cyst nematodes. *BMC Evolutionary Biology*, 12(221).

Ryu, T., Seridi, L., Moitinho-Silva, L., Oates, M., Liew, Y.J., Mavromatis, C., Wang, X., Haywood, A., Lafi, F.F., Kupresanin, M., Sougrat, R., Alzahrani, M.A., Giles, E., Ghosheh, Y., Schunter, C., Baumgarten, S., Berumen, M.L., Gao, X., Aranda, M., Foret, S., Gough, J., Voolstra, C.R., Hentschel, U. and Ravasi, T., 2016. Hologenome analysis of two marine sponges with different microbiomes. *BMC Genomics*, 17(158), pp.1–11.

Saito, T., Dohmae, N., Tsujimoto, M. and Takio, K., 2002. PCR cloning and heterologous expression of cDNA encoding a peptidyl-Lys metalloendopeptidase precursor of *Grifola frondosa*. *The Journal of General and Applied Microbiology*, 48(5), pp.287–292.

Saito, K., 2013. The epigenetic regulation of transposable elements by PIWI-interacting RNAs in *Drosophila*. *Genes & Genetic Systems*, 88(1), pp.9–17.

Sakarya, O., Kosik, K.S. and Oakley, T.H., 2008. Reconstructing ancestral genome content based on symmetrical best alignments and Dollo parsimony. *Bioinformatics*, 24(5), pp.606–612.

Salvesen, G.S. and Abrams, J.M., 2004. Caspase activation – stepping on the gas or releasing the brakes? Lessons from humans and flies. *Oncogene*, 23(16), pp.2774–2784.

Salzberg, S.L., 2017. Horizontal gene transfer is not a hallmark of the human genome. *Genome Biology,* 18(85).

Salzberg, S.L., White, O., Peterson, J. and Eisen, J.A., 2001. Microbial genes in the human genome: lateral transfer or gene loss? *Science*, 292(5523), pp.1903–1906.

Schaack, S., Gilbert, C. and Feschotte, C., 2010. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends in Ecology & Evolution*, 25(9), pp.537–546.

Schneider, S.E. and Thomas, J.H., 2014. Accidental genetic engineers: horizontal sequence transfer from parasitoid wasps to their lepidopteran hosts. *PLoS ONE*, 9(10), e109446.

Schnoes, A.M., Brown, S.D., Dodevski, I. and Babbitt, P.C., 2009. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Computational Biology*, 5(12), e1000605.

Schönknecht, G., Chen, W.-H., Ternes, C.M., Barbier, G.G., Shrestha, R.P., Stanke, M., Bräutigam, A., Baker, B.J., Banfield, J.F., Garavito, R.M., Carr, K., Wilkerson, C., Rensing, S.A., Gagneul, D., Dickenson, N.E., Oesterhelt, C., Lercher, M.J. and Weber, A.P.M., 2013. Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science*, 339(6124), pp.1207–1210.

Schönknecht, G., Weber, A.P.M. and Lercher, M.J., 2014. Horizontal gene acquisitions by eukaryotes as drivers of adaptive evolution. *BioEssays*, 36(1), pp.9–20.

Schröder, H.C., Boreiko, A., Korzhev, M., Tahir, M.N., Tremel, W., Eckert, C., Ushijima, H., Müller, I.M. and Müller, W.E.G., 2006. Co-expression and functional interaction of silicatein with galectin: matrix-guided formation of siliceous spicules in the marine demosponge *Suberites domuncula*. *The Journal of Biological Chemistry*, 281(17), pp.12001–12009.

Schröder, H.C., Grebenjuk, V.A., Binder, M., Skorokhod, A., Batel, R., Hassanein, H. and Müller, W., 2004. Functional molecular biodiversity: assessing the immune status of two sponge populations (*Suberites domuncula*) on the molecular level. *Marine Ecology*, 25(2), pp.93–108.

Schröder, H.C., Natalio, F., Shukoor, I., Tremel, W., Schloßmacher, U., Wang, X. and Müller, W.E.G., 2007. Apposition of silica lamellae during growth of spicules in the demosponge *Suberites domuncula*: biological/biochemical studies and chemical/biomimetical confirmation. *Journal of Structural Biology*, 159(3), pp.325–334.

Schultes, N.P. and Szostak, J.W., 1991. A poly(dA.dT) tract is a component of the recombination initiation site at the ARG4 locus in *Saccharomyces cerevisiae*. *Molecular and Cellular Biology*, 11(1), pp.322–328.

Schulz, F. and Horn, M., 2015. Intranuclear bacteria: inside the cellular control center of eukaryotes. *Trends in Cell Biology*, 25(6), pp.339–346.

Schulz, F., Lagkouvardos, I., Wascher, F., Aistleitner, K., ek, R.K.S. and Horn, M., 2014. Life in an unusual intracellular niche: a bacterial symbiont infecting the nucleus of amoebae. 8(8), pp.1634–1644.

Schuster, L.N. and Sommer, R.J., 2012. Expressional and functional variation of horizontally acquired cellulases in the nematode *Pristionchus pacificus*. *Gene*, 506(2), pp.274–282.

Schwartz, J.A., Curtis, N.E. and Pierce, S.K., 2014. FISH labeling reveals a horizontally transferred algal (*Vaucheria litorea*) nuclear gene on a sea slug (*Elysia chlorotica*) chromosome. *The Biological bulletin*, 227(3), pp.300–312.

Schwelberger, H.G., Kohlwein, S.D. and Paltauf, F., 1989. Molecular cloning, primary structure and disruption of the structural gene of aldolase from *Saccharomyces cerevisiae*. *European Journal of Biochemistry*, 180(2), pp.301–308.

Schwenteit, J., Bogdanović, X., Fridjonsson, O.H., Aevarsson, A., Bornscheuer, U.T., Hinrichs, W. and Gudmundsdottir, B.K., 2013a. Toxoid construction of AsaP1, a lethal toxic aspzincin metalloendopeptidase of *Aeromonas salmonicida* subsp. *achromogenes*, and studies of its activity and processing. *Veterinary Microbiology*, 162(2-4), pp.687–694.

Schwenteit, J., Gram, L., Nielsen, K.F., Fridjonsson, O.H., Bornscheuer, U.T., Givskov, M. and Gudmundsdottir, B.K., 2011. Quorum sensing in *Aeromonas salmonicida* subsp. *achromogenes* and the effect of the autoinducer synthase AsaI on bacterial virulence. *Veterinary Microbiology*, 147(3-4), pp.389–397.

Schwenteit, J.M., Breithaupt, A., Teifke, J.P., Koppang, E.O., Bornscheuer, U.T., Fischer, U. and Gudmundsdottir, B.K., 2013b. Innate and adaptive immune responses of Arctic charr (*Salvelinus alpinus,* L.) during infection with *Aeromonas salmonicida* subsp. *achromogenes* and the effect of the AsaP1 toxin. *Fish and Shellfish Immunology*, 35(3), pp.866–873.

Shafeeq, S., Kuipers, O.P. and Kloosterman, T.G., 2013. The role of zinc in the interplay between pathogenic streptococci and their hosts. *Molecular Microbiology*, 88(6), pp.1047–1057.

Shapiro, L.R., Scully, E.D., Straub, T.J., Park, J., Stephenson, A.G., Beattie, G.A., Gleason, M.L., Kolter, R., Coelho, M.C., De Moraes, C.M., Mescher, M.C. and Zhaxybayeva, O., 2016. Horizontal gene acquisitions, mobile element proliferation, and genome decay in the host-restricted plant pathogen *Erwinia tracheiphila*. *Genome Biology and Evolution*, 8(3), pp.649–664.

Shelomi, M., Danchin, E.G.J., Heckel, D., Wipfler, B., Bradler, S., Zhou, X. and Pauchet, Y., 2016. Horizontal gene transfer of pectinases from bacteria preceded the diversication of stick and leaf insects. *Nature*, 6(26388).

Shimizu, K., Cha, J., Stucky, G.D. and Morse, D.E., 1998. Silicatein alpha: cathepsin L-like protein

in sponge biosilica. *Proceedings of the National Academy of Sciences of the United States of America*, 95(11), pp.6234–6238.

Shin, J.-E., Lin, C. and Lim, H.N., 2016. Horizontal transfer of DNA methylation patterns into bacterial chromosomes. *Nucleic Acids Research*, 44(9), pp.4460–4471.

Shin, W., Boo, S.M. and Fritz, L., 2003. Endonuclear bacteria in *Euglena hemichromata* (Euglenophyceae): a proposed pathway to endonucleobiosis. *Phycologia*, 42, pp.198–203.

Sieber, K.B., Bromley, R.E. and Dunning Hotopp, J.C.D., 2017. Lateral gene transfer between prokaryotes and eukaryotes. *Experimental Cell Research,* 358, pp.421-426.

Siguier, P., Gourbeyre, E. and Chandler, M., 2014. Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiology Reviews*, 38(5), pp.865–891.

Siguier, P., 2006. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Research*, 34(90001), pp.D32–D36.

Sillitoe, I., Lewis, T.E., Cuff, A., Das, S., Ashford, P., Dawson, N.L., Furnham, N., Laskowski, R.A., Lee, D., Lees, J.G., Lehtinen, S., Studer, R.A., Thornton, J. and Orengo, C.A., 2015. CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Research*, 43, pp.D376–D381.

Simion, P., Philippe, H., Baurain, D., Jager, M., Richter, D.J., Di Franco, A., Roure, B., Satoh, N., Quéinnec, E., Ereskovsky, A., Lapébie, P., Corre, E., Delsuc, F., King, N., Wörheide, G. and Manuel, M., 2017. A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Current Biology*, 27(7), pp.958–967.

Simpson, T.L., 1984. The cell biology of sponges. New York: Springer-Verlag.

Sinzelle, L., Izsvák, Z. and Ivics, Z., 2009. Molecular domestication of transposable elements: From detrimental parasites to useful host genes. *Cellular and Molecular Life Sciences*, 66(6), pp.1073–1093.

Sloan, D.B., Nakabachi, A., Richards, S., Qu, J., Murali, S.C., Gibbs, R.A. and Moran, N.A., 2014. Parallel histories of horizontal gene transfer facilitated extreme reduction of endosymbiont genomes in sap-feeding insects. *Molecular Biology and Evolution*, 31(4), pp.857–871.

Slotkin, R.K. and Martienssen, R., 2007. Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics*, 8(4), pp.272–285.

Smith, E.E. and Malik, H.S., 2009. The apolipoprotein L family of programmed cell death and immunity genes rapidly evolved in primates at discrete sites of host-pathogen interactions. *Genome Research*, 19(5), pp.850–858.

Smith, M.W., Feng, D.F. and Doolittle, R.F., 1992. Evolution by acquisition: the case for horizontal gene transfers. *Trends in Biochemical Sciences*, 17(12), pp.489–493.

Smit, AFA, Hubley, R and Green, P., 1996-2010. RepeatMasker Open-3.0.

Smit, AFA, Hubley, R., 2008−2015. RepeatModeler Open-1.0.

Sogabe, S., Nakanishi, N. and Degnan, B.M., 2016. The ontogeny of choanocyte chambers during metamorphosis in the demosponge *Amphimedon queenslandica*. *EvoDevo*, 7(6).

Song, S.U., Gerasimova, T., Kurkulos, M., Boeke, J.D. and Corces, V.G., 1994. An *env*-like protein encoded by a *Drosophila* retroelement: evidence that *gypsy* is an infectious retrovirus. *Genes & Development*, 8(17), pp.2046–2057.

Sormacheva, I., Smyshlyaev, G., Mayorov, V., Blinov, A., Novikov, A. and Novikova, O., 2012. Vertical evolution and horizontal transfer of CR1 Non-LTR retrotransposons and Tc1/mariner DNA transposons in *Lepidoptera* species. *Molecular Biology and Evolution*, 29(12), pp.3685–3702.

Soucy, S.M., Huang, J. and Gogarten, J.P., 2015. Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, 16(8), pp.472–482.

Srivastava, M., Simakov, O., Chapman, J., Fahey, B., Gauthier, M.E.A., Mitros, T., Richards, G.S., Conaco, C., Dacre, M., Hellsten, U., Larroux, C., Putnam, N.H., Stanke, M., Adamska, M., Darling, A., Degnan, S.M., Oakley, T.H., Plachetzki, D.C., Zhai, Y., Adamski, M., Calcino, A., Cummins, S.F., Goodstein, D.M., Harris, C., Jackson, D.J., Leys, S.P., Shu, S., Ben J Woodcroft, Vervoort, M., Kosik, K.S., Manning, G., Degnan, B.M. and Rokhsar, D.S., 2010. The *Amphimedon queenslandica* genome and the evolution of animal complexity. *Nature*, 466(7307), pp.720–726.

Stanke, M., Tzvetkova, A. and Morgenstern, B., 2006., AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biology*, 7, pp.S11.1–8.

Starcevic, A., Akthar, S., Dunlap, W.C., Shick, J.M., Hranueli, D., Cullum, J. and Long, P.F., 2008. Enzymes of the shikimic acid pathway encoded in the genome of a basal metazoan, *Nematostella vectensis*, have microbial origins. *Proceedings of the National Academy of Sciences of the United States of America*, 105(7), pp.2533–2537.

Steele, M.I., Kwong, W.K., Whiteley, M. and Moran, N.A., 2017. Diversification of type VI secretion system toxins reveals ancient antagonism among bee gut microbes. *mBio*, 8(6), pp.e01630–17.

Steele, R.E., David, C.N. and Technau, U., 2011. A genomic view of 500 million years of cnidarian evolution. *Trends in Genetics*, 27(1), pp.7–13.

Steindler, L., Schuster, S., Ilan, M., Avni, A., Cerrano, C. and Beer, S., 2007. Differential Gene Expression in a Marine Sponge in Relation to Its Symbiotic State. *Marine Biotechnology*, 9(5), pp.543–549.

Strand, D.J. and McDonald, J.F., 1985. *Copia* is transcriptionally responsive to environmental stress. *Nucleic Acids Research*, 13(12), pp.4401–4410.

Syvanen, M. and Kado, C.I., 2002. Horizontal gene transfer. *Elsevier Science,* 2nd edn.

Takeichi, M., 1988. The cadherins: cell-cell adhesion molecules controlling animal morphogenesis. *Development*, 102(4), pp.639–655.

Tanaka-Ichiara, K. and Watanabe, Y. , 1990. Gametogenic cycle of *Halichondria okadai*. In Rützler, K., (ed.), *New Perspectives in Sponge Biology,* Smithsonian Institution Press: Washington, DC. pp.170–174.

Tatum, E.L. and Lederberg, J., 1947. Gene Recombination in the Bacterium *Escherichia coli*. *Journal of Bacteriology*, 53(6), pp.673–684.

Taylor, H.E., Khatua, A.K. and Popik, W., 2014. The innate immune factor apolipoprotein L1 restricts HIV-1 infection. *Journal of Virology*, 88(1), pp.592–603.

Thacker, R.W., Hill, A.L., Hill, M.S., Redmond, N.E., Collins, A.G., Morrow, C.C., Spicer, L., Carmack, C.A., Zappe, M.E., Pohlmann, D., Hall, C., Diaz, M.C. and Bangalore, P.V., 2013. Nearly complete 28S rRNA gene sequences confirm new hypotheses of sponge evolution. *Integrative and Comparative Biology*, 53(3), pp.373–387.

Thomas, J., Schaack, S. and Pritham, E.J., 2010. Pervasive horizontal transfer of rolling-circle transposons among animals. *Genome Biology and Evolution*, 2, pp.656–664.

Thomas, J. and Pritham, E.J., 2015. Helitrons, the eukaryotic rolling-circle transposable elements. *Microbiology Spectrum*, 3(4).

Thomas, J., Phillips, C.D., Baker, R.J. and Pritham, E.J., 2014. Rolling-circle transposons catalyze genomic innovation in a mammalian lineage. *Genome Biology and Evolution*, 6(10), pp.2595–2610.

Tolosano, E., Fagoonee, S., Morello, N., Vinchi, F. and Fiorito, V., 2010. Heme scavenging and the other facets of hemopexin. *Antioxidants & Redox Signaling*, 12(2), pp.305–320.

Torruella, G., de Mendoza, A., Grau-Bové, X., Antó, M., Chaplin, M.A., del Campo, J., Eme, L., Pérez-Cordón, G., Whipps, C.M., Nichols, K.M., Paley, R., Roger, A.J., Sitjà-Bobadilla, A., Donachie, S. and Ruiz-Trillo, I., 2015. Phylogenomics reveals convergent evolution of lifestyles in close relatives of animals and fungi. *Current Biology*, 25(18), pp.2404–2410.

Toussaint, A. and Merlin, C., 2002. Mobile elements as a combination of functional modules. *Plasmid*, 47(1), pp.26–35.

Timmis, J.N., Ayliffe, M.A., Huang, C.Y. and Martin, W., 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nature Reviews Genetics*, 5(2), pp.123–135.

Treco, D. and Arnheim, N., 1986. The evolutionarily conserved repetitive sequence d(TG.AC)n promotes reciprocal exchange and generates unusual recombinant tetrads during yeast meiosis. *Molecular and Cellular Biology*, 6(11), pp.3934–3947.

True, J.R. and Carroll, S.B., 2002., Gene co-option in physiological and morphological evolution. *Annual Review of Cell And Developmental Biology*, 18, pp.53–80.

Tskhovrebova, L. and Trinick, J., 2004. Properties of titin immunoglobulin and fibronectin-3 domains. *Journal of Biological Chemistry,* 279(45), pp.46351-46354.

Tsurumi, M. and Reiswig, H.M., 1997. Sexual versus asexual reproduction in an oviparous rope-form sponge, *Aplysina cauliformis* (Porifera; Verongida). *Invertebrate Reproduction & Development*, 32(1), pp.1–9.

Turnbull, M. and Webb, B., 2002. Perspectives on polydnavirus origins and evolution. *Advances in Virus Research*, 58, pp.203–254.

Tzfira, T., 2004. *Agrobacterium* T-DNA integration: molecules and models. *Trends in Genetics*, 20(8), pp.375–383.

Tzfira, T., Rhee, Y., Chen, M.H., Kunik, T. and Citovsky, V., 2000. Nucleic acid transport in plant-microbe interactions: the molecules that walk through the walls. *Annual Review of Microbiology*, 54, pp.187–219.

Urakami, H., Tsuruhara, T. and Tamura, A., 1982. Intranuclear *Rickettsia tsutsugamushi* in cultured mouse fibroblasts (L cells). *Microbiology and Immunology*, 26(5), pp.445–447.

Uzureau, S., Coquerelle, C., Vermeiren, C., Uzureau, P., Van Acker, A., Pilotte, L., Monteyne, D., Acolty, V., Vanhollebeke, B., Van den Eynde, B., Pérez-Morga, D., Moser, M. and Pays, E., 2016. Apolipoproteins L control cell death triggered by TLR3/TRIF signaling in dendritic cells. *European Journal of Immunology*, 46(8), pp.1854–1866.

Vadillo, M.A., Konstantinidis, E. and Shanks, D.R., 2016. Underpowered samples, false negatives, and unconscious learning. *Psychonomic Bulletin & Review*, 23(1), pp.87–102.

Vanhamme, L., Paturiaux-Hanocq, F., Poelvoorde, P., Nolan, D.P., Lins, L., Van Den Abbeele, J., Pays, A., Tebabi, P., Van Xong, H., Jacquet, A., Moguilevsky, N., Dieu, M., Kane, J.P., De Baetselier, P., Brasseur, R. and Pays, E., 2003. Apolipoprotein L-I is the trypanosome lytic factor of human serum. *Nature*, 422(6927), pp.83–87.

Vanhollebeke, B. and Pays, E., 2006. The function of apolipoproteins L. *Cellular and Molecular Life Sciences*, 63(17), pp.1937–1944.

Vanhollebeke, B., Truc, P., Poelvoorde, P., Pays, A., Joshi, P.P., Katti, R., Jannin, J.G. and Pays, E., 2006. Human *Trypanosoma evansi* infection linked to a lack of apolipoprotein L-I. *The New England Journal of Medicine*, 355(26), pp.2752–2756.

Vicient, C.M., Kalendar, R. and Schulman, A.H., 2001. Envelope-class retrovirus-like elements are widespread, transcribed and spliced, and insertionally polymorphic in plants. *Genome Research*,

11(12), pp.2041–2049.

Vogel, C., Teichmann, S. A. and Chothia, C., 2003. The immunoglobulin superfamily in *Drosophila melanogaster* and *Caenorhabditis elegans* and the evolution of complexity. *Development,* 130(25), pp.6317-6328.

Volff, J.-N., 2006. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *BioEssays*, 28(9), pp.913–922.

Walsh, A.M., Kortschak, R.D., Gardner, M.G., Bertozzi, T. and Adelson, D.L., 2013. Widespread horizontal transfer of retrotransposons. *Proceedings of the National Academy of Sciences*, 110(3), pp.1012–1016.

Wang, D. and Timmis, J.N., 2013. Cytoplasmic organelle dna preferentially inserts into open chromatin. *Genome Biology and Evolution*, 5(6), pp.1060–1064.

Wang, X. and Müller, W., 2011. Complex structures–smart solutions: formation of siliceous spicules. *Communicative & Integrative Biology*, 4, pp.684–688.

Wang, X., Schloßmacher, U., Wiens, M., Batel, R., Schröder, H.C. and Müller, W.E.G., 2012a. Silicateins, silicatein interactors and cellular interplay in sponge skeletogenesis: formation of glass fiber-like spicules. *FEBS Journal*, 279(10), pp.1721–1736.

Wang, X., Schrder, H.C., Wiens, M., Schlomacher, U. and Muller, W.E.G., 2012b. Biosilica: molecular biology, biochemistry and function in demosponges as well as its applied. 1st ed. *Advances in Sponge Science: Physiology, Chemical and Microbial Diversity, Biotechnology*, Elsevier Ltd., pp.231–271.

Wang, X., Wiens, M., Schröder, H.C., Hu, S., Mugnaioli, E., Kolb, U., Tremel, W., Pisignano, D. and Müller, W.E.G., 2010. Morphology of sponge spicules: silicatein a structural protein for bio-silica formation. *Advanced Engineering Materials*, 12(9), pp.B422–B437.

Watanabe, K., Yukuhiro, F., Matsuura, Y., Fukatsu, T. and Noda, H., 2014. Intrasperm vertical symbiont transmission. *Proceedings of the National Academy of Sciences*, 111(20), pp.7433–7437.

Waters, V.L., 2001. Conjugation between bacterial and mammalian cells. *Nature Genetics*, 29(4), pp.375–376.

Webster, N.S., Taylor, M.W., Behnam, F., Lücker, S., Rattei, T., Whalan, S., Horn, M. and Wagner, M., 2010. Deep sequencing reveals exceptional diversity and modes of transmission for bacterial sponge symbionts. *Environmental Microbiology*, 12(8), pp.2070–2082.

Weinthal, D.M., Barash, I., Tzfira, T., Gaba, V., Teper, D., Sessa, G. and Manulis-Sasson, S., 2011. Characterization of nuclear localization signals in the type III effectors HsvG and HsvB of the gall-forming bacterium *Pantoea agglomerans. Microbiology*, 157(5), pp.1500–1508.

Wenger, Y. and Galliot, B., 2013. Punctuated emergences of genetic and phenotypic innovations in Eumetazoan, Bilaterian, Euteleostome, and Hominidae Ancestors. *Genome Biology and Evolution*, 5(10), pp.1949–1968.

Wernegreen, J.J., 2012. Strategies of genomic integration within insect-bacterial mutualisms. *The Biological Bulletin*, 223(1), pp.112–122.

Werren, J.H., Richards, S., Desjardins, C.A. and Niehuis, O., 2010. Functional and evolutionary insights from the genomes of three parasitoid *Nasonia* Species. *Science*, 327(5963), pp.340–343.

Westcott, S. and Barker, K.R., 1976. Interaction of *Acrobeloides buetsehlii* and *Rhizobium leguminosarum* on Wando pea. *Phytopathology*, 66, pp.468-472.

Wheeler, D., Redding, A.J. and Werren, J.H., 2013. Characterization of an ancient lepidopteran lateral gene transfer. *PLoS ONE*, 8(3), e59262.

Wheeler, B.S., 2013. Small RNAs, big impact: small RNA pathways in transposon control and their effect on the host stress response. *Chromosome Research*, 21(6-7), pp.587–600.

Whelan, N.V., Kocot, K.M., Moroz, L.L. and Halanych, K.M., 2015. Error, signal, and the placement of Ctenophora sister to all other animals. *Proceedings of the National Academy of Sciences*, 112(18), pp.5773–5778.

Whelan, N.V., Kocot, K.M., Moroz, T.P., Mukherjee, K., Williams, P., Paulay, G., Moroz, L.L. and Halanych, K.M., 2017. Ctenophore relationships and their placement as the sister group to all other animals. *Nature Ecology & Evolution*, 1, pp.1737–1746.

Whitaker, J.W., McConkey, G.A. and Westhead, D.R., 2009. Prediction of horizontal gene transfers in eukaryotes: approaches and challenges. *Biochemical Society Transactions*, 37(4), pp.792-795.

White, F.F., Garfinkel, D.J., Huffman, G.A., Gordon, M.P. and Nester, E.W., 1983. Sequences homologous to *Agrobacterium rhizogenes* T-DNA in the genomes of uninfected plants. *Nature*, 301, pp.348–350.

White, F.F., Ghidossi, G., Gordon, M.P. and Nester, E.W., 1982. Tumor induction by *Agrobacterium rhizogenes* involves the transfer of plasmid DNA to the plant genome. *Proceedings of the National Academy of Sciences of the United States of America*, 79(10), pp.3193–3197.

Wicker, T., Sabot, F., Hua-Van, A. and Bennetzen, J.L., 2007. A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, 8, pp.973–982.

Wiens, M., Schröder, H.C., Wang, X., Link, T., Steindorf, D. and Müller, W.E.G., 2011. Isolation of the silicatein-α interactor silintaphin-2 by a novel solid-phase pull-down assay. *Biochemistry*, 50(12), pp.1981–1990.

Wijayawardena, B.K., Minchella, D.J. and DeWoody, J.A., 2013. Hosts, parasites, and horizontal gene

transfer. *Trends in Parasitology*, 29(7), pp.329–338.

Wilkinson, C.R. and Garrone, R., 1980. Ultrastructure of siliceous spicules and microsclerocytes in the marine sponge *Neofibularia irata* N. SP. *Journal of Morphology*, 166, pp.51–64.

Wilkinson, T.L., Koga, R. and Fukatsu, T., 2007. Role of host nutrition in symbiont regulation: impact of dietary nitrogen on proliferation of obligate and facultative bacterial endosymbionts of the pea aphid *Acyrthosiphon pisum*. *Applied and Environmental Microbiology*, 73(4), pp.1362–1366.

Wolf, Y.I. and Koonin, E.V., 2013. Genome reduction as the dominant mode of evolution. *BioEssays*, 35(9), pp.829–837.

Wozniak, R.A.F. and Waldor, M.K., 2010. Integrative and conjugative elements: mosaic mobile genetic elements enabling dynamic lateral gene flow. *Nature Reviews Microbiology*, 8(8), pp.552–563.

Wybouw, N., Balabanidou, V., Ballhorn, D.J., Dermauw, W., Grbić, M., Vontas, J. and Van Leeuwen, T., 2012. A horizontally transferred cyanase gene in the spider mite *Tetranychus urticae* is involved in cyanate metabolism and is differentially expressed upon host plant change. *Insect Biochemistry and Molecular Biology*, 42(12), pp.881–889.

Wybouw, N., Dermauw, W., Tirry, L., Stevens, C., Grbić, M., Feyereisen, R. and Van Leeuwen, T., 2014. A gene horizontally transferred from bacteria protects arthropods from host plant cyanide poisoning. *eLife*, 3(e02365).

Wybouw, N., Pauchet, Y., Heckel, D.G. and Van Leeuwen, T., 2016. Horizontal gene transfer contributes to the evolution of arthropod herbivory. *Genome Biology and Evolution*, 8(6), pp.1785–1801.

Xiao, H., Jiang, N., Schaffner, E., Stockinger, E.J. and van der Knaap, E., 2008. A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science*, 319(5869), pp.1527–1530.

Xie, X., Kamal, M. and Lander, E.S., 2006. A family of conserved noncoding elements derived from an ancient transposable element. *Proceedings of the National Academy of Sciences of the United States of America*, 103(31), pp.11659–11664.

Yamada, M., Tokumitsu, N., Saikawa, Y., Nakata, M., Asano, J., Miyairi, K., Okuno, T., Konno, K. and Hashimoto, K., 2012. Molybdophyllysin, a toxic metalloendopeptidase from the tropical toadstool, *Chlorophyllum molybdites*. *Bioorganic & Medicinal Chemistry*, 20(22), pp.6583–6588.

Yang, N. and Kazazian, H.H., 2006. L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. *Nature Structural & Molecular Biology*, 13(9), pp.763–771.

Yanai, I., Wolf, Y.I. and Koonin, E.V., 2002. Evolution of gene fusions: horizontal transfer versus independent events. *Genome Biology*, 3(5).

Yoshiyama, M., Tu, Z., Kainoh, Y., Honda, H., Shono, T. and Kimura, K., 2001. Possible horizontal transfer of a transposable element from host to parasitoid. *Molecular Biology and Evolution*, 18(10), pp.1952–1958.

Yue, J., Sun, G., Hu, X. and Huang, J., 2013. The scale and evolutionary significance of horizontal gene transfer in the choanoflagellate Monosiga brevicollis. *BMC Genomics*, 14(729).

Zambryski, P., Tempe, J. and Schell, J., 1989. Transfer and function of T-DNA genes from *Agrobacterium* Ti and Ri plasmids in plants. *Cell*, 56(2), pp.193–201.

Zaneveld, J.R., Nemergut, D.R. and Knight, R., 2008. Are all horizontal gene transfers created equal? Prospects for mechanism-based studies of HGT patterns. *Microbiology*, 154(1), pp.1–15.

Zeh, D.W., Zeh, J.A. and Ishida, Y., 2009. Transposable elements and an epigenetic basis for punctuated equilibria. *BioEssays*, 31(7), pp.715–726.

Zhang, D., de Souza, R.F., Anantharaman, V., Iyer, L.M. and Aravind, L., 2012. Polymorphic toxin systems: Comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics. *Biology Direct*, 7(18).

Zhang, Z. and Saier, M.H., Jr, 2009. A novel mechanism of transposon-mediated gene activation. *PLoS Genetics*, 5(10), e1000689.

Zhaorigetu, S., Wan, G., Kaini, R., Jiang, Z. and Hu, C.-A.A., 2008. ApoL1, a BH3-only lipid-binding protein, induces autophagic cell death. *Autophagy*, 4(8), pp.1079–1082.

Zhaxybayeva, O. and Doolittle, W.F., 2011. Lateral gene transfer. *Current Biology*, 21(7), pp.R242–R246.

Zhou, Q. and Wang, W., 2008. On the origin and evolution of new genes—a genomic and experimental perspective. *Journal of Genetics and Genomics*, 35(11), pp.639–648.

Zhu, L., Hope, T.J., Hall, J., Davies, A., Stern, M., Muller-Eberhard, U., Stern, R. and Parslow, T.G., 1994. Molecular cloning of a mammalian hyaluronidase reveals identity with hemopexin, a serum heme-binding protein. *The Journal of Biological Chemistry*, 269(51), pp.32092–32097.

Zhu, B., Lou, M.-M., Xie, G.-L., Zhang, G.-Q., Zhou, X.-P., Li, B. and Jin, G.-L., 2011. Horizontal gene transfer in silkworm, *Bombyx mori*. *BMC Genomics*, 12, pp.1–9.

Zielinski, F.U., Pernthaler, A., Duperron, S., Raggi, L., Giere, O., Borowski, C. and Dubilier, N., 2009. Widespread occurrence of an intranuclear bacterial parasite in vent and seep bathymodiolin mussels. *Environmental Microbiology*, 11(5), pp.1150–1167.

Zinder, N.D. and Lederberg, J., 1952. Genetic exchange in *Salmonella*. *Journal of Bacteriology*, 64(5), pp.679-699.

Appendices

General note

Some of the appendices are too large to include in printed form; therefore, these files are available online at CloudStor+ via the link specified in the appendix. Files are also available on request from the author, Simone Higgie at simone.higgie@uqconnect.edu.au

Appendix 2.1 AqHGTs containing domains related to enzymes, informational genes and mobile elements

File available online at CloudStor+ (https://cloudstor.aarnet.edu.au/plus/index.php/s/GzHzoWIy8mfqT2l)

Appendix 2.2 The 234 Pfam domains present in only one of the 576 AqHGTs

Clan	Accession	Name
CL0004	PF00840	Glyco_hydro_7
CL0015	PF00854	PTR2
CL0016	PF06293	Kdo
CL0018	PF07716	bZIP_2
CL0020	PF00514	Arm
CL0020	PF03224	V-ATPase_H_N
CL0020	PF13424	TPR_12
CL0021	PF00366	Ribosomal_S17
CL0022	PF12799	LRR_4
CL0023	PF00071	Ras
CL0023	PF02463	SMC_N
CL0023	PF00488	MutS_V
CL0023	PF01712	dNK
CL0023	PF00004	AAA
CL0023	PF01715	IPPT
CL0023	PF01637	Arch_ATPase
CL0023	PF10443	RNA12
CL0028	PF08538	DUF1749
CL0028	PF01764	Lipase_3
CL0028	PF00756	Esterase
CL0029	PF13640	2OG-FeII_Oxy_3
CL0029	PF10014	2OG-Fe_Oxy_2
CL0029	PF13532	2OG-FeII_Oxy_2
CL0029	PF02668	TauD
CL0029	PF06172	Cupin_5
CL0031	PF00782	DSPc
CL0034	PF04909	Amidohydro_2
CL0036	PF00701	DHDPS
CL0040	PF00587	tRNA-synt_2b
CL0044	PF12902	Ferritin-like
CL0050	PF13622	4HBT_3
CL0058	PF00150	Cellulase
CL0058	PF02449	Glyco_hydro_42
CL0058	PF00128	Alpha-amylase
CL0059	PF05592	Bac_rhamnosid
CL0059	PF03663	Glyco_hydro_76
CL0059	PF07944	Glyco_hydro_127
CL0061	PF02347	GDC-P
CL0061	PF01041	DegT_DnrJ_EryC1
CL0061	PF00202	Aminotran_3
CL0062	PF13906	AA_permease_C
CL0062	PF03845	Spore_permease
CL0063	PF16363	GDP_Man_Dehyd
CL0063	PF07992	Pyr_redox_2
CL0063	PF01593	Amino_oxidase
CL0063	PF00670	AdoHcyase_NAD
CL0063	PF05185	PRMT5
CL0063	PF01225	Mur_ligase
CL0063	PF08242	Methyltransf_12
CL0066	PF00340	IL1
CL0066	PF14200	RicinB_lectin_2
CL0066	PF05270	AbfB
CL0071	PF00300	His_Phos_1
CL0072	PF00788	RA
CL0078	PF09414	RNA_ligase
CL0088	PF01663	Phosphodiest
CL0092	PF00241	Cofilin_ADF
CL0098	PF02590	SPOUT_MTase
CL0099	PF00171	Aldedh
CL0110	PF00535	Glycos_transf_2
CL0110	PF01697	Glyco_transf_92
CL0110	PF03407	Nucleotid_trans
CL0115	PF04191	PEMT
CL0116	PF07137	VDE
CL0118	PF00294	PfkB
CL0123	PF08100	Dimerisation
CL0123	PF05225	HTH_psq
CL0123	PF13565	HTH_32
CL0123	PF13551	HTH_29
CL0124	PF00089	Trypsin
CL0125	PF13529	Peptidase_C39_2
CL0126	PF01400	Astacin
CL0131	PF07291	MauE
CL0131	PF07681	DoxX
CL0159	PF01345	DUF11
CL0172	PF10865	DUF2703
CL0172	PF01323	DSBA
CL0184	PF02537	CRCB
CL0184	PF02535	Zip
CL0186	PF01011	PQQ
CL0186	PF08450	SGL
CL0186	PF01436	NHL
CL0186	PF13854	Kelch_5
CL0186	PF13418	Kelch_4
CL0186	PF13360	PQQ_2
CL0194	PF03175	DNA_pol_B_2
CL0196	PF00035	dsrm
CL0197	PF02274	Amidinotransf
CL0202	PF08531	Bac_rhamnosid_N
CL0214	PF00627	UBA
CL0219	PF01612	DNA_pol_A_exo1
CL0221	PF14259	RRM_6
CL0229	PF01485	IBR
CL0236	PF14281	PDDEXK_4
CL0246	PF04223	CitF
CL0257	PF13718	GNAT_acetyltr_2
CL0257	PF00583	Acetyltransf_1
CL0257	PF13673	Acetyltransf_10
CL0263	PF13391	HNH_2
CL0266	PF00169	PH
CL0267	PF00411	Ribosomal_S11
CL0268	PF07602	DUF1565
CL0270	PF00180	Iso_dh
CL0272	PF00615	RGS
CL0291	PF05168	HEPN
CL0295	PF08700	Vps51
CL0329	PF03764	EFG_IV
CL0329	PF00288	GHMP_kinases_N
CL0347	PF04103	CD20
CL0361	PF12756	zf-C2H2_2
CL0369	PF09260	DUF1966
CL0369	PF16561	AMPK1_CBM
CL0378	PF00501	AMP-binding
CL0381	PF00753	Lactamase_B
CL0387	PF00186	DHFR_1
CL0390	PF01363	FYVE
CL0403	PF06314	ADC
CL0431	PF00235	Profilin
CL0479	PF13091	PLDc_2
CL0496	PF13400	Tad
CL0497	PF04399	Glutaredoxin2_C
CL0511	PF00098	zf-CCHC
CL0523	PF13961	DUF4219
CL0533	PF00156	Pribosyltran
CL0541	PF00017	SH2
CL0556	PF07610	DUF1573
CL0570	PF05922	Inhibitor_I9
CL0575	PF00297	Ribosomal_L3
No_clan	PF00669	Flagellin_N
No_clan	PF02183	HALZ
No_clan	PF11021	DUF2613
No_clan	PF13776	DUF4172
No_clan	PF13082	DUF3931
No_clan	PF07888	CALCOCO1
No_clan	PF06337	DUSP
No_clan	PF02048	Enterotoxin_ST
No_clan	PF05436	MF_alpha_N
No_clan	PF02268	TFIIA_gamma_N
No_clan	PF15303	RNF111_N
No_clan	PF16559	GIT_CC
No_clan	PF03479	DUF296
No_clan	PF02816	Alpha_kinase
No_clan	PF14283	DUF4366
No_clan	PF03359	GKAP
No_clan	PF01765	RRF
No_clan	PF15198	Dexa_ind
No_clan	PF04211	MtrC
No_clan	PF04051	TRAPP
No_clan	PF14253	AbiH
No_clan	PF05699	Dimer_Tnp_hAT
No_clan	PF10504	DUF2452
No_clan	PF01437	PSI
No_clan	PF06585	JHBP
No_clan	PF11337	DUF3139
No_clan	PF08472	S6PP_C
No_clan	PF05190	MutS_IV
No_clan	PF05192	MutS_III
No_clan	PF08251	Mastoparan_2
No_clan	PF06626	DUF1152
No_clan	PF08544	GHMP_kinases_C
No_clan	PF12213	Dpoe2NT
No_clan	PF12588	PSDC
No_clan	PF16562	HECW_N
No_clan	PF15169	DUF4564
No_clan	PF03476	MOSC_N
No_clan	PF03473	MOSC
No_clan	PF16124	RecQ_Zn_bind
No_clan	PF14619	SnAC
No_clan	PF11899	DUF3419
No_clan	PF04878	Baculo_p48
No_clan	PF07102	DUF1364
No_clan	PF16905	GPHH
No_clan	PF10880	DUF2673
No_clan	PF15604	Ntox15
No_clan	PF01442	Apolipoprotein
No_clan	PF05635	23S_rRNA_IVP
No_clan	PF14326	DUF4384
No_clan	PF11932	DUF3450
No_clan	PF00791	ZU5
No_clan	PF08719	DUF1768
No_clan	PF09730	BicD
No_clan	PF05837	CENP-H
No_clan	PF07200	Mod_r
No_clan	PF00252	Ribosomal_L16
No_clan	PF01847	VHL
No_clan	PF01504	PIP5K
No_clan	PF16115	DUF4831
No_clan	PF00122	E1-E2_ATPase
No_clan	PF04325	DUF465
No_clan	PF02140	Gal_Lectin
No_clan	PF05478	Prominin
No_clan	PF01608	I_LWEQ
No_clan	PF04513	Baculo_PEP_C
No_clan	PF01290	Thymosin
No_clan	PF05545	FixQ
No_clan	PF00088	Trefoil
No_clan	PF16511	FERM_f0
No_clan	PF09733	VEFS-Box
No_clan	PF16477	DUF5054
No_clan	PF08194	DIM
No_clan	PF00244	14_3_3_protein
No_clan	PF01425	Amidase
No_clan	PF16866	PHD_4
No_clan	PF11831	Myb_Cef
No_clan	PF00687	Ribosomal_L1
No_clan	PF14291	DUF4371
No_clan	PF15872	SRTM1
No_clan	PF08939	DUF1917
No_clan	PF00795	CN_hydrolase
No_clan	PF06221	zf-C2HC5
No_clan	PF11402	Antifungal_prot
No_clan	PF00082	Peptidase_S8
No_clan	PF07423	DUF1510
No_clan	PF08245	Mur_ligase_M
No_clan	PF02875	Mur_ligase_C
No_clan	PF16888	DUF5082
No_clan	PF08289	Flu_M1_C
No_clan	PF04117	Mpv17_PMP22
No_clan	PF05915	DUF872
No_clan	PF04156	IncA
No_clan	PF03698	UPF0180
No_clan	PF10590	PNPOx_C
No_clan	PF12794	MscS_TM
No_clan	PF15102	TMEM154
No_clan	PF10153	DUF2361
No_clan	PF05960	DUF885
No_clan	PF10495	PACT_coil_coil
No_clan	PF14979	TMEM52
No_clan	PF00412	LIM
No_clan	PF06092	DUF943
No_clan	PF03832	WSK
No_clan	PF11304	DUF3106
No_clan	PF16922	SLD5_C

Appendix 2.3 Differences in domain numbers and diversity between expressed and unexpressed AqHGTs

	Expressed AqHGTs (n=350)		Non expressed AqHGTs (n=226)		TOTALS
	Observed (no.)	Expected frequency	Observed (no.)	Expected frequency	
AqHGTs with at least one Pfam domain	329	315	190	204	519
AqHGTs with no Pfam domain	21	35	36	22	57
TOTALS	350	-	226	-	576
Domains (total no.)	540	-	270	-	810
Domain types (diversity)	271	-	122	-	350

Expression data is from a published genome-wide ontogenetic transcript dataset where expression was measured by CEL-Seq (Anavy et al. 2014; Levin et al. 2016). If a gene has at least one developmental stage with a normalised count of at least five, then it is considered expressed.

Appendix 2.4 Domain promiscuity and domain architecture conservation

Accession	Name	AqHGTs with at least one domain (no.)	Domain architectures in Pfam (no.)	Seqs. in Pfam (no.)	Ratio of architectures to seqs. in Pfam	Shared domain architectures in AqHGTs and Pfam	Unique architectures in AqHGTs and not in Pfam	% of shared domain architectures in AqHGTs
PF14521	Aspzincin_M35	59	11	275	1:25	1	6	14
PF07727	RVT_2	25	1415	2652	1:2	2	5	29
PF01048	PNP_UDP_1	25	225	6749	1:30	1	5	17
PF00665	rve	15	610	12 778	1:21	3	2	60
PF00271	Helicase_C	13	1484	55 624	1:37	3	4	43
PF15466	DUF4635	13	1	34	1:34	0	1	0
PF13358	DDE_3	12	109	3739	1:34	4	0	100
PF14214	Helitron_like_N	11	67	1397	1:21	1	3	25
PF02221	E1_DerP2_DerF2	10	28	1379	1:49	1	1	50
PF00069	Pkinase	9	4424	125 183	1:28	3	4	43
PF00531	Death	9	646	3298	1:5	0	2	0
PF00270	DEAD	8	1007	55 428	1:55	2	1	67
PF13847	Methyltransf_31	7	188	2449	1:13	1	0	100
PF04970	LRAT	7	20	1192	1:60	1	3	25
PF03160	Calx-beta	7	559	7291	1:13	1	1	50

An evaluation of Pfam domain promiscuity and the level of domain architecture conservation in the AqHGTs containing the most common 15 domains identified in the AqHGTs. The percentage of shared domain combinations found in these genes with those reported in the Pfam database ranges with a lower quartile of 21%, median 43%, upper quartile 55%.

Appendix 2.5 The ontogenetic expression profiles and domain architecture of the AqHGTs predicted to contain the 15 most common domains of the AqHGTs

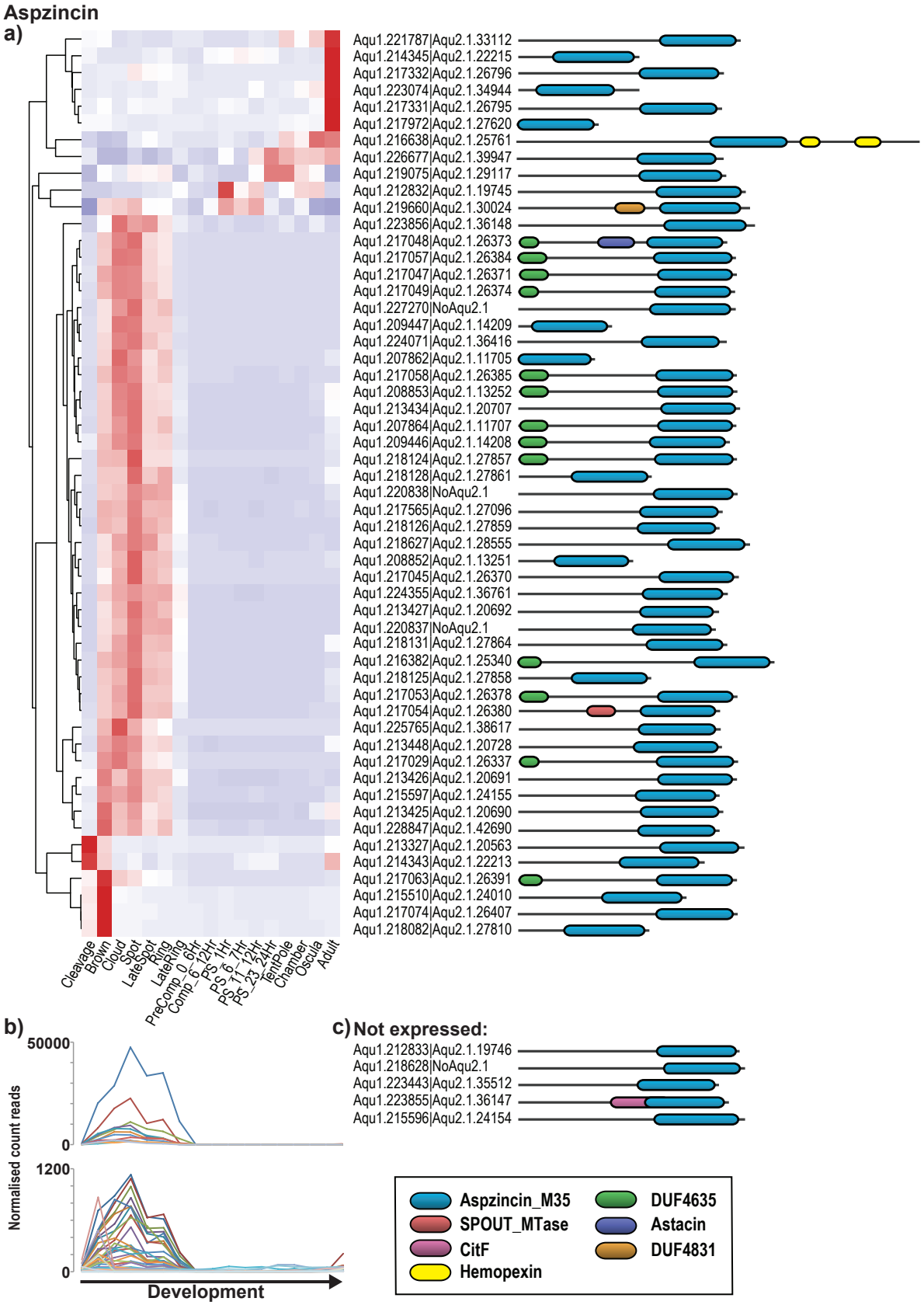
(7 parts)

(a) The expression profiles of gene models with at least one developmental stage with a normalised CEL-Seq count of at least five, with domain architectures shown on the right. Heat map scaled by row with genes organised by Euclidean distances in R Studio, Pretty Heatmaps package v 0.7.7. (b) The range in expression quantities of the genes shown in (a); note for the aspzincins, for scale purposes, the more highly expressed genes are presented in the top graph. (c) Shows the predicted domain architectures of the AqHGTs of the top domain groups that have normalised CEL-Seq counts of less than five for all developmental stages. Domains predicted by Pfam Batch Search and visualised in DoMosacis (Moore et al. 2014).

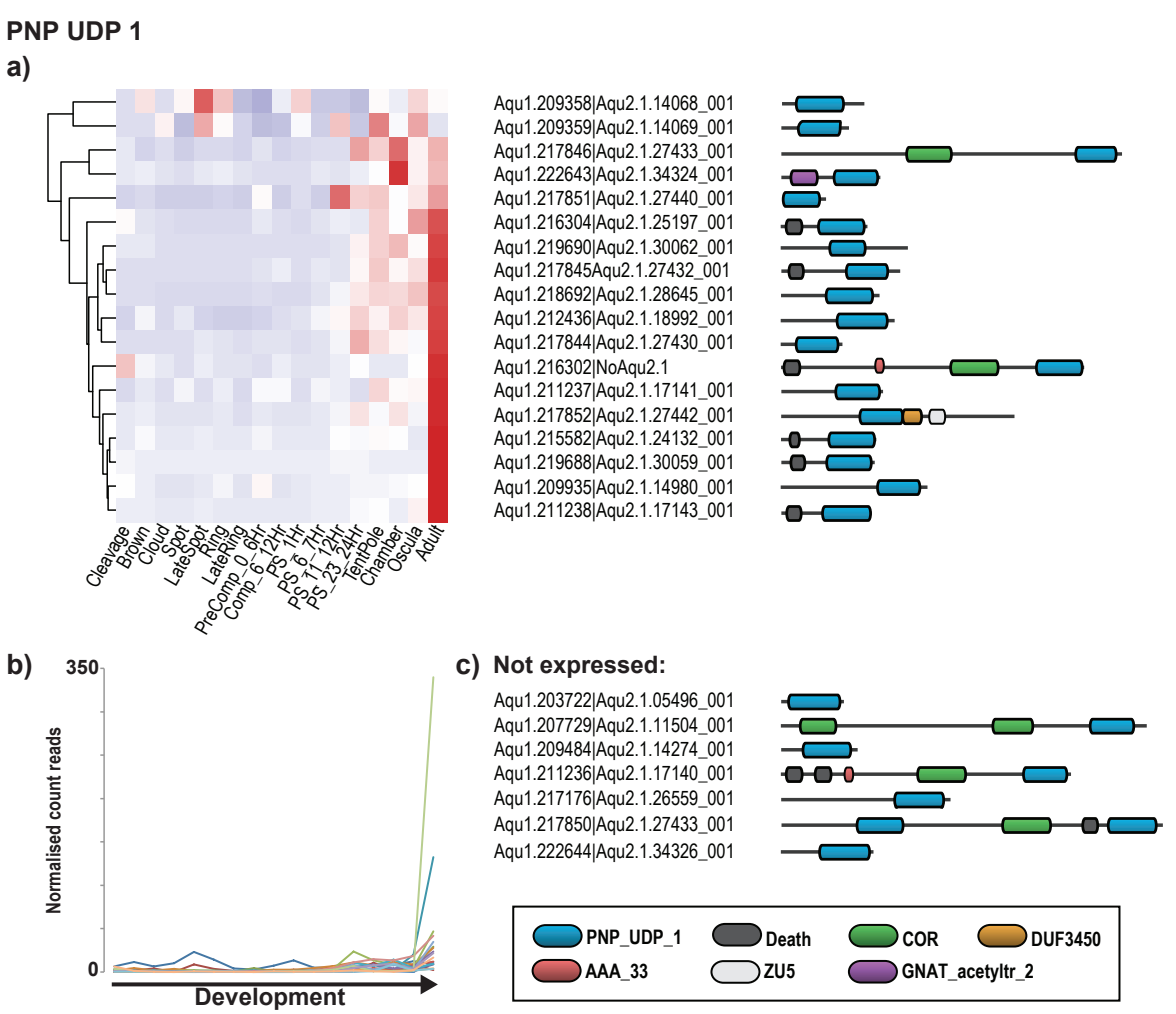
Note that for visual purposes, the Aqu2.1 gene model names are missing their common suffix "_001". Also, all 13 AqHGT members of domain group "DUF4635" are also aspzincins; these genes are all expressed and thus are all represented once only along with the other aspzincins. Further, all but one of the AqHGTs members of another domain group "Death" are also AqHGT_PNPs; these genes are shown with the AqHGT_PNPs and are not repeated later. The one AqHGT_Death that does not contain a PNP domain also contains a LRAT domain and is shown with the other LRATs.

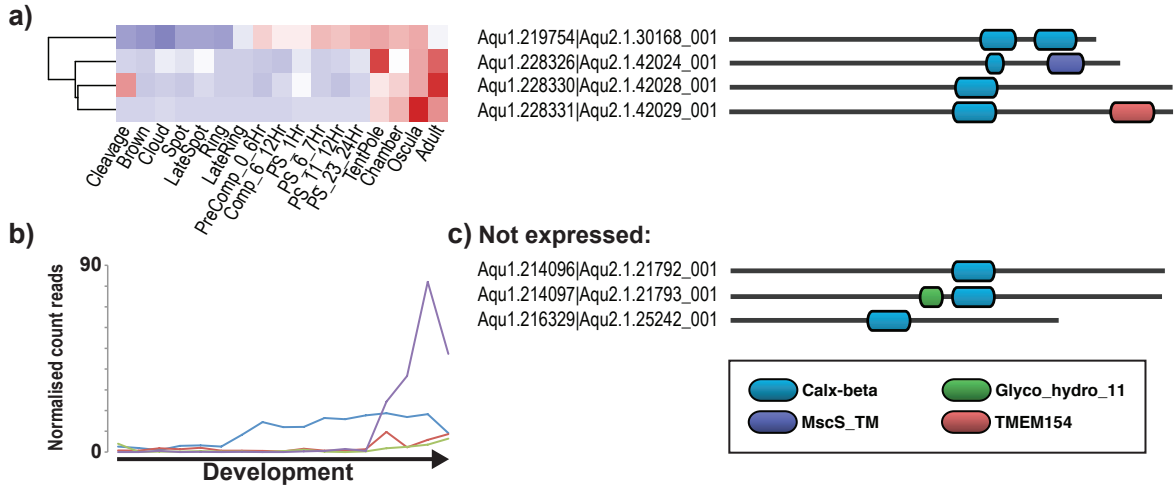
The developmental stages labelled on the heat maps are also appropriate to the x-axis of the graphs in (b). Embryogenesis occurs from "Cleavage" to the "Late Ring" stages. From this point, free-swimming larvae are released from the maternal brooding chamber but are not yet competent for settlement on a substrate preceding metamorphosis ("PreComp_0_6Hr"); typically they become competent 6-12 hours post emergence from the mother ("Comp_6_12Hr"). Following this are six time points of postlarval stages through which individuals settle on a substrate and undergo metamorphosis (stages abbreviated: 1 hour post settlement as "PS_1Hr", 6-7 hours post settlement as "PS_6_7Hr", 11-12 hours post settlement as "PS_11_12Hr", 23-24 hours post settlement as "PS_23_24Hr", tent pole as "TentPole", and chamber as "Chamber"). Finally, individuals become juveniles at the "Oscula" stage and are fully developed by "Adult" stage.

Appendix 2.5 The ontogenetic expression profiles and domain architecture of the AqHGTs predicted to contain the 15 most common domains of the AqHGTs *(Part 1 of 7)*

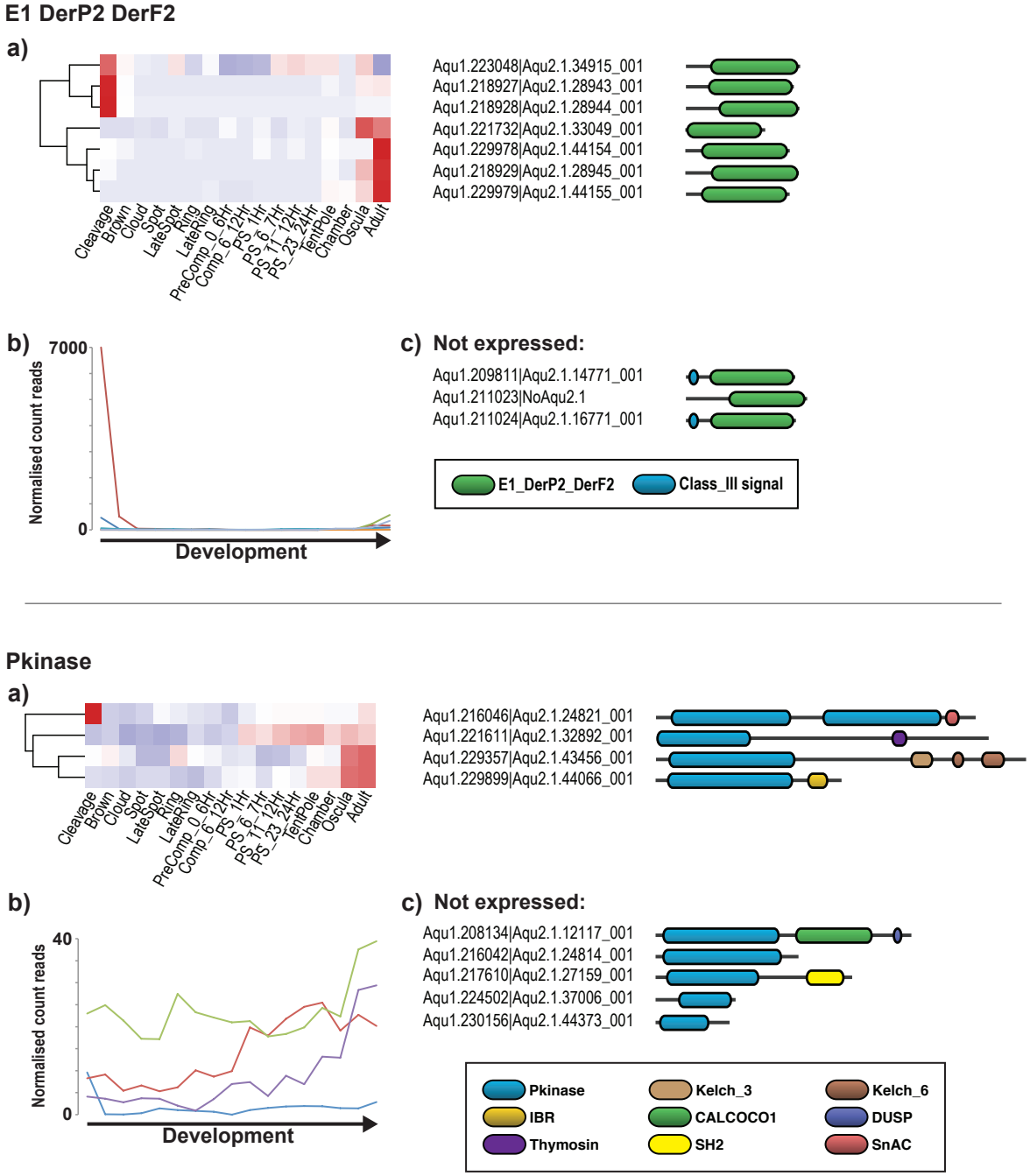


Appendix 2.5 The ontogenetic expression profiles and domain architecture of the AqHGTs predicted to contain the 15 most common domains of the AqHGTs *(Part 2 of 7)*

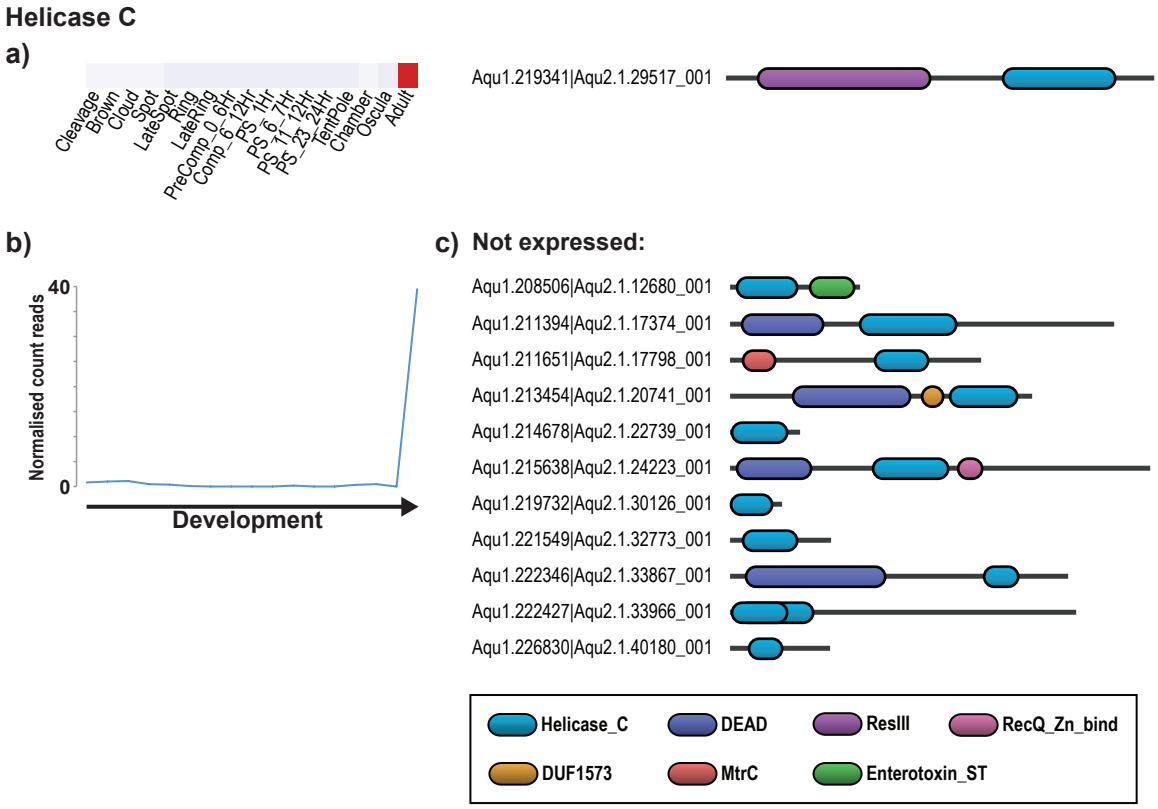


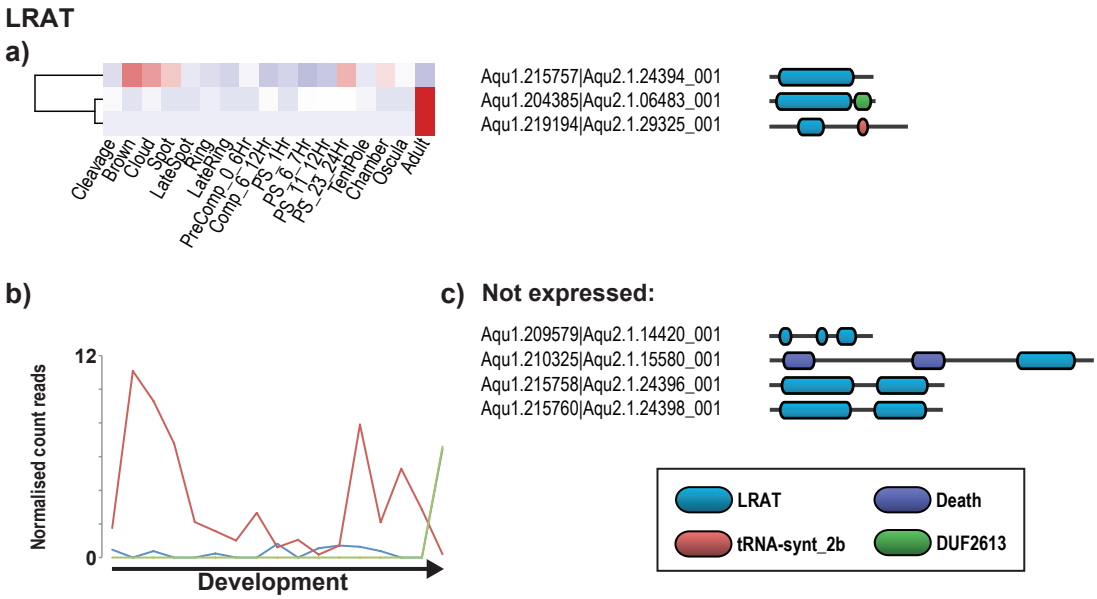


Appendix 2.5 The ontogenetic expression profiles and domain architecture of the AqHGTs predicted to contain the 15 most common domains of the AqHGTs *(Part 3 of 7)*

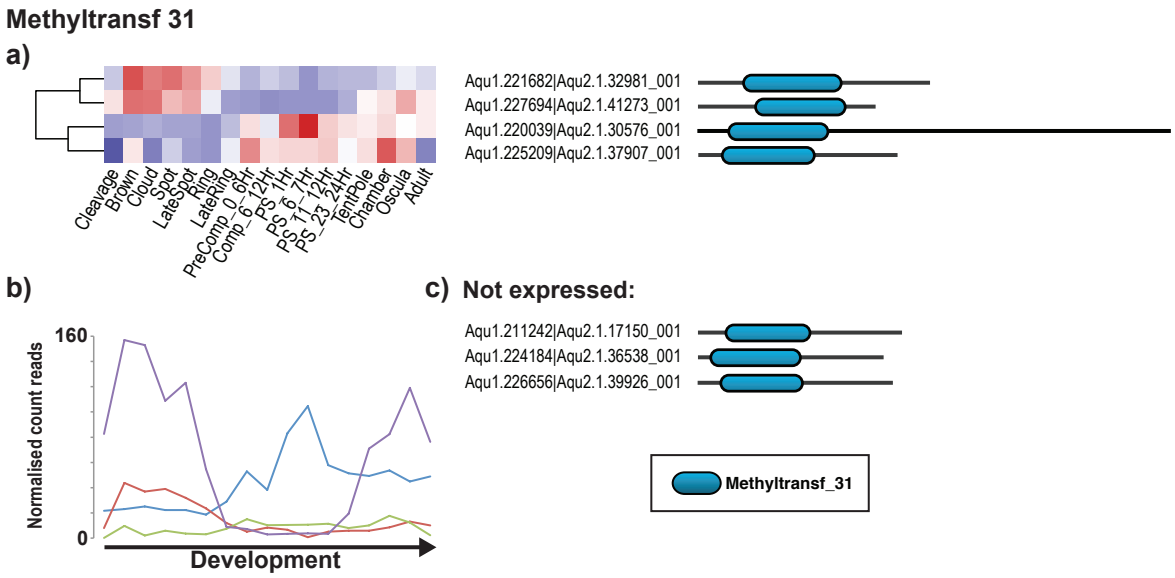


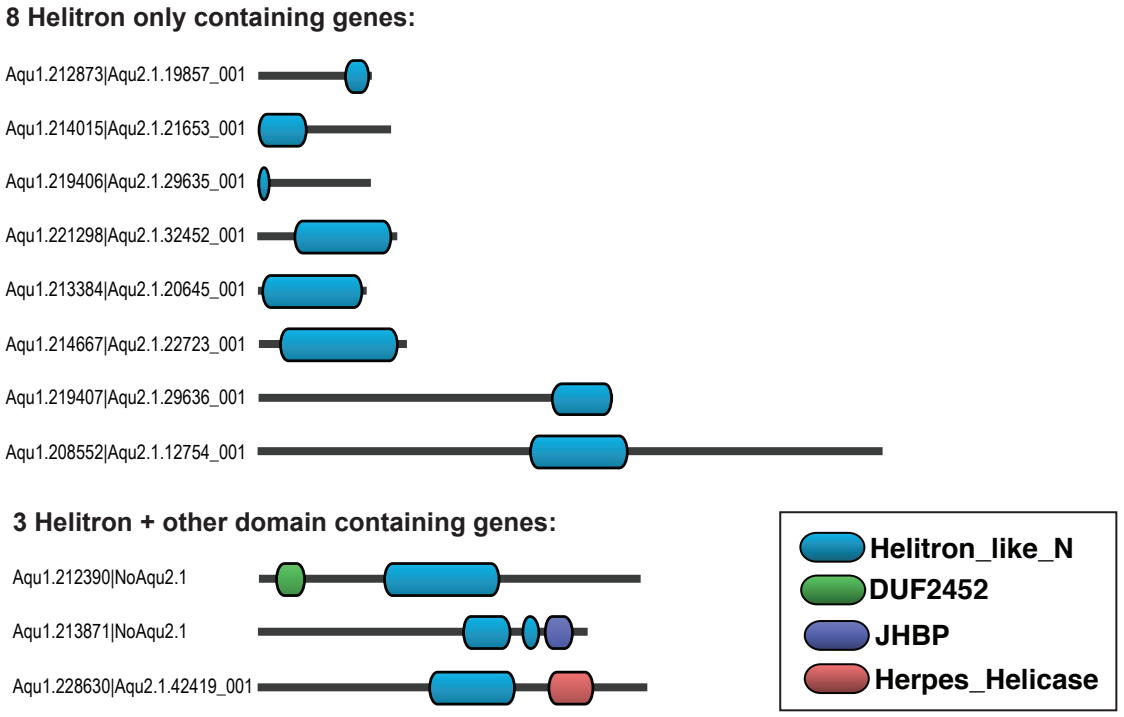
Appendix 2.5 The ontogenetic expression profiles and domain architecture of the AqHGTs predicted to contain the 15 most common domains of the AqHGTs *(Part 4 of 7)*



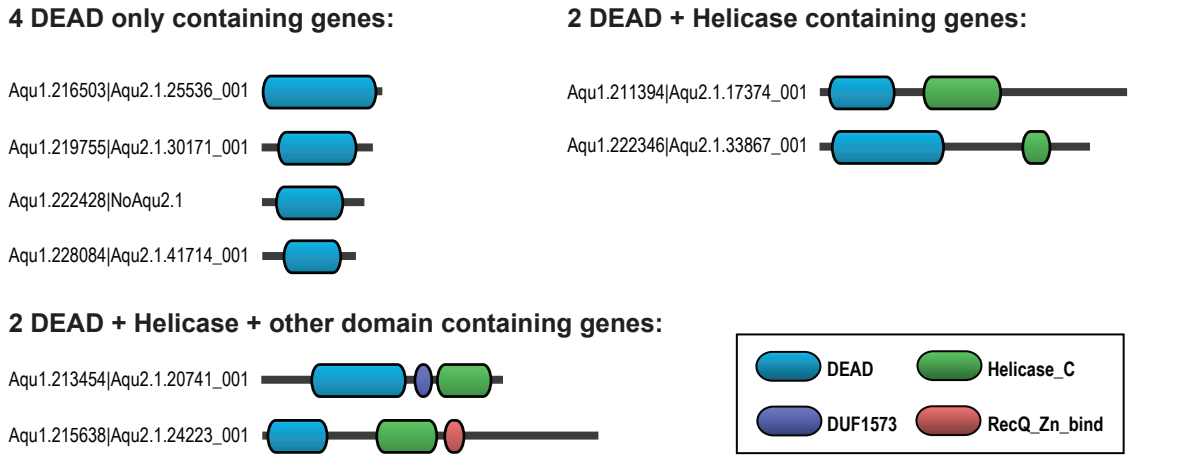


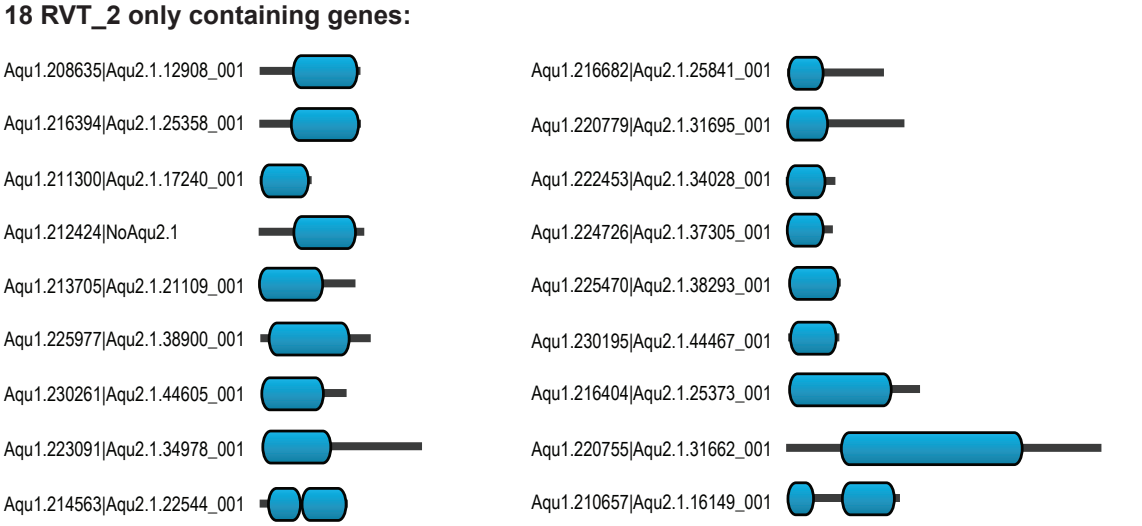
Appendix 2.5 The ontogenetic expression profiles and domain architecture of the AqHGTs predicted to contain the 15 most common domains of the AqHGTs *(Part 5 of 7)*

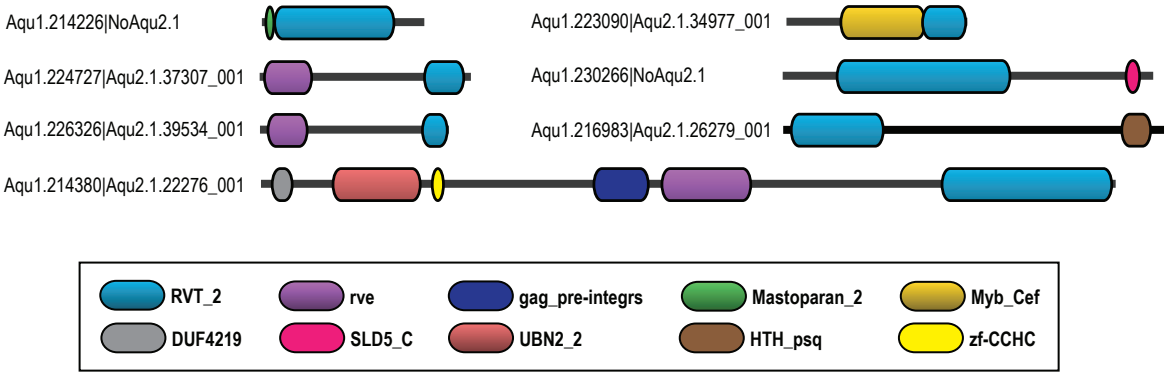




Appendix 2.5 The ontogenetic expression profiles and domain architecture of the AqHGTs predicted to contain the 15 most common domains of the AqHGTs *(Part 6 of 7)*







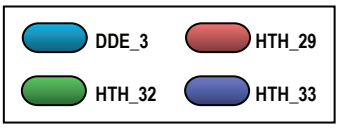
Appendix 2.5 The ontogenetic expression profiles and domain architecture of the AqHGTs predicted to contain the 15 most common domains of the AqHGTs *(Part 7 of 7)*

9 DDE_3 only containing genes:

Aqu1.210941|Aqu2.1.16634_001
Aqu1.216229|Aqu2.1.25068_001
Aqu1.218182|Aqu2.1.27947_001
Aqu1.221051|Aqu2.1.32063_001
Aqu1.229117|Aqu2.1.43155_001
Aqu1.216824|Aqu2.1.26055_001
Aqu1.219400|Aqu2.1.29623_001
Aqu1.224380|Aqu2.1.36788_001
Aqu1.221880|Aqu2.1.33317_001

3 DDE_3 + helix-turn-helix domain containing genes:

Aqu1.218299|Aqu2.1.28119_001
Aqu1.219378|Aqu2.1.29579_001
Aqu1.222910|Aqu2.1.34741_001



---

7 rve only containing genes:

Aqu1.224765|Aqu2.1.37403_001
Aqu1.214227|Aqu2.1.22027_001
Aqu1.211301|Aqu2.1.17241_001
Aqu1.214782|Aqu2.1.22896_001
Aqu1.207908|Aqu2.1.11776_001
Aqu1.226315|Aqu2.1.39514_001
Aqu1.220780|Aqu2.1.31696_001

4 gag_pre-integrs + rve containing genes:

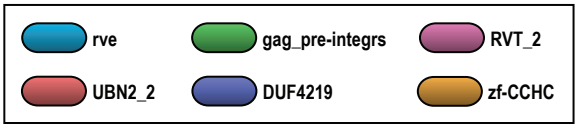
Aqu1.212498|Aqu2.1.19126_001
Aqu1.228616|Aqu2.1.42383_001
Aqu1.216982|NoAqu2.1
Aqu1.214884|Aqu2.1.23076_001

2 rve + RVT_2 containing genes:

Aqu1.226326|Aqu2.1.39534_001
Aqu1.224727|Aqu2.1.37307_001

2 rve + other domain containing genes:

Aqu1.224001|Aqu2.1.36306_001
Aqu1.214380|Aqu2.1.22276_001



Appendix 2.6 A taxonomic summary of the bacterial and fungal unique hits of the AqAspz BLASTp searching

	Phylum/Division	Class	Order
Fungi	Basidiomycota (n=34, 26%)	Agaricomycetes (n=34)	Agaricales (n=10)
			Auriculariales (n=1)
			Cantharellales (n=18)
			Russulales (n=1)
			Corticiales (n=2)
			Sebacinales (n=1)
			Polyporales (n=1)
	Ascomycota (n=6, 4.7%)	Sordariomycetes (n=2)	Hypocreales (n=2)
		Eurotiomycetes (n=4)	Onygenales (n=4)
Bacteria	Proteobacteria (n=77; 60%)	Gammaproteobacteria (n=48)	Xanthomonadales (n=10)
			Aeromonadales (n=13)
			Chromatiales (n=2)
			Oceanospirillales (n=1)
			Unclassified (n=2)
			Pseudomonadales (n=5)
			Vibrionales (n=7)
			Alteromonadales (n=8)
		Betaproteobacteria (n=17)	Burkholderiales (n=7)
			Neisseriales (n=10)
		Alphaproteobacteria (n=9)	Rhizobiales (n=4)
			Rhodobacterales (n=5)
		Deltaproteobacteria (n=3)	Myxococcales (n=3)
	Actinobacteria (n=10; 7.8%)	Actinobacteria (n=10)	Micromonosporales (n=7)
			Pseudonocardiales (n=1)
			Streptomycetales (n=2)
	Cyanobacteria (n=1; 0.8%)	Cyanobacteria (n=1)	Oscillatoriales (n=1)
	Bacteroidetes (n=1; 0.8%)	Chitinophagia (n=1)	Chitinophagales (n=1)

The top five hits of each AqAspz domain were collected from the 2017 NCBI nr database. Duplicates were removed.

Appendix 2.7 The taxonomy of the nonmetazoan sequences used for the aspzincin phylogenetic analysis

	Phylum/ Division	Class	Order	Species (accession)
Fungi	Basidiomycota	Agaricomycetes	Cantharellales	*Rhizoctonia solani* (ELU39650)
				Rhizoctonia solani (ELU44797)
				Rhizoctonia solani (CCO27676)
			Agaricales	*Pleurotus ostreatus* (KDQ24143)
			Corticiales	*Punctularia strigosozonata* (XP_007386987)
			Sebacinales	*Serendipita indica* (CCA71303)
	Ascomycota	Sordariomycetes	Hypocreales	*Drechmeria coniospora* (ODA82724)
Bacteria	Proteobacteria	Gammaproteobacteria	Vibrionales	*Vibrio aerogenes* (WP_073604269)
				Vibrio aerogenes (WP_073604268)
				Vibrio rhizosphaerae (WP_038177771)
			Chromatiales	*Chromatiales* bacterium (OED36449)
				Rheinheimera texasensis (WP_031571125)
			Pseudomonadales	*Pseudomonas fluorescens* (WP_047289146)
				Pseudomonas sp. ABAC61 (WP_058436019)
			Aeromonadales	*Aeromonas caviae* (ZP_08521271)
			Alteromonadales	*Shewanella loihica* (WP_011865045)
			Oceanospirillales	*Gynuella sunshinyii* (WP_044619391)
			Xanthomonadales	*Xanthomonas vesicatoria* (ZP_08179569)
		Alphaproteobacteria	Rhizobiales	*Agrobacterium arsenijevicii* (WP_045023548)
				Mesorhizobium australicum (WP_041163262)
			Rhodobacterales	*Roseobacter sp.* CCS2 (ZP_01751585)
		Betaproteobacteria	Burkholderiales	*Chitinimonas koreensis* (WP_028446313)
				Collimonas fungivorans (WP_014007414)
				Massilia sp. (WP_056341993)
			Neisseriales	*Chromobacterium violaceum* (WP_011137053)
		Deltaproteobacteria	Myxococcales	*Stigmatella aurantiaca* (WP_013377225)
	Actinobacteria	Micromonosporales	Micromonosporaceae	*Actinoplanes awajinensis* subsp. *mycoplanecinus* (KUL25300)
				Actinoplanes subtropicus (WP_051807411)
		Actinobacteridae	Actinomycetales	*Actinoplanes sp.* SE50/110 (WP_014691087)
	Cyanobacteria	Oscillatoriophycideae	Oscillatoriales	*Oscillatoria sp.* PCC 10802 (WP_017721273)

These 30 sequences represent the sequence diversity of the 129 sequences gathered and described in Appendix 2.6.

Appendix 2.8 Matrix and quartile distribution summaries of the aspzincin sequence identities

(A) The sequence identity matrix created from the aspzincin multiple alignment used to form the phylogenetic hypothesis in Figure 2.10. (B) The sequence identity data are divided into groups based on gene classes and the sequence identities within each group are compared across the groups using quartile distribution summaries. File available online at CloudStor+ (https://cloudstor.aarnet.edu.au/plus/index.php/s/GzHzoWIy8mfqT2l).

Appendix 2.9 Details of the three putative fusion AqAspzs that in fact are more likely the result of incorrect gene model prediction joining separate genes

Gene model	HGTracker classification	Putative domain architecture	Pfam taxonomic distribution	Blast results	Other gene model predictions	Expression data	Verdict
Aqu1.223843	Native: classified based on hits to the majority of the gene model, but not to the aspzincin at the end	PrmA Ribosomal protein L11 methyltransferase (PF06325) + aspzincin (exons 1-2 and 9)	PrmA: bacteria, Archaea, eukaryotes including animals, plants and fungi	BLASTp hits are animal sequences, but when blast only the last 172 aa that includes the aspzincin, get bacterial and fungal hits (top ten species: 8 bacterial and 2 fungal)	Two separate genes: genomescan.9105" contains the predicted PrmA and snap.54149 contains the predicted aspzincin	3' CEL-Seq reads map to the ends of the eighth and ninth exon, suggesting two separate genes	Alternative GM predictions and expression data suggests these are two separate genes (since aspzincin domain is in the ninth exon).
Aqu1.218127	Unclassified: ambiguous blast results	Partial aspzincin + partial ALG6, ALG8 glycosyltransferase family (PF03155) (exons 1 and 2-4)	ALG6: eukaryotes only (112 animals, 28 plants, 9 fungal)	Full GM: excellent hits to animal proteins: "probable dolichyl pyrophosphate Glc1Man9GlcNAc2 alpha-1,3-glucosyltransferase". Blast the first 76 aa: bacterial and fungal hits (top ten species: 5 bacterial and 5 fungal)	The most recent and transcriptomically guided gene models predict two separate genes: the full length glycosyltransferase domain in Aqu2.1.27856_001 and the aspzincin in Aqu2.1.27860_001	Transcript support for the two separate genes hypothesis, since none of the transcripts bridge the two	Two separate genes
Aqu1.219973	Unclassified: ambiguous blast results	Death (PF00531) + aspzincin (exons 2 and 12)	Death: animal only (137 species)	The aspzincin part of the gene model receives bacterial and fungal hits, and the 95 aas containing the Death domain receive animal hits	The most recent and transcriptomically guided gene models do not predict these two domains in the same gene; however they do not have any gene prediction for the aspzincin	Transcripts cover the Aqu2.1 gene model that excludes the aspzincin, no transcripts for the aspzincin sequence	More ambiguous than then the above two cases since the Aqu2.1 prediction is flawed because it does not predict the aspzincin; however likely two separate genes

Appendix 2.10 Bacterial species in the PNP phylogenetic analysis

Accession	Phylum	Class	Order	Species
WP_012143992 (replaces YP_001475390)	Proteobacteria	Gammaproteobacteria	Alteromonadales	*Shewanella sediminis* HAW-EB3
WP_012466292 (replaces YP_001943394)	Chlorobi	Chlorobia	Chlorobiales	*Chlorobium limicola* DSM 245
WP_013320650 (replaces YP_003885815)	Cyanobacteria	-	Oscillatoriales	*Cyanothece sp.* PCC 7822
WP_015349214 (replaces YP_007360638)	Proteobacteria	Deltaproteobacteria	Myxococcales	*Myxococcus stipitatus* DSM 14675
WP_002981345	Bacteroidetes	Flavobacteria	Flavobacteriia	*Chryseobacterium gleum*
WP_007042845	Proteobacteria	Gammaproteobacteria	Chromatiaceae	*Thiorhodococcus drewsii*
WP_017720792	Cyanobacteria	-	Oscillatoriales	*Oscillatoria sp.* PCC 10802

Appendix 2.11 Matrix and quartile distribution summaries of the PNP sequence identities

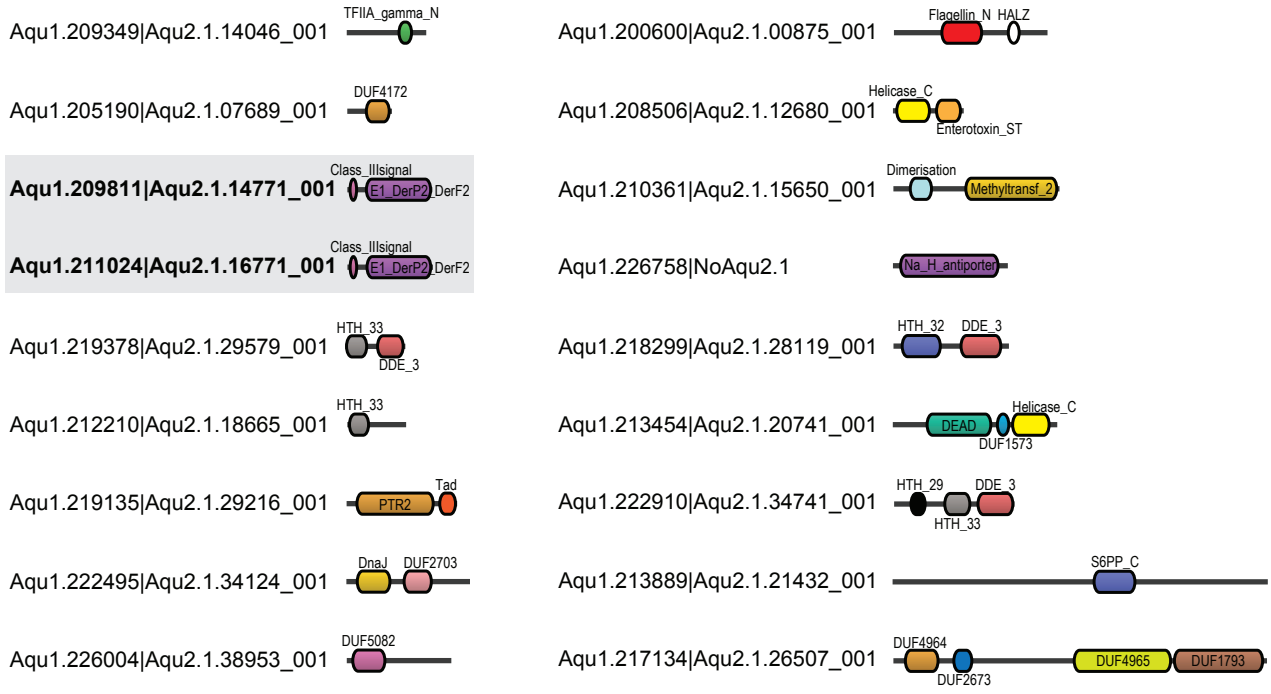
(A) The sequence identity matrix created from the PNP multiple alignment used to form the phylogenetic hypothesis in Figure 2.13. (B) The sequence identity data are divided into groups based on gene classes and the sequence identities within each group are compared across the groups using quartile distribution summaries. File available online at CloudStor+ (https://cloudstor.aarnet.edu.au/plus/index.php/s/GzHzoWIy8mfqT2l).

Appendix 2.12 Details of the putative fusion AqHGT_PNPs

Gene model	HGTracker classification	Putative domain architecture	Pfam taxonomic distribution (no. of species)	Other gene model predictions	Relevant gap in genome assembly?	Expression data	Verdict
Aqu1.204627	HGT (B)	**Death** + **Death** + PNP (exons 2, 3, 7 respectively)	Animals only (114)	No Aqu2.1	No	No expression data clarity on GM prediction	Possible fusion gene – no support for or against
Aqu1.205452	HGT (B)	**Death** + **Death** + PNP (exons 2, 4, 7 respectively)	Animals only (114)	No Aqu2.1	No	No expression data clarity on GM prediction	Possible fusion gene – no support for or against
Aqu1.206464	HGT (B)	PNP + **Death** (exons 5 and 8 respectively)	Animals only (114)	No Aqu2.1	No (one irrelevant gap)	CEL-Seq expression data after 5th exon	Separate genes
Aqu1.208731	Ambiguous (B)	COR + PNP + **Death** + PNP (exons 1-4 COR + PNP; exon 6 Death, exon 11 PNP)	Animals only (114)	COR + PNP in Aqu2.1.13072; Death in Aqu2.1.13069; last PNP not predicted by Aqu2.1.	Yes – gap after 4th exons	CEL-Seq expression data after 4th exon	Definite GM prediction problems, but still a possibility as the Aqu2.1 alternatives are not perfect (miss the last PNP). Two genes, one may be fusion still COR + PNP and Death + PNP.
Aqu1.211236	HGT (B)	**Death** + **Death** + COR + PNP (exons 2, 5, 8, 11 respectively)	Animals only (114)	Aqu2.1 supports Aqu1 prediction (Aqu2.1.17140)	Yes – gap after 5th exon	No CEL-Seq data clarity on GM predictions	Possible fusion gene – no support for or against
Aqu1.211238	HGT (B)	**Death** + PNP (exons 2, 6 respectively)	Animals only (114)	Aqu2.1 supports Aqu1 prediction (Aqu2.1.17143)	No	No CEL-Seq data clarity on GM predictions	Possible fusion gene – no support for or against
Aqu1.215582	HGT (B)	**Death** + PNP (exons 2, 6 respectively)	Animals only (114)	Aqu2.1 supports Aqu1 prediction (Aqu2.1.24132)	No	No CEL-Seq data clarity on GM predictions	Possible fusion gene – no support for or against
Aqu1.216302	HGT (B)	**Death** + COR + PNP (exons 1, 12, 14 respectively)	Animals only (114)	No Aqu2.1 for the Death; COR in Aqu2.1.25194 and PNP in Aqu2.1.25195. Augustus and Aqu0 support Aqu1 prediction	Large gap and small gap between 4th and 5th exons (i.e. between Death and COR)	No CEL-Seq data clarity on GM predictions	Possible fusion gene – no support for or against
Aqu1.216304	HGT (B)	**Death** + PNP (exons 2, 6 respectively)	Animals only (114)	Aqu2.1 does not predict the Death, and splits the rest of the gene into Aqu2.1.25198 and Aqu2.1.25197 (contains the PNP)	No gaps	No CEL-Seq data clarity on GM predictions	Possible fusion gene – no support for or against
Aqu1.217845	HGT (B)	**Death** + PNP (exons 1, 5 respectively)	Animals only (114)	Aqu2.1 supports the Aqu1 prediction	No	No CEL-Seq data clarity on GM predictions	Possible fusion gene – no support for or against
Aqu1.217850	HGT (B)	PNP + COR + **Death** + PNP (exons 6, 8, 9, 11 respectively)	Animals only (114)	Aqu2.1 does not predict the Death or the last PNP, Aqu2.1.27433 has the first PNP and COR	Yes – one large gap separating the Death from the last PNP	No CEL-Seq data clarity on GM predictions	Possible fusion gene – no support for or against
Aqu1.219688	HGT (B)	**Death** + PNP (exons 3, 6 respectively)	Animals only (114)	Aqu2.1 supports Aqu1 prediction (Aqu2.1.30059)	Yes – two small gaps between domains	CEL-Seq support for independent expression of the Death exon and the PNP exon	Separate genes
Aqu1.217852	HGT (B)	PNP + **ZU5** (exons 10, 13 respectively)	Animals (115), bacteria (5)	Aqu2.1 supports the Aqu1 prediction (Aqu2.1.27442)	No	No CEL-Seq data clarity on GM predictions	Possible fusion gene – no support for or against
Aqu1.204355	Unclassified (first ~200 aa animal-like, then bacterial-like)	PNP (exon 6)	-	Aqu2.1 prediction is different but supports the joining of the native and foreign sequence (Aqu2.1.06440)	No	Some light expression from the first two exons (animal-like)	Separate genes
Aqu1.214701	Unclassified (first ~200 aa animal-like, then bacterial-like)	Pkinase + PNP (exons 3-8, 14 respectively)	Animals (115), other eukaryotes (30), Plants (29), fungi (10), bacteria (456)	Aqu2.1 supports the joining of the native and foreign sequence (Aqu2.1.22779)	No	No expression data clarity on GM prediction	Possible fusion gene – no support for or against
Aqu1.217177	Unclassified (first ~200 aa animal-like, then bacterial-like)	Pkinase + PNP (exons 3-9, 15)	Animals (115), other eukaryotes (30), Plants (29), fungi (10), bacteria (456)	Different Aqu2.1 prediction – Aqu2.1. 26560 with Pkinase and Aqu2.1.26561 with PNP	No	No expression data clarity on GM prediction	Possible fusion gene – no support for or against

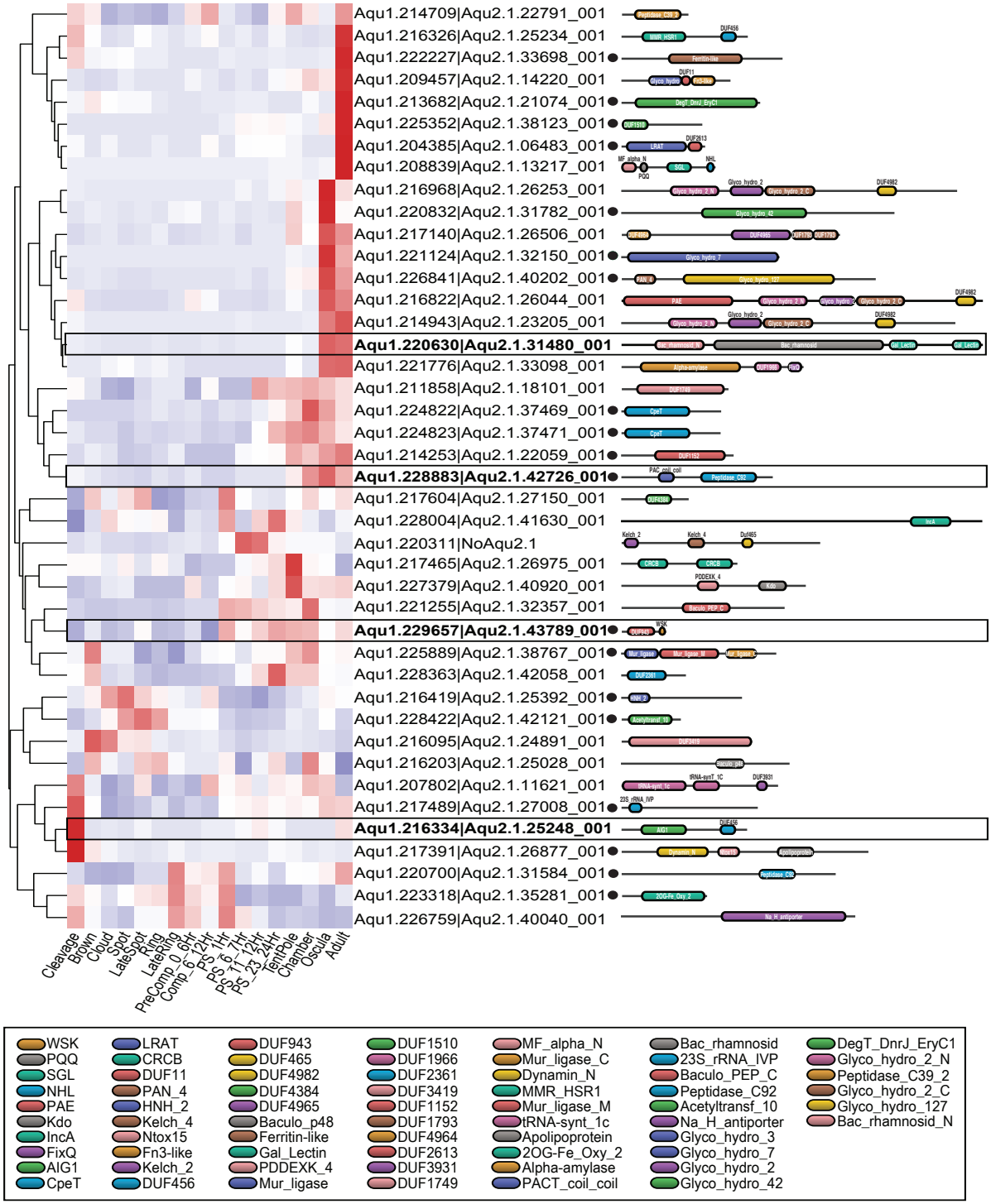
Appendix 2.13 Predicted domain architectures of the unexpressed AqHGT_NMs

None of these genes have at least one developmental stage with a normalised CEL-Seq count of at least five. Domain architectures were determined by searching the Pfam database v27.0. Two putative AqHGT_NM fusion genes are shaded in grey.



Appendix 2.14 The ontogenetic expression profiles and predicted domain architectures of the expressed AqHGT_NMs

Data presented for those genes that have at least one developmental stage with a normalised CEL-Seq count of at least five. Domain architectures were determined by searching the Pfam database v27.0. Four putative AqHGT_NM fusion genes are outlined with a black box. Heat map is scaled by row with genes organised by Euclidean distances (R Studio, Pretty Heatmaps package v 0.7.7). See Appendix 2.5 for an overview of the developmental stages.



Appendix 2.15 Details of the AqHGT_NM putative fusion genes

Gene model	HGTracker classification	Putative domain architecture (nonmetazoan in bold)	Nonmetazoan domain's taxonomic distribution		Other domain's taxonomic distribution		Other gene model predictions	Relevant gap in genome assembly?	Expression data	Verdict
			Pfam	BLASTp	Pfam	BLASTp				
Aqu1.209811| Aqu2.1.14771 _001	HGT (F)	**Class_IIIsignal +** E1_DerP2_DerF2 (exons 1 and 2-5)	Archaea & bacteria	Top 10 species include 8 bacteria and 2 fungi	Eukaryotes	Top 10 species are fungal	All support linking	No	None	Possible fusion gene – no support for or against
Aqu1.211024| Aqu2.1.16771 _001	HGT (F)	**Class_IIIsignal +** E1_DerP2_DerF2 (exons 1 and 2-4)	Archaea & bacteria	No results	Eukaryotes	Top 10 species include 8 fungi and 2 amoebas	All support linking	No	None	Possible fusion gene – no support for or against
Aqu1.216334| Aqu2.1.25248 _001	HGT (P)	AIG1 + **DUF456** (1 exon only)	Bacteria	All hits from 9 bacterial species	Eukaryotes	Top 10 species include 7 plants, 2 bacteria and 1 animal (BLASTp of first half of pep - includes AIG1)	All support linking	No	None	Possible fusion gene – no support for or against
Aqu1.220630| Aqu2.1.31480 _001	HGT (X)	**Bac_rhamnosid_N + Bac_rhamnosid** + Gal_Lectin + Gal_Lectin (1 exon only)	Bacteria	Top 10 species include 5 bacteria, 4 other eukaryotes (not animals, fungi or plants) and 1 archaea (BLASTp of first 850 aas; excellent hits)	Eukaryotes	Top 10 species include 9 plants and 1 other eukaryote (choanoflagellate) (BLASTp of last 285 aa)	All support GM prediction	No	Transcript support for GM prediction. CEL-Seq expression too	**Possible fusion gene – one transcript linking**
Aqu1.228883| Aqu2.1.42726 _001	HGT (B)	PACT_coil_coil + **Peptidase_C92** (exons 1 and 2-6)	Bacteria	Top 10 species include 9 bacteria and one other eukaryote (Cryptophyta) (BLASTp of last 279 aa)	Animal	Only one result, fungal (BLASTp of first 160 aa)	All support GM prediction	No	Transcript support for GM prediction. CEL-Seq expression too	**Possible fusion gene – one transcript linking**
Aqu1.229657| Aqu2.1.43789 _001	HGT (B)	**DUF943** + WSK (1 exon only)	Bacteria	Top 10 species include 8 bacteria and 2 archaea (BLASTp of first 96 aa)	Animal	No results (BLASTp last 32 aa)	All support GM prediction	No	Transcript support for GM prediction. CEL-Seq expression too	**Possible fusion gene – one transcript linking**

Appendix 2.16 General properties of the AqAspzs

File available online at CloudStor+ (https://cloudstor.aarnet.edu.au/plus/index.php/s/GzHzoWIy8mfqT2l).

Appendix 3.1 Sequence data sources used throughout Chapter 3

Species	Sequence type	Resource	Reference and/or NCBI accession
Monosiga brevicollis	Whole genome	http://genome.jgi-psf.org/Monbr1/Monbr1.info.html	King et al. 2008
Aphrocallistes vastus	Transcriptome	https://era.library.ualberta.ca/public/view/item/uuid:219b8059-cc22-4236-a62e-5fd63c4155d1/	Riesgo et al. 2014a *
Hyalonema populiferum	Transcriptome	http://figshare.com/articles/Error_signal_and_the_placement_of_Ctenophora_sister_to_all_other_animals/1334306	Whelan et al. 2015
Rossella fibulata	Transcriptome	http://figshare.com/articles/Error_signal_and_the_placement_of_Ctenophora_sister_to_all_other_animals/1334306	Whelan et al. 2015
Sympagella nux	Transcriptome	http://figshare.com/articles/Error_signal_and_the_placement_of_Ctenophora_sister_to_all_other_animals/1334306	Whelan et al. 2015
Amphimedon queenslandica	Whole genome	Aqu1s: http://metazoa.ensembl.org/Amphimedon_queenslandica/Info/Index and http://genome.jgi.doe.gov/AmpqueaRenierasp/AmpqueaRenierasp.download.html; Aqu2.1s http://amphimedon.qcloud.qcif.edu.au/downloads.html	Srivastava et al. 2010; Fernandez-Valverde et al. 2015
Chondrilla nucula	Transcriptome	http://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA225590	Riesgo et al. 2014a *
Crella elegans	Transcriptome	http://datadryad.org/resource/doi:10.5061/dryad.50dc6	Pérez-Porro et al. 2013 *
Ircinia fasciculata	Transcriptome	http://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA225586	Riesgo et al. 2014a *
Petrosia ficiformis	Transcriptome	http://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA162901	Riesgo et al. 2014a *
Pseudospongosorites suberitoides	Transcriptome	http://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA225580	Riesgo et al. 2014a *
Spongilla lacustris	Transcriptome	http://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA225591	Riesgo et al. 2014a *
Kirkpatrickia variolosa	Transcriptome	http://figshare.com/articles/Error_signal_and_the_placement_of_Ctenophora_sister_to_all_other_animals/1334306	Whelan et al. 2015
Latrunculia apicalis	Transcriptome	http://figshare.com/articles/Error_signal_and_the_placement_of_Ctenophora_sister_to_all_other_animals/1334306	Whelan et al. 2015
Ephydatia muelleri	Transcriptome	http://comparative.reefgenomics.org/datasets.html	Bhattacharya et al. 2016
Tethya wilhelma	Transcriptome	http://www.ncbi.nlm.nih.gov/sra/ERR216193/	ERR216193
Oscarella carmela	Whole genome	http://www.compagen.org/datasets/OCAR_T-PEP_130911	Nichols et al. 2012 *
Sycon coactum	Transcriptome	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA162899/	Riesgo et al. 2014a *
Corticium candelabrum	Transcriptome	http://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA162903	Riesgo et al. 2014a *
Clathria prolifera	Transcriptome	NA	Grice, Degnan, Fernandez-Valverde and Fernàndez-Busquets, unpublished data *
Leucosolenia complicata	Transcriptome	http://datadryad.org/resource/doi:10.5061/dryad.tn0f3/1	Fortunato et al. 2014
Sycon ciliatum	Whole genome	http://datadryad.org/resource/doi:10.5061/dryad.tn0f3/1	Fortunato et al. 2014
Xestospongia testudinaria	Transcriptome	https://www.ncbi.nlm.nih.gov/sra/?term=SRR1738101	SRR1738101
Cliona varians	Transcriptome	https://www.ncbi.nlm.nih.gov/sra/?term=SRR1391159	SRR1391159; Riesgo et al. 2014b
Stylissa carteri	Whole genome	http://sc.reefgenomics.org/download/	PRJNA254402; Ryu et al. 2016
Haliclona amboinensis	Transcriptome	http://compagen.zoologie.uni-kiel.de/datasets.html	Guzman and Conaco 2016
Haliclona tubifera	Transcriptome	http://compagen.zoologie.uni-kiel.de/datasets.html	Guzman and Conaco 2016
Niphatidae *indet.*	Transcriptome	NA	Gaiti, Kocot, and Degnan, unpublished data *

Further annotation of a dataset as described in Grice *et al.* (2017) is denoted by "*".

Appendix 3.2 Script for putative co-expression detection using a correlation matrix with probability significance values

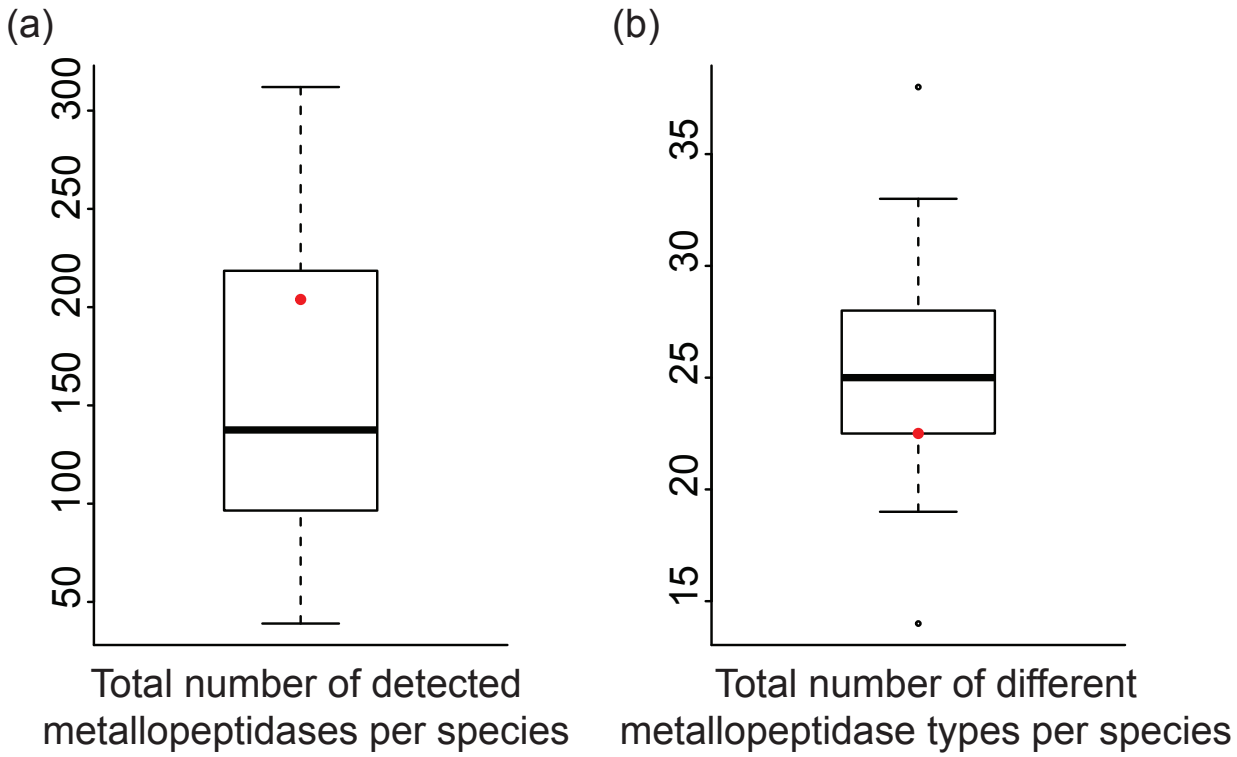
Script was implemented using R (R Core Team 2014) and was developed by Gaiti et al. (2015). File available online at CloudStor+ (https://cloudstor.aarnet.edu.au/plus/index.php/s/GzHzoWIy8mfqT2l).

Appendix 3.3 Script for the Pfam domain enrichment analysis

The analysis was performed using R (R Core Team 2014) using the following script that has minor adjustments made by W.L. Hatleberg from the original script developed by Chandran et al. (2009). File available online at CloudStor+ (https://cloudstor.aarnet.edu.au/plus/index.php/s/GzHzoWIy8mfqT2l).

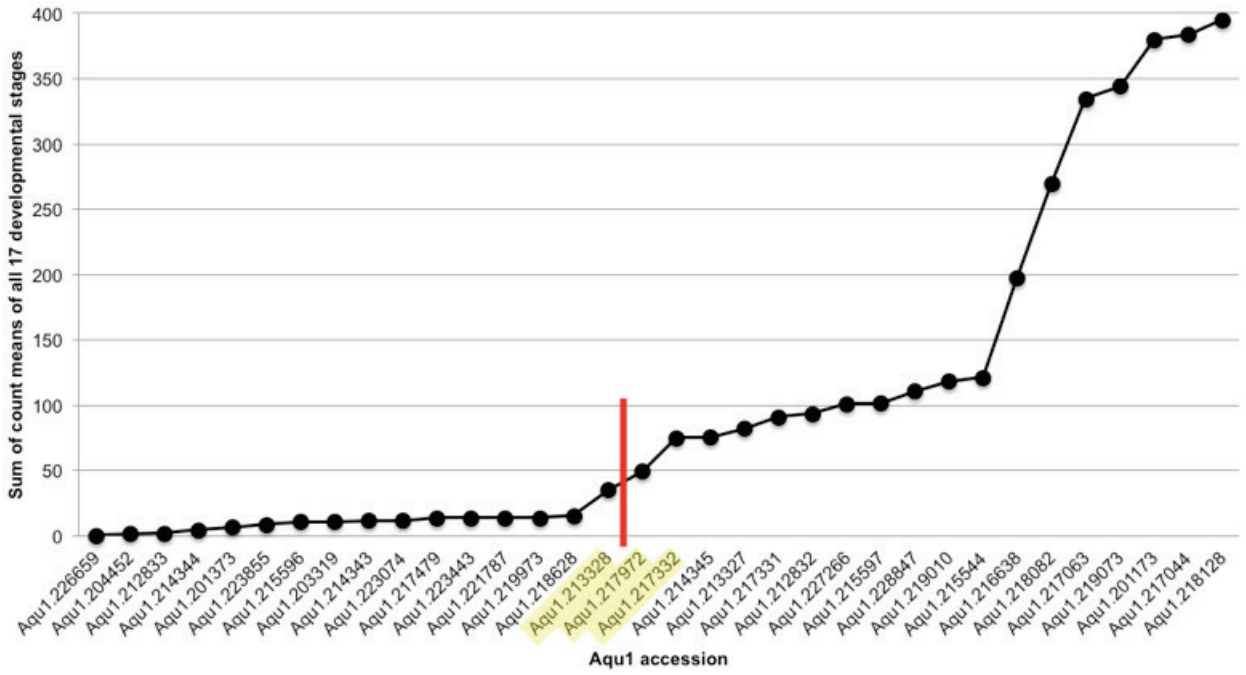
Appendix 3.4 Quartile summaries of the metallopeptidase repertoires of animal representatives

(a) The total number of metallopeptidases detected in each species analysed (n=28) and (b) the number of different types of metallopeptidases detected in each of the species. In both plots, the red point depicts the specific results for *A. queenslandica*.



Appendix 3.5 Manual assessment of which AqAspzs are unexpressed

Rather than use an arbitrary cut-off point, the sum of the read count means of each developmental stage was graphed for each of least expressed aspzincins. Aqu1.213328 (highlighted on the x-axis) has low expression throughout development with no peaks and a total sum of 35. Based on these data, this is the most highly expressed unexpressed AqAspz. Aqu1.217972 and Aqu1.217332 (highlighted) have total sums not that much greater (49 and 75 respectively); however, a significant part of those sums is from the adult stage, and such stage-specific up-regulation is likely to reflect meaningful expression. Those with greater total expression sums than 35 were thus considered expressed and those below were not.

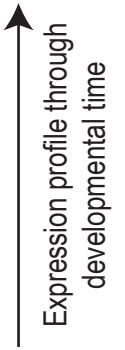


Appendix 3.6 Domains in *A. queenslandica* genes that have a significantly correlated expression profile with that of any of the AqAspzs

File available online at CloudStor+ (https://cloudstor.aarnet.edu.au/plus/index.php/s/GzHzoWIy8mfqT2l).

Appendix 3.7 Structural CATH class classifications

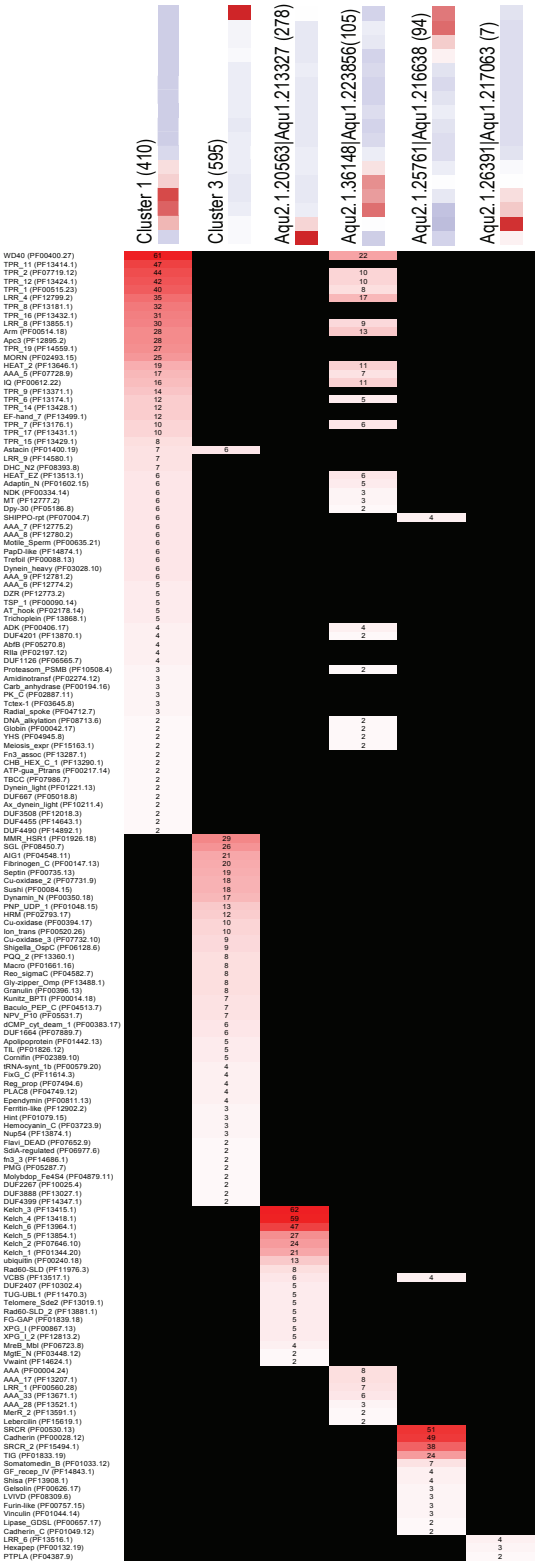
Classifications are for the *A. queenslandica* genes that have a significantly correlated expression profile with that of any of the AqAspzs.



CATH Class	Number of domains in CATH class	Cluster1 Correlates	Cluster2 Correlates	Cluster3 Correlates	Aqu2.1.20563Aqu1.213327 Correlates	Aqu2.1.25761Aqu1.216638 Correlates	Aqu2.1.24062 Aqu1.215544 Correlates	Aqu2.1.36148Aqu1.223856 Correlates	Aqu2.1.26391 Aqu1.217063 Correlates	Aqu2.1.20564 Aqu1.212832 Correlate	Aqu2.1.39947Aqu1.226677 Correlate	Aqu2.1.00805Aqu1.200550 Correlate
Mainly alpha	48121	29	13	21	31	19	80	32	0	0	100	0
Mainly beta	58944	22	50	25	52	44	20	19	14	0	0	0
Alpha beta	125772	46	38	48	31	35	0	49	71	100	0	0
Few secondary structures	3021	3	0	6	52	2	0	0	14	0	0	0
Dominant CATH Class		**Alpha beta**	**Mainly beta**	**Alpha beta**	**Mainly beta AND Few secondary structures**	**Mainly beta**	**Mainly alpha**	**Alpha beta**	**Alpha beta**	**Alpha beta**	**Mainly alpha**	**NA**
Total number of CATH predictions within genes		254	8	399	197	63	5	57	7	1	1	0
Total number of genes with at least one CATH prediction		201	7	295	141	47	4	49	4	1	1	0
Total number of genes interrogated for CATH predictions		410	14	595	278	94	6	105	7	1	1	1

Appendix 3.8 Domain enrichment in the 1512 *A. queenslandica* genes that have a significantly correlated expression profile with that of any of the AqAspzs

For reference, heatmaps showing the expression profile for each of the aspzincin cluster groups and uniquely expressed aspzincins are shown at the top.



Appendix 3.9 The HGTracker classification for *A. queenslandica* genes that correlate in ontogenic expression with that of any AqAspzs

Correlate gene(s)	Native	ForeignHGT	ForeignAS	ForeignFS	Unclassifiable	Sum
Cluster1Correlates (n=410)	280	3	0	0	52	**335**
Cluster2 Correlates (n=14)	6	0	0	0	2	**8**
Cluster3 Correlates (n=595)	136	26	16	2	122	**302**
Aspz Aqu2.1.20563Aqu1.213327 Correlates (n=278)	79	5	2	1	78	**165**
Aspz Aqu2.1.36148Aqu1.223856 Correlates (n=105)	82	0	0	0	10	**92**
Aspz Aqu2.1.25761Aqu1.223856 Correlates (n=94)	33	2	0	0	13	**48**
Aspz Aqu2.1.26391 Aqu1.217063 Correlates (n=7)	1	0	0	0	4	**5**
Aspz Aqu2.1.24062 Aqu1.215544 Correlates (n=6)	2	0	0	0	3	**5**
Aspz Aqu2.1.200805Aqu1.200550 Correlate (n=1)	1	0	0	0	0	**1**
Aspz Aqu2.1.20564 Aqu1.212832 Correlate (n=1)	0	0	0	0	1	**1**
Aspz Aqu2.1.39947Aqu1.226677 Correlate (n=1)	0	1	0	0	0	**1**

"Native" reflects animal-like genes originating from the last metazoan common ancestor and inherited by descent. "ForeignHGT" signifies a gene classified as nonmetazoan based on sequence similarity criteria yet that has been assembled into a scaffold also containing "native" genes. "ForeignAS" reflects genes deemed as nonmetazoan based on sequence similarity criteria and that are incorporated into ambiguous scaffolds that contain both foreign and taxonomically unclassified genes, thus could be comprised of native genes and HGTs, or all foreign genes and more likely a contaminating scaffold. "ForeignFS" reflects foreign genes on a scaffold comprising only foreign genes and thus is more likely a result of contamination in the genome. "Unclassifiable" signifies genes that get BLASTp hits, but those hits do not meet the criteria of HGTracker's taxonomic sequence similarity process. The sum column shows for each correlate group how many genes have HGTracker results; 36% of the correlates do not have a result either because they did not have any BLASTp hits at all so HGTracker could not classify them, or because their Aqu1 and Aqu2.1 gene model predictions are not within a 60% coverage threshold of each other.

Appendix 3.10 AqHGTs that correlate in ontogenetic expression with that of any AqAspzs

Correlate gene(s)	Gene count: HGTs/total	BLAST2GO annotations	Source of HGTs (#)
Cluster1 Correlates	3/410	secreted protein; 2 x methyltransferase	B(3)
Cluster3 Correlates	26/595	aldose 1-epimerase; ankyrin repeat-containing; atp-binding protein; ferritin-like family protein; 3 x helicase; ig-like domain-containing protein; indoleamine -dioxygenase; mac perforin domain-containing protein; membrane protein; ml domain; multidrug resistance protein 1-partial; nc domain protein; nhl repeat protein; 2 x nucleoside phosphorylase-like protein; perosamine synthetase; phospholipid transfer protein; 2 x purine or other phosphorylase family 1; secreted protein containing duf1549; type iii restriction protein res subunit; ubiquitin carboxyl-terminal hydrolase 47-like; wd repeat-containing protein 63- partial	F(2), X(3), B(21)
Aspz Aqu2.1.20563Aqu1.213327 Correlates	5/278	2 x phospholipid transfer protein; dna-n1-methyladenine dioxygenase; adp-ribosylglycohydrolase; atp gtp-binding protein	B(3), X(2)
Aspz Aqu2.1.25761Aqu1.223856 Correlates	2/94	mip family channel protein; glutaminase	X(2)
Aspz Aqu2.1.39947Aqu1.226677 Correlate	1/1	No result	B(1)

Taxonomic sources are “B” (bacterial-like), “F” (fungal-like) and “X” (clearly nonmetazoan but of uncertain specific classification).

Appendix 3.11 AqAspzs ranked by expression levels

Aqu1	Aqu2.1	Total expression (normalised read counts)	No. of aspzincins on the same scaffold	Secreted? (SP; NC; TM)	Conserved catalytic motif (HEXXH…D)?
Aqu1.226659	NoGM	0	4	SP	NO
Aqu1.204452	Aqu2.1.06590_001	1.244086802	1	NO	NO
Aqu1.212833	Aqu2.1.19746_001	1.428674658	2	SP	YES
Aqu1.214344	Aqu2.1.22214_001	4.068653333	3	NO	NO
Aqu1.201373	NoGM	6.12099517	1	SP	NO
Aqu1.223855	Aqu2.1.36147_001	8.481828514	4	NO	YES
Aqu1.215596	Aqu2.1.24154_001	10.43521184	2	SP	NO
Aqu1.203319	NoGM	10.50335072	1	**No SP; NC**	YES
Aqu1.214343	Aqu2.1.22213_001	11.2128524	3	NO	NO
Aqu1.223074	Aqu2.1.34944_001	11.21808586	1	NO	YES
Aqu1.217479	Aqu2.1.26989_001	13.17821882	1	SP	NO
Aqu1.223443	Aqu2.1.35512_001	13.27107629	1	**No SP; NC**	YES
Aqu1.221787	Aqu2.1.33112_001	13.30589622	1	**No SP; NC**	YES
Aqu1.219973	Aqu2.1.30497_001	13.48769347	1	SP	YES
Aqu1.218628	Aqu2.1.28557_001	15.23812602	3	SP	YES
Aqu1.213328	Aqu2.1.20564_001	34.62306909	2	SP	NO
Aqu1.217972	Aqu2.1.27620_001	48.94321978	1	NO	YES
Aqu1.217332	Aqu2.1.26796_001	74.63040349	2	NO	YES
Aqu1.214345	Aqu2.1.22215_001	75.30582647	3	SP	YES
Aqu1.213327	Aqu2.1.20563_001	82.04273356	2	SP	YES
Aqu1.217331	Aqu2.1.26795_001	90.51451571	2	**No SP; NC**	YES
Aqu1.212832	Aqu2.1.19745_001	93.2044618	2	SP	NO
Aqu1.227266	NoGM	100.7843761	2	NO, TM	NO
Aqu1.215597	Aqu2.1.24155_001	101.1262137	2	NO	YES
Aqu1.228847	Aqu2.1.42690_001	110.3088465	1	**No SP; NC**	YES
Aqu1.219010	Aqu2.1.29044_001	117.9628787	1	SP	YES
Aqu1.215544	Aqu2.1.24064_001	120.8861841	1	**No SP; NC**	YES
Aqu1.216638	Aqu2.1.25761_001	197.5962164	1	**SP**	YES
Aqu1.218082	Aqu2.1.27810_001	269.1258549	1	NO	YES
Aqu1.217063	Aqu2.1.26391_001	334.3196464	10	SP	YES
Aqu1.219073	Aqu2.1.29115_001	344.0011673	3	SP	YES
Aqu1.201173	NoGM	379.6900527	1	SP	YES
Aqu1.217044	Aqu2.1.26368_001	383.4251789	10	SP	NO
Aqu1.218128	Aqu2.1.27861_001	394.653226	6	NO	YES
Aqu1.219660	Aqu2.1.30024_001	497.4912502	1	SP	NO
Aqu1.217565	Aqu2.1.27096_001	520.61164	1	**No SP; NC**	YES
Aqu1.219070	Aqu2.1.29111_001	522.4374837	3	**No SP; NC**	YES
Aqu1.213448	Aqu2.1.20728_001	533.4483234	1	**No SP; NC**	NO
Aqu1.219075	Aqu2.1.29117_001	551.7997045	3	**No SP; NC**	YES
Aqu1.226677	Aqu2.1.39947_001	556.4901927	4	**No SP; NC**	YES
Aqu1.213425	Aqu2.1.20690_001	601.6460448	3	**No SP; NC**	YES
Aqu1.225770	Aqu2.1.38622_001	602.7418521	2	SP	YES
Aqu1.215510	Aqu2.1.24010_001	680.6217592	1	SP	YES
Aqu1.207864	Aqu2.1.11707_001	750.9827226	2	SP	NO
Aqu1.213426	Aqu2.1.20691_001	842.958037	3	SP	YES
Aqu1.200471	Aqu2.1.00695_001	999.7122893	1	NO	YES
Aqu1.208853	Aqu2.1.13252_001	1021.431143	2	SP	YES
Aqu1.223843	Aqu2.1.36133_001	1024.678458	4	NO	NO
Aqu1.207862	Aqu2.1.11705_001	1069.950412	2	**No SP; NC**	YES
Aqu1.223856	Aqu2.1.36148_001	1080.045419	1	SP	YES
Aqu1.213434	Aqu2.1.20707_001	1125.760537	1	**No SP; NC**	YES
Aqu1.217074	Aqu2.1.26407_001	1132.496188	1	SP	YES
Aqu1.223844	Aqu2.1.36134_001	1194.181499	4	SP	NO
Aqu1.224355	Aqu2.1.36761_001	1284.416379	1	NO	YES
Aqu1.217029	Aqu2.1.26337_001	1359.80345	1	SP	YES
Aqu1.218124	Aqu2.1.27857_001	1516.960068	6	SP	YES
Aqu1.218626	Aqu2.1.28554_001	1718.349775	3	SP	NO
Aqu1.200550	Aqu2.1.00805_001	1773.405267	1	**No SP; NC**	YES
Aqu1.208852	Aqu2.1.13251_001	2057.954139	2	NO	YES
Aqu1.218627	Aqu2.1.28555_001	2331.82246	3	SP	NO
Aqu1.218127	Aqu2.1.27860_001	2401.06193	6	NO; TM	YES
Aqu1.226660	Aqu2.1.39930_001	2438.910393	4	SP	NO
Aqu1.218125	Aqu2.1.27858_001	2721.034561	6	**No SP; NC**	YES
Aqu1.209446	Aqu2.1.14208_001	2888.585072	2	SP	YES
Aqu1.217054	Aqu2.1.26380_001	2968.343316	10	**No SP; NC**	YES
Aqu1.217058	Aqu2.1.26385_001	3017.944648	10	SP	YES
Aqu1.217053	Aqu2.1.26378_001	3509.962696	10	SP	YES
Aqu1.218131	Aqu2.1.27864_001	4138.983858	6	**No SP; NC**	YES
Aqu1.216382	Aqu2.1.25340_001	4347.751957	1	SP	YES
Aqu1.217045	Aqu2.1.26370_001	5081.638045	10	SP	YES
Aqu1.225765	Aqu2.1.38617_001	5722.607095	2	**No SP; NC**	YES
Aqu1.218126	Aqu2.1.27859_001	6232.332603	6	**No SP; NC**	YES
Aqu1.224071	Aqu2.1.36416_001	6770.0315	1	NO	YES
Aqu1.217047	Aqu2.1.26371_001	7420.32248	10	SP	YES
Aqu1.226658	Aqu2.1.39928_001	10257.77907	4	SP	NO
Aqu1.220838	NoGM	15046.22491	2	SP	YES
Aqu1.201735	Aqu2.1.02514_001	16026.91704	1	**No SP; NC**	YES
Aqu1.217057	Aqu2.1.26384_001	16277.30424	10	SP	YES
Aqu1.209447	Aqu2.1.14209_001	21141.49013	2	**No SP; NC**	YES
Aqu1.217049	Aqu2.1.26374_001	29002.32852	10	SP	YES
Aqu1.217048	Aqu2.1.26373_001	29929.16487	10	SP	YES
Aqu1.213427	Aqu2.1.20692_001	39778.97605	3	**No SP; NC**	YES
Aqu1.227270	Aqu2.1.40809_001	61447.42219	2	SP	YES
Aqu1.220837	NoGM	177433.49	2	**No SP; NC**	YES

Predicted secretion status (green signifies secreted, signal peptide presence/absence "SP"/"NO", non-classical secretion "NC"; or transmembrane helices predicted "TM"), and the conservation status of the essential residues in the catalytic motif (yes/no). Those genes above the bold line do not have a meaningful expression profile through development. Total expression reflects the sum of a gene's expression read counts from all 17 measured developmental stages.

Appendix 4.1 Summary of the ISFinder hits in the 48 HGT scaffolds used as a pilot test

IS hit locus in *A. queenslandica*	Hit identities	Hit e-value	IS family and group	IS name and host	IS length (nt)	IS ORFs (no., function, chemistry, strand)	Hit locus in IS	Hit locus in scaffold
Contig13085:55771-55793	23/23	0.008	IS1182,	ISCIbu1: *Clostridium butyricum*	1830	1; Transposase; DDE; +	Immediately after the transposase and on the other strand (-)	On –ve strand, between two native +ve strand Aqu genes. There is a transcript present and when I BLASTx it, I get good hits to four animal proteins, so possibly an unpredicted animal gene. No Pfam results.
Contig13328:121987-122010	24/24	0.004	ISLre2	*Thermoanaerobacter ethanolicus*	1596	1; Transposase, DDE, +	Within the transposase, same strand	Hit is in the 16th intron of a native gene (+ve strand). BLASTx the intron – no results (even with additional 80 bases either side).
Contig13409:110385-110411	27/27	9e-05	IS4, IS231	*Bacillus cereus*	2470	2; Transposase, DDE, +; Passaenger gene, +	Within the transposase, same strand	Hit is in an unclassified gene Aqu1.220849, no blast or Pfam results. Gene lies amongst bacterial HGTs, unclassified and natives.
Contig13467:304167-304199	31/33	0.002	IS200/IS605, IS1341	*Ferroplasma acidarmanus*	1514	1; Accessory gene, TnpB, +	Within the accessory gene but on the –ve strand	Hit is in the intron of unclassified Aqu1.223884/Aqu2.1.36176. BLASTx the intron – no results, nor from Pfam. BLASTp the protein and get good animal hits for the first half, the IS hit is in the second half which gets no blast hits. No Pfam domains detected.
Contig13474:58244-58271	27/28	0.009	IS30	*Spiroplasma citri*	8273	1; Transposase, DDE, +	Before the transposase	IS hit within Aqu1.224342|Aqu2.1.36742, a bacterial-like HGT. BLASTx this gene and get excellent hits to bacterial type IV secretion protein Rhs. BLASTp the whole IS, get hits to plectrovirus transmembrane proteins.
Contig13261:92624-92662	24/24	0.002	IS110	-	-	Irrelevant as hit only an AAA repeat. No information on this IS in IS database.	-	-

Hits retrieved via submission of whole scaffolds to ISfinder (e-value cut-off = 0.01; www-is.biotoul.fr; Siguier et al. 2006).

Appendix 4.2 Comparison of results from different e-value cut offs used for IS searching in the 48 HGT scaffolds used as a pilot test

E-value cut-off	No. of hits	E-value		Hit length		Identity (%)	
		Mean	St. dev.	Mean (nt)	St. dev.	Mean	St. dev.
10	1004	4.22	2.78	20.16	2.80	98.98	2.11
1	186	0.44	0.31	22.21	2.57	99.05	1.98
0.1	37	0.04	0.03	24.21	2.74	99.26	1.96
0.01	9	0.0004	0.0034	24.72	2.14	99.55	1.50

Hits retrieved via submission of whole scaffolds to ISfinder (www-is.biotoul.fr; Siguier et al. 2006).

Appendix 4.3 BLASTp details and domain content of the five best hits of AqHGT Aqu1.224342|Aqu2.1.36742_001

Protein name	NCBI Accession	Species	Protein length	BLASTp results (query coverage ; identity; e-value)	CDD domain architecture (domain accession; hit e-value)	Pfam domain architecture (domain accession; hit e-value)
Hypothetical protein	WP_080514117	*Vibrio campbellii*	1624 aa	86%; 33%; 0	RhsA (COG3209; 8.88e-03) Rhs repeat-associated core (TIGR03696; 1.05e-19); SseC (cl27103; 6.94e-05)	None detected.
Hypothetical protein VIBHAR_01674	ABU70643	*Vibrio campbellii* ATCC BAA-1116	1559 aa	82%; 32%; 0	Rhs repeat-associated core (TIGR03696; 8.79e-19); SseC (cl27103; 6.74e-05)	None detected.
Hypothetical protein	WP_081862750	*Chromobacterium haemolyticum*	1670 aa	86%; 32%; 2e-164	Rhs repeat-associated core (TIGR03696; 9.17e-19); Toxin0HDC (cl21448; 3.95e-08)	Toxin-HDC (PF15656; 1.2e-10)
Rhs repeat-associated core domain-containing protein	WP_017659411	*Geitlerinema* sp. PCC 7105	1597 aa	89%; 30%; 1e-144	Rhs repeat-associated core (TIGR03696; 3.53e-20	None detected.
Type IV secretion protein Rhs	WP_085717411	*Pseudomonas* sp. B28 (2017)	1597 aa	90%; 29%; 4e-140	RhsA (COG3209; 1.66e-13) Rhs repeat-associated core (TIGR03696; 3.06e-29); SseC (cl27103; 1.51e-05)	None detected.

Aqu1.224342|Aqu2.1.36742_001 was submitted against the January 2018 NCBI nonredundant database using the default search settings.

Appendix 4.4 Comparisons of the repetitive sequences surrounding unduplicated HGT and native genes +/- 2500 bp

Gene group	Gene group	DNA transposon	Retrotransposon	Low complexity	Simple repeats	Unknown TE	Satellite repeats	Total repeats
SUD HGTs	Young HGTs	0.943	0.458	0.745	0.328	0.266	0.997	0.175
SUD HGTs	LUD HGTs	0.388	0.706	0.104	0.3	0.349	0.7077	0.357
Young HGTs	LUD HGTs	0.527	0.392	0.15	0.926	0.163	0.997	0.132
SUD Natives	LUD Natives	0.161	0.0165	0.394	0.041	0.449	0.94	0.117
SUD HGTs	SUD Natives	0.609	0.045 ↓	0.093	0.312	0.948	0.344	0.625
SUD HGTs	LUD Natives	0.169	0.55	0.432	0.253	0.444	0.32658	0.225
Young HGTs	SUD Natives	0.769	0.224	0.166	0.621	0.294	0.997	0.16
Young HGTs	LUD Natives	0.125	0.574	0.473	0.145	0.451	0.997	0.366
LUD HGTs	SUD Natives	0.103	0.413	0.35	0.664	0.42	0.0898	0.602
LUD HGTs	LUD Natives	0.006 ↑	0.418	0.226	0.107	0.166	0.068	0.0848

Repeats are retrieved for each gene plus and minus 2500 bp either side from the TE reference library of *A. queenslandica.* Significance was determined with analysis of deviance tests with quasibinomial errors because of overdispersed errors. All *p*-values are raw and after corrections for multiple testing using the Benjamini-Hochberg method with a 20% false discovery rate, none are significant. Shading reflects raw *p*-values less than 0.05 and arrows reflect greater or lower proportion of the repeat class in the HGT sample, for those comparisons involving a HGT gene group and found to be possibly different (with raw *p*-values<0.05). SUD refers to strict criteria for a gene's putative unduplicated status, LUD reflects less strict criteria, and the young HGTs are 13 genes identified by Conaco *et al.* (2016) as resulting from more recent transfer events.

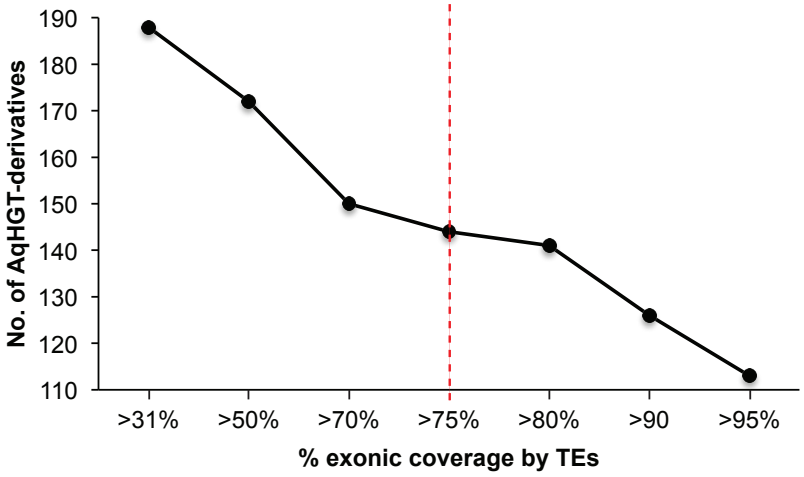
Appendix 4.5 Comparisons of the repetitive sequences surrounding unduplicated HGT and native genes +/- 20 kbp

Gene group	Gene group	DNA transposon	Retrotransposon	Low complexity	Simple repeats	Unknown TE	Satellite repeats	Total repeats
SUD HGTs	Young HGTs	0.507	0.685	0.875	0.117	0.914	0.308	0.52
SUD HGTs	LUD HGTs	0.532	0.246	0.0591	0.114	0.48	0.687	0.785
Young HGTs	LUD HGTs	0.752	0.258	0.0927	0.918	0.584	0.213	0.423
SUD Native	LUD Native	0.821	0.0657	0.354	0.0453	0.189	0.573	0.186
SUD HGTs	SUD Native	0.646	0.0322 ↓	0.156	0.0415 ↑	0.139	0.415	0.118
SUD HGTs	LUD Native	0.812	0.885	0.659	0.805	0.859	0.68	0.769
Young HGTs	SUD Native	0.111	0.168	0.352	0.707	0.347	0.063	0.11
Young HGTs	LUD Native	0.334	0.634	0.688	0.139	0.83	0.085	0.372
LUD HGTs	SUD Native	0.14	0.751	0.387	0.787	0.802	0.891	0.418
LUD HGTs	LUD Native	0.353	0.312	0.12	0.131	0.528	0.779	0.924

Repeats are retrieved for each gene plus and minus 20 kbp either side from the TE reference library of *A. queenslandica.* Significance was determined with analysis of deviance tests with quasibinomial errors because of overdispersed errors. All *p*-values are raw and after corrections for multiple testing using the Benjamini-Hochberg method with a 20% false discovery rate, none are significant. Shading reflects raw *p*-values less than 0.05 and arrows reflect greater or lower proportion of the repeat class in the HGT sample, for those comparisons involving a HGT gene group and found to be possibly different (with raw *p*-values<0.05). SUD refers to strict criteria for a gene's putative unduplicated status, LUD reflects less strict criteria, and the young HGTs are 13 genes identified by Conaco *et al.* (2016) as resulting from more recent transfer events.

Appendix 4.6 Cut-off point in exonic TE coverage for predicted AqHGT-derived TEs

Because of a consistent distribution of AqHGT-derivatives with exonic TE content between one and 99%, the broad defining criterion used to categorise putative TEs is that genes must contain at least 50% TE-derived sequences. To further refine this criterion to a specific and conservative cut-off point, the midway point between 50 and 100% was used to define putative TEs.



Appendix 4.7 Details of the 49 AqHGTs that are highly similar to bacterial ME-borne genes

File available online at CloudStor+ (https://cloudstor.aarnet.edu.au/plus/index.php/s/GzHzoWIy8mfqT2l)

Appendix 4.8 The GC content and intron numbers of putative HGT-derived, contaminant, ambiguous and native genes in *A. queenslandica*

Gene classifications made by HGTracker (reviewed in Chapter 1; Fernandez-Valverde et al. in preparation). Mean values: putative contamination genes 55 GC% and 0.6 introns; ambiguous genes 58 GC% and 0.9 introns; putative HGTs 41 GC% and 3 introns; native genes 41 GC% and 6 introns.